

Université catholique de Louvain - Université de Mons

PHD THESIS

to obtain the title of

PhD in Applied Sciences
Specialty : SPEECH PROCESSING

Defended by

Sandrine BROGNAUX

on December 14th 2015

Expressive speech synthesis: Research and system design with hidden Markov models

Thesis Advisors: R. BEAUFORT, T. DUTOIT and M.
SAERENS

Louvain School of Engineering (UCL) - Faculty of Engineering (UMons)

Jury:

Dr. Richard BEAUFORT	- Nuance Communications, Inc.
Dr. Thomas DRUGMAN	- Amazon, Inc.
Prof. Thierry DUTOIT	- Université de Mons (UMons)
Prof. Julia HIRSCHBERG	- Columbia University in the City of New York
Prof. Piet MERTENS	- KU Leuven, University of Leuven
Dr. Vincent PAGEL	- Acapela Group
Prof. Marco SAERENS	- Université catholique de Louvain (UCL)
Prof. Xavier SIEBERT	- Université de Mons (UMons)
Prof. Anne Catherine SIMON	- Université catholique de Louvain (UCL)

*To Martin, my cute little son,
whose smiles, laughs and babblings
are the loveliest forms of expressivity.*

I have no special talents. I am only passionately curious.

Albert Einstein (1879-1955)

The only way to do great work is to love what you do.

Steve Jobs (1955-2011)

Abstract

While text-to-speech has long been centered on the production of an intelligible message of good quality, interest has recently shifted to the generation of more natural and expressive speech. This comes as an answer to the widespread criticism stating that current speech synthesizers lack fundamental human components. This thesis tackles that issue by considering three fundamental stages of HMM-based speech synthesis: the phonetic and prosodic annotations of the training corpus and their automatic alignment with the speech signal.

We first propose a systematic step-by-step study of HMM-based phonetic alignment in which the models are directly trained on the corpus to align. Based on a detailed analysis of the errors made by this technique, we developed three fully-automatic improvement methods which are shown to significantly improve the alignment of highly variable and expressive corpora.

We then define a two-level prosodic annotation of expressive corpora, describing accentual patterns and changes in speaking style. The integration of this manual annotation in the synthesis of sports commentaries positively impacts the naturalness of the expressivity. We also present an automatic annotator of accentual patterns in French and show that its integration in synthesis contributes to the naturalness of the voice.

Finally, our study points out that the choice for phonetic variants in French is influenced by the speaking style and that their consideration in the synthesis of sports commentaries improves the naturalness of the message. It indicates that phonetic changes should be considered, both at training and synthesis stages.

Keywords: HMM-based Speech Synthesis, Expressive Speech, Phonetics, Prosody, Phonetic Alignment, Speaking Style Adaptation

Acknowledgements

*Surround yourself with people who take
their work seriously, but not themselves,
those who work hard and play hard.*
Colin Powell (1937- ...)

The completion of 4 years of research naturally goes together with acknowledgements to all the people who have supported me, collaborated with me and advised me in multiple ways.

My first words obviously go to my three ‘promoters’ who proved to be very complementary in the supervision of my work. This thesis project would have never been launched without the invaluable help of **Dr. Richard Beaufort**. He was the one to convince me that PhD research was the new challenge I was looking for and he proofread and corrected many project drafts to achieve a high-level proposal which was accepted by the FNRS. **Prof. T. Dutoit**, for his part, proved to be an available supervisor, quickly answering all my questions and setting up meetings whenever I felt the need for expert advice... which is rather impressive for someone supervising about 15 PhD students at once! His humility and open-mindedness have been a real source of inspiration for me, guiding me like a "mystery light". Finally, **Prof. M. Saerens** has also been very present and always there to encourage me and answer to my numerous questions about statistical techniques. He also frequently advised me to take it easy and even to insisted that I take some rest when I was too busy in my work to realize I could take some time off. Many shared lunches have uncovered his sincere kindness and positive mind. While this three-party configuration may have seemed rather complex at first, it offered me an ideal research context in which I was left with much freedom to create but also advice to be reassured.

My thanks are next addressed to each member of my jury, as they all, in their own way, played a key role in the writing of this thesis. I am extremely grateful to **Dr. Thomas Drugman**, who acted as a real fourth supervisor (never too many...) as he took the time to organize weekly meetings with me during more than a year, giving me at the time the motivation and support I needed to keep faith with my

project. He is the one who taught me how to write a good quality academic paper and who drove me to present my research in so many international conferences during these 4 years. But more than all this, he has become a real friend who I trust and can turn to, which is priceless. **Prof. Anne Catherine Simon** can be seen as a major initiator of my project as she was probably the first one to ever talk to me about prosody, and to arouse my wish to go further in that field. Throughout my thesis, she has been an available expert, always pleased to give precious advice. **Prof. Piet Mertens** has also played a key role in the realization of my research project, mainly through his personal implication in the development of the local prosody annotator described in Chapter 10. He stood out as an extremely available and reactive collaborator with whom it was a real pleasure for me to work. Through more than 75 pages of e-mail exchanges (!!!), he provided me with very insightful advice while being personally involved in the research. **Dr. Vincent Pagel**, for his part, first came into play when supervising my Master’s internship at Acapela. He communicated to me his tremendous passion for text-to-speech along with his acute interest in prosody. If I remember correctly, he was the very first one to talk to me about the need for expressive speech synthesis and research in that field... I guess it did not fall on deaf ears! Throughout my PhD research, he always kept in touch, taking interest in my work and giving me advice anytime I asked for it. Finally, my deepest gratitude goes to **Prof. Julia Hirschberg** who welcomed me for a few months in her lab and treated me just as one of her own PhD students, making me feel immediately at home. From my very first day in New York, she initiated weekly meetings and took the time to comment on my work and to propose insightful research directions. She was also one of the first to congratulate me on my very new pregnancy and proved to be acutely understanding when morning sickness delayed my lab arrivals... She is an extremely smart and humane person, who is able to underline what’s best in each of her collaborators, with a steadily positive mind. I was extremely lucky to meet such a great and humble expert during my PhD.

This thesis has been realized in a co-tutoring framework (which meant two offices for me) in the Circuit Theory and Signal Processing lab (TCTS, UMons) and the Cental lab (UCL). This gave me the chance to mix with twice as many colleagues.

A majority of my research time was spent at the UCL, where I was lucky to share my office with my colleague **Sophie**. First collaborating on the Train&Align tool, Sophie and I soon became more than simple office roomies. Throughout my PhD, she encouraged me, proofread many paper drafts and gave me a helping hand whenever I was stuck. Our weekly shared coffee breaks significantly helped in keeping me on track and motivated. Most of all, Sophie turned into a close friend and a trusted confidant and this is probably worth more than a thousand accepted papers... For his invaluable technical help, I also warmly thank **Hubert** who is a real computer library by himself. His extreme availability and kindness clearly ought to be acknowledged here. My thanks also go to **Thomas** who offered precious advice in statistics and with whom I finally got to write a paper very recently. His acute sense for technical details and his ‘conceptual’ mind clearly were an asset for collaboration. I also want to thank my colleague **Amélie** who encouraged me throughout the thesis and kept faith, until the

end, in my ability to finish the writing in time despite the very short timing. Besides, all four of them proofread parts of this thesis and provided me with very judicious comments, which constituted precious help at the very end of such a tremendous work. In addition, I am indebted to **Cédrick**, who offered me a roof in his lab, despite my total lack of official connection to the Cental. I am also grateful to some present or past ‘Centaliens’ (some already quoted) and ‘Earlytracksiens’ for much party time spent outside the lab, in front of a nice beer or some tasty dinner: many thanks to **Amélie, Denis, Laurence, Jean-Léon, Louis, Patrick, Thomas, Sophie**, etc. Your love of life and your tireless desire to celebrate it offered me a real breath of fresh air outside work. Finally, I am grateful to all other previous and current Centaliens for the nice lunches and discussions we have shared and for their steady participation in my perception tests: thanks to **Adrien, Anais, Babette, Charlotte, Claude, Damien, Laetitia, Michele, Nadja, Nicolas, Noémi, Olga, Oto, Serge, Sasha and Stéphanie (X2)**. It’s been a real pleasure for me to share with you that small slice of my professional life.

I also take the opportunity to thank all the other PhD students in the UCL Arts Faculty, with whom I have shared lunch breaks or drinks at the now renowned "PhD lunches" or outside in the city. You are too many to cite but I would like to address special thanks to **Barbara, Julie, Natassia and Samantha**. **Marie-Catherine** (alias MC) is also one of them but deserves a separate mention. I met MC 10 years ago, in first year of my Bachelor’s degree in Germanic languages. Since then, we have followed parallel paths, both of us obtaining an FNRS grant in 2011. MC is a colleague with whom I have shared research concerns and written papers and is a great collaborator. But MC is, first and foremost, a very close friend who I want to thank for her infallible support and trust.

I also spent some time at my office at the TCTS in Mons, randomly moving from desk to desk at the whim of incoming and leaving colleagues. There, I was lucky to meet **Benjamin**. My very first collaborations, in the framework of the Sportic project, were mainly realized in collaboration with Benjamin who had the patience to teach me how to use (and compile!) the HTS library. I am very grateful for his technical help throughout my thesis and really enjoyed our shared lunches at the cité. I am also grateful to all my other colleagues at the TCTS with whom I had the chance to share very nice discussions, at lunch or at conferences far abroad : **Alexis, Husseyn, Jérôme, Joëlle, Justine, Maria** (especially for our great Mustang trip in the Keys!), **Nathalie, Nicolas** (alias Soso), **Onur, Stéphane, Zach**, etc. I hope to stay in touch, now that my new job brings me back to Mons...

I would also like to acknowledge all the other colleagues I have met across the world, in conferences, internships or research collaborations. My thanks go to **Erica** and **Laura** for their warm welcome at Columbia and tons of good ideas on how to discover New York City, to **Rasmus** (CSTR) for his interest in my work and the resulting fruitful collaboration, but also to **John, Julie B., J-P, Finnian** and many other research experts with whom I have enjoyed partying on conference nights around one (or several... I must admit) beer(s). Finally, I must address my infinite thanks to **Mathieu Avanzi**, who proved to be the best collaborator a researcher can dream of. He is a passionate and hard-working linguist and I really enjoyed our multiple discussions on and outside work. Thanks to his extreme reactivity and very

smart advice, he played a crucial role in the writing of this thesis.

My thanks then go to all my friends, who have closely followed my PhD project, from my first doubts about embarking on this perilous adventure to my deep contentment facing the accomplished work. They are too many to list them all but I must acknowledge here that their help in participating in an endless number of perception tests without a single complaint is rather remarkable. Special thanks to **Coline**, **Nath**, **Pauline** and **Thomas** for their infallible support and sincere friendship.

I am also thankful to **my family-in-law** for its acute interest in my work and extremely warm welcome into their household. I would also like to thank my **extended family** for its support all along this journey.

My deepest gratitude then obviously goes to **my parents** who played a crucial role in making me the curious and passionate adult I am today. From a very early age, they gave me an unquenchable thirst for knowledge and never failed to show me how life is amazing if we take time to take a closer look at it. Thank you, mum, for teaching me how to read as soon as I asked for it, for taking the time to answer to my endless questions, and for reading so many bed-time stories when I did not want to sleep. Thank you, daddy, for communicating your love for physics and technical details, your passion for travel and your flawless trust in my ability to succeed in any project I undertook. Finally, I want to thank you both for your infinite love and extreme availability, both in good and more difficult times. I could not have dreamt of a better home to gain the required self-confidence and willingness to realize this project and many other to come...

My last words are naturally addressed to my husband and soulmate **Laurent**, who has shared my life for more than 9 years now. He is obviously the one who has most closely followed my thesis journey, backing me when my first paper was repeatedly rejected and sharing my joy facing my first (and many next) accepted papers. Through his unconditional love, he never failed to encourage me, and a single look sufficed to make me feel proud of what I had achieved. He (nearly) never complained about my late returns at home and even gave me the opportunity to follow my dreams by accompanying me for my internship in New York City. I am extremely lucky to have had him by my side throughout this great adventure, acting as a constant reminder of what I considered to be an essential priority, *i.e.* building a rich personal life as a fulfilled friend, spouse and mother... Laurent sure gave me the most beautiful gift I could ever have hoped for to finish this thesis on a perfect note: our wonderful son Martin, who has brought a gust of happiness in our life...

This thesis has been realized in a co-tutoring framework within the Circuit Theory and Signal Processing (TCTS) lab of the Faculté Polytechnique de Mons (UMons) and the Cental and ICTEAM at the Université catholique de Louvain, under a research fellow FNRS grant.

Contents

1	General introduction	1
1.1	General background	2
1.1.1	Why synthesizing expressive speech ?	2
1.1.2	What is meant by ‘expressive’ speech ?	4
1.1.3	Which are the interconnections between expressive and spontaneous speech ?	5
1.1.4	What are the major characteristics of spontaneous expressive speech ?	5
1.2	HMM-based speech synthesis	6
1.2.1	Speech synthesis: A brief history	6
1.2.2	Comparing unit-selection and HMM-based speech synthesis	7
1.3	HMM-based expressive speech synthesis	8
1.4	Contributions of the thesis	9
1.5	Structure of the thesis	14
I	Automatic phonetic alignment of expressive speech corpora	17
2	Training on the corpus to align : Train&Align	21
2.1	Introduction and context	24
2.2	HMM-based phonetic alignment: State of the art	27
2.3	Development protocol	31
2.3.1	Development database	31
2.3.2	Evaluation metrics	31
2.4	Description of our method: Train&Align	32
2.5	Evaluation	38
2.5.1	Speech material	38
2.5.2	Results with Train&Align	39
2.5.3	Comparison with state-of-the-art models	41
2.5.4	Influence of the corpus size	42
2.5.5	Effect of bootstrap	43
2.6	Analysis of errors made by <i>Train&Align</i>	44
2.7	Conclusion	47

CONTENTS

3	Improving HMM-based alignment	49
3.1	Introduction	51
3.2	Existing improvement methods of HMM-based forced alignment	52
3.3	Proposed improvement methods	53
3.3.1	Use of a Voice Activity Detection algorithm (<i>VAD</i>)	53
3.3.2	Augmentation of the feature vector (<i>AddFeat</i>)	54
3.3.3	Exploitation of the time-reversed sound (<i>Reverse</i>)	56
3.4	Evaluation	58
3.4.1	Improvement methods	58
3.4.2	Advantage of <i>T&A2</i> in small corpora	60
3.4.3	Comparison with state-of-the-art models	61
3.5	Conclusion	61
4	Train&Align: A new online tool for automatic phonetic alignment	63
4.1	Introduction	65
4.2	Overview of the tool	66
4.3	Perspectives	71
II	Definition of a three-level expressivity framework for the synthesis of sports commentaries	73
5	Prosodic and phonetic improvements for HMM-based expressive TTS: A theoretical and practical introduction	77
5.1	Introduction	80
5.2	Phonetics in speech synthesis : The Phonetizer	82
5.2.1	The phonetization module	82
5.2.2	Integrating phonetic variation in speech synthesis	85
5.3	Prosody in speech synthesis: The Prosody Generator	88
5.3.1	Phrasing, intonation and accentuation	88
5.3.2	Existing models of prosody	93
5.3.3	Integrating prosody in text-to-speech	101
5.3.4	The question of global prosody features	109
5.4	HMM-based expressive speech synthesis: The DSP	111
5.4.1	HMM-based speech synthesis, spectral and prosodic modeling .	111
5.4.2	Expressive speech synthesis: State of the art	118
5.5	Available front-ends : The development of eLite-HTS	125
5.5.1	Architecture of eLite	126
5.5.2	eLite generation of the HTS file	128
5.5.3	Specifications of the web service	129
5.5.4	Future work	130
5.6	A concrete application of expressive speech synthesis: The Sportic project	131
5.6.1	Outline of the project	131
5.6.2	Concept-to-speech synthesis	133
5.6.3	Corpus design	135
5.6.4	Outline of the next chapters	136

6	Modeling Prosody for the Synthesis of Sports Commentaries	139
6.1	Introduction	142
6.2	Proposed prosody annotation protocol	145
6.2.1	Description of the two-layer prosody annotation	145
6.2.2	Inter-annotator rate	149
6.3	Acoustic analysis of the prosodic labels	149
6.3.1	Analysis of the local annotation	150
6.3.2	Analysis of the global annotation	152
6.4	Application to speech synthesis	154
6.4.1	Methodology	155
6.4.2	Integration of the local labels	155
6.4.3	Integration of the global labels	157
6.4.4	Integration of both local and global labels	162
6.4.5	Discussion of the evaluation framework	163
6.5	Conclusion and perspectives	164
7	Improving the local prosodic annotation: One or several emphatic stresses ?	167
7.1	Introduction	169
7.2	Statistical analysis of emphatic stresses	170
7.2.1	Extraction of acoustic features	170
7.2.2	Dimensionality reduction	171
7.2.3	Clustering	172
7.2.4	Correlation between clusters and linguistic contexts	174
7.3	Speech synthesis: A perceptual study	176
7.3.1	Evaluation protocol	176
7.3.2	Results	176
7.4	Conclusion	177
8	Automatic prediction of the expressive prosody annotation	179
8.1	Introduction	181
8.2	Automatic prediction of local labels from text and sound	183
8.3	Automatic prediction of global labels from text and sound	186
8.3.1	Which minimal unit should be chosen ?	186
8.3.2	Extraction of acoustic features	189
8.3.3	Automatic prediction of macro-prosodic labels: Method	190
8.3.4	Integrating the automatic annotation in speech synthesis: A subjective evaluation	193
8.4	Automatic prediction of prosodic labels from text only: Perspectives	196
8.5	Conclusion	197
9	Impact of phonetic variations on the synthesis of sports commentaries	199
9.1	Introduction and state of the art	201
9.2	Corpus design	204
9.3	Methodology for the phonetic analysis of the corpus	206
9.4	Global phonetic analysis of the corpus	208

CONTENTS

9.4.1	Overall proportion of phonetic changes	208
9.4.2	Schwa elision	209
9.4.3	Final schwa pronunciation	209
9.4.4	Final liquid elision	211
9.4.5	Liaison	212
9.4.6	Summary	213
9.5	Phonetic variations : Focus on sports commentaries	213
9.6	Correlations between phonetic and rhythmic features	214
9.7	Impact of phonetic variation on synthesized speech	217
9.7.1	Experimental protocol	217
9.7.2	Evaluation and discussion	217
9.8	Conclusion and perspectives	219
 III Prosodic and phonetic improvements of HMM-based speech synthesis: Further developments		223
10 Automatic annotation of an autosegmental description of French prosody		227
10.1	Introduction	229
10.2	Developing an automatic prosody annotator	232
10.2.1	Development corpus	234
10.2.2	Annotation protocol	234
10.2.3	Choice for an automatic prominence detector	239
10.2.4	An alternative option: Training an AuToBI classifier on a French annotated corpus	241
10.3	Integration of the local annotation in speech synthesis	242
10.3.1	Description of the corpus	242
10.3.2	Method	243
10.3.3	Evaluation and discussion	245
10.4	Reducing the set of labels	247
10.4.1	Annotation protocol	247
10.4.2	Perceptive evaluation	250
10.5	Conclusion	251
 11 Modeling spontaneous speech: The integration of phonetic variants in speech synthesis		253
11.1	Introduction	257
11.2	A multifactorial analysis of phonotactic phenomena in French: Impact of the speaking style	259
11.2.1	Introduction	259
11.2.2	Overview of the factors affecting French pronunciation	261
11.2.3	Material	263
11.2.4	Overview of the impact of speaking style on phonotactic variation	264
11.2.5	An in-depth multifactorial analysis of post-obstruent liquid elision	268

CONTENTS

11.2.6 Conclusion	272
11.3 Spontaneous speech synthesis in English: Investigating the impact of the phonetization of the training corpus	273
11.3.1 Introduction	273
11.3.2 Material: A parallel corpus	275
11.3.3 Phonetic analysis of the corpus	276
11.3.4 Integrating pronunciation variation in forced alignment	278
11.3.5 Synthesis evaluation	282
11.3.6 Conclusion	284
12 General conclusion	287
12.1 Conclusion and contributions	287
12.2 Limitations and perspectives	294
Bibliography	299
Appendices	341
A HTS description file	343
B List of eLite phonetic symbols	345
C Distribution of liaisons in the CPROM-PFC corpus	347
D List of Combilex phonetic symbols	349
E The set of hand-written rules exploited to populate the lattices	353
F PER differences between the improved systems and the Gold Stan- dard annotation	355
G Our publications and related talks	357
G.1 Regular papers in journals	357
G.2 Papers in conference proceedings	357
G.3 Talks and posters (with abstract)	359
G.4 Invited talks (without abstract)	359
G.5 Technical reports	359

CONTENTS

List of Figures

1.1	Scientific papers dealing with (expressive) TTS, from 1990 until today.	3
1.2	General structure of a HMM-based speech synthesizer.	10
1.3	Plan of the thesis.	12
2.1	Example of aligned sound transcription in Praat.	24
2.2	Schematic representation of HMM-based forced alignment.	28
2.3	Message encoding.	28
2.4	HCompV command for flat start initialization.	29
2.5	Graphical representation of the Viterbi algorithm.	30
2.6	Model corresponding to the phoneme /O/.	33
2.7	Silence model and short-pause model.	34
2.8	Correct alignment rates, with varying corpus sizes.	43
2.9	Average relative improvements over T&A, with various bootstrap sizes.	45
2.10	Example of silence and liquid misalignment in the automatic alignment.	46
3.1	Correct alignment rates with a varying VAD threshold.	55
3.2	Average relative improvement over T&A with additional acoustic features.	57
3.3	Relative improvement over T&A for the improvement methods, with varying corpus sizes.	60
4.1	Train&Align graphical interface (Registration, step 1)	67
4.2	Train&Align graphical interface (Registration, step 2)	68
4.3	Train&Align graphical interface (Home)	69
4.4	Train&Align graphical interface (Train and align - Stage 1)	69
4.5	Train&Align graphical interface (Train and align - Stage 2)	70
4.6	Train&Align graphical interface (Train and align - Stage 3)	71
5.1	A simple but general functional diagram of a TTS system.	80
5.2	The functional diagram of a fairly general text-to-speech.	81
5.3	Retrieval of the pronunciation of < a > in < behave > by trie search.	84
5.4	Elementary typology of accentual categories.	91
5.5	A functional model for the process of generating F0 contours.	95
5.6	Mertens' tones distribution in maximal intonational phrase	100
5.7	Different stages of prosody integration in speech synthesis.	102

LIST OF FIGURES

5.8	Liberman's "chinks and chunks"	104
5.9	Block diagram of HMM-based speech synthesis system (HTS).	112
5.10	Feature vector.	112
5.11	Decision tree for state duration models.	114
5.12	Synthesis part of the system.	115
5.13	CMLLR and its related algorithms.	117
5.14	Combined algorithm of the linear regression and MAP adaptation.	118
5.15	Relations between the 3 modules of eLite in the MLDS.	126
5.16	Example of NLP analysis for sentence in TextGrid format.	128
5.17	Flowchart of eLite-HTS.	130
5.18	Outline of the Sportic project	132
5.19	Flowchart of the audiovisual game summary application.	133
5.20	Recording conditions of the Sportic corpus.	136
6.1	Speech styles in sports commentaries.	144
6.2	The global labels on a dimensional scale.	149
6.3	Dynamic representation of the pitch on a small part of the corpus.	154
6.4	Preference scores of the synthesis for the integration of the local labels.	157
6.5	Synthesis of a sentence with focus.	158
6.6	Averaged intelligibility MOS scores.	161
6.7	Averaged segmental quality MOS scores.	161
6.8	Averaged quality MOS score for the integration of the global labels.	162
7.1	Scree plot for the 10 first PCA components.	171
7.2	Ward's dendrogram on the first four PCA components of our data.	172
7.3	Value of the silhouette of the K-means on four PCA components.	173
7.4	K-means clustering on our data.	174
7.5	Percentage of preferences for each model in the three comparison pairs.	177
8.1	Correlation between linguistic segments and prosodic changes	188
8.2	Scree plot on the ten first PCA components.	191
8.3	Subjective preferences in terms of naturalness of the synthesized voice.	195
8.4	Perceived excitation level of each sub-genre.	196
9.1	Percentage of elided schwas in middle word position.	210
9.2	Percentage of words, not ending in -e, pronounced with a final schwa.	211
9.3	Percentage of words pronounced with a final schwa.	212
9.4	Percentage of elided schwas.	215
9.5	Percentage of prominent syllables.	216
9.6	Preferences for NLP and real phonetization.	218
9.7	Preferences for NLP and real phonetization with at least 2 schwa elisions.	218
9.8	Preferences for NLP and real phonetization for each type of variation.	220
10.1	Preference rates in terms of expressivity	245
10.2	Preference rates in terms of resemblance to the original voice	246
10.3	Preference rates with reduced set of labels in terms of expressivity	250

LIST OF FIGURES

10.4 Preference rates with reduced set of labels in terms of resemblance to the original voice	251
11.1 Distribution of the liaisons in the corpus.	268
11.2 WPFOL deletion rate as a function of the following segment.	271
11.3 WPFOL deletion rate as a function of the speaking style.	272
11.4 Segment of the same sentence produced spontaneously or as read prompt.	279
11.5 Boxplot of the side-by-side MOS test.	283

LIST OF FIGURES

List of Tables

1.1	Comparison between unit-selection and HMM-based speech synthesis.	7
2.1	Languages covered by various existing tools for automatic phonetic alignment.	25
2.2	Development data set features.	31
2.3	Studied parameters which are tuned on our development corpus. . . .	34
2.4	Alignment rates with different sizes and target rates of the speech frames.	35
2.5	Different model configurations.	36
2.6	Classes used to determine the context with tied-state triphones.	36
2.7	Alignment rates with monophones, triphones and tied triphones. . . .	37
2.8	Alignment rates with optimization of the number of training iterations.	38
2.9	Evaluation data set features.	39
2.10	Correct alignment rates of T&A on our evaluation database.	40
2.11	Correct alignment rates with speaker-independent models and T&A. . .	42
2.12	For specific phoneme classes, percentage of global and specific errors. .	45
3.1	Example of systematic errors made by <i>T&A</i>	57
3.2	Relative improvement over T&A for <i>VAD</i> , <i>AddFeat</i> , <i>Reverse</i> and <i>T&A2</i> . .	59
3.3	Correct alignment rates with speaker-independent models and T&A2. . .	62
5.1	Tonal labels of ToBI.	98
5.2	Pitch levels in Mertens' model of French prosody.	99
5.3	Correspondance between eLite and HTS layers.	129
5.4	Main requests of the web service.	131
6.1	List of local labels.	147
6.2	Inter-annotator confusion matrix for the local annotation of the corpus.	150
6.3	Inter-annotator confusion matrix for the global annotation of the corpus.	150
6.4	Average acoustic values of the local labels in the corpus.	151
6.5	Average prominence levels assigned by PromGrad.	152
6.6	Average acoustic values of the global labels in the corpus.	152
6.7	Transition probability matrix for global prosody annotation.	153
6.8	Average duration of each sub-genre in the corpus.	153
6.9	Total duration of the various sub-genres in our database.	159
6.10	Preference scores for the integration of both global and local labels. . .	164

LIST OF TABLES

7.1	Acoustic characteristics of the clusters.	174
7.2	Correlation between the clusters and linguistic contextual information.	175
7.3	Acoustic realization of emphatic stresses at the end of RGs or not.	176
8.1	Correct prediction rates with SVM.	185
8.2	Confusion matrix between manual annotation and automatic reclassification, with Gaussian mixture models.	192
8.3	Confusion matrix between manual annotation and automatic reclassification, with logistic regression.	192
8.4	Number of labels for each annotation.	192
9.1	Distribution of the sub-corpora in the three studied situational traits.	205
9.2	Summary of the phonetic changes for the three situational traits.	209
9.3	Phonetic changes for sports commentaries and neutral read speech for TTS.	214
10.1	Mapping of Polytonia labels to our autosegmental annotation.	238
10.2	Prominence annotation rates.	241
10.3	Number of occurrences of each ToPPos-Fr annotation label in our corpus.	243
10.4	Corresponding labels for each function assigned to final stresses.	249
11.1	Distribution of words ending with a post-obstruent in the corpus.	267
11.2	Predictors with a significant effect on WFPOL behavior in the corpus.	270
11.3	Overall differences between the manual and automatic annotations.	277
11.4	Overall differences between manual and (improved) automatic annotations.	281
11.5	Adjusted p-values for the Wilcoxon signed-rank test.	284
B.1	List of phonemes used in eLite	346
D.1	List of phonemes used in Combilex	349
F.1	PER differences between the improved systems and the Gold Standard annotation.	355

Abbreviations and acronyms

ANN	Artificial Neural Networks
AP	Accentual Phrase
ASR	Automatic Speech Recognition
AM	Autosegmental and Metrical
CART	Classification And Regression Tree
CGD	Chirp Group Delay
CI	Confidence Interval
CMLLR	Constrained Maximum Likelihood Linear Regression
CRF	Conditional Random Field
CS	Communicative Situation
CTS	Context To Speech
dB	Decibel
DCG	Definite Clause Grammar
DP	Dynamic Programming
DPPT	Differential Phase Peak Tracking
DSM	Deterministic plus Stochastic Model
DSP	Digital Signal Processing/Processor
DTW	Dynamic Time Warping
EM	Expectation Maximization
ER	Error Rate
F0	Fundamental Frequency
FIR	Finite Impulse Response

LIST OF TABLES

FFT	Fast Fourier Transform
FST	Finite State Transducer
GEE	Generalized Estimated Equation
GLMM	Generalized Linear Mixed Model
GMM	Gaussian Mixture Model
GPS	Global Positioning System
HC	Higher Class
HMM	Hidden Markov Model
HSMM	Hidden Semi Markov Model
HTK	Hidden Markov Model Toolkit
HTS	HMM-based Speech Synthesis System (H -Triple S)
Hz	Hertz
IP	Intonational Phrase
IPA	International Phonetic Alphabet
IPU	InterPausal Unit
MAP	Maximum A Posteriori
MGCC	Mel-Generalized Cepstral Coefficients
MFCC	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLDS	Multi-Layer Data Structure
MLLR	Maximum Likelihood Linear Regression
MLSA	Mel Log Spectrum Approximation
MOS	Mean Opinion Score
MSD	Multi-Space Distribution
NLG	Natural Language Generation
NLP	Natural Language Processing/Processor
PCA	Principal Component Analysis
PDFs	Probability Density Functions

LIST OF TABLES

RF	Random Forest
POS	Part Of Speech
RG	Rhythmic Group
RP	Received Pronunciation
RQ	Research Question
SPTK	Speech Signal Processing Toolkit
SRH	Summation of the Residual Harmonic
ST	Semi-Tone
SUS	Semantically Unpredictable Sentences
SVM	Support Vector Machine
T&A	Train & Align
ToBI	Tones and Break Indices
TTS	Text To Speech
UG	Unification Grammar
VAD	Voice Activity Detection
VFS	Vector Field Smoothing
WC	Working Class
WFPOL	Word Final Post Obstruent Liquid

LIST OF TABLES

Chapter 1

General introduction

Contents

1.1	General background	2
1.1.1	Why synthesizing expressive speech ?	2
1.1.2	What is meant by ‘expressive’ speech ?	4
1.1.3	Which are the interconnections between expressive and spontaneous speech ?	5
1.1.4	What are the major characteristics of spontaneous expressive speech ?	5
1.2	HMM-based speech synthesis	6
1.2.1	Speech synthesis: A brief history	6
1.2.2	Comparing unit-selection and HMM-based speech synthesis	7
1.3	HMM-based expressive speech synthesis	8
1.4	Contributions of the thesis	9
1.5	Structure of the thesis	14

1.1 General background

Speech is a fantastic means of communication which makes us, humans, so unique. It plays a crucial role in any man-to-man interaction and fulfills a wide range of functions, from pure information transfer to the expression of emotions or affect. For about 50 years now, technology has focused on this very complex form of human communication, in an attempt to produce and recognize speech sounds. Speech synthesis, in particular, aims at the pronunciation of an unknown text by a machine. It forms part of our everyday life and can be found in a wealth of practical applications, including GPSs, public-space announcements, vocal prostheses, electronic cards and video games. In 2011, LT-Innovate indicated that the speech technology and service markets accounted for nearly 6 billion dollars worldwide, with a high expected growth in the following years. The interest for research in the domain is therefore striking.

1.1.1 Why synthesizing expressive speech ?

While the interest has long been centered on the production of an intelligible message of good quality, it has recently shifted to the generation of more natural and expressive speech. In the last ten years, the attention paid to expressive and emotional speech has rapidly grown, making it a key current research topic (see Figure 1.1¹). One of the first marks of this new interest is found in the 2000 ISCA Tutorial and Research Workshop (ITRW) which focused on speech and emotion. Since then, it has been a recurrent topic in international speech conferences. In 2014, more than 300 scientific papers referred to the notion of expressive/emotional speech synthesis and 27 contained these terms in their title.

We believe that this sudden rise in the attention paid to expressive speech synthesis may be explained by the convergence of multiple factors:

- Digital technologies have achieved tremendous improvements in the imitation of humans by machines. A wide range of functions, previously fulfilled by humans, have been robotized. On screen, synthesized images have also reached an extremely high quality, the distinction between synthesized and real images becoming hard to make. In that framework, synthesized voice is supposed to display human characteristics. While people are gradually accepting to be in increased contact with machines, they expect them to behave like humans would. In that sense, the voice should sound natural and express emotions and affect as normal people do in conversation. In 2006, Campbell pointed out that the discrepancies that still remained between a synthetic voice and a natural human voice precluded its large-scale marketing.
- Speech synthesis has achieved, for more than a decade, a very high level of intelligibility. With the development of unit-selection synthesis (see below),

¹Computed on Google Scholar search engine, with the queries : expressive/emotional speech synthesis/text-to-speech/TTS and speech synthesis/text-to-speech/TTS, respectively.

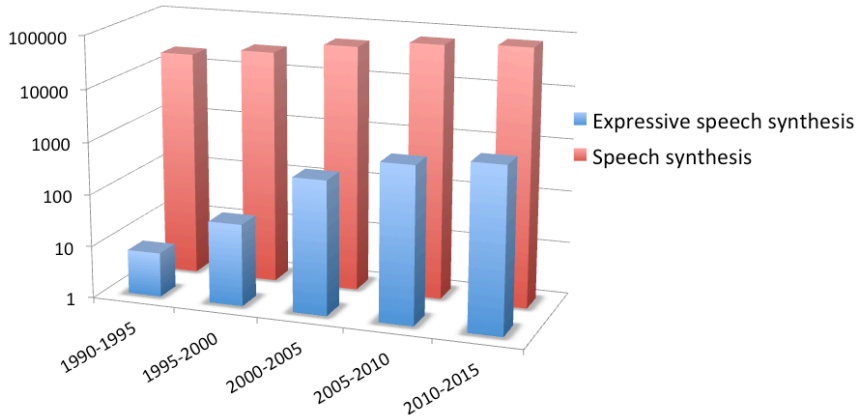


Figure 1.1 - Number of scientific papers dealing with expressive speech synthesis or speech synthesis in general, from 1990 until today (on a logarithmic scale), computed with Google Scholar.

synthesized speech has also integrated basic neutral prosodic patterns, which have resulted in high quality neutral reading voice. There is now room for improvement in terms of naturalness and expressivity of that voice. In fact, commercialized speech synthesizers are often criticized as unnatural because of the very high monotony of the voice (Kim et al., 2004). They are said to be not representative of the way people speak in real life (Campbell, 2006).

- In the last decade, a new synthesis technique based on Hidden Markov Models (HMM) (Yoshimura et al., 1999; Black et al., 2007; Zen et al., 2009) has gradually been adopted by the research community. One of the major advantages of this technique is that it allows for an increased degree of flexibility of the synthesized voice. This obviously facilitates the integration of more complex intonation patterns in view of expressive speech synthesis.

Numerous applications could benefit from expressive speech synthesis. Montero et al. (1998) insist on the importance of generating non-monotonous expressive speech for people using vocal prostheses. Another application mentioned by Burkhardt and Stegmann (2009) is the creation of believable characters for emotional chat and gaming. They also highlight the importance of producing naturally-sounding artificial agents in man-machine applications². Beller (2010) also emphasizes the role that expressive speech synthesis could play in the artistic field. Film directors, dubbing studios or composers should be highly interested in such technologies. Finally, the recent take-up of audiobooks also triggers a need for an automatic expressive reading of narrative texts.

²Links to available emotion-aware applications are listed at <http://emotionalapplications.syntheticspeech.de/>. Regrettably, the page has not been updated since 2007.

1.1.2 What is meant by ‘expressive’ speech ?

The notion of expressivity is rather unstable and has been used in a wide range of contexts to relate to sometimes very different concepts. This term usually covers many facets of speech naturalness, including emphasis, focus, speaking style but also the expression of emotions and attitudes. **Emphasis** and **focus** are directly related to the notion of stress in French (see Chapter 5) and concern the highlighting of speech segments. Political speech is often considered as a typical example of emphatic speech. This also shows the role played by the **speaking style** (*i.e.* the context in which we speak) in that framework. **Emotions** and **attitudes** differ in the fact that they are more centered on the speaker and the interlocutor and less on the semantic content of the message itself. The distinction between attitudes and emotions is hard to make and rather controversial in the literature. In the framework of this thesis, we will consider that attitudes, unlike emotions, do not exist outside the communication context: a speaker can be in a specific internal emotional state (*e.g.* sad or angry) but its attitude will only appear in man-to-man interactions (*e.g.* ironic or seducing)³. Studies on emotions in speech are usually based on two major classification frameworks, *i.e.* basic/palette emotion theories (Ekman, 1992) which define a fixed set of basic emotions⁴ and dimensional theories (Mehrabian and Russel, 1974; Russell, 1980) which classify emotions according to several axes, usually arousal and valence.

In the framework of this thesis, we propose our own definition of expressive speech. Expressivity is described "ex negative", by indicating what it is not (Schuller et al., 2011). In that perspective, **expressive speech is considered as not "non-expressive", *i.e.* non neutral** (Beller, 2010; Batliner et al., 2010). This provides an extended meaning which covers all the aforementioned concepts which largely interact.

In line with these considerations, is regarded as expressive any aspect of speech that makes it more naturally sounding and suited for a specific communicative situation other than read non-emotional laboratory speech (that can be seen as a typical example of what we designate as ‘neutral speech’)⁵. Expressive speech is considered to provide information beyond the propositional content of the uttered sentences and is seen to be characterized by context-dependent speaking styles, emotions and attitudes and to feature specific prosodic events like emphasis and focus. This thesis aims at tackling these multiple aspects of expressive speech and emphasizes the interest of spontaneous speech in that context, for the reasons detailed below.

³The reader interested in an in-depth discussion of the concepts of emotion and attitude in the framework of prosody is referred to Rillard (2014).

⁴In line with this theory, the MPEG4 standard defines 6 basic emotions: anger, disgust, fear, sadness, joy and surprise.

⁵The resulting objective of synthesizing expressive speech is therefore to make the message sound more believable in a specific communicative situation.

1.1.3 Which are the interconnections between expressive and spontaneous speech ?

Interestingly, **spontaneous speech** synthesis is usually regarded as fitting in research on expressive speech⁶. We believe that this can be explained by multiple connexions between both types of speech:

- Spontaneous speech is rarely completely neutral, as pointed out by Campbell (2006) who indicated that "more than half of the utterances used in daily interaction have minimal propositional content and [...] function instead to establish speaker-listener relationships and to express the speaker's affective states".
- Spontaneous speech, like expressive speech, can be seen as relevant in the quest for naturalness, as it is the type of speech we are most used to in everyday communication (Laan, 1997). Perceptual experiments in Dall et al. (2014) pointed out that sentences pronounced in a spontaneous context were assessed as significantly more natural than the exactly same sentences read by the same speakers.
- "Natural" expressivity can only be found in spontaneous speech, expressivity encountered in read speech being acted. Acted speech has often been criticized as displaying stereotypical or exaggerated realizations of the expression that highly differ from real speech (Campbell, 2000; Iida et al., 2000).

In line with these considerations, the core of our research focuses on a specific application of **spontaneous expressive speech synthesis**: the synthesis of live sports commentaries.

1.1.4 What are the major characteristics of spontaneous expressive speech ?

Spontaneous expressive speech is characterized by both *what* is said, and *how* it is said. In speech synthesis, the text to pronounce is known, and the focus is then on *how* to say it⁷.

Spontaneous expressive speech is mainly characterized by 3 major speech components, which are detailed in Chapter 5:

- **Phonetics** is concerned with the sequence of pronounced speech sounds. In phonology (linguistic analysis of speech sounds) and speech synthesis, such speech sounds are specified at the level of linguistically-distinctive speech segments, called phonemes. Slightly different speech sounds may be contextual or free variants (allophones) of one and the same phoneme. Phonetic modifications in French include for instance [ə] elisions and liaisons, *i.e.* the pronunciation

⁶The 7th Speech Synthesis Workshop (SSW7) in 2010, for instance, encouraged publications on "spontaneous/expressive" speech synthesis.

⁷This is in opposition to Concept-to-Speech (CTS) synthesis in which the propositional content of the sentences also needs to be generated by a Natural Language Component (See Chapter 5).

of a latent final consonant before a vowel or mute ‘h’⁸. Spontaneous speech, in particular, has been shown to influence pronunciation by triggering a higher rate of elisions and less realized liaisons (Lucci, 1983; Hansen, 1991; Léon, 1992; Fougeron and Frauenfelder, 2001).

- **Prosody** can be defined as "the study of various acoustic parameters as fundamental frequency (F0), intensity, duration" (Mounin, 1974). The variation of these parameters covers multiple phenomena like intonation, accentuation and phrasing (Taylor, 2009). A wide range of studies have indicated that accentual patterns in expressive speech differ from neutral speech in terms of stress density, distribution and realization (Fónagy, 1979; Paeschke et al., 1999; Paeschke and Sendlmeier, 2000).
- **Voice quality** is defined as the characteristic auditory coloring of an individual speaker’s voice (Laver, 1980). It refers to what is commonly known as the ‘timbre’ (*e.g.* a soft, creaky or breathy voice). Voice quality has sometimes been regarded as an additional dimension of prosody (Campbell and Mokhtari, 2003) but will be considered as a different speech component in the framework of our research. It has been shown to be complementary to prosody in the expression of emotions (Hirose et al., 2000) and Gobl and Ni Chasaide (2003).

The three speech components should therefore be considered when synthesizing spontaneous expressive speech.

1.2 HMM-based speech synthesis

1.2.1 Speech synthesis: A brief history

During the last 40 years, speech synthesis techniques have considerably evolved. The first widely-marketed technique, **formant** or **rule-based synthesis**, appeared in the 70’s and was implemented in the widespread Klatt synthesizer (Klatt, 1980). It however suffered from a low signal quality, notably because of the complexity to heuristically model co-articulation, *i.e.* the transition between phonemes.

More recent techniques have benefited from the apparition of large-size speech corpora, thereby moving from knowledge-based to data-based approaches. Two major techniques emerged in the 80’s: **unit-concatenation, or diphone synthesis** (Moulines and Charpentier, 1989; Dutoit et al., 1996) and **unit-selection synthesis** (Sagisaka et al., 1992; Hunt and Black, 1996). These techniques take advantage of signal dynamics stored in the units of the database, such as co-articulation, which clearly improves the speech quality. In concatenation synthesis, sub-word units (usually diphones) that are stored in a database are concatenated to generate the required sentence. One sample of each diphone being stored, these samples still need

⁸It should be noted that this definition is a simplistic reduction of the high complexity of liaison phenomena in French which can imply more than the pronunciation of a consonant and may require vowel changes (*e.g.* un beau séjour *versus* un bel endroit). For a detailed description of liaisons in French and the wide range of forms they can take, see Mertens (2005). As further pointed out later in the thesis, the realization of liaisons depends on a large number of factors and is notably directly influenced by the syntactic links between the lexical units.

to be processed to correspond to the expected prosodic features. Conversely, in unit selection synthesis, the database contains several samples of each unit. At synthesis time, the most appropriate unit is selected. As each unit is available in a wide range of prosodic contexts, less modification of the signal is required at synthesis. It is however impossible to cover the large number of potential prosodic patterns which means that the ideal unit cannot always be found.

Speech synthesis has thus oscillated between speech signal modeling and the use of speech data. Signal modeling (*e.g.* formant synthesis) presents the advantage of requiring very little space but produces a poorly natural voice. Conversely, the exploitation of speech databases in the framework of unit-selection speech synthesis allows for the generation of speech that can sometimes hardly be differentiated from a human voice but requires much memory. The last state-of-the-art technique, *i.e.* **HMM-based speech synthesis** (Yoshimura et al., 1999; Black et al., 2007; Zen et al., 2009), tries to strike a balance, offering a good-quality signal without requiring a large storage space. It is now widely used in research, even if unit-selection synthesis is still currently the leader on the market.

1.2.2 Comparing unit-selection and HMM-based speech synthesis

Parametric speech synthesis, usually HMM-based, presents multiple advantages compared to unit-selection synthesis (see Table 1.1). We briefly outline its major assets and drawbacks (for more details, see Chapter 5).

Table 1.1 - *Comparison between unit-selection and HMM-based speech synthesis.*

	Unit selection	HMM-based
Small footprint	X	✓
Very good quality	✓	(X)
Stability	X	✓
Flexibility	X	✓

HMM-based speech synthesis generates speech from acoustic models learned on large speech databases, which allows producing a quality voice without storing the database after the training stage. It requires therefore much **less storage space**, which facilitates its integration in embedded systems like smartphones.

However, because the signal has to be generated by a vocoder from its parametric representation, the overall quality of the speech is known to be a little lower than that of unit-selection synthesis, the signal being usually described as "**buzzy**". A wealth of research has focused on that issue, giving rise to improvements of the vocoder, such as STRAIGHT (Kawahara et al., 1999; Kawahara and Morise, 2011) or the Deterministic plus Stochastic Model (Drugman and Dutoit, 2012).

Unlike unit-selection synthesis, HMM-based speech synthesis is known to be more

stable. Because the generated acoustic parameters are averaged on a large set of similar units, annotation errors are smoothed out, especially if a large database is used for training. On the contrary, a single annotation error in unit-selection synthesis may ruin the naturalness of the synthesized sentence. Besides, unit-selection synthesis relies on a "hit or miss" strategy; if a phonetic or prosodic context is under-represented in the database, it may also significantly impair the resulting synthesis. In HMM-based synthesis, the clustering of acoustically-similar models alleviates that issue.

Finally, HMM-based synthesis presents the additional advantage of being more **flexible**. The characteristics of the synthesized speech can be changed by transforming the HMM parameters appropriately (*i.e.* cepstral coefficients, F0 and duration), with little signal degradation. The transformation of the parameters can rely on multiple techniques like adaptation, interpolation, eigenvoice or multiple regression (Zen et al., 2007a), requiring only a small amount of data. This allows for changing not only speaker individuality but also speaking style and emotional expression (Yoshimura et al., 1999). Unlike unit-selection synthesis, the generated speech does not necessarily resemble the speaking style recorded in the database (Zen et al., 2007a).

Overall, HMM-based speech synthesis presents the advantage of being highly flexible while requiring few storage space. This increased flexibility explains its use in expressive speech synthesis research today and justifies its choice in the framework of this thesis.

1.3 HMM-based expressive speech synthesis

The need for a more "natural" synthesized voice has quickly given rise to a wealth of studies on expressive speech synthesis. While many attempts to produce expressive speech have been carried out with formant (Murray and Arnott, 1995; Montero et al., 1998), unit-concatenation (Schröder, 2004; Burkhardt, 2005) or unit-selection TTS systems (Iida et al., 2000; Black, 2003), most current studies rely on HMM-based text-to-speech because of its increased flexibility.

A detailed state of the art of studies on HMM-based expressive speech synthesis is discussed in Chapter 5. We here provide a brief summary of the major trends and their drawbacks.

Besides the basic training on a corpus of a specific emotion called *style-dependent model* (Yamagishi et al., 2005), the methods can be classified into three main trends:

1. **The use of adaptation techniques:** A generic model is trained on neutral speech (Yamagishi et al., 2004) or an average model is trained on various emotions or speaking styles (Qin et al., 2006; Eyben et al., 2012) and that model is adapted to a specific type of expressivity.
2. **The modification of the tree-based clustering:** Additional contextual features, related to the various expressive styles, are integrated in the tree-based

clustering of the models (Yamagishi et al., 2004; Tsuzuki et al., 2004; Yamagishi et al., 2005; Eyben et al., 2012).

3. **Post-processing techniques:** F0 and duration values are directly modified (Hirose et al., 2004, 2005).

In the framework of this thesis, we address three major criticisms to existing studies:

CRITICISM 1 Most of the aforementioned studies (with the exception of Eyben et al. (2012)) use one corpus per basic emotion, read by a professional actor. This poses in our view two main issues:

- Basic emotions (*e.g.* happy, sad, angry) may be seen as a simplistic reduction of the expressive naturalness lacking in speech synthesizers. In line with Campbell (2006), we rather believe that synthesizers should include other types of expressivity such as various speaking styles, attitudes, or emphatic speech.
- Considering each (stereotypical) emotion in isolation neglects macro prosodic changes which occur in real speech, as we shift from one emotion to another, with various degrees of intensity.

CRITICISM 2 Expressive speech is often exclusively generated via adaptation or training on corpora with the targeted expressivity, without considering more local (or accentual) prosodic phenomena which are specific to expressive speech, such as emphatic stress.

CRITICISM 3 Most studies focus on modifications at the prosodic and voice quality levels only and clearly neglect potential phonetic modifications.

The motivations behind these criticisms and how current research has tried to address them are detailed in Chapter 5.

1.4 Contributions of the thesis

The current PhD research focuses on the **improvement of HMM-based expressive speech synthesis in French**. To that aim, we postulate that speech synthesis should benefit from the integration of (i) detailed linguistic analyses of the language (here French) and (ii) specificities of expressive and spontaneous speech. The field of natural language processing (NLP), in general, has largely evolved from expert-based to more data-based approaches. Recent research has been increasingly attempting to limit expert intervention by assuming that all required information is contained in the data itself. A notable example of this new trend is Simple⁴All (Watts et al., 2013), a large European project initiated in 2011 which proposes a new framework to train HMM-based synthetic voices, with little or no expert supervision, based on machine learning techniques only. In this thesis, we try to strike a balance, as we believe that statistical methods can benefit from linguistic knowledge, while

GENERAL INTRODUCTION

remaining fully automatic.

To better understand in which components of the synthesizer we will explore certain modifications, we present in Figure 1.2 the general workflow of an HMM-based speech synthesizer. Two major stages can be distinguished: the training and the synthesis. **At training**, a large speech corpus is used to train the models. The front-end, or Natural Language Processing (NLP) component automatically transforms speech and its orthographic transcription into a non-ambiguous representation, including (i) the phonetic transcription, (ii) a description of the linguistic context (which will serve for the generation of prosodic patterns) and (iii) their time-alignment with speech. Contextual features typically include the segmentation of the transcription in syllables, words and rhythmic groups, with associated information like parts of speech (POS) and stress status. Based on the output of the front-end, the back-end or Digital Signal Processor (DSP) trains context-dependent HMM-models of each phoneme. **At synthesis**, only text is provided to the front-end. A similar linguistic analysis is carried out, except for the alignment which does not need to be computed as no sound is provided. From this explicit description, the trained HMM-models are used to generate speech parameters which will be transformed into a speech signal by the back-end component.

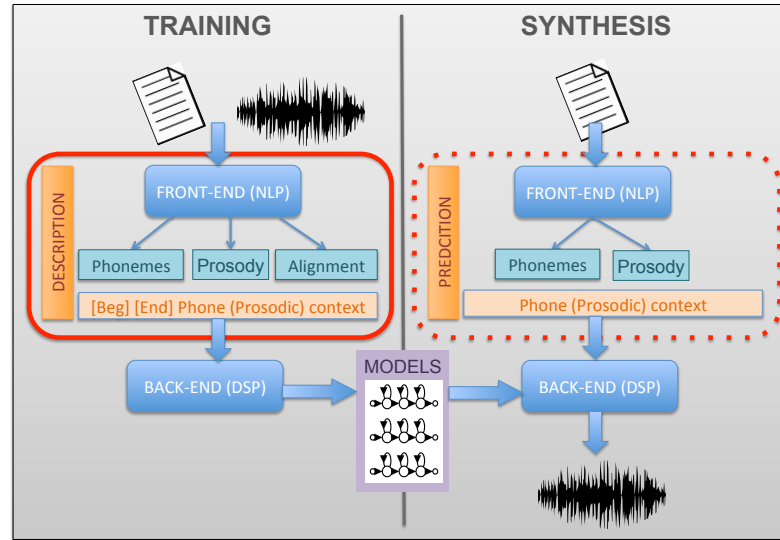


Figure 1.2 - General structure of a HMM-based speech synthesizer. Red frames indicate the components that are modified in the framework of the current research.

The current thesis mainly focuses on modifications of the front-end at training by dealing with:

- The phonetic annotation of the training corpus
- The prosody annotation of the training corpus
- Their time-alignment with the corresponding speech signal

While our evaluation is first based on a manual annotation of the training corpus, the current research also targets the automatization of the phonetic and prosodic annotation of the training corpus.

In addition, other stages of the synthesizers are considered in our research:

- We investigate how these modifications of the annotation of the training corpus can be integrated in the back-end at training.
- As this annotation (with the exception of the alignment) will also need to be generated at synthesis, our annotation protocol is developed to facilitate the prediction of the labels for any new text to synthesize. We provide ways to achieve annotation prediction at synthesis and present preliminary results albeit this does not constitute the core of our study.

To address the aforementioned criticisms, several strategies are proposed:

CRITICISM 1 To avoid the use of stereotypical acted emotions (typically one corpus per emotion), the core of our research focuses on the synthesis of live sports commentaries. To that end, our corpus presents various advantages compared to corpora used by state-of-the-art studies:

- It is spontaneous⁹.
- It contains ‘natural’ expressivity, as the speaker was not prompted which emotion or attitude to produce (Schuller et al., 2011)¹⁰. Various types of expressivity (*e.g.* emotions, emphatic speech) follow each other, depending on the state of the speaker, thereby providing an ideal framework for the study and synthesis of macro-prosodic changes.
- The corpus was recorded in high quality conditions, which allows for its use in the training of an HMM-based synthesizer.

In a second stage, our observations are confirmed on other corpora *i.e.* *audiobooks* and *conversational speech*. While audiobooks are read speech, they contain a wide range of expressivity variations and are therefore

⁹Spontaneous is here opposed to read speech. Our corpus however does not fit in Rillard’s spontaneous category (2014) which he defines as containing productions that are not constrained by speech laboratory settings.

¹⁰It should however be noted that, while sports commentaries are expressive, they may be seen as much more expressive than ordinary non-professional conversational speech.

GENERAL INTRODUCTION

suited for application of prosody modifications. Conversational speech, on the contrary, displays the major characteristics of spontaneous speech in which pronunciation changes, in particular, are known to play a key role. We believe that both these corpora contain interesting features which should be modeled in view of more natural synthesis

The use of such corpora also poses the question of their automatic alignment with the speech signal. We indeed believe that their high degree of variability may impair their alignment with generic speaker-independent models of the language, as provided by most available alignment tools.

CRITICISM 2 To deal with accentual prosodic phenomena, we develop a new prosody annotation protocol, considering phonological aspects of French and specificities of expressive speech. The objective is to propose a simple annotation, easy to integrate in speech synthesis, thereby answering Wightman's criticism (2002) who describes prosody in speech synthesis as a "messy quagmire". This protocol is integrated in speech synthesis and evaluated through perception tests.

CRITICISM 3 The role played by phonetic modifications in the naturalness of spontaneous expressive speech synthesis is evaluated through perception tests. Preliminary studies are also carried out to facilitate the prediction of post-phonetic modifications at synthesis.

Figure 1.3 proposes a graphical overview of the thesis, indicating the chapters dealing with each level of annotation (at training and/or synthesis) and providing further information about the corpora exploited for evaluation.

PART			PART I			PART II					PART III	
Chapter			C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
TRAINING	Annotation level	Phonetics				State-of-the-art and introduction PART II				✓		✓
		Prosody/Voice quality					✓	✓	✓		✓	
		Alignment	✓	✓	✓							✓
	Automatizes the annotation		✓	✓	✓				✓		✓	✓
SYNTHESIS	Proposes avenues on annotation prediction from text		NA	NA	NA				✓		✓	✓
STYLE OF SYNTHESIS	Spontaneous		✓	✓	✓		✓	✓	✓	✓		✓
	Expressive		✓	✓	✓		✓	✓	✓	✓	✓	(✓)
	Description of the training corpora		15 corpora varying in speaking style and language				Sports commentaries in French			Audiobook in French		Conversational speech in French and English

Figure 1.3 - Table of the structure of the thesis, indicating the levels of annotation considered (at training and/or synthesis) and the type corpus used for the evaluation. NA = not applicable.

In relation to the three aforementioned annotation levels (*i.e.* phonetic annotation,

prosody annotation¹¹ and their alignment with speech), three major hypotheses are formulated. These hypotheses and related research questions (RQ) are listed below, with an indication of the chapters (C) dealing with them.

HYPOTHESIS 1. The automatic alignment of (spontaneous) expressive corpora by generic HMM models of the language poses alignment problems and could be automatically improved

RQ 1.	Does training the models directly on the corpus to align improve the results ?	C1
RQ 2.	From the analysis of the typical alignment errors, is it possible to develop automatic improvement methods ?	C2
RQ 3.	Does the quality of the alignment significantly impact the resulting synthesis ?	C10

HYPOTHESIS 2. The integration of a prosody annotation considering phonological aspects of French and specificities of expressive speech should enhance HMM-based expressive speech synthesis in French

RQ 4.	How many annotation levels should be defined, associated with which level of linguistic segmentation (<i>e.g.</i> phoneme, syllable, word, rhythmic group) ?	C5, C6
RQ 5.	How should this annotation be integrated in the DSP ?	C5
RQ 6.	Is it possible to automatize the prosody annotation of the training corpus ?	C7, C9
RQ 7.	Would that annotation be easily predictable for new sentences to synthesize ?	C7, C9
RQ 8.	Would this protocol be generalizable to different types of expressive speech ?	C9

HYPOTHESIS 3. The consideration of phonetic modification phenomena in the annotation of the training corpus and for new sentences to synthesize should improve the naturalness of HMM-based spontaneous expressive speech synthesis

¹¹It should be noted that integrating prosodic and phonetic modifications indirectly adapts voice quality features too, as voice quality features, F0 and duration are modeled in a unified framework in HMM-based speech synthesis.

RQ 9.	Can specific phonotactic phenomena be associated with expressive and spontaneous speech ?	C8, C10
RQ. 10	Which modifications most improve the resulting expressive speech synthesis ?	C8
RQ 11.	Is the correct phonetization of the training corpus still important if the same phonetization errors are made at synthesis ?	C10

The main objectives of the current research are :

- To test the aforementioned hypotheses and answer the related research questions. For each practical chapter, an evaluation of the results is provided, either through objective comparisons with manual annotation (for speech alignment) or through subjective perceptual assessment of the resulting synthesis (for the integration of phonetic and prosody annotation).
- To develop tools allowing for HMM-based expressive speech synthesis in French. These tools are made available online to facilitate research on the topic, and consist in a language-independent speech aligner (Train&Align) and a front-end for HMM-based expressive and spontaneous speech synthesis in French (eLite-HTS). These tools are detailed in Chapters 4 and 5, respectively.

1.5 Structure of the thesis

The thesis is structured as follows:

- **Part I** focuses on the first hypothesis, *i.e.* the automatic phonetic alignment of expressive corpora.

We show that existing tools, which provide the user with pre-trained speaker independent models of the language, perform poorly on (spontaneous) expressive corpora. We also point at their major drawback *i.e.* the fact that they offer a limited number of models, and that some corpora in rare languages cannot be aligned because no corresponding model is provided. We point out that both issues may be alleviated by training new models directly on the corpus to align. This technique is developed and evaluated in **Chapter 2**, relying on a database of 15 corpora varying in size, language and speaking style. An in-depth analysis of the typical alignment errors made by this algorithm is also carried out.

Based on this analysis, we propose in **Chapter 3** three improvement methods that are fully automatic and do not require the use of manually-annotated data. A widely-known technique, *i.e.* the augmentation of the feature vectors, is explored. Two innovative methods are also developed: the use of a voice activity detection (VAD) algorithm to better initialize the silence models and the computation of average boundaries based on the alignment of the reversed sound.

Chapter 4 closes this first part by describing a new alignment tool we have developed, based on the technique detailed in Chapter 2: Train&Align. That tool is available online and is now used by several research centers at the international scale.

- **Part II** investigates the second and third hypotheses, *i.e.* the potential improvement of prosodic and phonetic generation in HMM-based spontaneous expressive speech synthesis.

An introduction to expressive speech synthesis is provided in **Chapter 5**. It offers a theoretical background on phonetics, prosodic models, and their integration in speech synthesis. A detailed state of the art of expressive HMM-based expressive speech synthesis is also provided. In addition, the chapter presents two main resources that were partly developed in the framework of the thesis and will be used throughout Part II (and III): a front-end for French and a corpus of sports commentaries (*Sportic*).

Regarding prosody, a two-level annotation protocol is proposed in **Chapter 6**. A local tier relates to accentual phenomena and is linked to the syllable level. A global prosody annotation classifies speech segments into distinct sub-genres, or speaking styles, defined in terms of valence and arousal. That annotation is integrated in the framework of speech synthesis of sports commentaries. The chapter offers an analysis of the annotation protocol and the evaluation of the resulting synthesis.

The accentual prosody annotation is further investigated in **Chapter 7**. Specifically, it is explored whether the various acoustic realizations of emphatic stress should be explicitly distinguished in the annotation or whether the annotation of emphasis with a single label does not impair the resulting synthesis.

Chapter 8 provides preliminary studies on how to predict the two-tier prosody annotation from sound and text (at training stage) or from text only (at synthesis stage). As manually annotating is a time-consuming task, several automatic approaches are investigated. The main objective of this chapter lies in the automatic annotation of the training corpus with global prosodic labels, based on classification methods. The possibility of exploiting that automatic annotation to correct the manual annotation is also explored.

Finally, **Chapter 9** concentrates on phonetic issues and proposes a detailed analysis of phonetic changes influenced by 3 binary ‘traits’ characterizing the situation of communication, *i.e.* spontaneous/read, expressive/neutral and media/non media. It also assesses the role played by phonetic variation in the naturalness of the resulting synthesized voice. The relative impact exerted by the different phonotactic phenomena on spontaneous expressive speech synthesis is further explored.

- **Part III** analyses the possible generalization of the observations made in Part II to other corpora containing other types of expressive data (*e.g.* audiobooks and conversational speech) and other languages. This part benefits from various collaborations with international experts we have met during the PhD.

Chapter 10 presents a method to automatically annotate local prosodic features on any corpus in French. Based on an autosegmental approach of prosody, we propose to combine the output of existing automatic annotators (*i.e.* prominence detector, tonal annotator and synthesis front-end) to produce a high quality annotation of accentual patterns in French. The integration of this annotation in HMM-based speech synthesis of audiobooks is assessed through perception tests.

Chapter 11 focuses on phonetic changes in the framework of spontaneous speech synthesis. The first part of the chapter proposes an in-depth analysis of the impact exerted by various linguistic and extralinguistic factors on phonetic variation in French. This should provide a preliminary theoretical background for the design of post-phonetic rewrite rules for spontaneous speech synthesis. The second part relies on a parallel corpus (read/spontaneous) in English. It investigates the widely accepted consistency assumption, which states that phonetization errors in the training corpus should not impair the resulting synthesis if the same errors are made at synthesis. This assumption is questioned in the framework of spontaneous speech synthesis, in which we show that large discrepancies are found between the automatic phonetization of the corpus and its real pronunciation.

Finally, **Chapter 12** concludes the thesis and presents some perspectives.

Part I

Automatic phonetic alignment of expressive speech corpora

This first part of the thesis deals with the automatic phonetic alignment of speech sounds with their corresponding phonetic transcription. As pointed out in the introduction, this constitutes one of the major outputs of the synthesis front-end at training stage, when annotating the training corpus. The most commonly used alignment procedure relies on the modeling of speech with hidden Markov models. It is referred to as HMM-based forced alignment. Most available alignment tools provide the user with HMM-based speaker-independent models of the language, trained on large multi-speaker databases. These models are supposed to offer an overall representation of the language that is not specific to one speaker or speaking style and can be used to align any new corpus. In the framework of our thesis, we question the performance of such generic models on highly variable speech, be it expressive and/or spontaneous.

In **Chapter 2**, we compare this technique with a widely-used method in which the models are directly trained on the corpus to align, to improve the agreement between training and alignment stages. This technique offers the additional advantage of being applicable to any language and speaking style without the need for a large or manually-annotated database. Through a systematic step-by-step study, the role played by various training parameters (*e.g.* models configuration, size and overlap of the speech frames and number of training iterations) on the alignment accuracy is assessed on a large database of 15 corpora varying in size, language (some being under-resourced) and speaking style.

Based on an in-depth analysis of the typical errors made by this technique, we develop in **Chapter 3** three fully-automatic improvement methods which do not require the use of manually-aligned data. Besides the use of supplementary acoustic features, we propose two innovative techniques: (i) an initialization of the silence model based on a Voice Activity Detection (VAD) algorithm and (ii) a consideration of the forced alignment of the time-reversed sound. These methods are developed and evaluated on our database.

Finally, we present in **Chapter 4** an alignment tool that we have developed, called Train&Align, which is based on the method described in **Chapter 2**. It makes it possible to align any new corpus in any language, provided that the sound and its phonetic transcription are provided. An illustrated presentation of the tool is given, along with a detailed description of its functionalities. Train&Align is available

online and should be useful to the research community. It is currently used by several research laboratories at the international scale.

Chapter 2

Training on the corpus to align : Train&Align

Contents

2.1	Introduction and context	24
2.2	HMM-based phonetic alignment: State of the art	27
2.3	Development protocol	31
2.3.1	Development database	31
2.3.2	Evaluation metrics	31
2.4	Description of our method: Train&Align	32
2.5	Evaluation	39
2.5.1	Speech material	39
2.5.2	Results with Train&Align	40
2.5.3	Comparison with state-of-the-art models	41
2.5.4	Influence of the corpus size	42
2.5.5	Effect of bootstrap	43
2.6	Analysis of errors made by <i>Train&Align</i>	45
2.7	Conclusion	47

Abstract

Speech segmentation refers to the problem of determining the phoneme boundaries from an acoustic recording of an utterance together with its orthographic transcription. Several automatic phonetic alignment tools have been proposed in the literature. They usually rely on pre-trained speaker-independent HMM models to align new corpora. Their drawback is that they cover a very limited number of languages and might not perform properly for different speaking styles (*e.g.* spontaneous or expressive). This chapter focuses on a particular case of Hidden Markov Model (HMM)-based forced alignment in which the models are directly trained on the corpus to align. The obvious advantage of this technique is that it is applicable to any language or speaking style and does not require manually-aligned data. Through a systematic step-by-step study, the role played by various training parameters (*e.g.* models configuration, number of training iterations) on the alignment accuracy is assessed. The evaluation is carried out on 12 corpora of different sizes, languages (some being under-resourced) and speaking styles. Comparisons indicate that this technique provides as good or better results than using speaker-independent models of the language. The study also points out that even a small neutral corpus of a few minutes can be exploited to train a model that will provide high-quality alignment. Yet, the alignment rates obtained for some expressive corpora or corpora in under-resourced languages are rather low. To alleviate such issues, a manually-aligned part of the corpus can also be used as bootstrap to improve the model quality. A significant improvement of the alignment performance was found when using only 10 seconds of bootstrapping data, especially for the expressive corpus in which an increase of about 30 % (with a 20 ms tolerance threshold) was achieved.

This chapter is based upon the following publications:

- Sandrine Brogniaux, Sophie Roekhaut, Thomas Drugman & Richard Beaufort. (2012). Automatic Phone Alignment. A Comparison between Speaker-Independent Models and Models Trained on the Corpus to Align. *Lecture Notes in Computer Science*, 7614, 300-311.
- Sandrine Brogniaux, Sophie Roekhaut, Thomas Drugman & Richard Beaufort. (2012). Train&Align: A New Online Tool for Automatic Phonetic Alignment. *IEEE Spoken Language Technology Workshop (SLT)*, Miami (USA).
- Sandrine Brogniaux, Sophie Roekhaut, Thomas Drugman & Richard Beaufort (2014). Train&Align: Un outil d'alignement phonétique automatique disponible en ligne. *Journées d'étude de la parole (JEP)*, Le Mans (France).

- Sandrine Brognaux & Thomas Drugman. (2015). HMM-based Speech Segmentation: Improvement of Fully Automatic Approaches, Accepted for publication in the *IEEE Transactions on Audio, Speech and Language Processing*.

Many thanks to Prof. C. Wellekens for his insightful advice and J.-P. Goldman and B. Bigi for their help and enthusiasm regarding this study. We are also very thankful to Acapela Group, Trinity College Dublin and Prof. van Niekerk who kindly offered their corpora for this study.

2.1 Introduction and context

Large speech corpora play a major role both in linguistic research and speech technologies. A peculiarity of the use of such data is that the sound is rarely studied alone (Dister and Simon, 2008). Orthographic and phonetic transcriptions of the audio files are usually required. Phonemes, in particular, should be time-aligned with the sound. A precise correspondence between the sound and specific speech segments is essential to allow for prosodic or phonetic analyses. The alignment precision of corpora used to train speech synthesizers or recognizers also determines the quality of the resulting systems (Hunt and Black, 1996; Zen et al., 2009; Jelinek, 1997).

Many alignment tools offer the possibility to define various transcription levels (*e.g.* WaveSurfer (Sjölander and Beskow, 2000), Praat (Boersma, 2001) and ELan (Wittenburg et al., 2006)). These can then be manually aligned with the sound (see Figure 2.1). However, such a task exhibits two major drawbacks. First it is time-consuming, requiring 130 (Kawai and Toda, 2004) to 800 (Schiel et al., 1998) times the sound duration. For corpora of several hours as used in speech technologies, the resulting manual annotation time would become prohibitive, which is economically impracticable. A second issue lies in the consistency of the alignment, especially if several annotators work on a same corpus. Even with trained phoneticians, high consistency is rarely achieved when several annotators collaborate (Ljolje et al., 1997).

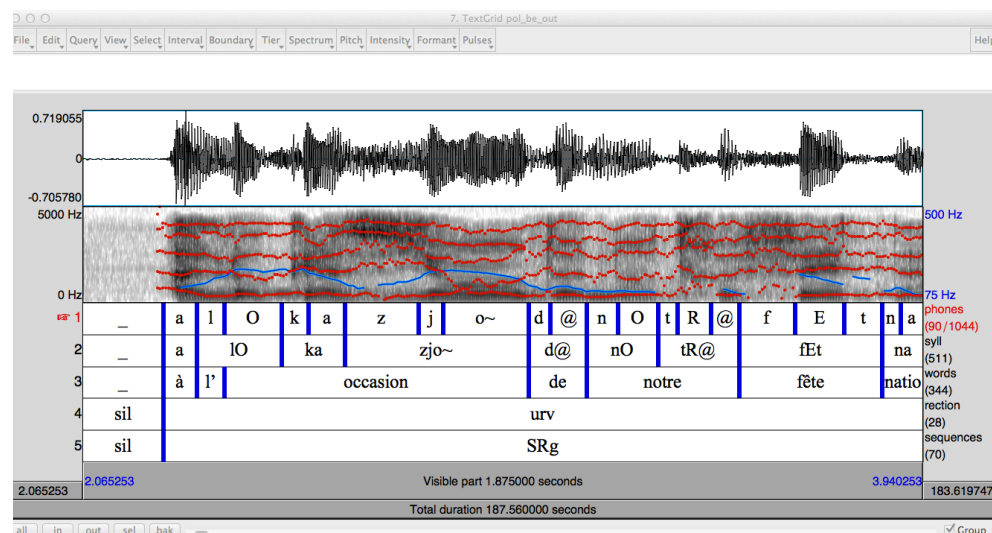


Figure 2.1 - Example of aligned sound transcription in Praat.

To alleviate these issues, automatic alignment tools have been developed (*e.g.* The Aligner (Wightman and Talkin, 1997), nAlign (Sjölander, 2003), P2FA (Yuan and Liberman, 2008), SPRAAK (Demuyne et al., 2008), IrcamAlign (Lanchantin

et al., 2008), SailAlign¹ (Katsamanis et al., 2011), EasyAlign (Goldman, 2011) and SPPAS (Bigi and Hirst, 2012))². They offer a consistent and reproducible alignment at reduced cost. The task they perform is known as ‘linguistically constrained segmentation’ or ‘forced alignment’. The phonetization of the text (possibly with phonetic variants, see Bigi and Hirst (2012)) is supposed to be known and only the time boundaries of the phonemes have to be determined. For this purpose, acoustic modeling based on Hidden Markov Models (HMMs), relying on speech recognition techniques, has been shown to achieve the best results (Adell et al., 2005; van Niekerk and Barnard, 2009).

Most existing alignment tools provide the user with pre-trained speaker-independent HMMs of several languages, trained on large multi-speaker databases. These models account for an overall realization of the language which is not specific to one speaker or speaking style. The set of provided models being limited, only a reduced set of languages can be aligned (see Table 2.1). Furthermore, the models highly depend upon the training corpus. As shown in the next sections of this chapter, these models may produce low-quality alignment of expressive or spontaneous corpora when they are trained on neutral read speech. Some phonemes may also be improperly aligned if they are under-represented in the training corpus (as for semi-vowels in Goldman and Schwab (2011)).

Table 2.1 - *Languages covered by various existing tools for automatic phonetic alignment.*

Tool	Language
EasyAlign	French, Spanish, Portuguese, Taiwan Min
SPPAS	French, English, Italian, Chinese
P2FA	American English
SailAlign	American English
SPRAAK	Northern and Southern Dutch
IrcamAlign	French and English
nAlign	Swedish
The Aligner	American English

A possible way to solve these issues is to train the models directly on the corpus to align, thereby providing a better agreement between the training and alignment stages. This technique exhibits the advantage of applying to any language or speaking style without the need for manually-aligned data. Besides, training the models on the speaker to align was proven to be highly profitable in speech recognition (Leggetter and Woodland, 1995). It also addresses the widespread criticism stating that HMM-based alignment tends to be language specific and requires a

¹SailAlign (Katsamanis et al., 2011) proposes a particular technique relying on speech recognition to alleviate issues due to imprecise transcriptions. It also proposes to adapt the model on the corpus to align which reduces the mismatch between training and alignment data.

²HTK (Young et al., 2000) and Kaldi (Povey et al., 2011) are speech recognition toolkits which also provide methods to perform automatic phonetic alignment

high amount of training data (Van Santen and Sproat, 1999; Scharenborg et al., 2010).

The aim of this chapter is to investigate the quality of such alignments. This is done in comparison with the use of available speaker-independent models provided by recent alignment tools. It focuses on the evaluation of the alignment when modifying various training settings: context-dependent models, number of training iterations, size and overlap of the speech frames and corpus size. Finally, it proposes a detailed analysis of typical errors found in the automatic alignment that will serve as a basis to develop refinements of the conventional HMM-based alignment in the next chapter. Our study offers the advantage to be based on a large database, made of 15 corpora in French, English and rather under-resourced languages such as Faroe or Gaelic. Our database also contains a wide range of speaking styles (*e.g.* read vs. spontaneous speech, neutral vs. expressive speech).

It should be noted that our study does not address two well-known challenges of speech alignment, *i.e.* the processing of long speech files and the use of imperfect sound transcriptions. The evaluation is here performed on clean speech, with manually-verified phonetic transcriptions, as used for speech synthesis purposes.

Regarding the use of long audio segments, various solutions have been brought up by Goldman (2011); Moreno et al. (1998) and Prahallad et al. (2007). Goldman (2011) proposes to segment utterances in a preliminary step based on the punctuation marks in the transcription and a pause detection tool. Moreno et al. (1998) presents an iterative method that makes use of speech recognition to map text sequences with speech segments. This method was integrated in SailAlign (Katsamanis et al., 2011). Prahallad et al. (2007) propose modifications of the standard forced alignment to work on pre-segmented speech data blocks of 30 seconds and sequences of maximum 125 words.

Several studies have also focused on the alignment of speech corpora with imprecise transcriptions (*e.g.* Braunschweiler et al. (2010); Moreno and Alberti (2009)). In Moreno and Alberti (2009) it is proposed to make use of a factor automaton to represent all substrings from a string. It is trained on the transcripts and encodes a highly constrained language model. This model allows for modeling of word deletions and substitutions. Braunschweiler et al. (2010) propose to simply discard sentences for which the recognized transcription differs from the text book.

More recently, Stan et al. (2016) also proposed to tackle these two issues with the aim of automatically producing corpus alignment of any language without the need for expert knowledge. They offer a good overview of the state of the art. They however focused on sentence alignment only.

The chapter is organized as follows. Section 2.2 proposes an overview of the state of the art of existing alignment techniques. Our development protocol is presented in Section 2.3. Section 2.4 provides a detailed description of our method based on training on the target corpus to align. Results of our experiments are then shown in Section 2.5, providing an assessment of the proposed approach as well as a comparative evaluation with state of the art techniques based on speaker-independent models. Typical alignment errors made by this automatic technique are detailed in Section

2.6. Finally, Section 2.7 concludes and discusses further work.

2.2 HMM-based phonetic alignment: State of the art

Several techniques have been proposed to automatically provide the segmentation of speech files. Among these, we can essentially distinguish between methods based on Dynamic Time Warping (DTW) algorithms (*e.g.* Wagner (1981); Malfrère and Dutoit (1997)) and those using HMMs, alone (*e.g.* Brugnara et al. (1993); Sjölander (2003); Toledano and Gómez (2002)) or in combination with artificial neural networks (ANN) (*e.g.* Hosom (2002); Deroo et al. (2003)). DTW-based alignment, presented by Malfrère and Dutoit (1997), relies on speech synthesis techniques. It synthesizes the speech from the phonetic transcription. Subsequently, it attempts to align the generated speech signal with the original sound. This allows one to determine the position of the phoneme boundaries.

Conversely, HMM-based approaches make use of speech recognition paradigms. Additional methods also include the use of acoustic rate of change or discontinuity of the signal (Brandt, 1983; Scharenborg et al., 2010). These latter techniques are said to be unsupervised as they rely on the sound only and do not require any phonetic transcription. They automatically identify potential phoneme boundaries based on acoustic features. Their main drawback is that they usually either over or under-detect the number of phonemes, which makes it hard, in a second stage, to map the segments to linguistic units.

While DTW algorithms were shown to provide acceptable results (Malfrère and Dutoit, 1997), HMM-based acoustic modeling has been pointed out as being the most reliable technique for automatic phonetic alignment (Adell et al., 2005; van Niekerk and Barnard, 2009; Paulo and Oliveira, 2004). It is currently the most widely-used technique for forced alignment. The interested reader is referred to Hosom (2000) for an insightful review on automatic speech segmentation.

HMM-based forced alignment uses methods derived from speech recognition. It is ‘linguistically constrained’, which means that the transcription of the sound files is required as input. This transcription may contain phonetic variants, the most likely being selected during the alignment stage (Wightman and Talkin, 1997; Bigi and Hirst, 2012). The global working of HMM-based phonetic alignment can be subdivided in two stages: training and alignment (see Figure 2.2).

In both stages, the sound is first parametrized. Feature vectors are extracted on frames of the signal for which a size and a target rate are defined (see Figure 2.3). A finite number of values are extracted for each small slice of the signal (generally around 10 ms). These parameters, typically Mel Frequency Cepstral Coefficients (MFCC), are representative of the configuration of the vocal tract when the sound was produced. They should provide a most accurate representation of the signal.

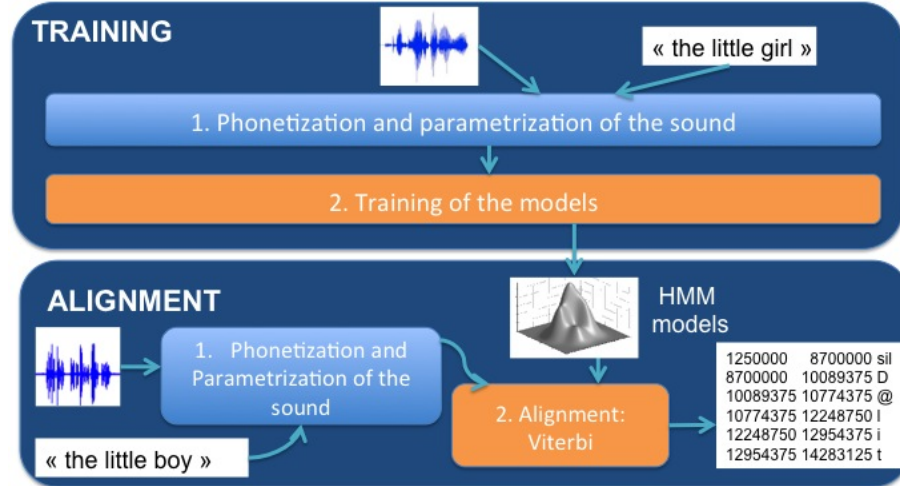


Figure 2.2 - Schematic representation of HMM-based forced alignment.

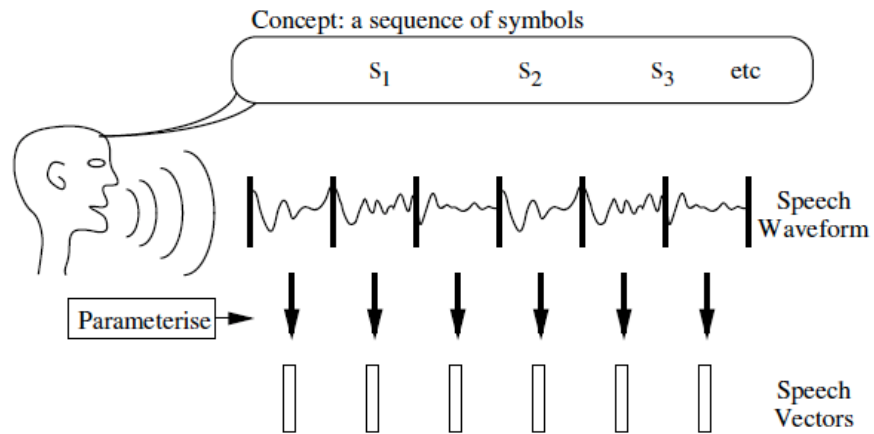


Figure 2.3 - Message encoding (Young et al., 2000).

The training stage allows for the building of context-(in)dependent³ HMMs of each phoneme. Two different initialization stages are proposed, depending upon the provided phonetic transcription.

If some part or the whole transcription is time-aligned (referred to as *bootstrap*), the manually-aligned corpus is exploited to improve the initialization stage. An iterative procedure determines the value of the parameters. In a first stage, the training data is uniformly segmented. Each model is matched with the corresponding data segments and the means and variances are estimated. In the second and successive stages, this segmentation is replaced by a Viterbi alignment (Forney, 1973). The ini-

³Through the use of monophones, triphones or tied-state triphones as detailed below.

tial parameter values computed by Viterbi are then further re-estimated by a Baum-Welch procedure (Baum et al., 1970). Baum-Welch makes use of the well-known expectation-maximization (EM) algorithm to find the parameter values, based on a max-likelihood criterion.

Conversely, if no manually-aligned part of the corpus is available, the phonetic transcription is first uniformly aligned with the sound, with a so-called ‘flat-start initialization’. In that case, all states of all models are initialized as equal (see Figure 2.4).

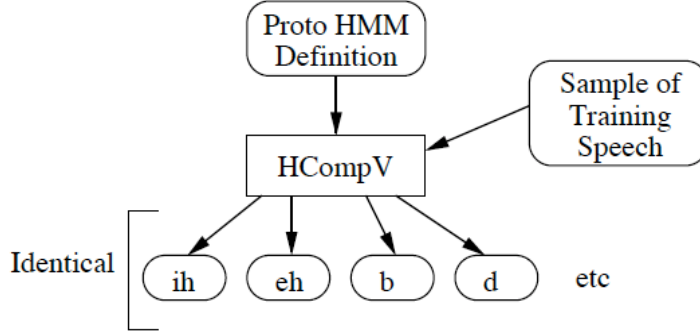


Figure 2.4 - *HCompV* command for flat start initialization (Young et al., 2000).

In the alignment stage, the models are used to align a corpus (possibly identical to the training corpus) with the Viterbi algorithm, which provides the best path among the network of possible transitions (see Figure 2.5). For a detailed description of training and alignment formulae, see Young et al. (2000). The Hidden Markov Model Toolkit (HTK) (Young et al., 2000) provides an implementation of HMM and methods for speech recognition. Most existing alignment tools and studies (Goldman, 2011; Yuan and Liberman, 2008; Bigi and Hirst, 2012; Kawai and Toda, 2004; Cangemi et al., 2011) are based on this toolkit or similar toolkits like Julius (Lee et al., 2001). Our study also relies on HTK.

Most studies in the literature have investigated the alignment obtained when training the models on a large database and using these language-dependent models to align other corpora (Chen et al., 2004b; Wightman and Talkin, 1997; Goldman, 2011; Bigi and Hirst, 2012; Cangemi et al., 2011). Depending upon the study, the training stage relies on aligned or non-aligned data. Chen et al. (2004b) offers an insightful comparison of the alignment rates when using various available speaker-independent models. The performance of speaker-dependent models was analyzed in Kawai and Toda (2004) and Charonnat et al. (2008). In both studies, however, the model is trained on *aligned data* of one speaker and used to align some *other part* of a corpus of the same speaker. Similar tests are performed by Sjölander (2001) who trains the models on some aligned-part of the multi-speaker Waxholm test set (Bertenstam et al., 1995) and tests on another part of the corpus. This improves the

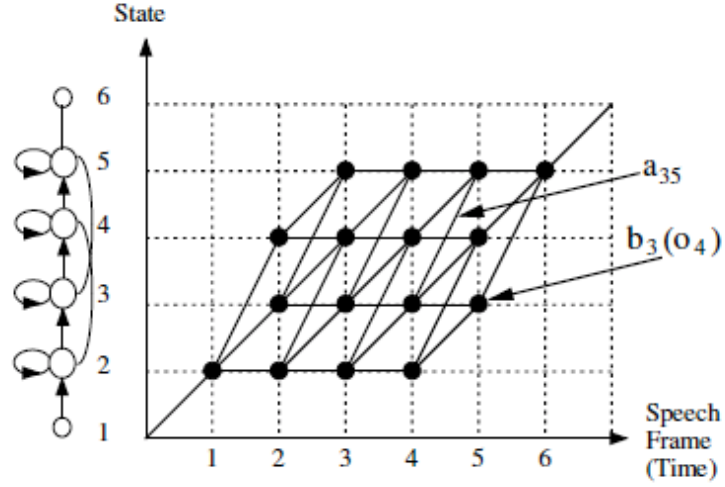


Figure 2.5 - Graphical representation of the Viterbi algorithm (Young et al., 2000).

quality of the alignment because of a better agreement between the model and the corpus to align. However, it requires some part of the corpus to be manually-aligned. This is time-consuming, and hence costly.

While training directly on the corpus to align alleviates the need for large transcribed corpora of the language, the performance of this method has been the topic of very few studies (see van Niekirk and Barnard (2009); Ljolje et al. (1997)). In van Niekirk and Barnard (2009), the use of this method for under-resourced languages was investigated. However, it was not compared with the results obtained when aligning the same corpus with existing speaker-independent models. This method was also shown to provide interesting results for English corpora (Ljolje et al., 1997) but no evaluation was made in that study of the obvious role played by the size of the corpus. Moreover, the alignment rates were not compared to results achieved when using available models of the language trained on large amounts of data.

This chapter helps to provide answers to these questions. Specifically, it investigates the performances of the widespread alignment procedure in which the models are directly trained on the corpus to align. Multiple parameters are considered in the development of our method which is then evaluated on a large database of 12 corpora varying in size, speaking style and language (some being under-resourced). The alignment rates are compared to those obtained with existing speaker-independent models of the language. The performances of the aligner are assessed on expressive and spontaneous speech, in particular, which display greater variation and constitute key corpora in the framework of our thesis.

Table 2.2 - *Development data set features.*

Corpus	Language	Speaking style	Duration (minutes)
Will (Creature)	English	Read/Expressive	13
Antoine (Sad)	French	Read/Expressive	12
Margaux (Happy)	French	Read/Expressive	12

2.3 Development protocol

Our study first proposes a development stage, which provides an in-depth description of our method and focuses on the tuning of its parameters. To that end, a development database made up of 3 corpora varying in language and speaking style is used. The developed method is then evaluated on a larger database in Section 2.5.

2.3.1 Development database

The proposed technique is developed on clean and read speech, as used for speech synthesis purposes. Our development set is presented in Table 2.2. It is made of 3 corpora, kindly provided by Acapela Group SA, and sampled at 22050 Hz. Each corpus has a total duration (silences included) of 12 or 13 minutes and contains 131 to 200 speech files. The aim is to try to cover various features of speech: different types of expressivity and speaking style (sad, happy and little creature, a monster-like voice), two languages (English and French) and three different speakers (of both genders). The phonetic transcriptions were checked and manually aligned with the sound. The French and English corpora are respectively annotated with 39 and 56 phonetic symbols.

2.3.2 Evaluation metrics

The evaluation of the methods is based on the comparison between the automatic and the manual segmentation. In this chapter, the results are presented relying on the boundary-based measure (Adell et al., 2005; Goldman and Schwab, 2011), also referred to as the “percent agreement” (Hosom, 2002), which is used in most studies. It computes the percentage of boundaries that are correct, with a certain tolerance threshold ranging from 10 to 40 ms. This measure will be referred to as the *correct alignment rate* in the remainder of the thesis.

We make use of an additional measure to assess the alignment quality: the relative improvement over the baseline Train&Align, as described in Section 2.4. The so-called *relative improvement* is here defined as the relative reduction of the alignment error rate at a certain tolerance threshold. Denoting $ER(x)$ the error rate made by method x , using a given tolerance threshold (varying from 10 to 40 ms), the second metric is called the relative improvement of method x over technique y and is defined as:

$$Rel\_Imp(x, y) = \frac{ER(x) - ER(y)}{ER(x)} \quad (2.1)$$

$Rel\_Imp(x, y)$ is then positive if y outperforms x , while negative values indicate a degradation of the alignment performance. This measure is also used by Hosom (2009).

It should be noted that several studies have pointed out the high degree of inter-rater disagreement, with sometimes large discrepancies between human-made alignments (Wesenick and Kipp, 1996). Generally, results are provided within a certain tolerance threshold on the timing error. Usually, 20 ms constitutes a limit above which the inter-annotator agreement rate is fairly high. Using this threshold, Goldman (2011) reported inter-annotator rates of about 81 % and 79 % for the alignment of a French and of an English corpus respectively. Rates between 88 % and 90 % were obtained on an Italian corpus in Cusi et al. (1991). On the TIMIT corpus, 93.5 % agreement were obtained (Hosom, 2009). In Pitt et al. (2005), the average distance between boundaries of human annotators was shown to be 16 ms. A value of 20 ms is also considered as an acceptable limit for speech synthesis purposes (Adell et al., 2005). Throughout our experiments, we provide the performance measures using thresholds of 20 and 40 ms (errors exceeding 40 ms can be seen as gross errors).

2.4 Description of our method: Train&Align

The experiments in this chapter make use of the standard HMM-based alignment as described in Section 2.2. The only special characteristic of our technique is that the models are directly trained on the corpus to align. This method, called Train&Align (T&A), was further developed into a tool freely available online⁴, presented in Chapter 4. Its specific implementation, along with the tuning of its parameters, is further described in the next paragraphs.

As input, the method only requires the speech signal and its phonetic transcription⁵; no manually-aligned part of the corpus is needed. In a first stage, the entire (unaligned) corpus to align and its phonetic transcription are used to train a new language model. The correct phonetization of the sound (as determined by a human annotator) is provided, with no variant. Acoustic vectors are made of 39 features, *i.e.* 12 Mel-Frequency Cepstral Coefficients (MFCCs) and the log-energy, along with their delta and acceleration coefficients. For delta coefficients, a context window of five frames is considered, as suggested in HTK (Young et al., 2000). The MFCCs are automatically computed by HTK from the Fast Fourier Transform (FFT)-based log spectra. The FFT makes use of a Hamming window and the signal has first order preemphasis applied using a coefficient of 0.97. For more information about feature

⁴http://cental.fltr.ucl.ac.be/train_and_align/

⁵Silences should ideally be annotated, but a short pause model allows for the detection of non-annotated silences (see below). Throughout this chapter and the next, the manually-checked phonetic transcription of the corpora, including silences, is provided as input.

extraction, see Young et al. (2000).

The models are left-to-right 5 state monophones with no skip and three emitting states (see Figure 2.6). Each state of the model is associated with a set of parameters. They offer an acoustic representation of the state of the corresponding phoneme. They also provide the probabilities to move from one state to another.

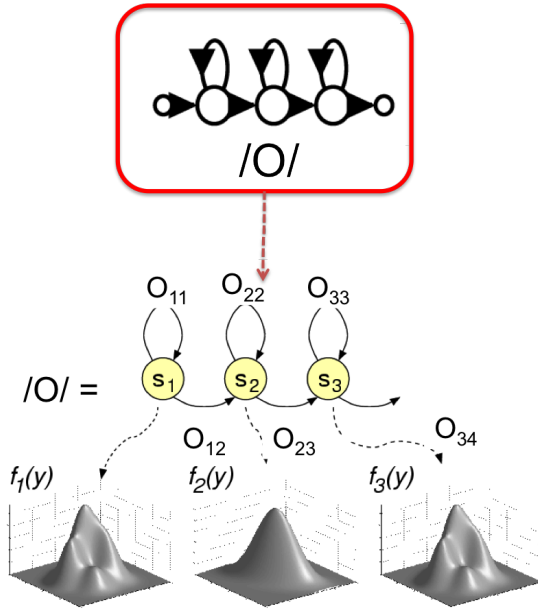


Figure 2.6 - Model corresponding to the phoneme /O/, inspired by (Young et al., 2000).

The models are initialized with a so-called ‘flat start’, *i.e.* a uniform segmentation of the speech signal. For silences, a standard widely-used configuration is applied (see Figure 2.7). Two specific models are added: a silence model (‘sil’) represents silent pauses and allows for a direct transition from second to fourth state and back to better model duration variations. Silences can be indicated in the phonetic transcription. A short-pause tee model (‘sp’) is also implemented and automatically inserted between words. It allows for automatic detection of non-annotated silences. It has one emitting state tied to the center state of the silence model. *T&A* computes a commonly-used two-pass training, which allows for an embedded feature re-estimation: three iterations of the Baum-Welch algorithm (Baum et al., 1970) are applied before the introduction of the ‘sp’ model and five iterations after, as proposed in Young et al. (2000). Training and alignment are performed with the HTK toolkit (Young et al., 2000).

Based on that standard alignment method, the tuning of several training param-

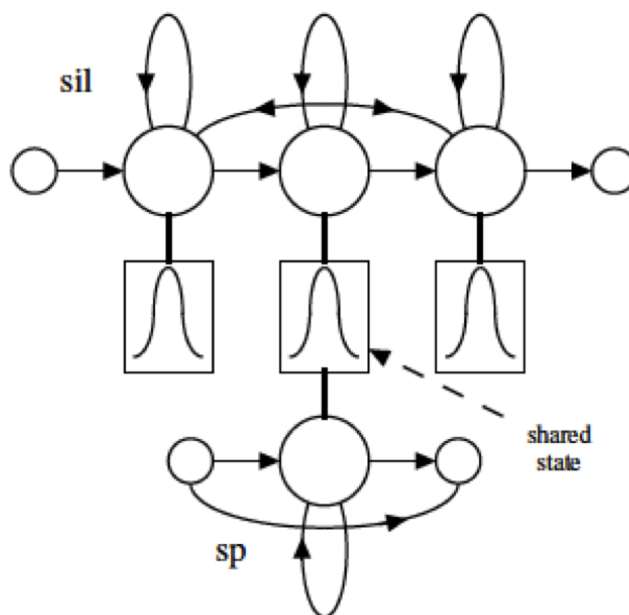


Figure 2.7 - Silence model and short-pause model (Young et al., 2000).

eters was investigated (See Table 2.3). Specifically, the size and overlap of the speech frames, the use of context-dependent models, and the number of training iterations were further analyzed.

Table 2.3 - Studied parameters which are tuned on our development corpus.

Parameter	Studied values
1. Size and overlap of speech frames	Window sizes from 10 to 30 ms with or without overlap
2. Use of context-dependent models	Monophones, triphones and tied-state triphones
3. Number of training iterations	From 2 to 40 iterations at first and/or second training stages

Regarding the **size of the speech frames** considered to extract the feature vectors, five options were investigated with overlapping and non-overlapping frames varying from 10 to 30 ms (see Table 2.4). When the target rates equal the frame size, there is no overlap. Our results show that frames of 10 ms with no overlap provide the best alignments, with an average absolute improvement of the boundary-based alignment rate of about 12 %, with a 20 ms threshold compared to frames of 30 ms

with a shift of 10 ms, as commonly used in automatic speech recognition (ASR). While this may look contradictory to ASR standards, it can be explained by the fact that narrower windows with no overlap avoid mixing data belonging to different phonemes. Conversely, ASR usually makes use of highly context-dependent models, mixing data helping in the recognition process as they are not concerned with the precise location of the phoneme boundaries.

Table 2.4 - *Alignment rates for monophone alignment with different sizes (S) and target rates (T) of the speech frames (in ms), with a 20 ms and 40 ms tolerance threshold, for the development database.*

20 ms					
Corpus	S:30 T:20	S:30 T:10	S:20 T:20	S:20 T:10	S:10 T:10
Will (Creature)	58.14	47.41	61.26	55.55	63.13
Antoine (Sad)	65.91	66.86	70.28	73.50	75.40
Margaux (Happy)	61.39	64.15	70.90	78.23	83.71
40 ms					
Corpus	S:30 T:20	S:30 T:10	S:20 T:20	S:20 T:10	S:10 T:10
Will (Creature)	81.35	77.81	82.96	81.71	84.27
Antoine (Sad)	86.31	85.44	86.74	85.79	86.71
Margaux (Happy)	88.50	92.82	90.31	94.30	95.08

The use of context-dependent models was also investigated.

Usually, each phoneme is represented by one model, called a monophone (Table 2.5 (1)). Three to five states represent the different stages of the realization of the phoneme: the transition and the stabilization phases.

However, the models can also be associated with phonemes in context, regarding the phonemes on the left and on the right (Table 2.5 (2)). They are called triphones and allow modeling the coarticulation phase. Their use, however, can be problematic: a lot of data is required to allow for a good representation of each triphone. Besides, the augmentation in the number of models also results in an increase in processing time.

A solution may be found in the use of tied-state triphones: the phonetic context of each phone is no longer modeled in terms of phonemes but in terms of classes (Table 2.5 (3)). These classes are articulatory characteristics, as listed in Table 2.6. It should be noted that the possibility of considering the phonetic context is an advantage of our method. Indeed, the use of pre-existing speaker-independent models makes it harder to use triphones. In pre-existing models, all the triphones of the language should be present. If the corpus to align contains new triphones, the alignment process fails. However, the phonetic context coverage of the training corpus usually differs from the coverage of the target corpus, even if the training corpus is rather large. Obviously, this problem does not arise when the model is trained on the corpus to align. For pre-existing models, a particularly large

corpus would be required to model every triphone. This issue has been pointed out by Bigi and Hirst (2012). In that case, a clustering procedure may be used to map unavailable triphones onto models present in the corpus (Lanchantin et al., 2008).

Table 2.5 - *Different model configurations.*

	Models	Examples
(1)	Monophones	[a]; [u]
(2)	Triphones	[b-a+f]; [p-a+v]; [j-u+z]; [w-u+Z]
(3)	Tied-State Triphones	[occlusive - a + fricative]; [semi-vowel-u+fricative]

Table 2.6 - *Classes used to determine the context with tied-state triphones.*

Classes	Values
Type	Vowel/Consonant/Semi-vowel
Place of articulation	Bilabial/Labiodental/Alveolar/Palatal/...
Manner of articulation	Plosive/Fricative/Liquid/...
Voicing	Voiced/Unvoiced

Results in Table 2.7 show that the effect of context-dependent models is highly dependent on the corpus. While they improve the alignment of Antoine (Sad), triphones tend to degrade the alignment of Will (Creature) and tied-state triphones have a negative effect on the alignment Margaux (Happy). Interestingly, Margaux is also the best-aligned corpus with monophones. We may therefore wonder whether the use of tied-state triphones would be more useful only when the monophone initial alignment is relatively low, but this would need to be confirmed on more data. Interestingly, there seems to be a tradeoff, tied-state triphones carrying out, for all three corpora, improvement for finer resolutions, while monophones could be more suited at coarser resolution (here with the exception of Antoine (Sad)).

Earlier studies have indicated that the use of context-dependent models may not be suited for forced alignment (Ljolje and Riley, 1993; Kawai and Toda, 2004; Bürki et al., 2008). A possible justification was that the context-dependent models are always trained within a specific context, which implies that they may not learn to correctly discriminate between the phoneme itself and its context (Toledano and Gómez, 2002). Our results partly confirm this hypothesis. Besides, the use of such context-dependent models slows down the training process ⁶. For these reasons, the remainder of this chapter will focus on the use of monophone models only.

⁶The processing time was shown to more than double with triphones instead of monophones, tied-state triphones requiring intermediate processing durations.

Table 2.7 - *Alignment rates for monophone (Mono) alignment and relative improvement (in %) with triphones (Tri) and tied-state triphones (Tied), with a 20 ms and 40 ms tolerance threshold, for the development database.*

Tolerance	20 ms			40 ms		
	Mono	Tri	Tied	Mono	Tri	Tied
Will (Creature)	63.13	-2.32	3.12	84.27	0.84	1.67
Antoine (Sad)	75.40	1.64	3.39	86.71	3.46	7.36
Margaux (Happy)	83.71	2.92	2.47	95.08	-0.37	-3.35
Average		0.75	2.99		1.31	1.89
Standard Deviation		2.73	0.47		1.96	5.36

As previously mentioned, the standard alignment technique typically includes 3 iterations for the first training stage and 5 for the second, *i.e.* after the insertion of the short-pause tee models. However, it is encouraged in Young et al. (2000) to **optimize that number of training iterations**. We investigate here, on the development dataset, the role played by the number of iterations on the alignment rates.

A first round of experiments showed that only the number of iterations of the second training stage should be optimized. All three corpora were then aligned with a varying number of iterations in that stage, ranging from 2 to 40. This indicated high variability across the corpora, the best alignment rates being reached after 17 to 31 iterations. To account for this influence of the corpus, we investigated the possibility of dynamically optimizing the number of embedded training iterations based on the log-likelihood per frame of the training data. This technique relies on the fact that Pearson’s tests revealed strong positive correlation scores (average Rho of 0.93) between the log-likelihood of the model and the alignment rates at 40 ms, consistently across all three corpora: if the log-likelihood stagnates, it can be interpreted as an indication that the training should be stopped.

Based on experimental tests, the minimum increase in log-likelihood was set to 0.001, as the alignment performance curve seems to flatten past that level. The results also drove us to set the maximum number of iterations to 35. This dynamic optimization of the number of training iterations provides consistent improvement across all development corpora and for all tolerance thresholds (see Table 2.8). Average relative improvement reaches 7.08 % at 20 ms and 11.50 % at 40 ms.

This development stage indicated that our *T&A* alignment method provides the best performances on the development database with:

- 10 ms speech frames with no overlap
- monophone models

Table 2.8 - *Alignment rates for standard HMM-based alignment (Base) and relative improvement (in %) with optimization of the number of training iterations (Opti), with a 20 ms and 40 ms tolerance threshold, for the development database.*

Tolerance	20 ms		40 ms	
	Base	Opti	Base	Opti
Will (Creature)	63.13	10.77	84.27	6.89
Antoine (Sad)	75.40	9.01	86.71	11.26
Margaux (Happy)	83.71	1.46	95.08	16.36
Average		7.08		11.50
Standard Deviation		4.95		4.74

- a dynamic optimization of the number of training iterations

This configuration will be used in the remainder of the chapter. In the next section, this method is assessed on a large evaluation database varying in size, language and speaking style. This also allows for testing the performance of the alignment on under-resourced languages as this is an advantage of our method.

2.5 Evaluation

This section shows the results obtained by our alignment method on the evaluation database (presented in Subsection 2.5.1). Our $T&A$ alignment is evaluated in Subsection 2.5.2. It is then compared to alignment rates obtained by state-of-the-art available models of the corresponding languages in Subsection 2.5.3. In Subsection 2.5.4, the role played by the size of the corpus is evaluated. The impact of the size of the bootstrapping corpus when one is exploited is further studied in Subsection 2.5.5.

2.5.1 Speech material

To assess the performance of the various techniques, 12 corpora are used (see Table 2.9). They vary in terms of language, size, and speaking style. Most are read speech, with high recording quality, used for speech synthesis. The use of Sportic allows for an analysis of the results on spontaneous speech. Each corpus contains one speaker, male or female, except for Woggle which consists of recordings from 5 female speakers. This corpus is characterized by a high level of variability, containing also 5 different emotional states (*e.g.* happy, sad, angry). All corpora are classified as neutral or expressive. The expressive tag contains different kinds of expressivity: emotions with different valences (happy, sad, angry and afraid) and specific attitudes/speaking styles (bad guy and sports commentaries). The advantage of our basic $T&A$ technique being to apply to any language, the method is also tested on under-resourced languages like Faroe and Gaelic. All corpora were manually phone-aligned by experts.

⁷Kindly provided by Acapela Group SA

Table 2.9 - *Evaluation data set features.*

Corpus	Language	Speaking style	Duration (minutes)
Will (Bad guy) ⁷	English	Read/Expressive	12
Will (Neutral) ⁷		Read/Neutral	14
Woggle (Dellaert et al., 1996)		Read/Expressive	51
Antoine (Happy) ⁷	French	Read/Expressive	12
Margaux (Sad) ⁷		Read/Expressive	12
Marie		Read/Neutral	108
Sportic (Brognaux et al., 2013) (described in Section 5.6.3)		Spontaneous/Expressive (sports commentaries)	15
Faroe ⁷		Read/Neutral	21
Gaelic ⁸		Read/Neutral	8
Afrikaans (van Niekerk and Barnard, 2009)		Read/Neutral	22
Setswana (van Niekerk and Barnard, 2009)		Read/Neutral	47
Isizulu (van Niekerk and Barnard, 2009)		Read/Neutral	20

2.5.2 Results with Train&Align

Our baseline approach with training on the corpus to align ($T&A$, as described in Section 2.4) is first applied to all corpora. The resulting correct alignment rates are shown in Table 2.10⁹.

Rates around 80 % with a 20 ms threshold are reached for all of our corpora in French and in some under-resourced languages. This is comparable to observed inter-annotator agreement rates in Goldman (2011). Low alignment rates for Woggle, can be explained by its high degree of expressivity and diversity, as it contains 5 speakers and 5 different emotions. Unlike the other models, the one trained on Woggle does not capture a specific speaking style of a speaker, which might explain the lower alignment rates. The size of the sub-corpora containing only one speaker and one emotion being too small to allow for a correct model training, the splitting of Woggle results in even lower alignment rates. However, tests show that, for an identical corpus size, not mixing the emotions provides higher alignment quality and that this distinction is more important than that of training on a single speaker only, but containing several emotions. This tends to indicate that pronunciation is more emotion-specific than speaker-specific; this should however be further analyzed. When training on data from one speaker and

⁸Kindly provided by the Phonetics and Speech Laboratory, Trinity College, Dublin

⁹We also analyzed the results with context-dependent models and different speech frame sizes and overlap. The results confirmed those obtained on the development corpora, with better alignment with speech frames of 10 ms with no overlap and high variability when using context-dependent models. Interestingly, tied-state triphone models were however shown to provide slightly better results than triphone models. The results presented here rely on monophone models, for the aforementioned reasons.

Table 2.10 - *Correct alignment rates (in %) of T&A on our evaluation corpora in English (En), French (Fr) or under-resourced languages (O).*

Tolerance threshold		10 ms	20 ms	30 ms	40 ms
En	Will (Bad guy)	47.04	72.14	83.08	88.19
	Will (Neutral)	50.86	78.86	89.72	94.64
	Woggle	44.95	65.23	80.07	88.82
Fr	Antoine (Happy)	59.08	82.70	89.14	92.59
	Margaux (Sad)	59.38	81.12	88.99	92.26
	Marie	60.06	84.42	92.93	96.63
	Sportic	65.01	82.53	90.31	94.22
O	Faroe	48.45	75.18	87.95	93.74
	Gaelic	77.34	89.98	94.25	96.05
	Afrikaans	45.66	73.50	88.38	93.11
	Isizulu	46.78	72.18	85.22	90.94
	Setswana	47.29	70.51	84.71	90.31

one emotion, results show that the ‘neutral’ part of the corpus is the best aligned. This points at the degraded results of our method on expressive speech¹⁰. This, however, is not reflected in the alignment of *Sportic* and may be highly corpus-related.

The optimization of the number of training iterations was shown in Section 2.4 to improve the alignment and was therefore included in the baseline. While its application allows for an average relative improvement of 8 % (40 ms threshold) of the alignment of the evaluation dataset, it performs much better on highly expressive corpora like Woggle (+16 %) and Sportic (+ 45 %). This is particularly interesting in the framework of our study, *i.e.* for the alignment of spontaneous expressive corpora in view of HMM-based speech synthesis. Conversely, it achieves lower performances on Marie (-3 %), which contains read neutral speech.

¹⁰We also wondered whether this difference between the alignment of the ‘neutral’ sub-corpus and that of the other emotions might rely on the fact that the emotions are less articulated. This relationship between articulation degree and emotions has been previously pointed out by Beller (2007) who showed that the degree of articulation is closely correlated with the activation level of the emotion. This hypothesis was tested by computing the formant values of the central phonemes /ae/, /I/ and /U/ in the different sub-corpora, with the "To Formant (burg)" Praat function (Boersma, 2001) to plot the vocalic triangles. The degree of articulation was computed as the vocalic space, *i.e.* surface of the respective triangles, as in Picart et al. (2010). We observed that, as in Beller (2007), more activated emotions display a higher degree of articulation. This seems to influence the alignment rate as the less articulated emotions (*i.e.* sad and happy) are also those with the lowest alignment rates.

2.5.3 Comparison with state-of-the-art models

As previously mentioned, most automatic alignment tools provide the user with pre-trained speaker-independent acoustic models of the language (*e.g.* EasyAlign (Goldman, 2011), SPPAS (Bigi and Hirst, 2012), P2FA (Yuan and Liberman, 2008)). These models are supposed to offer a generic representation of the language and can be used to align any new corpus in the corresponding language. In opposition to our method, the corpus to align is not used in the model training stage.

1. **EasyAlign** provides a model for French but not for English. Its French model was trained on “30 minutes of unaligned multi-speaker speech for which a verified phonetic transcription was provided” (Goldman, 2011). The model consists of monophones.
2. **SPPAS** provides models for both French and English. SPPAS French model used here is a monophone model from SPPAS-1.6. It was trained on 8 hours of phonetically transcribed but not aligned spontaneous speech, made of 7h30 of the CID corpus (Bertrand et al., 2008) and 5 minutes from the grenelle corpus (Bigi, 2012). The model uses 2 minutes of bootstrap from radiophonic data (Europe1) which were manually phonetized and aligned (Portes, 2004). For more information on the training of that model, see Bigi (2014). SPPAS English model is the model of July 2011 provided by VoxForge. It contains about 85 hours of multi-speaker read speech. The corpus was automatically phonetically transcribed but not aligned. The model consists of triphones.
3. **P2FA** only provides an English model. It was trained on 25.5 hours of word-aligned speech from the Scotus corpus. This corpus consists of oral arguments from the Supreme Court of the United States. The model is made of monophones. It is worth noting that P2FA model depends on the lexical stress level of the phoneme. Three levels are considered: no stress, primary and secondary stress. Each vocalic phoneme is associated with three models. To exploit the full capacity of the system, all the phonetic transcriptions were stress-annotated when aligning with P2FA model.

Results of the comparison are shown in Table 2.11. For the corpus in French, interestingly, Train&Align is observed to clearly outperform SPPAS and EasyAlign models across all measures. The gain compared to SPPAS goes up to 13 % with a 20 ms tolerance threshold. The overall rate is fairly high. It is comparable to inter-annotator rates reported in Goldman (2011).

For the corpus in English, only P2FA provides results better than Train&Align, at least for thresholds inferior to 40 ms. The explanation for this may be twofold. First, the English model used by P2FA was trained on more than 25 hours of manually-word aligned speech. To provide such results for every language, it should be repeated on a very large number of corpora. This is economically impracticable, especially to cover all (even rare) languages. However, it is bound to produce better results. It is striking to notice that Train&Align offers slightly inferior but still

Table 2.11 - Correct alignment rates (in %) achieved by available speaker-independent models and by our method (*T&A*).

Tolerance threshold	10 ms	20 ms	30 ms	40 ms
A French neutral corpus: Marie (Colotte and Beaufort, 2005)				
<i>SPPAS</i>	45.76	71.28	83.06	89.84
<i>EasyAlign</i>	52.54	77.54	87.72	92.11
<i>T&A</i>	60.06	84.42	92.93	96.63
An English expressive corpus: Woggle (Dellaert et al., 1996)				
<i>VoxForge</i>	37.3	65.2	82.07	88.69
<i>T&A</i>	44.95	65.23	80.07	88.82
<i>P2FA</i>	46.68	69.74	81.2	87.46

comparable results while using no preexistent model trained on aligned data. The second explanation for the higher alignment rate of P2FA results from the use of acoustic models associated with various levels of stress. The Woggle corpus being an expressive corpus, it contains many emphatic stresses. It is well-known that these stresses usually fall on the same position as lexical stresses. This could also explain the better achievements of P2FA. The overall low alignment rates for the corpus in English are clearly due to its high acoustic variability.

On the whole, it should be noted that Train&Align provides alignment rates better or comparable to most other alignment tools. In addition, it has the advantage of being applicable to any language, as shown in Table 2.10, for Faroe, Gaelic and several African languages.

The acoustic models being directly trained on the corpus to align in *T&A*, the size of the corpus plays an obvious role in the model quality. It can therefore be wondered whether the rather high alignment rates obtained for Marie can be explained by the size of the corpus. More than 100 minutes of speech are used to train the models. This provides a fair amount of occurrences for each phoneme, which is not possible for all databases. This question is now addressed in Section 2.5.4.

2.5.4 Influence of the corpus size

The influence of the corpus size on the alignment performance was analyzed by varying the size of the various corpora. This evaluation was performed on the three largest corpora of our evaluation dataset which have each a duration greater than 30 minutes (initial and final silences excluded). The resulting correct alignment rates are shown in Figure 2.8. It shows that the alignment rate plateaus for corpora larger than 5 minutes. A similar curve was observed with a 40 ms tolerance threshold. For Setswana, 2 minutes are enough to reach the correct alignment maximum. For corpora smaller than 2 minutes, the alignment performance rapidly degrades. It should be noted that the alignment of the 2-minute Woggle corpus is, comparatively to the two other corpora, of rather low quality. This may be explained by the fact that properly learning the higher variability of the corpus (various emotions and

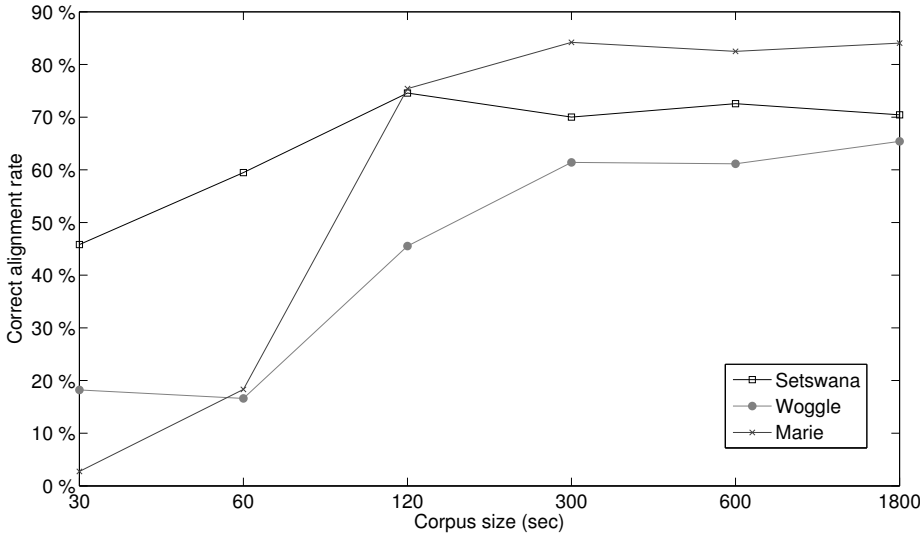


Figure 2.8 - Correct alignment rates (in %) with a 20 ms tolerance threshold on the evaluation set, with varying corpus sizes.

speakers) requires larger datasets. On the whole, a few minutes of neutral speech seem to be sufficient to train and align a new corpus, which confirms findings in van Niekerk and Barnard (2009) but is much lower than the limit of 128 sentences fixed by Brugnara et al. (1993). This may be explained by the fact that Brugnara et al. (1993) does not make use of speaker-dependent models, the training being performed on a corpus different from the corpus to align.

A second question to investigate is the size of the speech files provided to Train&Align. Most existing alignment tools rely on an initial segmentation stage to prevent large audio files from being aligned (Goldman, 2011; Bigi and Hirst, 2012). To better alleviate this issue, Train&Align makes use of pruning methods to reduce the processing time for longer files. Experiments with Train&Align showed that files up to one minute can be used without needing a segmentation beforehand. Large files of more than a few minutes, however, should be cut to avoid very long processing time. It should be noted that files of more than 1 minute containing very few silences can be misaligned if they are not segmented.

2.5.5 Effect of bootstrap

Section 2.5.2 showed that some expressive corpora and datasets from low-resourced languages are rather poorly aligned with a standard HMM-based alignment when training on the corpus to align. A possible way to alleviate this issue is by improving the initialization of the models based on some manually-aligned part of the corpus

(Brugnara et al., 1993; Jarifi et al., 2008). Instead of a ‘flat start’ initialization, the manually-aligned corpus helps to improve the initialization stage as described in Section 2.2.

We investigate here which size of the bootstrapping corpus is required and which improvement may be expected. For this experiment, we gradually increased the size of the bootstrap (from 10 to 600 seconds, initial and final silences excluded) while evaluating on a fixed portion of the corpus (2 minutes) to avoid biases due to a varying evaluation corpus. The evaluation set was never included in the bootstrap data. The monophones are initialized on manually-aligned data only if at least three occurrences of the phoneme are available. If not, average values (for means and variances) are assigned to the model, similarly to a ‘flat start’ initialization. Figure 2.9 interestingly points out that the use of a bootstrap part of the corpus, as small as 10 seconds, leads to a relative improvement of about 25 to 35 %. While the use of a larger bootstrap slightly improves the quality with a 20 ms threshold, the curve rapidly flattens, the use of a bootstrap corpus larger than 30 seconds being essentially unnecessary. As expected, the use of bootstrap data is especially effective on expressive corpora like Woggle, which was poorly aligned with the initial *T&A* technique. Improvement of about 53 % is here observed with only 30 seconds of bootstrap data, for both 20 and 40 ms thresholds. High improvement is also found on some corpora in under-resourced languages: the use of 30 seconds of bootstrap improves the alignment of Setswana, in particular, by 56 % with a 20 ms threshold, to reach alignment rates of more than 87 %. As previously mentioned, it is reported in Kawai and Toda (2004) that the manual alignment process takes about 130 times real-time. For 30 seconds, about 1 hour would be required for a human annotator, which seems highly feasible. Besides, the annotator can rely on the initial alignment without bootstrap and only correct the erroneous boundaries. This should reduce the processing time.

2.6 Analysis of errors made by *Train&Align*

In order to develop methods improving HMM-based phonetic alignment in the next chapter, an in-depth analysis of the errors most frequently made by *T&A* was carried out in this last section (see Table 2.12). This analysis relied on our development database as presented in Section 2.3 to allow for an evaluation of our improvement methods in the next chapter on our evaluation database. We focused on errors with a timing difference greater than 40 ms which correspond to high incongruencies unlikely to be produced by human annotators. In this analysis, we distinguished between two measures. The *global error rate* regards the percentage of all errors that this specific phoneme class accounts for. The *specific error rate* relates to the percentage of transitions of that class which are erroneous. This distinction is important as it shows, for example, that transitions between vowels are often prone to errors (in 38.31 % of the cases) but that they account for a smaller global error rate (*i.e.* 25.13 %), being less frequent than most other transitions in the corpora.

This study drove us to distinguish between 4 typically problematic transitions.

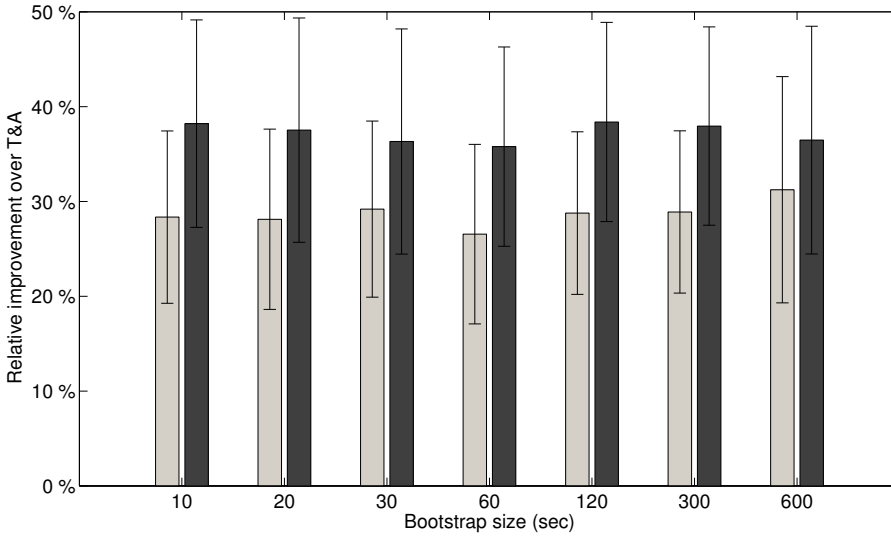


Figure 2.9 - Average relative improvements over *T&A*, for all evaluation corpora, with 20 (light grey) and 40 ms (dark grey) tolerance thresholds, using a bootstrapping corpus varying in size, along with their 95 % confidence interval.

Table 2.12 - For specific phoneme classes, percentage of global and specific errors (for errors > 40 ms), computed on the development database.

Phoneme class	Global errors (%)	Specific errors (%)
Silences	61.61 %	49.79 %
Vowel-vowel	25.13 %	38.31 %
Approximants	13.89 %	7.61 %
Plosives	6.69 %	2.17 %

The first relates to transitions to and from silences, which account for more than 60 % of the total amount of errors higher than 40 ms. In fact, 49.79 % of the silences are erroneously aligned with that threshold. This issue has been pointed out by other studies (Wightman and Talkin, 1997; Jarifi et al., 2008; Kawai and Toda, 2004; Nguyen and Espesser, 2004), which indicated low alignment rates for silence boundaries. Kawai and Toda (2004) also noticed low alignment rates on silences in their corpus, even between human annotators. They explained that phenomenon by the fact that it is not always easy to precisely define the beginning and end of voicing from waveforms. In fact, final phonemes often undergo devoicing, which makes it harder to find the precise beginning of the subsequent silence. Regarding the initial boundary of silences, the issue is probably due to the presence of the breath intake. Interestingly, other segmentation algorithms like unsupervised methods based on

the acoustic rate of change were also shown to produce similar errors (Scharenborg et al., 2010). Such misalignment of silences is also found when aligning our corpus with existing models such as models from EasyAlign (Goldman, 2011) and P2FA (Yuan and Liberman, 2008). This misalignment is even more striking provided that our development set contains corpora designed for speech synthesis, *i.e.* with no background noise. This issue may then be even worse for corpora recorded in noisier conditions.

Other error types that were found to be frequent are, in decreasing order, vowel-to-vowel transitions, approximants and plosives. For all three categories, pairs containing a silence were excluded, silences inducing a high amount of errors. Approximants include lateral approximants (like [l]), non-lateral approximants (like [ɹ]), and semi-vowels (like [j], [w] or [ɥ]). Semi-vowels were shown to be problematic in Goldman and Schwab (2011); Nguyen and Espesser (2004) and Kawai and Toda (2004), and also present issues for human annotators, probably due to their high degree of co-articulation. Vowel-to-vowel transitions were also pointed out as prone to errors in Brugnara et al. (1993); Wesenick and Kipp (1996); Ljolje et al. (1997) and Cangemi et al. (2011). This transition was also shown to be highly problematic for human labelers (Wesenick and Kipp, 1996). Regarding plosives, they were also misaligned in Demuynck and Laureys (2002). For approximants, especially the initial boundary tends to be problematic in our corpus. Conversely, final boundaries of plosives seem to be more prone to errors, the explosion stage being harder to model. Figure 2.10 shows an example of typical misalignment.

While the optimization of the number of training iterations was shown to improve the alignment rate, it should be noted that it does not impact a specific type of error. All classes of errors tend to be reduced with better improvement for some phoneme classes, depending on the corpus.

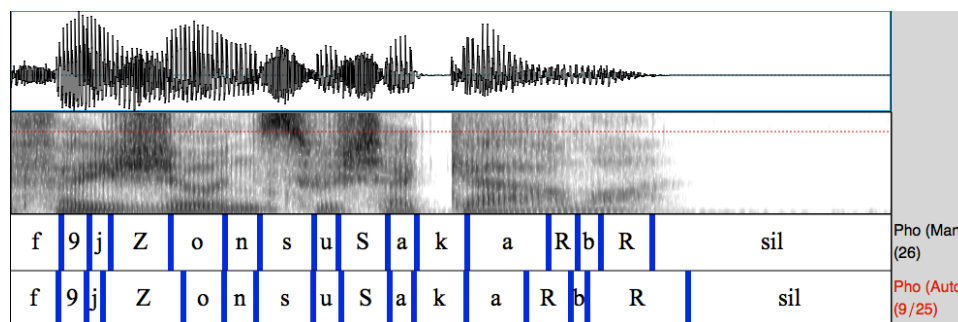


Figure 2.10 - *Example of silence and liquid misalignment in the automatic alignment (Pho (Auto)) of one speech file of the development database.*

As shown in Section 2.5.5, a possible way to reduce the amount of errors consists in using some manually-aligned data as bootstrapping data so as to produce a better initialization of the models. It requires, however, some manual intervention. In order

to stick to a fully automatic method, the next chapter presents three improvement strategies driven from the aforementioned typical errors.

2.7 Conclusion

In this chapter, we proposed a systematic, step-by-step study of a particular case of HMM-based speech segmentation, *i.e.* when training the models directly on the corpus to align. A development database was first used to tune different parameters (*e.g.* use of context-dependent models and number of training iterations) to obtain our baseline model, Train&Align ($T\mathcal{E}A$). It showed that the best results were obtained when extracting the features on speech frames of 10 ms with no overlap. Regarding the number of training iterations, we proposed a dynamic optimization which led to an average improvement of the correct alignment rate reaching 7 %, with a 20 ms tolerance threshold. Finally, the use of context-dependent models was investigated. It showed that their impact on the alignment quality varies across corpora and should not always be advised. For that reason, our $T\mathcal{E}A$ technique relies on monophone models.

$T\mathcal{E}A$ was then evaluated on a large dataset containing 12 speech corpora, varying in speaking style, size and language. It showed that using a model trained on the target corpus yields nearly as good or even better results than using available speaker-independent models of the language. These available models were those provided by recent alignment tools: EasyAlign, SPPAS and P2FA. Improvements of about 7 % (20 ms threshold) in the alignment quality of a French neutral corpus can be observed with our method, compared to the best aligner, *i.e.* EasyAlign. This can be explained by the fact that the model better captures the characteristics of the target corpus. On the English-speaking expressive corpus, only P2FA outperforms our method, by about 4 % for 20 ms but only 1 % for 30 ms. This may be due to their training corpus that consists of more than 25 hours of word-aligned speech. It was also shown that T&A reaches good alignment rates on under-resourced languages like Faroe or Gaelic.

A second aim of our study was to investigate the role played by the size of the database or by the use of bootstrap data on the performance of HMM-based phonetic alignment. The conclusions we drew in that respect were that:

- The performance seems to plateau beyond 5 minutes of training data, which implies that fairly high alignment capabilities are possible for small databases of only a few minutes.
- Manually-annotated bootstrapping data can be used to enhance the initialization of the models. Using only 10 seconds of such data generates a relative improvement of 28 % at a tolerance threshold of 20 ms and increasing the amount of bootstrapping data does not seem to significantly improve the alignment.

Our results however indicated that some corpora exhibiting great expressive variability achieved poor alignment rates if no bootstrapping data was used. We

therefore provided an in-depth analysis of the typical and systematic errors made by *T&A* to propose specific refinement techniques in the next chapter. This pointed at the poor alignment of silences, vowel-to-vowel transitions, semi-vowels and plosives.

It should be noted that our T&A method was further developed into a new online automatic alignment tool: Train&Align¹¹ which is presented in Chapter 4. Unlike other user-friendly alignment tools that do not grant access to the training phase and only provide the user with a set of speaker-independent models to align new corpora, Train&Align trains new models directly on the corpus to align. It implements all options presented in this chapter, *i.e.* the consideration of the phonetic context, the modification of the size and overlap of window frames and the use of bootstrap data. Besides, it offers an automatic evaluation of the alignment if some manually-aligned part of the corpus is provided.

¹¹http://cental.fltr.ucl.ac.be/train_and_align

Chapter 3

Improving HMM-based alignment

Contents

3.1	Introduction	51
3.2	Existing improvement methods of HMM-based forced alignment	52
3.3	Proposed improvement methods	53
3.3.1	Use of a Voice Activity Detection algorithm (<i>VAD</i>)	53
3.3.2	Augmentation of the feature vector (<i>AddFeat</i>)	54
3.3.3	Exploitation of the time-reversed sound (<i>Reverse</i>)	56
3.4	Evaluation	58
3.4.1	Improvement methods	58
3.4.2	Advantage of <i>T&A2</i> in small corpora	60
3.4.3	Comparison with state-of-the-art models	61
3.5	Conclusion	61

Abstract

The automatic phonetic alignment of large speech corpora is of crucial importance, both for linguistic research and speech technologies. In the previous chapter, we highlighted the usefulness of HMM-based alignment in which the models are directly trained on the corpus to align. This achieved good-quality results on most neutral corpora but it was seen that alignment rates are generally poorer for expressive speech and corpora in some under-resourced languages. Based on the analysis of the errors commonly made by this technique, we investigate in this chapter the use of additional fully automatic strategies to improve the alignment. Beside the use of supplementary acoustic features, we explore two novel approaches: an initialization of the silence models based on a Voice Activity Detection (VAD) algorithm and the consideration of the forced alignment of the time-reversed sound. The evaluation is carried out on 12 corpora of different sizes, languages (some being under-resourced) and speaking styles. It aims at providing a comprehensive study of the alignment accuracy achieved by the different versions of the speech segmentation algorithm depending on corpus-related characteristics. Our results indicate that the insertion of additional features outperforms both other strategies. The performance of *VAD*, however, is shown to be notably striking on very small corpora, correcting more than 60 % of the errors greater than 40 ms. Finally, the combination of the three improvement methods is also pointed out as providing the highest alignment rates, with very low variability across the corpora, regardless of their size. This combined technique is shown to outperform available speaker-independent models, improving the alignment rate by 8 to 10 % absolute.

This chapter is based upon the following publication:

- Sandrine Brognaux & Thomas Drugman. (2016). HMM-based Speech Segmentation: Improvement of Fully Automatic Approaches, *IEEE Transactions on Audio, Speech and Language Processing*, 24(1):5-15.

Many thanks to Prof. C. Wellekens for his insightful advice.

3.1 Introduction

Large speech databases play a major role in speech technologies. The sound needs to be phonetically transcribed and the phonemes usually need to be time-aligned with the sound. Manually aligning being a tedious and time-consuming task, automatic alignment tools have been developed (*e.g.* Easyalign (Goldman, 2011), SPPAS (Bigi and Hirst, 2012)). As mentioned in Chapter 2, these tools provide the user with speaker-independent models that highly depend on the training corpus. When trained on neutral speech, these models tend to produce low-quality alignment of expressive corpora. Besides, they only provide a limited set of models that makes it impossible to align some languages. A possible way to solve this issue is to train the models directly on the corpus to align, thereby providing a better agreement between the training and alignment stages. This method was investigated in Chapter 2 and was shown to achieve alignment rates comparable to those of available speaker-independent models. It presents the advantage of applying to any language or speaking style without the need for manually-aligned data. That fully-automatic alignment, however, also remains of poor quality on expressive corpora and on speech from some under-resourced languages.

A possible way to reduce the amount of errors consists in using some manually-aligned data as bootstrapping data so as to produce a better initialization of the models (Brugnara et al., 1993). This technique was investigated in Subsection 2.5.5 and was shown to significantly improve the alignment. It requires, however, some manual intervention. Based on typical errors found in the automatic alignment (see Subsection 2.6), we propose in this chapter some refinements of the conventional HMM-based alignment. The proposed improvement methods were developed so that they do not require training on manually-aligned data. The development and evaluation is based on our large database, made of 15 corpora in French, English and rather under-resourced languages such as Faroe or Gaelic. Similarly to tests in Chapter 2, the evaluation is performed on clean, mostly read speech, with manually-verified phonetic transcriptions, as used for speech synthesis purposes. Here again, the results are compared to alignment rates obtained with publicly available models of the languages, as used by most existing tools.

The chapter is organized as follows. Section 3.2 presents an overview of existing strategies to improve HMM-based alignment proposed in the literature. Our three improvement methods, along with their parameter settings, are described in Section 3.3. In Section 3.4, they are evaluated on our large evaluation database and compared with the use of available speaker-independent models. Finally, Section 3.5 concludes the chapter.

3.2 Existing improvement methods of HMM-based forced alignment

Several studies have tried to improve conventional HMM-based forced alignment. Different post-processing methods have been developed to refine the produced alignment, using various statistical techniques (*e.g.* Toledano and Gómez (2002), Sethy and Narayanan (2002), Lo and Wang (2007) and Adell et al. (2005)). However, they usually depend on the training of a second model (*e.g.* a regression tree (Adell et al., 2005), Gaussian Mixture Models (GMMs), HMMs (Sethy and Narayanan, 2002) or support vector machines (SVM) (Lo and Wang, 2007)) which requires a manually-annotated corpus. Heuristic rules can also be used to modify the boundaries as a post-process, making use of the average deviation for each pair of phonemes (Demuynck and Laureys, 2002). In a similar way, it is proposed in Jarifi et al. (2008) to train, on a manually-aligned corpus, boundary models depending on the right and left context of each phoneme to improve the alignment. They also suggest making use of an automatic detection of discontinuities in the signal, based on Brandt (1983).

The combination of various HMM-based alignments trained with different parameters has also been investigated (Park and Kim, 2007). Here again, manually-aligned data is required for the training. For approximate transcriptions, a slightly modified version of HMM-based alignment was also proposed in Katsamanis et al. (2011), proceeding in multiple iterations and combining recognition and alignment methods.

Most of the aforementioned techniques are supervised and therefore require manually-aligned data for training. The annotation process can be tedious when aligning rare languages for which little data are available. Furthermore, manual alignment is known to be very time-consuming and consequently costly (Kawai and Toda, 2004; Schiel et al., 1998). Contrary to the methods described above, the approach we proposed in Chapter 2 is unsupervised and does not require any labelled data. A flat-start initialization is first considered on the target corpus, and the alignments are expected to converge across the training iterations. This technique is used as a baseline in the remainder of this chapter. In Chapter 2, however, this method was shown to achieve rather poor results on some corpora, especially for highly expressive speech.

An analysis of the typical errors made by a standard HMM-based alignment was provided in Section 2.6. Based on this analysis, we propose in this chapter three improvement techniques specifically designed to alleviate these alignment flaws. Conversely to most existing improvements of HMM-based forced alignments, these methods exhibit the advantage of not requiring the use of manually-aligned data. A first widely-known improvement lies in the augmentation of the feature vector. The addition of new acoustic characteristics to the feature vector has already been investigated in Brugnara et al. (1993); Young et al. (2000); Hosom (2000, 2002, 2009) and Mporas et al. (2008). In Brugnara et al. (1993), for instance, features related to spectral variation have been added and led to a slight increase in the alignment rates. Hosom (2002) proposed using an additional set of acoustic-phonetic

features consisting of: intensity discrimination (computed as the change in intensity with a 40 ms window), voicing, fundamental frequency, glottalization and burst detection. Hosom (2000) offers a more detailed description of the results of this integration of additional features and shows that it yields average relative improvement of about 8 %, with a 20 ms tolerance threshold. The impact of each feature independently is however not assessed. Mporas et al. (2008) proposed to choose the best features for each phone transition. The improvement of their alignment rates, however, is rather small (0.7 % for a 20 ms tolerance threshold). Finally, Lo and Wang (2007) also exploit additional features like burst degree and sub-band energy in the SVM, which is trained in a second stage, to improve the HMM-based alignment.

Beside this use of additional features, we also propose two novel strategies: a better initialization of the silence model based on a VAD algorithm and the use of the alignment of the time-reversed sound to provide smoothed boundary estimations. These refinement techniques are detailed in Section 3.3. The objective of this chapter is both to confirm the validity of the first widespread method and to compare it with the two proposed approaches. The combination of all three techniques is also further analyzed as well as its possible application to various types of corpora (varying in size, style and language).

3.3 Proposed improvement methods

Three improvement methods are investigated in this chapter: the use of a voice activity detection algorithm (Subsection 3.3.1), the augmentation of the feature vector (Subsection 3.3.2) and the exploitation of the time-reversed sound (Subsection 3.3.3). They are all implemented as complementing the standard HMM-based forced alignment presented in Chapter 2. This standard alignment (*Train&Align*) is obtained when training the acoustic models directly on the corpus to align. *Train&Align* relies on acoustic vectors made of 39 features, *i.e.* 13 MFCC parameters, along with their delta and acceleration coefficients. These parameters are computed on 10 ms-long windows with a shift of 10 ms. The models are monophone models and the number of training iterations is dynamically optimized with a strategy proposed in Section 2.4. For the three improvement methods, our development protocol is similar to that described in Chapter 2. The parameter setting of the three improvement techniques is also computed on our development database, made of 3 corpora and including different types of expressivity, speaking style, two languages and three different speakers.

3.3.1 Use of a Voice Activity Detection algorithm (*VAD*)

Seeing that silences are often prone to alignment errors, we propose to improve their initialization. As no manually-aligned part of the corpus is provided, our standard HMM alignment (*T&A*) relies on a ‘flat start’ uniform initialization. This means that each phone model is first assigned average values (for means and variances) and uniformly aligned with the sound. The proposed refinement

aims at modifying this initialization stage. Sohn’s voice activity detection (VAD) algorithm (Sohn et al., 1999) is first applied to the sound files and allows detecting non-speech segments. These segments are then used to initialize silence models only. All other phonemes are initialized with the ‘flat start’ strategy, exactly as in *T&A*. This particular use of the VAD exhibits the advantage not to fix the silence boundaries, which would be problematic as VAD algorithms tend to over- or under-detect silences notably in plosion stages or noisy pauses. For that reason, these VAD-detected boundaries cannot be directly used to allow for a better uniform segmentation of the other phonemes. It allows, however, the silence models to be better initialized and helps to iteratively converge to better silence boundaries. Besides, our method provides an initialization of the silence model specific to the corpus and its recording conditions, conversely to the use of a generic silence model. This may also help for corpora recorded in noisier conditions.

Sohn’s VAD algorithm generates trajectories of posterior probabilities about the presence or not of speech activity. In order to draw a binary decision, a threshold has to be applied on these trajectories, below which the sound is assumed not to contain speech. With a null threshold, all frames are regarded as voice and no signal section is exploited to train the silence models. The alignment is then that of the original *T&A*. Experiments on our development database showed that the performance of the alignment reaches a plateau from a threshold of 0.3 (see Figure 3.1). For the remainder of our experiments, we set the threshold to 0.8 which provides the best alignment rates on average. This achieves an average relative improvement over *T&A* of 23.29 % on the development set at a 40 ms threshold.

Interestingly, the alignment rates achieved by *VAD* 0.8 are very similar, regardless of the optimization of the number of training iterations, which was shown to improve *T&A* alignment in Section 2.4 and which was therefore integrated in the baseline. When discarding this optimization from the baseline, the relative improvement of *VAD* reaches 37 % on average with a 40 ms threshold. This indicates that the *VAD* also helps to compensate for a lack of training iterations: if the initialization is better performed, less embedded training iterations are required. Another interesting observation is that *VAD* especially reduces the amount of gross errors (*i.e.* errors greater than 30 ms) while it has less effect on finer errors.

3.3.2 Augmentation of the feature vector (*AddFeat*)

It is widely acknowledged that part of the errors made by automatic phonetic alignment techniques comes from the fact that humans make use of additional cues, which are not represented in the MFCC coefficients, to decide the precise location of the boundaries (Demuynck and Laureys, 2002). This may especially be the case for vowel-to-vowel transitions or transitions to and from semi-vowels which are very hard to model. For that reason, we considered the possibility of adding up to 9 supplementary acoustic features on top of the MFCC coefficients.

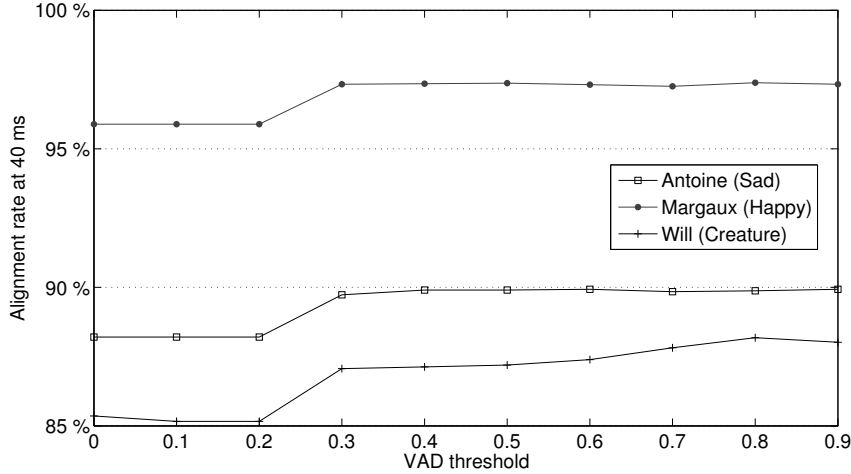


Figure 3.1 - Correct alignment rates (in %) with a varying VAD threshold with a 40 ms tolerance threshold, on the development set.

These additional features are:

- the total perceptual loudness (Peeters, 2004)
- the frequencies of the first five formants estimated by the Differential-Phase Peak Tracking (DPPT (Bozkurt et al., 2004)) technique
- the fundamental frequency F0 and a measure of periodicity based on the Summation of the Residual Harmonic (SRH (Drugman and Alwan, 2011)) algorithm
- a measure of turbulence using the Chirp Group Delay (CGD) function which was shown in Drugman et al. (2011) to highlight irregularities of phonation

For each of the proposed additional features, we have calculated its coefficient of correlation with the MFCCs (static and dynamic values) using the canonical correlation method. This analysis was based on the speech segments of the development dataset. The results are as follows: perceptual loudness (0.93), F1 (0.63), F2 (0.48), F3 (0.33), F4 (0.29), F5 (0.17), F0 (0.57), periodicity (0.74), turbulence (0.47). For some features, this correlation is quite high. That is the case of the perceptual loudness (which was expected) and surprisingly of the periodicity measurement. Note however that even if a feature is redundant with an existing set, it might still convey relevant complementary information. As further described, that is the case of the loudness, which despite its high correlation is shown to carry out an interesting improvement.

The contribution of each feature, individually, is shown in Figure 3.2. High variability is observed across the corpora and only the addition of loudness clearly improves the alignment of all datasets. Interestingly, fundamental frequency is seen

to carry out little improvement or even degradation of the alignment rate. This differs from Saito (1998)’s results which indicate that changes in fundamental frequency are good indicators of phonetic boundary between voiced consonants and vowels. They investigate the use of fundamental frequency to refine the boundaries obtained through a basic dynamic programming (DP) alignment and find it yields improved alignment rate. In that study, however, dynamics of the fundamental frequency are exploited while we do not consider delta values (see below). This should be further investigated in future work.

Various feature combinations have also been explored. While the addition of both loudness and periodicity is shown to provide the best relative improvement on average, their combination with turbulence leads to a reduced variability across the corpora. With that configuration, the initial alignment is improved by 10 to 35 %, with a 40 ms threshold, for all three development corpora. This combination will be exploited in the remainder of the chapter and referred to as *AddFeat*. This amounts to a total of 42 coefficients.

Here again, the interaction between the addition of new features and the dynamic optimization of the number of training iterations is worth discussing. With a standard fixed number of iterations, all individual features are shown to provide relative improvement over the baseline. The minor impact played by features such as formants or F0 seems however to disappear with an optimized number of training iterations.

Seeing that the initial set of parameters contains delta and acceleration coefficients, the insertion of these differential coefficients for the 9 additional features was also considered. However, the analysis of the results on our corpora shows that the consideration of these derivatives does not bring any relevant complementary information when included in addition to their static version. The 9 supplementary features are therefore added to the feature vector without their derivatives.

3.3.3 Exploitation of the time-reversed sound (*Reverse*)

An in-depth analysis of the alignment errors with *T&A* revealed systematic errors, some phonemes being often misaligned with a too early or too late boundary. Examples of such errors can be observed in Table 3.1. It shows, for example, that final [j] boundaries tend to be predicted too late while final silence boundaries tend to be too early. As can be observed, several aforementioned classes of errors are here concerned (*i.e.* plosives, semivowels and silences). This type of errors occurring in the same direction for specific transitions has been pointed out by Adell et al. (2005); Bürki et al. (2008) and Cangemi et al. (2011). In Bürki et al. (2008), [ə] tended to be left-shifted in their corpus. Cangemi et al. (2011) also find the displacement of plosives to be systematic in their study. On the whole, we notice that a majority of boundaries are predicted too early.

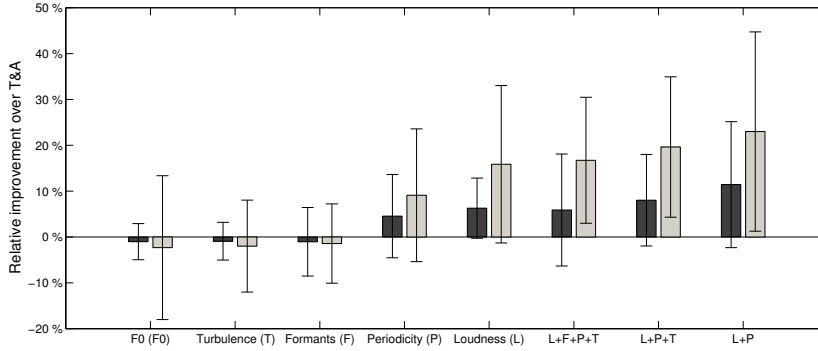


Figure 3.2 - Average relative improvement (in %) over $T\&A$ with additional acoustic features with 20 and 40 ms tolerance thresholds, on the development set, along with their 95 % confidence interval.

Table 3.1 - Example of systematic errors made by $T\&A$.

Boundary	Percentage of cases	Average deviation
Late boundaries		
Initial boundary of silences	74.6 %	47.80 ms
Final boundary of [j]	70.85 %	7.62 ms
Early boundaries		
Final boundary of silences	82.60 %	28.90 ms
Initial boundary of [t]	69.19 %	3.84 ms

To reduce such errors, we propose to time-reverse the sound and its phonetic transcription. The initial feature vector of 39 MFCC-based features (and possibly additional features) is extracted on this time-reversed copy of the sound. $T\&A$ is then used to align the reversed and the original corpora. Both alignments are then exploited to compute average boundary locations. This sound reversal essentially modifies transition probabilities of the HMM. The segmentation of the sound also begins from the end of the file, which might modify the content of the respective frames of signal. The assumption behind this method is that, for highly probable boundaries, the alignment of both corpora should provide similar results which will not influence the resulting alignment. Conversely, for uncertain boundaries, computing the average between both alignments should provide smoothed estimations, thereby reducing errors with a high tolerance threshold. To the best of our knowledge, this technique has never been investigated in previous studies.

3.4 Evaluation

This section shows the results obtained by our three improvement methods of the automatic HMM-based forced alignment on our evaluation database (presented in Subsection 2.5.1). The results of the evaluation of each improvement method are discussed in Subsection 3.4.1. Then, the impact of the size of the corpus is investigated in Subsection 3.4.2. Finally, we offer the comparison between our best alignment method and alignment rates obtained by state-of-the-art available models of the corresponding languages in Subsection 3.4.3.

3.4.1 Improvement methods

This section compares the contribution of each refinement method independently, with the parameters defined in Section 3.3. The relative improvements over the baseline ($T\mathcal{E}A$) computed on our evaluation database are shown in Table 3.2.

Interestingly, all techniques are seen to carry out improvement, especially at a 40 ms tolerance threshold. As expected, they help correcting gross errors, unlikely to be produced by human annotators. While both *VAD* and *AddFeat* achieve an average improvement of more than 13 % with a 40 ms threshold, *AddFeat* is shown to be much more consistent across the corpora, positively impacting the alignment of all corpora but one. Both *VAD* and *Reverse* display a quite high inter-corpus variability. This variability for *VAD* is mostly due to its low performance on under-resourced languages. When looking at English and French corpora only, the average improvement exceeds 20 % at a 40 ms tolerance threshold. Some under-resourced languages seem to be less sensitive to the benefits of this proposed method. In-depth linguistic analyses of these languages may provide greater insight into their reaction to the various improvement methods.

As previously mentioned, the optimization of the number of training iterations in the baseline allows for an average relative improvement of about 8 % at a 40 ms tolerance threshold. Experiments on the evaluation set, without this dynamic adaptation of training iterations confirmed the observations made on the development corpora. Here again, *VAD* is shown to compensate for a lack of number of training iterations as it provides similar alignment rates without this training optimization, the relative improvement exceeding then 20 % at 40 ms threshold on average, and 30 % on French and English corpora only. Conversely, *AddFeat* seems to be rather complementary to the optimization of the number of training iterations, their combination providing higher alignment rates.

We also investigated the improvement achieved by combining the methods. It is worth noting that the combination of all three improvement methods, hereafter called $T\mathcal{E}A2$, provides the best results. The improvement over *VAD* or *AddFeat* alone is statistically significant (respectively, $p < 0.001$ and $p < 0.05$ with paired t-tests). This method is also highly consistent, carrying out improvement for all evaluation corpora at 40 ms, regardless of their style and language. At 20 ms, only the alignment of

Table 3.2 - *Relative improvement (in %) over T&A with 20 and 40 ms tolerance thresholds for VAD, AddFeat, Reverse and the combination of the three methods (T&A2), for all evaluation corpora, along with the average and standard deviation of the improvement.*

Tolerance threshold	20 ms				40 ms			
Improvement method	VAD	Reverse	AddFeat	T&A2	VAD	Reverse	AddFeat	T&A2
Will (Bad guy)	0.09	2.59	20.03	19.69	34.62	4.28	43.99	40.73
Will (Neutral)	0.95	-0.19	5.15	5.15	6.77	-33.83	14.29	13.53
Woggle	30.26	24.60	18.66	37.83	30.60	20.72	21.23	36.94
Antoine (Happy)	0.48	0.16	5.94	20.22	12.36	-6.37	4.87	45.32
Margaux (Sad)	25.61	3.51	17.43	40.31	50.59	-2.38	25.42	63.18
Marie	15.95	13.61	19.68	26.25	17.53	31.25	18.93	23.92
Sportic	2.06	13.57	0.98	0.65	-11.15	9.51	11.15	9.51
Faroe	-18.00	7.19	5.63	14.42	-10.00	12.57	10.27	22.70
Gaelic	-73.62	-15.96	2.61	-85.02	5.79	1.65	10.74	12.40
Afrikaans	3.79	-7.79	9.10	22.04	-11.10	6.44	3.82	1.67
Isizulu	36.47	8.21	0.17	38.39	26.51	39.37	-2.89	46.33
Setswana	12.77	-6.91	14.36	11.54	11.50	1.53	10.81	22.75
Average	3.07	3.55	9.98	12.62	13.67	7.06	14.39	28.25
Standard deviation	28.55	10.97	7.60	33.28	19.48	18.74	12.13	18.28

Gaelic is degraded, due to the poor performances of *VAD* on that corpus, which reduces the correct alignment rate at 20 ms by about 8.5 %. This can be partly explained by the fact that Gaelic is the best-aligned corpus with *T&A* and that this baseline alignment contains very few erroneous silence boundaries. On average, *T&A2* reaches high improvement rates, with a decrease of nearly 30 % of the errors greater than 40 ms. Interestingly, *T&A2* allows reaching relative improvement at 40 ms threshold that is similar to that obtained when using bootstrapping data, at least for some corpora of our evaluation set. It should be noted that other combinations have been tested, especially the combination of *VAD* with *AddFeat* only, as *Reverse* was seen to achieve few improvement alone. Interestingly, *Reverse* is shown to contribute to the improvement of *T&A2*, increasing the relative improvement by 3.9 % at 40 ms, on average.

A detailed analysis of the errors shows that, as expected, *VAD* contributes to the reduction of the amount of errors related to silences. While, on average, 29.7 % of the silence boundaries of our evaluation set are erroneously aligned with an error greater than 40 ms with *T&A*, this rate reduces to 21.1 % when applying the *VAD* refinement. *AddFeat* is shown to reduce both the errors related to silences and to vowels (by respectively 7 and 4 % on average).

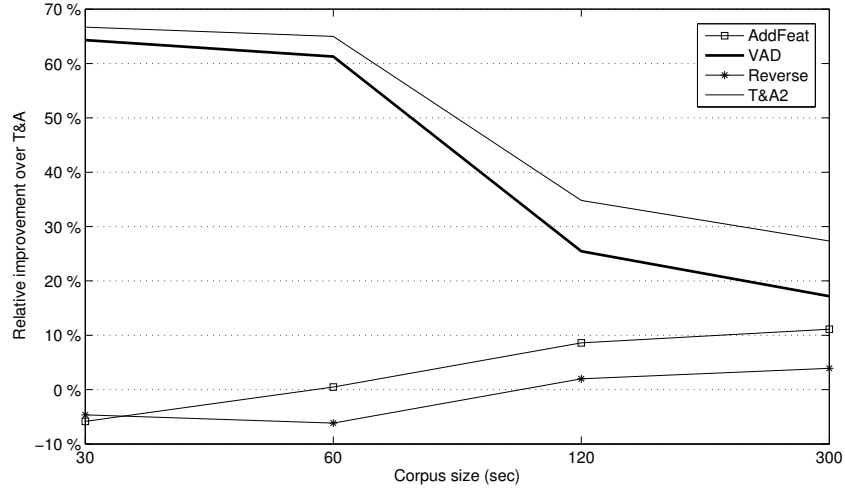


Figure 3.3 - Relative improvement (in %) over $T\&A$, with a 40 ms tolerance threshold, for the different improvement methods, with varying corpus sizes, on the evaluation dataset.

3.4.2 Advantage of $T\&A2$ in small corpora

To investigate the role played by the size of the corpus on the performance of the various improvement methods, the size of the corpora was modified to range between 30 and 300 seconds. Section 2.5.4 showed that corpora shorter than 120 to 300 seconds tend to be poorly aligned. This experiment sets out to investigate whether our improvement methods allow alleviating that issue. The results for 40 ms tolerance thresholds are shown in Figure 3.3. Similar patterns are found with a 20 ms tolerance threshold.

Both *Reverse* and *AddFeat* are shown to be ineffective on short corpora. Their positive impact, however, gradually improves with the size of the corpus. Their use for corpora shorter than 2 minutes is highly inadvisable. What should be highlighted, however, is that *VAD*, conversely to the other methods, produces significantly higher alignment rates, especially on very small corpora for which its relative improvement over the baseline ($T\&A$) becomes dramatic. This finding is of utmost importance as it allows achieving fair alignment rates on corpora for which a standard alignment performs very poorly. An example of such improvement can be observed on the 30-second version of the French expressive corpus Antoine (Happy), which shows improvements in terms of alignment rates from 7.2 % and 11.7 % (for 20 and 40 ms thresholds respectively), to 71.5 % and 88.9 % with the use of *VAD*. Consistency is found across all small corpora, with a minimum relative improvement of 32.2 % and 15.8 % for 40 ms errors, with corpora of 30 and 60 seconds respectively. Interestingly, the combined method $T\&A2$ achieves the best

results, even on the smaller corpora. This seems to indicate that, while *AddFeat* and *Reverse* are ineffective when considered separately, their combination to *VAD* still contributes to alignment improvement. It should however be noted that *T&A2* is only shown to significantly outperform *VAD* with corpora of 120 and 300 seconds ($p < 0.05$ with paired t-test).

3.4.3 Comparison with state-of-the-art models

As previously mentioned, most automatic alignment tools provide the user with pre-trained speaker-independent acoustic models of the language (*e.g.* EasyAlign (Goldman, 2011), SPPAS (Bigi and Hirst, 2012), P2FA (Yuan and Liberman, 2008)). In Chapter 2, we showed that, for medium-size corpora, training the acoustic models directly on the corpus to align achieves comparable alignment rates. We compare here the performance of these speaker-independent models with our proposed combined method: *T&A2*.

This comparison is carried out with two available French models (EasyAlign (Goldman, 2011) and SPPAS (Bigi and Hirst, 2012)) and two English models (VoxForge used by SPPAS (Bigi and Hirst, 2012) and P2FA (Yuan and Liberman, 2008)). Results are shown in Table 3.3. Compared to the best speaker-independent model, an absolute increase in the correct alignment rate of more than 10 % with a 20 ms threshold is observed for the French neutral corpus. For the English expressive database (Dellaert et al., 1996), we pointed out in Chapter 2 that P2FA achieved slightly better results than the *T&A* alignment. A possible cause was that P2FA models were trained on more than 25 hours of word-aligned speech, which required a considerable annotation time. We show here that the combination of the three improvement methods (*T&A2*) outperforms P2FA by more than 8 % absolute with a 20 ms tolerance threshold, while remaining fully automatic. It should also be noted that *T&A2* is shown to be effective on some rare languages, as pointed out for Faroe and Isizulu. The absolute improvement over *T&A* reaches in fact more than 10 % at a 20 ms threshold for Isizulu.

3.5 Conclusion

This chapter has presented three improvements techniques for the standard HMM-based phonetic forced alignment presented in Chapter 2. Based on the analysis of common errors made by this method, we proposed the integration of three types of refinement techniques:

- the use of a VAD to get a better initialization of the silence model
- the exploitation of the time-reversed sound to reduce systematic errors
- the addition of features complementary with the conventional MFCCs

Table 3.3 - Correct alignment rates (in %) achieved by available speaker-independent models (when available) and by our combined method ($T\mathcal{E}A2$), on four corpora of our evaluation dataset.

Tolerance threshold	10 ms	20 ms	30 ms	40 ms
A French neutral corpus: Marie (Colotte and Beaufort, 2005)				
<i>SPPAS</i>	45.76	71.28	83.06	89.84
<i>EasyAlign</i>	52.54	77.54	87.72	92.11
<i>T&A</i>	60.06	84.42	92.93	96.63
<i>T&A2 (proposed)</i>	65.76	88.51	94.84	97.44
An English expressive corpus: Woggle (Dellaert et al., 1996)				
<i>VoxForge</i>	37.3	65.2	82.07	88.69
<i>T&A</i>	44.95	65.23	80.07	88.82
<i>P2FA</i>	46.68	69.74	81.2	87.46
<i>T&A2 (proposed)</i>	54.01	78.38	88.92	92.95
Rare languages: Faroe				
<i>T&A</i>	48.45	75.18	87.95	93.74
<i>T&A2 (proposed)</i>	52.71	78.76	90.00	95.16
Rare languages: Isizulu				
<i>T&A</i>	46.78	72.18	85.22	90.94
<i>T&A2 (proposed)</i>	58.46	82.86	91.82	95.14

The contribution of each of these components was studied separately on a large dataset of 12 speech corpora, varying in speaking style, size and language. All three methods were shown to improve the alignment and the VAD-based technique turned out to be the most advantageous on small databases. The resulting algorithm, integrating the three refinement methods and called $T\mathcal{E}A2$, was shown to achieve the highest results. This algorithm was finally compared to state-of-the-art speaker-independent alignment techniques. Across all our experiments, $T\mathcal{E}A2$ was observed to achieve an improvement, sometimes by a substantial margin. With a tolerance threshold of 20 ms, the absolute improvement over the best existing approach is 11 % on a neutral French corpus, and 8 % on an expressive English corpus. The improvement over $T\mathcal{E}A$ is also appreciable, with notably an absolute improvement ranging from 3.5 to 10.5 % on two corpora in under-resourced languages.

Chapter 4

Train&Align: A new online tool for automatic phonetic alignment

Contents

4.1	Introduction	65
4.2	Overview of the tool	66
4.3	Perspectives	71

Abstract

In this chapter, we present Train&Align, a new tool for automatic phonetic alignment available online. Based on the method developed in Chapter 2, it proposes to train the models directly on the corpus to align. This makes it applicable to any language and speaking style. In this chapter, we provide an illustrated presentation of the tool and of its functionalities. Train&Align constitutes a rich resource for the research community in linguistics and speech technologies. It is currently used by several laboratories at the international scale.

This chapter is based upon the following publications:

- Sandrine Brognaux, Sophie Roekhaut, Thomas Drugman & Richard Beaufort. (2012). Train&Align: A New Online Tool for Automatic Phonetic Alignment. *IEEE Spoken Language Technology Workshop (SLT)*, Miami (USA).
- Sandrine Brognaux, Sophie Roekhaut, Thomas Drugman & Richard Beaufort (2014). Train&Align: Un outil d’alignement phonétique automatique disponible en ligne. *Journées d’étude de la parole (JEP)*, Le Mans (France).

We would like to warmly thank Sophie Roekhaut and Hubert Naets for their collaboration in the development and maintenance of the online tool.

4.1 Introduction

As mentioned in the previous chapters, several tools have been developed to automatically align speech with its phonetic transcription. They can be classified in two groups, regarding the simplicity of their interface and the potential control of some parameters they offer to the user:

HMM-based recognition libraries, *e.g.* HTK (Young et al., 2000) or Julius (Lee et al., 2001). These toolkits propose a set of methods that can be used to perform speech recognition (model training, model adaptation, recognition, and so on). As automatic speech alignment can be seen as a specific application of speech recognition in which the transcription of the sound is known, these method can be easily adapted to fit that purpose. The advantage is that they offer a wide range of options to optimize the training (*i.e.* different model configurations, speech frame sizes, various feature vectors, adaptation methods, and so on). However, they present several drawbacks. First, their use requires programming skills which most linguists do not have (Bigi and Hirst, 2012). The methods proposed by these toolkits should be integrated in a broader script that provides the expected input files. Besides, mastering such toolkits is time-consuming. As a result, most users stick to the basic version of the various methods and do not attempt to tune the training and alignment parameters (Cangemi et al., 2011). Not exploiting the full capacity of the methods obviously induces a lower quality of the resulting alignments.

User-friendly tools integrating existing libraries. Several tools have proposed an encapsulation of the aforementioned libraries into a user-friendly environment: The Aligner (Wightman and Talkin, 1997), nAlign (Sjölander, 2003), P2FA (Yuan and Liberman, 2008), SPRAAK (Demuynck et al., 2008), IrcamAlign (Lanchantin et al., 2008), SailAlign (Katsamanis et al., 2011), EasyAlign (Goldman, 2011), SPPAS (Bigi and Hirst, 2012). The advantage is that they require less programming skill and can be easily mastered in a short time. Some provide a graphical interface (*e.g.* EasyAlign and SPPAS) and other only offer generic methods that can be called in command line (*e.g.* P2FA, Spraak, The Aligner and SailAlign). For the latter, some programming skills are usually still required. In addition to standard HMM-based speech alignment, some of these tools also offer additional functionalities such as text phonetization (*e.g.* The Aligner, EsayAlign, SPPAS, IrcamAlign, nAlign), text syllabification (*e.g.* EasyAlign, SPPAS), the integration of phonetic variants (*e.g.* The Aligner, EasyAlign, SPPAS, IrcamAlign), the specific processing of long (possibly noisy) speech files (*e.g.* SailAlign, nAlign) or speech recognition (*e.g.* SailAlign, SPRAAK, IrcamAlign). These tools usually rely on the aforementioned HMM-based recognition libraries. To the best of our knowledge, IrcamAlign, nAlign and The Aligner are not publicly available, unlike the other aforementioned tools.

As previously mentioned, these tools present three major weaknesses:

- They provide the user with speaker-independent models. However, these models are supplied for a very limited number of languages only (see Table 2.1 in

Chapter 2).

- Most of these tools grant access to the alignment stage only. It is impossible, or very intricate, for the user to train new models, *e.g.* for an unsupported language. The issue also applies to different language varieties. For example, aligning Belgian or Canadian French on the basis of a model trained on another variety, *e.g.* "standard" French, is unlikely to produce an accurate alignment. A second issue is that the training parameters cannot be tuned. A pre-existent model is provided to the user and cannot be modified, with respect to the corpus to align.
- The provided models have usually poor generalization abilities. Ideally, they should be generic enough to produce high-quality alignment of different speaking styles: neutral speech, expressive speech, spontaneous speech, and so on. However, corpora with various emotions or speaking styles (*e.g.* sports commentaries or political speech) can have a significantly different acoustic variation that will result in a low-quality alignment. Indeed, the model provided to the user is strongly related to the corpus used for the training. It also means that, if some phonemes were rare or mis-represented in the training corpus, they will be prone to alignment errors. This issue was pointed out in Goldman and Schwab (2011) where semi-vowels were found to be badly aligned because of a possible under-representation of these phonemes in the training corpus.

The proposed Train&Align¹ tool alleviates the aforementioned issues. Like EasyAlign and SPPAS, it offers a user-friendly graphical interface implementing HTK methods. Unlike EasyAlign which only works on Windows O.S., our tool is proposed as an online application that can be accessed from any platform. It also offers the possibility to train new models. Specifically, it implements the technique presented in Chapter 2, the acoustic models being trained directly on the corpus to align. Finally, our tool also implements various training options that allow one to improve the quality of the alignment. The configuration of the models can be modified to take into account the phonetic context. The size and overlap of the speech frames can be defined by the user. With the bootstrapping option, a manually-aligned part of the corpus can also be exploited to improve the quality of the model.

This chapter presents this new online automatic phonetic alignment tool. An overview of the application is proposed in Section 4.2. Future developments are also detailed in Section 4.3.

4.2 Overview of the tool

Train&Align is a new online automatic phonetic alignment tool available at http://cental.fltr.ucl.ac.be/train_and_align. It was developed in collaboration with

¹The alignment tool was developed in Perl, based on the C-library HTK. The graphical interface was developed in PhP and Javascript.

Mrs Sophie Roekhaut and M. Hubert Naets. Its uniqueness is that it does not rely on pre-existent models of each language. The models are directly trained on the corpus to align. As a result, it can be used to align any language or speaking style. Besides, it was shown, in the previous subsections, to perform comparably or better than most existing tools. It is completely compatible with Praat formats, as it both accepts TextGrids as input and generates the aligned TextGrids as output.

The user should first register to access the tool (see Figure 4.1 and 4.2). This allows us to obtain personal information about each user. The user is also asked to confirm that he has obtained an HTK license on the HTK website.

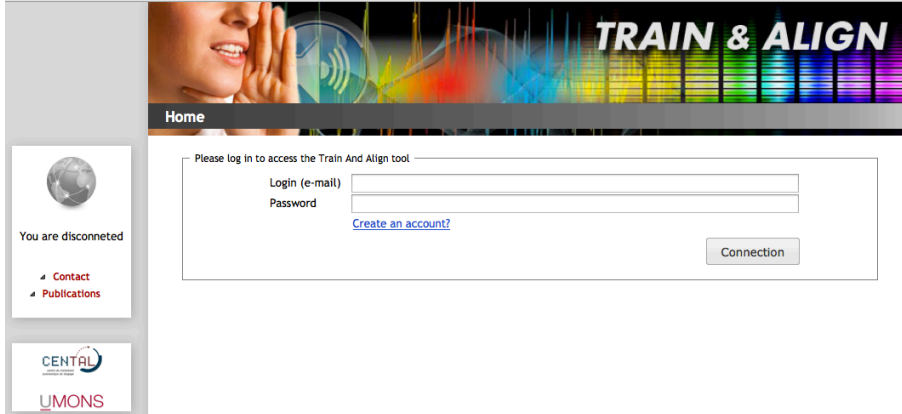


Figure 4.1 - *Train&Align graphical interface (Registration, step 1)*

The Train&Align graphical interface is shown in Figure 4.3. A complete tutorial is also available on the home page. The tool is divided into three stages.

The first stage (Figure 4.4) is the corpus uploading stage. The user is asked to name his corpus and to upload it, along with the list of phonemes corresponding to its phonetic transcription. This allows the tool to check that no transcription errors are found in the files. The training corpus directory should consist of two subdirectories, the first one containing the Audio files (in *.wav format) and the second containing their phonetic transcription. Several formats are accepted for that transcription, thereby allowing the user to provide both *.txt files or *.TextGrid files. If no correspondence between transcriptions and audio files is given by means of TextGrid, the audio files should ideally not be longer than 1 minute. Files from 2 to 5 minutes can still be accepted by the tool but the alignment process will then take much longer (from 1 to more than 10 hours) and the alignment might be of poorer quality. The training directory will be used to train the models which will align the corpus itself. A list of the phoneme characteristics should also be provided if the user wishes to use the "tied triphone" option (see above). This list should contain the various features of each of the phonetic symbols (*e.g.* voicing or place of articulation). There is no fixed number of possible features, as the user can choose

TRAIN&ALIGN: A NEW ONLINE TOOL FOR AUTOMATIC PHONETIC ALIGNMENT

TRAIN & ALIGN

Create an account

Fill in the form

Personal data

First name

Last name

Date of birth

Language

User information

E-mail

Password

Re-type the password

Licence HTK

☐ I confirm that I have registered on the [HTK](#), [Speech Recognition Toolkit](#) site

ok

Home

Figure 4.2 - *Train&Align graphical interface (Registration, step 2)*

those he believes to most acoustically distinguish between his phonetic symbols. This first stage incorporates a series of checks so as to ensure a correct working of the future training.

The interest of accepting TextGrid as input files is that Train&Align can benefit from a previously made correspondence between the transcription and the sound. If chunks of text have already been aligned with the sound, these boundaries will not be changed and only boundaries inside those chunks will be computed. Another advantage is that, if other annotation tiers are present in the TextGrid, they will also be synchronized with the aligned phonemes in the output alignment files.

The second stage (Figure 4.5) is dedicated to the parametrization of the training and of the resulting alignment. The model configuration can be determined, *i.e.* the use of monophones, triphones or tied triphones. The size and rate of the windows used to extract the acoustic parameters on the sound files can also be tuned. Finally, three options are offered to the user:

- **Train without bootstrap:** This is the standard "train & align" function.
- **Train with bootstrap:** This function allows the user to provide a manually-aligned (small) part of the corpus. This transcription is required to be in a

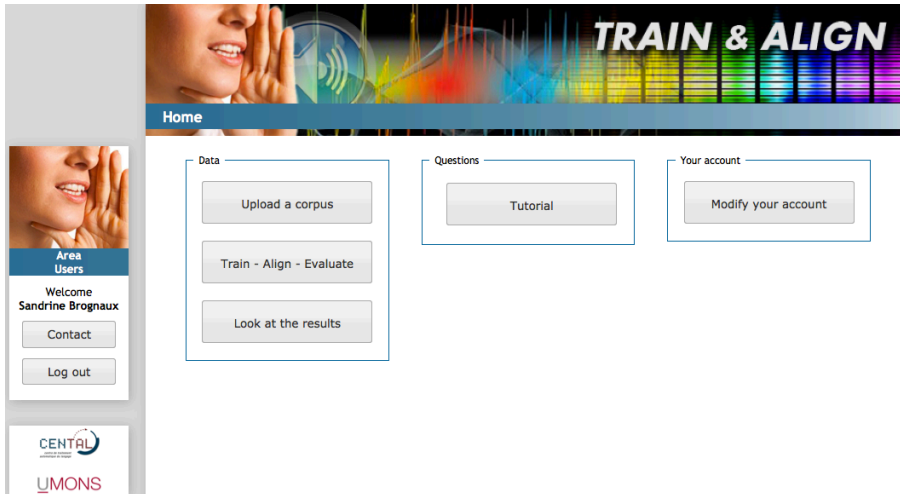


Figure 4.3 - *Train&Align graphical interface (Home)*

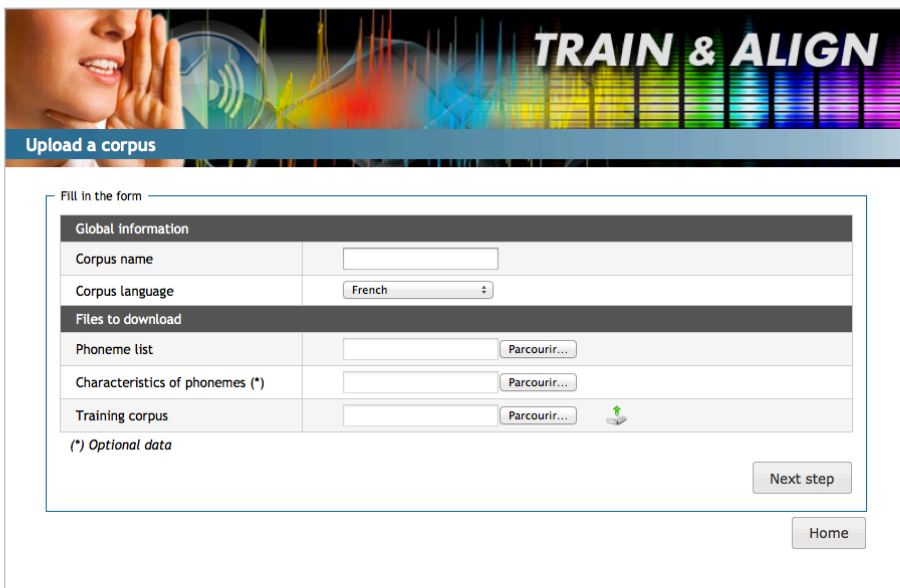


Figure 4.4 - *Train&Align graphical interface (Train and align - Stage 1)*

TextGrid format, with a tier corresponding to the aligned phonemes. This allows for a better initialization of the acoustic models at training stage, which usually results in a better alignment of the corpus.

- **Evaluate:** Another advantage of Train&Align is that it makes it possible to evaluate the quality of the produced alignment. If a manually-aligned part of the

TRAIN&ALIGN: A NEW ONLINE TOOL FOR AUTOMATIC PHONETIC ALIGNMENT

corpus is provided, it can be compared to the automatic alignment to compute alignment rates (*i.e.* the boundary-based measures, with a threshold ranging from 10 to 40 ms). This file should be provided in a TextGrid format, like a bootstrap file. The bootstrapping option comes with a specific evaluation process. Because manually-aligned data is required for both the bootstrap and the evaluation, a five-fold cross-validation can be performed. To choose this option, the same corpus should be provided for both the evaluation and the bootstrap. In that case, 4/5 of the bootstrapping dataset is used for the bootstrap and 1/5 is exploited for the evaluation, this operation being repeated 5 times such that all the annotated data is used for the evaluation. Conversely, if the evaluation corpus is different from the bootstrapping corpus, the whole bootstrapping corpus is used to improve the initialization and the final alignment is then tested on the evaluation corpus.

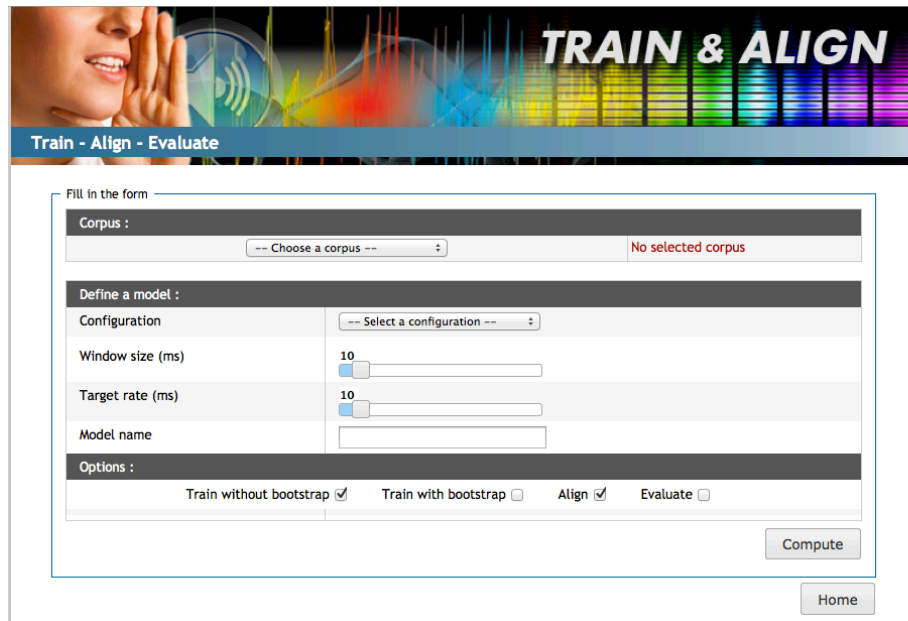


Figure 4.5 - *Train&Align graphical interface (Train and align - Stage 2)*

Finally, the last stage (Figure 4.6) displays the alignment rates (if an evaluation corpus was provided) and offers the possibility to download the aligned speech files. The alignment is provided in a *.TextGrid format with an aligned phonetic tier (as in EasyAlign, SPPAS and P2FA). All the experiments realized by the user are summarized in this same window. This is very useful to make comparisons between different training options (*e.g.* bootstrap or not, triphones or monophones). The advantage of this division of the tool into three independent stages is that the training corpus does not need to be uploaded several times. Indeed the second stage can be accessed directly and the corpus can be chosen at that time, among all the corpora uploaded by the user during his previous visits. Similarly, the results page can be accessed

directly to consult the results of the different tests. Even if the user logs out, the tests that were launched still continue until they reach the alignment. The results can therefore be checked by the user during his next visit, if the computational time is a little longer.

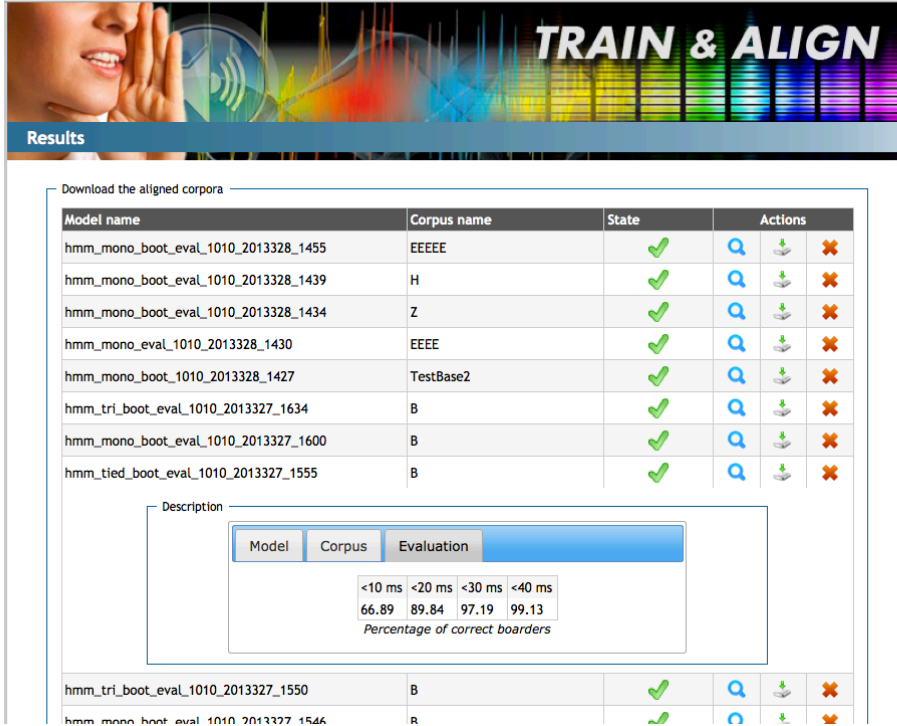


Figure 4.6 - Train&Align graphical interface (Train and align - Stage 3)

4.3 Perspectives

Several improvements should be carried out to the tool in the future:

- More input formats should be accepted. Several *.txt formats and *.TextGrid formats should be considered.
- The alignment being problematic for long speech files, the tool should provide an automatic segmentation of the speech files and of their corresponding phonetic transcription. If no phonetic transcription has been made yet, the sound files would be cut by the tool (on silences) and provided to the user so that he can phonetically transcribe all the files separately.
- Other improvements methods relying on the techniques developed in Chapter 3 should also be integrated in the tool.

TRAIN&ALIGN: A NEW ONLINE TOOL FOR AUTOMATIC PHONETIC ALIGNMENT

- Finally, the aim is to integrate Train&Align in a broader tool including eLite-HTS which is described in Section 5.5. This would allow for the alignment of the speech directly from text as eLite-HTS proposes an automatic phonetization stage. Besides, the aligned transcription could be automatically enriched with other annotation tiers, as eLite-HTS also computes syllabification, parts of speech and rhythmic groups.

In the future, the objective could also be to divide the graphical interface into two separate sections: the training and the alignment. The user could then choose a model for the alignment. This model could be trained on the corpus to align or could be a pre-existing model that the user trained on other corpora. Train&Align would then provide a useful platform for comparison of alignment rates between various models. At the long run and with user agreement, models for several languages could be made available on the website. This would facilitate the alignment of corpora that are too small and variable to reliably train a new model.

Part II

Definition of a three-level expressivity framework for the synthesis of sports commentaries

This second part of the thesis constitutes the core of our research. The synthesis of spontaneous expressive speech is investigated through HMM-based TTS of basketball commentaries. The advantage of this study is that it relies on a spontaneous corpus which contains ‘natural’ expressivity. Beside, this corpus was obtained in high recording quality with no background noise, which allows for its use in the training of a quality speech synthesizer. We mainly focus on modifications of three voice parameters which play a crucial role in expressive speech, *i.e.* prosody, voice quality and phonetic realization.

Chapter 5 provides a theoretical overview of expressive speech synthesis, with a major focus on HMM-based techniques, prosody, voice quality and phonetic modeling. It additionally presents two major resources we partly developed in the framework of this thesis and which will be used throughout Part II, *i.e.* a synthesis front-end for French and a corpus of sports commentaries. It constitutes a general introduction for the next chapters.

Prosody (and voice quality) is modified in **Chapter 6**. In that respect, we propose a two-level prosody annotation aiming at covering both local (*i.e.* accentual) phenomena and more global speaking-style changes.

The adequacy of the local annotation is further investigated in **Chapter 7** in which we explore multiple ways to model emphatic stress in HMM-based speech synthesis.

In **Chapter 8** we examine different classification methods to automatically predict our two-tier prosody annotation from text and/or acoustics. The possibility of exploiting this automatic annotation to improve the manual annotation is also explored.

Finally, **Chapter 9** concentrates on phonetic issues and investigates the part played by phonetic variation in expressive speech. Subjective evaluations of the speech synthesis of sports commentaries with, and without these variations are carried out to assess the impact of each type of variation on the naturalness of the resulting synthesis.

Chapter 5

Prosodic and phonetic improvements for HMM-based expressive TTS: A theoretical and practical introduction

Contents

5.1	Introduction	80
5.2	Phonetics in speech synthesis : The Phonetizer	82
5.2.1	The phonetization module	82
5.2.2	Integrating phonetic variation in speech synthesis	85
5.3	Prosody in speech synthesis: The Prosody Generator	88
5.3.1	Phrasing, intonation and accentuation	88
5.3.2	Existing models of prosody	93
5.3.3	Integrating prosody in text-to-speech	101
5.3.4	The question of global prosody features	109
5.4	HMM-based expressive speech synthesis: The DSP	110
5.4.1	HMM-based speech synthesis, spectral and prosodic modeling	110
5.4.2	Expressive speech synthesis: State of the art	118
5.5	Available front-ends : The development of eLite-HTS	125
5.5.1	Architecture of eLite	126
5.5.2	eLite generation of the HTS file	128
5.5.3	Specifications of the web service	130
5.5.4	Future work	131
5.6	A concrete application of expressive speech synthesis: The Sportic project	131
5.6.1	Outline of the project	131

PROSODIC AND PHONETIC IMPROVEMENTS FOR HMM-BASED
EXPRESSIVE TTS: A THEORETICAL AND PRACTICAL INTRODUCTION

5.6.2	Concept-to-speech synthesis	133
5.6.3	Corpus design	134
5.6.4	Outline of the next chapters	136

Abstract

This chapter offers a theoretical and practical introduction to Parts II and III which concentrate on the integration of phonetic and prosodic modifications in view of spontaneous expressive speech synthesis. We first provide a theoretical background on phonetics, prosody and how they are integrated in speech synthesis systems. Special attention is paid to prosodic models in French which will serve as basis for the development of our prosody annotation protocol. We also present in this chapter two major resources which will be exploited throughout Part II. A front-end component for French, called eLite-HTS has been developed in collaboration with other researchers to facilitate HMM-based speech synthesis in French. It offers the possibility of integrating a prosody annotation and a manually-checked phonetic transcription, which permits its use in spontaneous expressive speech synthesis. We also present a large corpus of sports commentaries which will be exploited as a core training corpus in the framework of our research. It presents the advantage of being spontaneous and recorded in high quality conditions, while displaying "natural" expressivity.

Section 5.5 is partly based upon the following publications:

- Sophie Roekhaut, Sandrine Brognaux, Richard Beaufort & Thomas Drugman, eLite-HTS: Un outil TAL pour la génération de synthèse HMM en français, *Journées d'étude de la parole (JEP) - Démonstrations*, Le Mans (France), 2014.
- Sophie Roekhaut, Sandrine Brognaux, Richard Beaufort & Thomas Drugman, eLite-HTS: A NLP tool for French HMM-based speech synthesis, *Interspeech - Show&Tell*, Singapore, 2014.

Many thanks to Dr. S. Audrit and Dr. B. Picart for their implication in the recording of the Sportic corpus.

5.1 Introduction

Speech synthesis relies on a complex sequence of modules to transform the text into an accurate phonetic and acoustic realization. Whatever the speech synthesis technique (*e.g.* unit concatenation, unit selection or HMM-based speech synthesis), synthesizers can basically be divided in two main parts: the natural language processor (NLP), and the digital signal processor (DSP), as shown in Figure 5.1. The first deals with the linguistic analysis of the text and is also called the "front-end" component. The second is concerned with acoustic aspects of the synthesis and is often referred to as the "back-end".

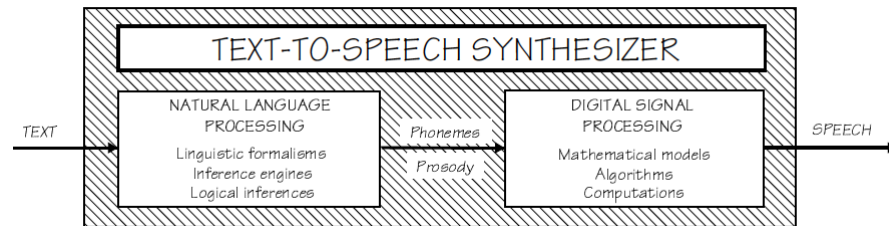


Figure 5.1 - A simple but general functional diagram of a TTS system (Dutoit, 1997a).

While the DSP differs depending on the synthesis technique, the natural language component is rather equivalent across the various techniques. For a detailed description of this component, see Dutoit (1997b) and Boite et al. (2000).

The NLP analyzes the text to transform it into a non-ambiguous representation including its phonetic transcription and prosodic information. This component is designed for a specific language and integrates linguistic resources. It usually consists of 4 main stages (see Figure 5.2):

- **The preprocessor** aims at normalizing the text. It mainly provides a segmentation of the text into utterances as well as words and rewrites numbers and abbreviations.
- **The morphosyntactic analyzer** produces the sequence of parts-of-speech from the string of words of the sentence. This is usually done with the use of a lexicon and the exploitation of the context (*e.g.* with n-grams).
- **The phonetizer** relies on both previous levels to determine a sequence of phonemes from text. Parts-of-speech notably help for the disambiguation of homophones (*e.g.* *est* in French which can be pronounced [ɛ] or [ɛst], for the verb or noun, respectively).
- **The prosody generator** exploits both morphosyntactic information and the phonetic sequence to determine the prosodic realization of the sentence, mainly

through the prediction of prosodic groups, accentuation and/or F0 and duration values.

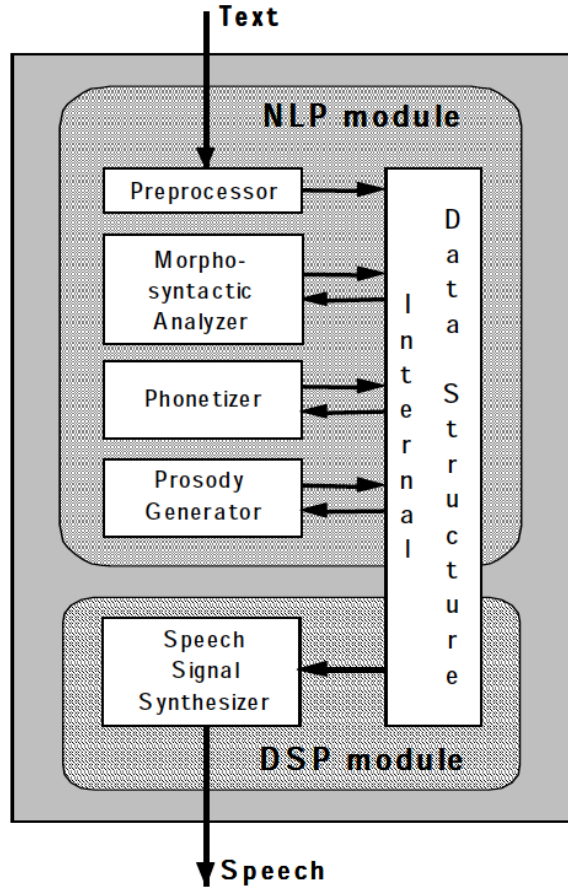


Figure 5.2 - *The functional diagram of a fairly general text-to-speech (Dutoit and Stylianou, 2003).*

While both phonetization and prosody generation depend on the preceding analysis of the text, it should be noted that these two blocks are also largely intertwined. The phonetic sequence is required to permit text syllabification, which is very useful for prosody generation, notably for accentuation prediction. However, phonetization itself may also be influenced by the prosodic organization of the sentence. In French, a well-known example of that interaction is the "liaison" phenomenon, by which a latent final consonant may be pronounced when followed by a vowel. In fact, liaisons are known to occur much less frequently at the boundary between two intonational phrases. This interaction explains the existence, in some French NLPs (*e.g.* eLite-HTS as presented in Section 5.5), of a last additional module, a **post-phonetic module** which modifies the phonetic strings based

on prosodic information. The output of the NLP component consists in both the sequence of phonemes (see Section 5.2) and a prosody representation (see Section 5.3).

This chapter focuses on three major components of the speech synthesizer: the phonetizer, the prosody generator and the DSP (here in the framework of HMM-based speech synthesis). Special emphasis is put on how prosodic and phonetic modifications can be used in speech synthesis in order to produce a naturally-sounding realization of expressive speech in French. It aims at providing a theoretical framework for the next chapters which will deal with prosodic and phonetic modeling to synthesize a specific type of expressive speech, *i.e.* sports commentaries.

The structure of the chapter is dictated by the organization of a speech synthesizer. Section 5.2 proposes a description of the phonetizer and the integration of phonetic variants in speech synthesis. The prosody generator is described in Section 5.3 along with a description of prosody and prosodic models in French. The DSP of an HMM-based speech synthesizer is summarized in Section 5.4 with a specific focus on expressive speech synthesis. In Section 5.5, we present a front-end that we developed, in collaboration with other researchers, and which will be used in the remainder of the thesis. Finally, Section 5.6 introduces a practical application of expressive speech synthesis, *i.e.* the synthesis of live sports commentaries, which will constitute the topic of this part of the thesis.

5.2 Phonetics in speech synthesis : The Phonetizer

This section focuses on the processing of phonetics in speech synthesis and provides a general outline of the phonetizer module and how it deals with phonetic variation. For a recent and more complete description of phonetics and phonology in general, see Roach (2009); Carr (2012) and Ladefoged and Johnson (2014) for English and Nguyen et al. (2005) and Léon (1992) for French.

5.2.1 The phonetization module

As shown in Figure 5.2 the text to synthesize first goes through a number of modules which provide, among others, the corresponding **sequence of phonemes**. These phonemes do not correspond to precise acoustic realizations as it is well known that a single phoneme can be pronounced in many different ways, according to contextual features (*e.g.* surrounding sounds, speaker idiosyncrasies, speech rate) (Taylor, 2009). These various realizations of a single phoneme are called allophones and link phonemes to phones, which are closer to the acoustic realization. While phones are usually annotated with the International Phonetics Alphabet (IPA), which contains about 100 symbols, there is no universally-agreed protocol for phonemic annotation. For speech synthesizers, however, ASCII characters are often used to define a new set of symbols (*e.g.* TIMIT ASCII character set from American English (Garofolo et al., 1990)).

Speech phonetizers can differ in their level of representation, ranging from broad to narrow transcription. In broad transcriptions, the symbols represent a canonical

or phonemic form whereas narrow symbols are closer to phonetics. While the latter is more precise, the use of too many symbols results in a fragmentation of the training data which leads to sparsity. Phonemic representations offer the advantage of relying on a smaller set of symbols but only give an approximate representation of the pronunciation, which cannot cover all differences (Taylor, 2009).

It is well known that the pronunciation of words may differ markedly from their orthographic transcription. For that matter, English and French are particularly complex languages in which the rule "one letter = one phoneme" can hardly be applied (Dutoit, 1997b). Therefore, most phonetizing modules of these two languages mainly rely on phonetic dictionaries. Dictionary look-ups alone are however not adequate for several reasons:

- Dictionaries usually do not include morphological variations (*i.e.* plurals, conjugated forms, and so on.)
- Dictionaries do not permit one to distinguish between heterophonic homographs (such as *est* in French, which can be pronounced [ɛ] or [ɛst], for the verb or noun, respectively)
- Dictionaries do not cover the very large set of possible words (which may contain, for instance, new words or proper names)
- Dictionaries do not cover pronunciation modifications due to the context

The first two issues are solved by the use of morpho-syntactic information provided by the previous modules of the NLP (see Figure 5.2). Morphological forms can then be used to browse the dictionary. For dictionaries with morphological entries, morpho-phonemic rules are added to account for inflection, derivation and compounding. Most heterophonic homographs can be distinguished based on their parts-of-speech (homographs differing in terms of semantics only are obviously much harder to disambiguate). One of the most widely-used phonetic dictionaries for English TTS is the freely available CMU Pronouncing Dictionary (Weide, 1998) which contains pronunciations of about 120,000 words. A well-known phonetic dictionary in French is Brulex (Content et al., 1990) which provides the pronunciation of about 36.000 words¹. Morphological dictionaries include, for example, the dictionary exploited in the MITalk system (Allen et al., 1987) covering up to 12,000 morphemes.

Beside the use of dictionaries, many TTS exploit the possibility of computing simple letter-to-sound rules and just store exceptions in a dedicated dictionary. One of the advantages of this technique is that it highly reduces the required storage space. Rules are mostly of the following type:

$$a \Rightarrow [b]/l\_r \quad (5.1)$$

¹Fonilex (Mertens and Vercammen, 1998) is a similar dictionary which contains the phonetic transcription of the most frequent word forms of Dutch as spoken in Flanders.

in which segment *a* is phonetized as *b* in a specific left (*l*) and right (*r*) context.

Such rules may be hand-derived by linguist experts or learned on aligned phonetic dictionaries. They can be implemented by transduction machines including finite state transducers and multi-level rewrite rules (see Dutoit (1997b) for a detailed description of these techniques). As a single letter can be pronounced in many different ways, these rules should rely on disambiguative information originating from different linguistic levels (*e.g.* graphemic context, syllable structure, part-of-speech, stress patterns). These rules should be ordered from the most particular to the most general. Tries have also been exploited to represent letter-to-sound rules, as shown in Figure 5.3. They offer the advantage of easily representing left and right contexts. Finally, letter-to-sound rules were also represented with neural networks, which model implicit rules, not understandable to humans (*e.g.* architectures like NETTALK (Sejnowski and Rosenberg, 1987)).

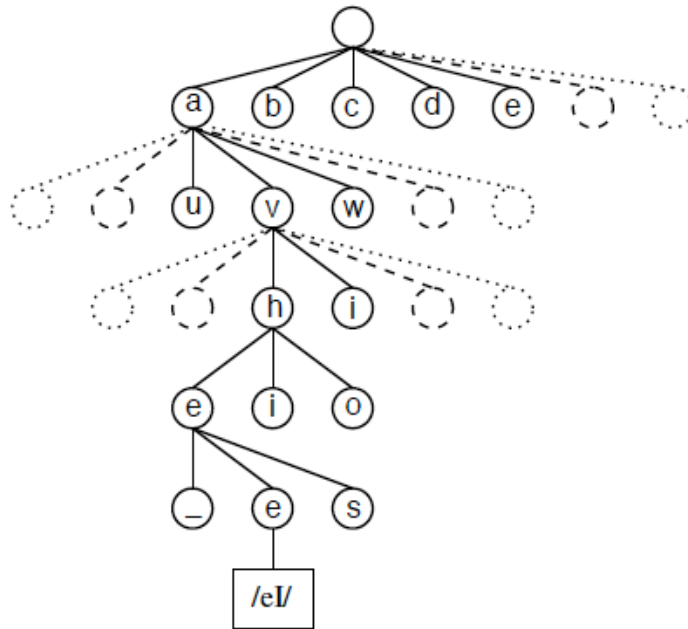


Figure 5.3 - Retrieval of the pronunciation of *< a >* in *< behave >* by trie search (Daelemans and Van den Bosch, 1993).

Letter-to-sound rules are also used to deal with out-of-vocabulary words. Specific processing of proper names may be facilitated by an etymological analysis of the nouns.

In a second stage, a phonetic post-processing is generally applied to account for some phonetic coarticulatory smoothing phenomena (*e.g.* elisions, liaisons) which are due to the phonetic context (Dutoit, 1997b). This distinction between phonetic and post-phonetic rules can be compared to the notion of lexical and post-lexical rules in Miller (1998b). While the first relates to the pronunciation that is specific to a word, post-lexical rules deal with pronunciation changes that are specific to the context. Post-phonetic rewrite rules are usually very similar to the aforementioned letter-to-sound rules but include phonetic contextual factors. They permit the consideration of modifications of word pronunciations according to the pronunciation of the previous and following ones (*e.g.* liaison phenomena).

As most NLP modules for TTS produce a phonemic representation of the text, they disregard modifications at the purely phonetic level. These modifications include changes due to the phonetic context, such as assimilation (*e.g.* *assimilation de nasalité* pronounced [asimilasjõnnazalite] instead of the expected [asimilasjõdnazalite]) or palatalization (*e.g.* the initial *t* of *tree* being pronounced [tʃ] instead of [t]) which are automatically produced to facilitate the speech flow (Dutoit, 1997b). However, as such modifications come from the close phonetic context, they should be integrated in the corresponding models (for HMM) or units (for unit-selection) at training, and therefore be adequately reproduced at the synthesis stage (Jurafsky et al., 2001a). This partly echoes the general assumption of consistency stating that, if identical phonetization errors are produced at training and at synthesis, these should not affect the quality of the produced pronunciation.

It should however be noted that:

- this mainly holds for *phoneme modifications* due to contextual features considered by the models (or units)
- this hardly generalizes to issues of *phoneme insertions* and *deletions*
- this may be questionable if multiple gross errors are made in a small speech context

This will be further investigated in Chapter 11.

On the whole, the phonetizer integrates most mandatory phonetic variations due to contextual features through post-processing letter-to-sound rules. The question that still remains, however, is how to deal with *optional* phonetic modifications. This is addressed in the next section.

5.2.2 Integrating phonetic variation in speech synthesis

In French, optional variations include four main types of phonetic modifications which will be discussed in the next sections:

1. **Schwa elisions:** These are one of the most intricate phonetic variations in French and relate to the optional pronunciation of a [ə] in the middle of a word (*e.g.* *petite* pronounced [ptit]).

2. **Final schwas:** In French, final schwas may or may not be pronounced at the end of words. We include in this category what we will refer to as ‘epenthetic schwas’, *i.e.* the insertion of a final schwa for words not ending in ‘-e’, like the word *match* being pronounced [matʃə] (Hansen, 1991).
3. **Final liquid elisions:** This category relates to two phenomena. The first is the elision of the final liquid when preceded by an obstruent (*e.g.* *peut-être* pronounced [pøtɛt]) (de Reuse, 1987)². The second is a typical elision of the [l] in the pronoun *il(s)* [il] (meaning *he* or *it*).
4. **Liaisons:** This phenomenon relates to the pronunciation of a latent consonant at the end of a word when followed by a vowel or a mute h (*e.g.* *ils vont au cinéma* pronounced [ilvõ(t)osinema], *pas ouvert* pronounced [pa(z)uvɛR]).

It should be noted that some of these variations may induce other modifications, in particular on the voicing state of the contextual consonants (*i.e.* assimilation phenomena). We may, for instance, observe the devoicing of the obstruent when the liquid is elided (*e.g.* *prendre* pronounced [prât] and not [prâd]) or the voicing of the consonant if the subsequent [ə] is elided (*e.g.* *ce ballon* pronounced [zbalo] instead of [sbalə]).

In English, optional phonetic modifications also include phoneme elisions, *e.g.* the deletion of final /t/ and /d/ (Labov, 1972), the suffix -ing which can be pronounced [ɪn] in casual speech (Labov, 2006)³, schwa elision before /r/ (Kim et al., 2004) and cluster simplifications (*e.g.* "impacts" pronounced without the "t" (Kim et al., 2004)). English also displays a high amount of vowel reductions to schwa.

The realization of words with optional phonetic variants depends on a wide range of contextual factors and many linguistic studies have investigated the role played by speech rate (Lacheret-Dujour, 1991; Fougeron and Frauenfelder, 2001; Bürki et al., 2011), word frequency (Fougeron and Frauenfelder, 2001; Fougeron et al., 2001), word probability (Jurafsky et al., 2001b), degree of articulation (Picart et al., 2010), and also sociolinguistic factors like origin (Martinet, 1971; Hambye, 2005), socio-economic status (Encrevé, 1988; Armstrong, 1996; Léon, 1992) and age of the speaker (Hambye, 2005). Few linguistic analyses however, have studied the influence exerted on phonetic variations by the degree of expressivity in speech or by the communicative situation in general (*e.g.* TV news, political speech, text reading, sports commentaries). Yet, the potential interaction between both levels is widely acknowledged (Simon et al., 2009; Bürki et al., 2011) and was shown to be very influential for prosody (Simon et al., 2009; Roekhaut et al., 2010; Pršir et al.,

²In such contexts, the liquids are in a weak position (Dell, 1995), and present a non-syllabic sonority peak more important than the preceding fricative or plosive consonant (Côté, 2004; Putska, 2011). Therefore, they violate the Sonority Sequencing Principle, which dictates that onsets must rise in sonority while codas must fall in sonority (Selkirk, 1986). As a consequence, the liquids are often deleted, giving "ouvre" pronounced [uv] instead of [uvR], and "semble" pronounced [sa~b] instead of [sa~bl]).

³The English -ng final-words pronunciation was recently compared to word-final post-obstruent liquid deletion in French in Boughton (2014)

2013). Because communicative situations cannot be easily ranked on a single scale, it should be insightful to look at the role played by diverse dimensions it contains (*e.g.* media/non media, expressive/non expressive). Interestingly, such questions have not been addressed and only the phonetic differences between read and spontaneous speech have aroused great interest (Lucci, 1983; Hansen, 1991; Fougeron and Frauenfelder, 2001). Regarding speaking style-specific phonetic variations, Van Bael et al. (2004) offers one of the rare analyses on Dutch by studying pronunciation in read speech, public lectures and telephone dialogues.

Much research on automatic **speech recognition** has focused on the modeling of optional phonetic variants (*e.g.* Ostendorf et al. (1996); Riley et al. (1999); Wester (2003); Adda-Decker (2007) for a review, see Strik and Cucchiaroni (1999); Jurafsky and Martin (2009)). Specific attention has been paid to the integration of variants typical of spontaneous speech, as speech recognizer applications rarely deal with read speech. The complexity of recognizing spontaneous speech was emphasized by Weintraub et al. (1996) who showed that the performance of the recognizer was much lower on conversational (word error rate (WER) = 53 %) than on read speech (WER = 29 %) using identical sentences, speakers and microphones. This degradation of the recognition could be due to a wide range of characteristics of spontaneous speech, including phonetic modifications but also the degree of articulation and the variety of the prosodic contours. Saraclar et al. (2000) however showed that the phonetic realization plays a crucial role in that respect and indicated, through experimental tests, that the word error rate significantly drops when the spontaneous variants of the pronounced words are integrated in the dictionary. The main issue which ASR has to face is however that including all variants of all words implies an extra confusability which degrades the Viterbi decoding (Jurafsky and Martin, 2009). The main interest has then lain in the search for the right balance between pronunciation coverage and lexicon sparsity. Ostendorf et al. (1996) investigated further than the binary distinction between read and spontaneous speech by training "style specific" models for ASR. Speaking style is here considered as a "hidden mode" which is automatically determined based on clustering of regions correctly recognized and those with a high error rate. This clustering mainly relies on prosodic features including speaking rate, normalized fundamental frequency and sound-to-noise ratio. Style-specific acoustic models can then be learned, which solves the problem of lexicon sparsity.

In comparison, very few studies have investigated the integration of optional phonetic variants in **speech synthesis**. As previously mentioned, most mandatory phonetic modifications (which are mostly due to phonetic and/or syntactic contextual factors) are modeled in existing TTS systems, through (post-processing) letter-to-sound rules. However, they are trained to produce a phonetic transcription corresponding to standard read speech. For optional variations, the most likely variant is generally produced, independently of the communicative situation.

A very insightful but rather unknown work on phonetic variation modeling for TTS is that of Miller (Miller et al., 1997; Miller, 1998b,a). His study aims at the

modeling of individual pronunciations, *i.e.* pronunciations that are specific of a speaker we want to imitate. He proposes to train a neural network to model phonetic variations, based on a window of nine phones and linguistic contextual information. This system provides a clear improvement of the predicted phonetics but presents, in our view, two main drawbacks. First, the whole study is based on neutral read speech of a single speaker, which should therefore contain little phonetic variation compared to spontaneous or expressive speech. Second, the resulting predicted pronunciations are not assessed when integrated in speech synthesis. No evaluation is made of the improvement of the resulting synthesized voice. Similar studies investigating the modeling of speaker-specific phonetic variants include Kim et al. (2004) and Bennett and Black (2005). Here again, the studies are based on read speech as recorded for neutral speech synthesis. Existing systems have also been developed to model phonetic modifications for speech synthesis of specific regional varieties (Fitt, 1997; Briony and Isard, 1997). The integration of phonetic variants in the synthesis of expressive speech has aroused surprisingly little interest. This is further discussed in Subsection 5.4.2.

5.3 Prosody in speech synthesis: The Prosody Generator

Prosody plays a crucial role in human communication and fulfills multiple functions, be they linguistic (*e.g.* sentence segmentation, hierarchical discourse structuring, focus) or paralinguistic (*e.g.* the expression of emotions and attitudes). In Dutoit (1997b), it is emphasized that the impression of naturalness of a synthesized voice is mainly due to "the richness of melodic contours and the quality of the rhythmic patterns", *i.e.* two major manifestations of prosody. An adequate integration of prosody in speech synthesis is therefore of utmost importance. The main issue is that there is no widely-agreed description or representation system. Facing the very large number of prosodic schools and theories, this question is challenging and has led to a considerable amount of research. For a detailed review of the integration of prosody in speech synthesis, see Dutoit (1997b); Boite et al. (2000); Taylor (2009) and Jurafsky and Martin (2009). We offer here a summary of the main trends with a special focus on French and on the integration of tonal approaches, which are further applied in this thesis.

5.3.1 Phrasing, intonation and accentuation

Taylor (2009), distinguishes between three main components of prosody: phrasing, prominence and intonation. A similar distinction is made in Jurafsky and Martin (2009) who speak of prosodic structure, prosodic prominence and tune. While we believe that prominence is more of a perceived prosodic phenomenon that could be seen as related to the two other manifestations, we will call this component 'accentuation' instead. In a first stage, these components are studied here in a theory-independent framework (as possible as it may be). For each component, we

then further describe its configuration in the case of French.

Phrasing

Phrasing aims at the segmentation of the speech flow. Some words naturally group together. In Jurafsky and Martin (2009), it is shown how different levels of phrases can be observed in a sentence like:

[[I wanted] [to go] [to London]], [[but I went][to Paris]]

Similar grouping is observed in French (Mertens, 1993, 2006). Prosodists generally recognize two structuring levels (Post, 2000; Jun and Fougeron, 2000):

- Minor units, *i.e.* accentual phrases (AP)
- Major units, *i.e.* intonational phrases (IP)

APs are considered as the smallest tonal unit in French and can contain more than one word. It is the domain of the primary and secondary stress (see below). They refer to what others have called phonological words, rhythmic words or rhythmic units. Their formation is based on syntactic relationships and more specifically on the dependency relations between a word carrying the stress and a clitic word (without stress). This clitic will form a prosodic unit with its syntactic head.

The IP regroups several APs and is marked by a major continuation rise or final fall. It has also been said to rely on semantic and pragmatic constraints (Selkirk, 1986). For a discussion of these two phrasing levels, see D'Imperio et al. (2007). The reader interested in further details on prosodic phrasing in French is also referred to Delais-Roussarie (1995).

The interactions between syntactic and prosodic phrasing are very complex and debated in a wealth of research (Rossi et al., 1981; Mertens et al., 2001b; Delais-Roussarie et al., 2011; Avanzi, 2012). They depend on many factors like phonetic and acoustic constraints but also on the speaker's choice. Prosody is also often seen as flatter than the syntactic structure which may contain many embedded levels (Mertens et al., 2001b).

Accentuation

Accentuation refers to the set of rules governing the position and realization of stresses in the language. Stress can be seen as the "bringing into relief of a syllable inside a word" (Garde, 1968). The notion of prominent syllables, *i.e.* syllables standing out from their environment by virtue of a certain number of perceptual and acoustic parameters (Terken, 1991) is crucial in that respect. Mertens (1991) states that a syllable "is prominent when it stands out from its context due to a local difference for some prosodic parameter. Prominence is continuous (not categorical) and contributions of multiple parameters can interact". The relation

between prominence and stress may however be complex (Mertens, 1991). A main point to make is that the accentual patterns highly depend on the language. French is a stress-fixed language, which means that the lexical stress has no phonological distinctive function (Rossi, 1980). It always falls on the last syllable of the word.

Many theories of French accentuation have been proposed (Lacheret-Dujour and Beaugendre, 1999). They classify stresses according to their structural characteristics, their acoustic realization or their function. Relations are also usually given between the three. From a *structural* point of view, a widespread distinction is made between **primary** and **secondary** stress. The first is seen as an actualization of an underlying lexical stress while the second fulfills enunciative (*e.g.* focus), expressive and/or rhythmic functions. While primary stresses fall on the last syllable of content words, secondary stresses are usually assigned to the initial syllable, or the second for words with an initial vowel and non consonantic linkage (Di Cristo and Hirst, 1996; Lacheret-Dujour and Beaugendre, 1999). There are, however, exceptions to that rule depending on the phonetic and semantic context. Regarding their acoustic realization, primary stresses are usually characterized by a lengthening of the syllable (and a potential pitch movement) while secondary stresses are more realized in terms of pitch and intensity (Lacheret-Dujour and Beaugendre, 1999). Primary stresses tend to be followed by a pause whereas secondary stresses are often preceded by a pause. Unlike primary stresses in specific syntactic configurations, the presence of a secondary stress is never essential and mainly depends on the speaker’s choice.

While Fónagy (1979) considers that there is only one type of non-final group stress, referred to as secondary stress and which serves as a rhythm regulator, most other studies define different categories of secondary stresses⁴. Rossi (1985) distinguishes between 3 types of secondary stresses according to their function: rhythmic stress, focus stress and emphatic stress. In a similar view, Di Cristo (1999, 2000) proposes describing secondary stresses according to their function, thereby proving a *functional* categorization of the French accentual framework (see Figure 5.4). This categorization will be further exploited in the remainder of this study. While lexical and nuclear stresses can be viewed as primary stresses, rhythmic, intensification or contrast stresses are various types of secondary stresses.

The distinction between lexical and nuclear stress is made considering their respective domain, *i.e.* the prosodic word or the syntactic group. The nuclear stress is usually considered as the last stressed nucleus of an intonational phrase (IP). While lexical stress in free stress languages is supposed to be morphologically or lexically distinctive and thereby belongs to the lexical entries of a dictionary, we have mentioned that this does not apply to French, which is a stress-fixed language. In that sense, most linguists agree on the fact that the French accentuation is post-lexical, *i.e.* that it is realized at a level that is above the word. Lexical stress may then be rather considered as a ‘pre-nuclear’ stress which refers to any preceding accent in the intonational phrase (D’Imperio et al., 2007). In that respect, it determines the boundaries of accentual phrases (AP).

⁴For a typology of initial stresses and their acoustic correlates, see Astesano (1999)

Regarding emphatic stress, Di Cristo considers contrast stress to have an informative objective. It can span over a large (*e.g.* the phrase) or a small domain (*e.g.* a syllable or a word). Intensification stress mainly plays an expressive (*e.g.* expression of emotions and feelings) or impressive role (to arouse emotions in the interlocutor). These two functions are also described in Séguinot (1976) and in Léon's phonostylistic functional model 2005. Both functions are of interest in the framework of our study. Intensification stress has also been shown to have a different distribution than nuclear stress. Séguinot (1976) indicates that mostly adjectives and adverbs are assigned an intensification stress while nuclear stress generally falls on nouns or verbs which delimit accentual phrases.

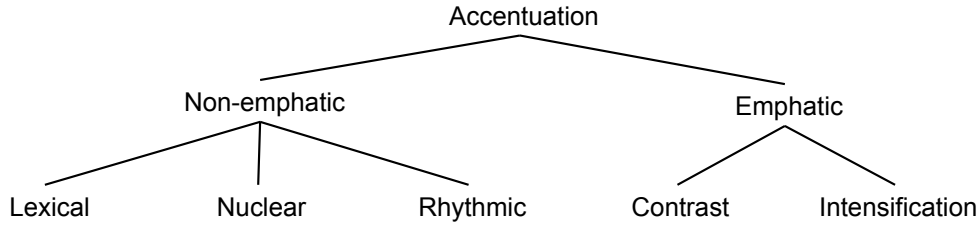


Figure 5.4 - *Elementary typology of accentual categories (Di Cristo, 2000).*

This distinction between emphatic and non-emphatic stresses is somehow linked to other distinctions made by other authors. Martinet (1965) distinguishes between passive and active functions, the latter essentially referring to emphatic stresses. Communication theorists also draw a distinction between the primary context (that is purely linguistic) and the secondary content which is linked to expression (Léon, 2005). In the same perspective, Caelen-Haumont and Bel (2003) consider two vectors of prosody. The first vector allows expressing the syntactic organization of the sentence. The second "is the one to express subjectivity and affectivity". In that respect, they draw a distinction between a structural (*cf.* non-emphatic) axis and an affective (*cf.* emphatic) axis. We will stick to the non-emphatic/emphatic terminology in the remainder of the thesis.

It should be noted that this distinction between emphatic and non emphatic stresses (as exploited in this thesis) is based on function and not on the position of the stresses. As mentioned by Rossi (1985) not all non-final stresses are emphatic stresses (see Di Cristo's rhythmic stress). He mentions notably the "melodic ictus" which can fall on any syllable and which is due to physiological and rhythmic constraints.

Intonation

Intonation is sometimes used as a synonym of prosody, sometimes just as a reference to the tonal and melodic aspects of prosody (Rossi, 1999). It is adopted here in its

restricted definition and is seen as related to pitch contours only, *i.e.* the rise and fall of the fundamental frequency over time. The same prosodic phrasing with stresses on identical syllables can still differ prosodically by displaying different tunes (Jurafsky and Martin, 2009). French is said to be an intonational language, *i.e.* a language in which pitch is used to signal syntactic and/or semantic distinctions at the phrase or at the sentence level (Gandour, 1978). Modality (*e.g.* declarative or interrogative) is one of the semantic distinctions that is brought by intonation (Grammont, 1971). Most French models focus on intonation and are further detailed in Section 5.3.2.

Relationships between the three prosodic components

It should be noted that phrasing, intonation and accentuation are heavily intertwined in French prosody and can hardly be studied independently:

- Intonation patterns are directly related to phrasing. In Rossi et al. (1981), for instance, each intonational phrase is assigned a specific pitch contour. In Mertens’ model (1987), the intonational contours are also directly related to the hierarchical structuring of the prosodic phrases.
- As carefully explained in Bardiaux (2014), there is also a syncretism between French intonation and accentuation (Carton, 1974; Di Cristo, 1999) that has led some authors to postulate the absence of accentuation in French (Rossi, 1980) or at least to consider that accentuation does not play a distinctive role in the segmentation of prosodic phrases (Vaissière, 1991; Beckman, 1992). Some authors also state that stress alone is not sufficient to mark the end of a prosodic phrase and that only the pitch contour is essential (Mertens, 1987). As shown in Subsection 5.3.2, several prosodic models offer a simultaneous annotation of both components (*e.g.* Mertens (1987); Silverman et al. (1992)).
- Accentuation in French is directly linked to phrasing as primary stresses have a demarcative function in the delimitation of prosodic phrases (Martin, 1979). The sequence of nuclear stresses determines the structure of intonational phrases (IPs). Prenuclear stresses can also be considered as indicators of the boundaries of accentual phrases (APs) (D’Imperio et al., 2007).

Prosody in expressive speech

As previously mentioned, prosody plays a crucial role in the rendering of expressive speech. This can be witnessed in the two last prosody components, *i.e.* accentuation and intonation. While producing prosody from **phrasing** only is said to convey a neutral, acceptable prosody of the sentence (Monaghan, 1989), the production of expressive speech requires the integration of other prosodic events, especially emphatic stresses and specific intonational contours.

In terms of **accentuation**, the discrepancy between neutral and expressive speech is clearly indicated by Di Cristo’s theory (1999, 2000) which distinguishes between emphatic and non-emphatic stresses. Part of the expressivity is notably borne by

the so-called "intensification stresses". On the other hand, contrast stresses can also be used to put the focus on a specific element, with an implicit contrast, thereby conveying an expressive content. Fónagy (1979, 1949) shows that emphatic speech is characterized by a higher accentual piling, adjacent syllables of a word being often accentuated simultaneously. He also observes that emphatic speech triggers more accentuation on enclitic words. Paeschke et al. (1999) analyzed the proportion of stressed syllables in emotional speech and highlighted that a higher proportion of syllables are stressed in angry speech, the lower proportion being found in neutral, bored and sad speech. Paeschke and Sendlmeier (2000) also analyzed the accentual realization in 6 acted emotions in German. They studied the duration of accents (sentence accent, word accent and final accent of the sentence) and the duration and steepness of the rising and falling parts of the accents. They found all three criteria to significantly discriminate between the emotions. For example, accents in fearful sentences are significantly shorter than in all other emotions. The steepness of F0 rise clearly discriminates between two types of emotions: sad, neutral and bored (with little steepness) and fearful, angry and happy (twice as steep as the first group). This can be seen as a correlation between F0 steepness of accents and the arousal level.

On the whole, most results tend to indicate that accentual patterns in expressive speech differ from neutral speech in terms of accent density, distribution and realization.

Regarding **intonation**, several studies have investigated the existence of pitch contours specific to (or adapted to) expressive speech. In their corpus of spontaneous expressive speech in French, Caelen-Haumont and Bel (2003) observe a very rich diversity of intonational patterns with a large variability of maximal and minimal pitch values. Paeschke and Sendlmeier (2000) also analyzed pitch movements across emotions in German and showed, for example, that pitch contour is much flatter in fear than in happiness or anger. Similar observations were made in Portuguese by Rillard (2014) (based on de Moraes and Rilliard (ress)) who points out an expansion of the amplitude of F0 contours in joyful and angry speech, and reduced contours in sad speech.

5.3.2 Existing models of prosody

A major problem of prosody modeling lies in the search for an adequate compromise between a certain degree of abstraction (*i.e.* to represent the knowledge of the speaker) and the phonetic concrete realization (*i.e.* the physical manifestation). Because of the complexity of prosodic configurations, no commonly agreed model has been adopted, as a counterpart to the International Phonetic Alphabet (IPA) for phonetics⁵. A multitude of models have been proposed to account for one or several of the three aforementioned components of prosody. This section does not aim to provide an exhaustive survey of existing models but highlights those which

⁵It should be noted that the IPA itself proposes a set of symbols specific to the encoding of suprasegmental phenomena (including stress, lengthening and phrase boundaries). This labelling framework is however largely unexploited by current studies.

have been applied to speech synthesis, with a special focus on models designed for French. For a detailed description of existing French models, see Lacheret-Dujour and Beaugendre (1999); Rossi (1999); Martin (2009). Our subsection is also based on Bardiaux (2014) who provides an insightful summary of prosodic models.

Prosody can be represented at different levels, varying in terms of degree of abstraction. The description can be acoustic (*i.e.* through the measurable physical realization), perceptive (*i.e.* as heard by the average listener) or linguistic. For the latter, prosody is considered as a sequence of abstract units that have a specific communicative function or that fulfill syntactic requirements. We will rely on this categorization, as proposed in Dutoit (1997b) and Mertens (1987), to offer a state of the art of existing prosody models.

Acoustic models

Acoustic models consider prosody as a sequence of F0 values. They aim at providing a pure description of prosodic phenomena, without interpretation, through an objective description of the signal. One of the most famous models of that type, which has been extensively applied to speech synthesis, is **Fujisaki's intonation model** (Fujisaki and Kawai, 1982; Fujisaki, 1992) (see Figure 5.5). It aims at providing an accurate description of the F0 contour that simulates the human production mechanisms. Through the use of finite-impulse-response (FIR) filters, it proposes to modify the F0 curve with two types of commands: phrase and accent commands, which offer a representation of prosodic groups and accented syllables. It makes use of impulses to produce phrase shapes and step functions to produce accent shapes. The parameters of the model correspond to the amplitude and timing of the different commands. It was specifically designed for Japanese but has been applied to other languages, including French (Bailly, 1989).

Another model that we include in this category is Taylor's **Tilt model** (Taylor, 1994, 2000), while it differs in many aspects from Fujisaki's model. Tilt was initially developed for English and was specifically designed for engineering purposes. In that sense, it is not based on biological features of speech and it neglects purely linguistic concerns. It proposes to represent the F0 contour as a series of events which represent accents and boundaries. It is similar in that sense to autosegmental and metrical (AM) models as presented below. Unlike AM models, however, it does not specify a finite set of discrete labels but assigns continuous values to the events. Six parameters ought to be defined: the shape of the event, the amplitude and duration of its rise and fall component and its alignment with the verbal component. The parameters are trained on a hand-labelled corpus.

Perceptual models

The problem of acoustic models is that "they cannot guarantee that the details that are smoothed out are not actually audible and that the ones that remain can really be heard" (Dutoit, 1997b). To answer that criticism, perceptual models have

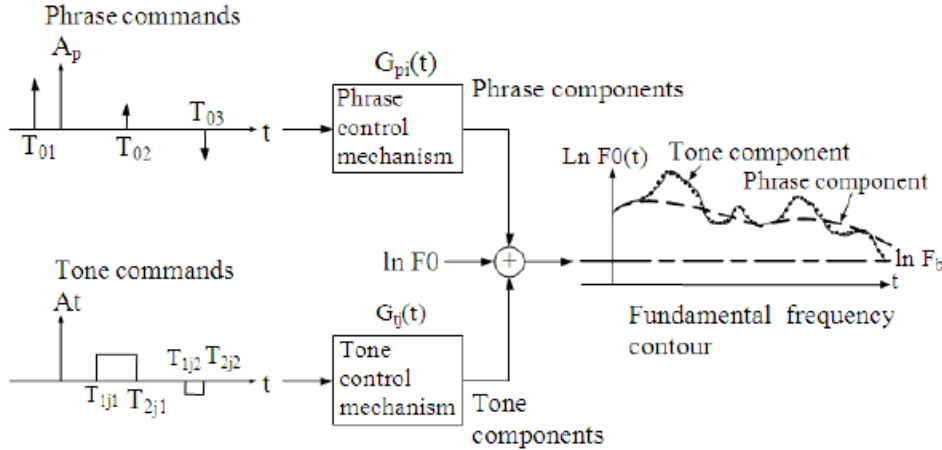


Figure 5.5 - A functional model for the process of generating F_0 contours (Fujisaki, 2004).

been developed to restrict the representation. Unlike acoustic models, which rely on parameters such as F_0 , duration and energy, perceptive models focus on perceptually relevant features only. In that sense, pitch for example, is seen in terms of semi-tones, which are more representative of the logarithmic perception of frequency.

Mertens (1987) and d'Alessandro and Mertens (1995) propose an automatic stylization of the pitch contour based on perception. For each syllable, the observed F_0 is stylized into one or more parts, which may be level, rising or falling, depending on whether or not the corresponding pitch change exceeds the glissando threshold, and hence is perceived as a changing pitch. These parts are called tonal segments. Unlike linguistic models, the set of pitch contours is larger and is not directly related to specific functions.

With similar objectives, the **IPO⁶ model** ('t Hart et al., 1990)⁷, which falls within the **Dutch school**, lies on a method of analysis by synthesis, in which sentences are resynthesized with a stylized pitch contour to check their equivalence to the original intonation. It goes one step further however by defining a limited set of standardized patterns and a grammar for their grouping into standardized pitch contours. It is in that sense closer to linguistic models described hereunder.

⁶Instituut voor PerceptieOnderzoek

⁷The IPO model has been the dominant model for 20 years until about 1990, when dominance was taken over by autosegmental models (see below).

Linguistic models

Finally, the last category relates to linguistic models. In fact, two different acoustic realizations or perceptions could be simple variants of a single function or meaning. To model this aspect, a prosodic vocabulary (of intonational events) and a grammar to organize them is needed. This is achieved by linguistic models. They aim at representing an intermediate level between the quantitative, acoustic level and linguistic functions. Di Cristo (1999) insists on the importance of such an intermediate phonological representation to avoid a direct mapping between acoustics and functions. It can basically be distinguished between two types of linguistic models: global (or holistic) and local (or tonal) models. Given the very large number of existing models, we put a special emphasis on models specifically designed for French and especially focus on Mertens' model (1987) which is exploited as a basis in the framework of our study.

Global approaches consider the intonational contour as the minimal unit. The speech flow is segmented in a sequence of pitch contours. Most of these models are inspired by the **British school** (Halliday, 1967; Crystal, 1969; O'Connor and Arnold, 1973). The British school based its model on "rises" and "falls" to describe intonation. The contour is made of different parts, the pre-head (P), the head (H), the nucleus (N) and the tail (T), according to the following configuration:

$$(P) (H) N (T)$$

where parentheses indicate optional parts. The most important part of the contour is the nucleus, which bears the accent and which can be assigned various contour configurations (*e.g.* fall, rise-fall, fall-rise). Several models of French intonation rely on a similar description in terms of pitch contours. We focus here on some of the most well known.

- **Delattre (1966)** defines a set of 10 fundamental pitch contours, on a 4-level grid, at the rhythmic group level. These contours can be distinguished by slope, curvature, initial and final levels. They rely on perceptive and functional criteria (*e.g.* finality, question, exclamation, parenthesis).
- **Martin (1979)** bases his model on minimal intonational units called prosodic words (corresponding to accentual phrases). Similarly to Delattre (1966), each unit is assigned a basic contour. The difference lies in the fact that the contours are defined in terms of binary phonological criteria (*e.g.* regarding their amplitude and length), which account for 6 basic contours. Based on this, he proposes a generative grammar of intonation. Interestingly, this model was further developed in Martin (1997, 1998) and phonological criteria were added to account for pitch contours of specific emotions and attitudes (*e.g.* doubt, surprise).
- **Rossi et al. (1981)** propose a model of intonation based on perceptually-relevant contours, through a stylization of intonation. The sentence is segmented

in intonational groups and each group is assigned a specific terminal contour. A set of 6 contours is defined (minor/major continuation, major/minor conclusion, appellative continuation and intonational parenthesis). Compared to Delattre (1966), intonational phenomena are more related to accentuation, the model being based on Di Cristo (1975) who defined anchoring points of the intonational phrases according to initial (attack) and final stresses (tonic).

Unlike global models, tonal approaches propose a discretization of the intonation continuum in a sequence of tones, which are phonological abstractions for the target points (Dutoit, 1997b). These target points are usually related to syllables, which differs from global approaches which are based on larger prosodic phrases. Most of these approaches are inspired by the **autosegmental and metrical (AM) theory** as named by Ladd (1996), which is based on Pierrehumbert's study (1980) and which was formalized in **ToBI** (Silverman et al., 1992; Beckman and Hirschberg, 1994; Pitrelli et al., 1994). This theory proposes a hierarchical prosodic structuring based on prominent syllables, which are considered as the head of prosodic components. Unlike other prosody models which mainly focus on intonation, AM models integrate all three prosodic components, *i.e.* phrasing, accentuation and intonation.

ToBI is the most famous annotation scheme for English prosody and has been widely used in speech technologies. ToBI annotation framework is made of four symbolic levels:

- the orthographic tier (with orthographic and/or phonetic transcription)
- the tone tier
- the break-index tier
- the miscellaneous tier (including comments about laughing, cough, and so on)

The two tiers directly linked to prosody are the tone tier and the break index tier. The tone tier offers a discrete representation of the intonation in terms of two tone levels: high (H) and low (L). These labels are associated with diacritics which indicate accentual characteristics. Regarding accentuation types, ToBI distinguishes between "phrasal tones" (which delimit phrasing boundaries) and "pitch accents" (inside intonational phrases). Three types of boundary tones are defined, according to their relation to the intonational phrase: intermediate (H- and L-), final (H% and L%) and initial (%H). They are directly related to the break-index tier (see below). Pitch accents are defined in terms of their pitch contour. Each label is supposed to be assigned to a specific function (see Table 5.1). Stars indicate which tone is directly associated with the accented syllable.

The break-index tier is based on Price (1991) which assigns a level ranging from 0 to 4 to phrasing boundaries. Degree "0" corresponds to no boundary, "1" occurs between two different prosodic words, "3" corresponds to an intermediate phrase boundary and "4" is assigned to boundaries of full intonational phrases. Break index "2" is slightly more complex and corresponds to (i) a strong disjuncture with a

Table 5.1 - *Tonal labels of ToBI (Pitrelli et al., 1994).*

Pitch accents		
Label	Intonative contour	Function
H*	Simple High	Canonical declarative
L*	Simple Low	Yes-no question
L+H*	Rising to high from low	Contrastive focus
L*+H	"Scooped" late rise accent	Pragmatic uncertainty
H+!H*	Fall onto stress	Pragmatic inference

(virtual) pause but no tonal mark or (ii) a disjuncture that is weaker than expected at a clear intermediate or intonational phrase boundary.

Originally designed for American English, ToBI has been extended to other languages, including French (*e.g.* Jun and Fougeron (1995); Post (2000) and more recently Delais-Roussarie et al. (2015)). Most tonal French models used in speech technology, however, slightly differ from ToBI and are based on INTSINT (Hirst and Di Cristo, 1998) or Piet Mertens' model (1987) which are described hereunder. In Lacheret-Dujour and Beaugendre (1999), these two models are considered to be the only comprehensive annotation schemes for French prosody.

INTSINT (International Transcription System for Intonation) (Hirst and Di Cristo, 1998; Campione et al., 2000) is a multilingual prosody annotation system in line with the autosegmental and metrical approach. It is meant to be "theory-neutral" in providing what could be seen as an equivalent to the IPA for prosody. In that sense, the model is considered to be more phonetic (and less phonological) than ToBI (Taylor, 2009). The model relies on a first stage, realized by the MOMEL algorithm (Hirst and Espesser, 1993) which provides an automatic stylization of the pitch. Unlike aforementioned perceptive studies, this stylization is carried out on acoustic properties only and does not integrate perceptual factors. Through a smoothing of the F0 curve, supposed to erase most micro-prosodic variations⁸, the algorithm automatically detects target points, which are considered as anchors to the pitch contours. To each of these targets, INTSINT assigns a symbol, which can be absolute and related to the speaker's range (*i.e.* 'mid', 'top' or 'bottom') or relative to the preceding symbols (*i.e.* 'higher', 'lower', 'same'). Accents can also be marked as 'downstep' or 'upstep'. In addition, INTSINT also proposes a grouping of tone sequences into intonational phrases. It should be noted that INTSINT mainly provides an annotation of intonation and does not directly deal with accentual phenomena (unlike Piet Mertens' model presented hereunder).

Finally, **Mertens** (1987; 1990; 2001b; 2006; 2008) also proposes a model to annotate accentuation and intonation in spontaneous French. Unlike INTSINT which offers a description of the F0 curve closer to the acoustics, Mertens' model only

⁸Microprosody refers to "short-term variations that can be measured (provided very accurate instrumentation is available), but that cannot be perceived (as such) and clearly have no function in intonation (although they may contribute to segmental identification)" (Dutoit, 1997b)

describes events that are considered to be relevant from a perceptive point of view. A second difference lies in the fact that the system also describes pitch movements. Syllables at positions in prosodic structure are assigned a tone, which specifies the structural position (final stress, initial stress, beginning or end of unstressed segments) and the pitch level. These labels are defined as ‘intonational morphemes’. They differ from ToBI’s labels which are only assigned to accented syllables and phrase boundaries.

The model relies on four *pitch levels* (see Figure 5.2). The number of levels is based on (i) the top and bottom of the speaker pitch range, (ii) the presence of major and (iii) minor pitch intervals. The extra high (H+) and extra low (L-) levels represent the maximal and minimal height in the speaker’s pitch range. In that sense, they rarely occur in the annotation. The size of a major interval as defined between a tone and the next is determined in a relative manner, based on the pitch range of the sentence, and is typically of about 5 semi-tones (ST). Beside the four basic levels, the system also allows for the raising (/) or lowering (\) of a tone, which defines intermediate tones. The interval with the basic tone is then usually smaller than 3 semi-tones. This direct relationship between the tones and semi-tones, and the resulting set of 8 tones, explains why Mertens’ model is often considered as more phonetic than ToBI, which relies on two tone levels only and can be seen as more phonological in that sense (Bardiaux, 2014). Finally, it should be said that, similar to INTSINT, these tones are relative: they depend both on the speaker’s pitch range and on the height of the preceding syllables.

Table 5.2 - *Pitch levels in Mertens’ model of French prosody (Mertens, 1987).*

Value	Notation
Extra high	[H+]
Raised high	[/H], [/h]
High	[H], [h]
Lowered high	[\\H], [\\h]
Raised low	[/L], [/l]
Low	[L], [l]
Lowered low	[\\L], [\\l]
Extra low	[L-], [l-]

The information on pitch level is combined with information on the position (of the syllable) within prosodic structure. Mertens distinguishes between two stress types: final and initial stresses. Final stresses can be realized with a syllable lengthening and can be either static or dynamic. For that reason, they are assigned two pitch levels. These two levels are said to be sufficient as no complex movement is usually observed on final stress in standard French. Initial stresses are static and are therefore represented with one pitch level only. The corresponding syllable is usually not lengthened and is often preceded by a small pause. Syllables corresponding to final or initial stresses are assigned pitch levels in capital letters (*e.g.* HH for final stress or H for initial stress), while those in unaccented position are labelled with

PROSODIC AND PHONETIC IMPROVEMENTS FOR HMM-BASED
EXPRESSIVE TTS: A THEORETICAL AND PRACTICAL INTRODUCTION

lowercase letters (*e.g.* h). The latter being usually static, they are represented by a single pitch level. A sequence of unaccented syllables can be modeled with just two pitch levels, *i.e.* from the first and last syllable of the sequence. These two syllables are connected with the diacritic symbol "...". It should also be noted that the model allows for register changes which are marked $\uparrow$ and $\downarrow$ and give rise to a new interpretation of L and H levels (Mertens, 1987). These changes occur between intonational phrases (see below).

The model also proposes a grouping of the syllables in terms of intonational phrases and a hierarchical structuring of these phrases. Mertens (2004b) distinguishes between accentual phrases (based on the natural structuring with clitic and non-clitic words), and intonational phrases, which are defined as a sequence of one or more syllable with the last non-schwa syllable bearing a final stress. This is in line with the description of intonation provided in Subsection 5.3.1 which indicates that nuclear stresses are final and serve as boundaries to prosodic phrases. The interested reader should however be careful with terminological issues as intonational phrases in Mertens (1987, 1990, 2006, 2008) can be considered as counterparts of APs as described in Subsection 5.3.1. In Mertens' model, accentual phrases are therefore potential phrases, which correspond to the maximal dividing of the syntactic structure, whereas intonational phrases are the effectively realized groups. Based on a corpus-based analysis of French spontaneous speech, he defines the possible morpheme sequences of intonational phrases (See Figure 5.6). Most elements are optional and the single required morpheme to form an intonational phrase is the final stress.

unstr	AI	unstr	AF	appendix
l	l	l	L-L-	l...l-
h	h	h	H+H+	
	L		HL-	
			H/H	h...h
			/HH	
			\HH	
			HL	
			LH	
			HH	
			/LL	
			LL	
			\LL	

Figure 5.6 - *Tone distribution in maximal intonational phrase (Mertens, 2006). AF = final stress, AI = initial stress, unstr = unaccented syllable.*

The hierarchical structuring of the intonational phrases is based on the height of the final stress which determines a boundary level. The association tone of final stress / degree of boundary follows the relative height (*i.e.* $L < \dots < H < /H < \dots$) with the exception of the L-L- level which is considered to determine the maximal boundary (see Mertens (1987, 1990); Mertens et al. (2001b)). The respective intonational groups are then broken away or grouped according to the boundary level between them. The grouping is based on the following rule: if, in the contour hierarchy, the morpheme of the accented syllable of a phrase is below that of the following phrase, the first phrase constitutes a sub-unit of the second. Otherwise, the two intonational phrases are said to belong to two independent prosodic units. These different levels of prosodic phrasing can be seen as the counterpart of ToBI's accentual and intonational phrases. The difference, however, lies in the accentual models of both studied languages (*i.e.* French and English). While primary stress and intonational boundary do not coincide in English (which justifies the need for a distinction between pitch accent and boundary stresses), there is a syncretism between these phenomena in French. This explains why primary stresses are sufficient in French to describe grouping in intonational phrases.

Finally, it should be noted that besides melodic and accentual differences between the intonational morphemes, they are also distinguished by their semantics. As pointed out in Mertens (2012), each prosodic form is seen to fulfill a specific function. This is further described in Mertens (2008). Morpheme L-L- indicates a final contour, while HL, for instance, indicates focus. This correspondence between intonational morphemes and specific functions is further described and exploited in Chapter 10.

5.3.3 Integrating prosody in text-to-speech

The integration of prosody in speech synthesis has been the topic of a wealth of research (see Taylor (2009) and Dutoit (1997b) for a detailed description). The general framework of prosody prediction can be seen as shown in Figure 5.7. We propose here to distinguish between three types of generation processes based on the level of prediction (from higher to lower level), *i.e.* phrasing and accentuation features, prosodic labels or acoustic values. The advantages of the three techniques depend on the type of synthesis (unit concatenation, unit selection or HMM) and are detailed below.

- **TYPE 1: The prediction of acoustic information** from text usually includes F0 values and phoneme durations. In **diphone concatenation**, they can be directly used to modify the signal. This usually results in signal degradation inherent to this type of synthesis. While a similar technique can be used for **unit selection** synthesis, acoustic values can also be used to select the best unit in the database of available units. This requires however computing a target cost to define the distance between predicted and available values. Finally, in **HMM-based** speech synthesis, acoustic values can also be used to modify the signal. However, because HMM TTS relies on parametric modeling and not on a set of pre-recorded units, this prediction of target acoustic values

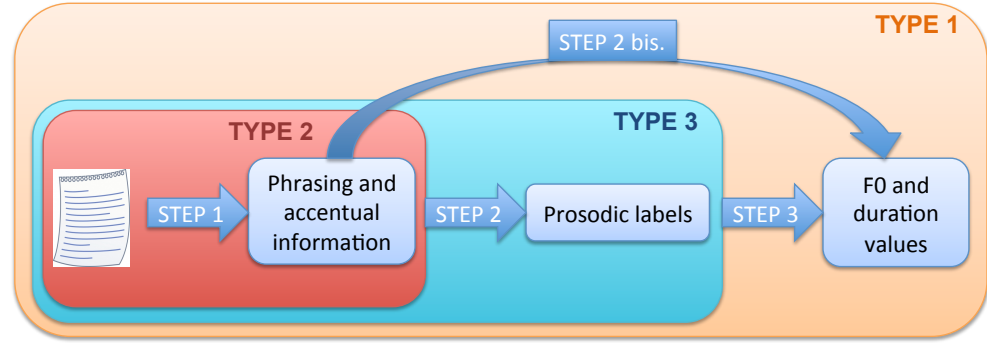


Figure 5.7 - *Different stages of prosody integration in speech synthesis, for three generations of speech synthesizers.*

does not degrade the signal, which is a clear advantage compared to diphone concatenation.

On the whole, predicting acoustic values allows a very precise prosody prediction. This requires going through 2 or 3 steps (see Figure 5.7) to transform the text into these values, which may be rather complex. Besides, in unit-selection, a target cost function needs to be implemented to account for the differences between the predicted F0 and duration values and the acoustic values of available units. This is not straightforward as it should ideally take into account perceptual criteria, *i.e.* the fact that some differences are heard more than other but also the fact that some acoustically different patterns may carry a similar function.

- **TYPE 2: The prediction of prosodic labels** from text essentially relies on existing linguistic models as described in Section 5.3.2. The rationale behind this technique is that it is assumed that linguistic features are more correlated with symbolic prosody than with the acoustic realization. Tonal models are especially used in speech synthesis. Because of the restricted set of possible labels, a direct mapping can be made in **unit selection** with units with the expected label. For unavailable units however, a target cost still requires being defined. This issue can be partly solved by attempting a maximal reduction of the number of exploited labels. For **HMM-based** speech synthesis, the prosodic label information can be directly integrated in decision trees which are exploited to distinguish between the models (see Section 5.4). In that sense, various models of a single phoneme are available and depend on contextual criteria, be they phonetic, syntactic or prosodic. The advantage of HMM-based techniques is that they do not require explicitly defining a target cost if no unit with the expected prosodic context is available in the training set. As further detailed in Section 5.4, an automatic clustering of models with similar acoustic realizations provides a generalization to unseen events.
- **TYPE 3: The prediction of accentual and phrasing information** from

text is the first step of all three types of prosody generation. Some systems, however, propose to rely on this information only without predicting prosodic labels or acoustic values. In **unit selection**, a mapping can then be done between the target linguistic context and that of existing units in the database. The main flaw compared to techniques of TYPE 2 is that the amount of contextual information can be rather large (*e.g.* position according to current/previous/next intonational phrase, position of the previous/next stress). It therefore increases the dimensionality of the feature space and all possible configurations cannot be stored in the database (Taylor, 2009). Target costs still need to be computed and their computation can be rather complex. The integration of accentual and phrasing information directly in **HMM-based** speech synthesis is similar to the integration of prosodic labels. In fact, both prosody levels are usually included in current systems. In that case, the model automatically learns, at training stage, which information best discriminates between the various realizations of a single phoneme.

The most popular models in speech synthesis are acoustic (Subsection 5.3.2) and tonal models (Subsection 5.3.2). While acoustic models allow for a rather straightforward prediction of F0 and duration values from phrasing and accentual information (TYPE 1 with step 2.bis), tonal models can be exploited alone (TYPE 2) or can be exploited to predict acoustic values (TYPE 1 with step 2). The three steps are detailed below. After a general summary of existing techniques, attention is paid to studies that have integrated them in French speech synthesis.

Finally, it should be pointed out that a clear advantage of synthesizers of TYPE 2 or 3 is that they allow for a prosody prediction that goes beyond F0 and duration prediction. It is well known that prosodic manifestations also influence other aspects of the signal like voice quality and intensity. Predicting only stress/phrasing or prosodic labels should allow for a correct realization of these other levels as they should be present in the respective training segments.

Predicting phrasing and accentual information (Step 1)

While the exact interaction between syntactic and prosodic phrasing is extremely complex (Simon et al., 2009; Delais-Roussarie et al., 2011; Avanzi, 2012), most automatic prosody generators still need to rely on syntactic information to predict **prosodic phrasing**. This can benefit from the fact that melodic phrases still often coincide with morpho-syntactic groups (Dutoit, 1997b). Most speech synthesizers rely in fact on this unique source of information to generate prosodic patterns. The prediction of phrasing is indeed much easier than that of prominence that includes the notion of focus and depends on semantical and pragmatical information which are much harder to predict (Boite et al., 2000). One or several levels of phrasing can be predicted.

Various approaches have been proposed. As usually in NLP, they range from rule-based methods to statistical techniques.

Deterministic approaches are usually based on punctuation and on the distinction between clitics (*i.e.* function words) and non-clitics (*i.e.* content words). They are referred to as "deterministic content function punctuation (DFCP) algorithms" (Taylor, 2009). Prosodic groups are determined from a local, left-to-right, low level analysis, which can be realized in real-time. Liberman and Church (1992) proposed an approach based on "chinks and chunks", shown in Figure 5.8.

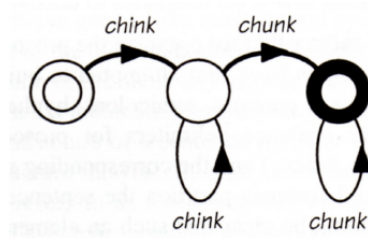


Figure 5.8 - Liberman's "chinks and chunks" (Dutoit, 1997b).

In this technique, a minor prosodic phrase is considered to be a sequence of chinks followed by a sequence of chunks. While chinks and chunks are roughly defined as function and content words, respectively, Liberman and Church (1992) propose to refine the model by slightly modifying this correspondence. While object pronouns are classified as chunks, conjugated verbs are seen as chinks. This allows for a better functioning of the algorithm which can be observed on the following English sentence:

Content/Function words	Chinks and Chunks
I asked	I asked them
Them if they were going home	If they were going home
To Idaho	To Idaho
And they said yes	And they said yes
And anticipated	And anticipated one more stop
One more stop	Before getting home
Before getting home	

As prosodic groups are said to be composed of a non-clitic word and all clitic elements depending on it (Mertens, 2004b), this algorithm considers that all clitics are on the left of the non-clitic, which is often correct but does not apply to all syntactic constructions. Beside, such approaches have been said to over-predict the number of prosodic boundaries (Taylor and Black, 1998). More complex heuristics have been proposed, as in Boula de Mareüil and d'Alessandro (1998) which relies on a set of rules determining different levels of boundaries, based on an automatic parsing of the text in terms of noun sequences, verb sequences and transjunctives (*i.e.* prepositions, conjunctions, etc). Other heuristics are based on a morpho-syntactic analysis of the sentence and rely on augmented transition networks, definite-clause grammars (DCGs) or unification grammars (UGs) (*e.g.* Allen et al. (1987); Traber

(1993)). These grammars allowing for the processing of a syntactic tree of the sentence, the degree of boundary between two words is then defined according to the height of two words common ancestor. In the Elan TTS (Boula de Mareüil and Soulage, 2001), about 150 rules are designed to predict prosodic boundaries based on the categories of the three following words, with the use of a dependency grammar.

Data driven techniques include classification and regression trees (CART) based on contextual criteria like neighboring parts-of-speech, punctuation and accentuation (*e.g.* Wang and Hirschberg (1992)) or more complex syntactic information (*e.g.* Koehn et al. (2000)). HMM-based approaches have also been proposed (Taylor and Black, 1998). More recently, CART were also used to predict style-specific phrase breaks in Parlikar and Black (2011). They make use of longer-distance contextual features (such as the number of words before the end of the sentence) and prosodic features based on a grammar that is learned to predict shallow prosodic parses from text.

Most French synthesizers rely on an adaptation of the "chinks and chunks" system (*e.g.* Malfrère and Dutoit (1997); Colotte and Beaufort (2005)). The use of morpho-syntactic parsing was included in Gaudinat and Wehrli (1997) and Obin et al. (2010b) and further integrated in the Mingus synthesizer (Mertens et al., 2001b).

Prominence information can also be predicted from text. Most studies rely on the prediction of lexical stresses only, while relying on rhythmic constraints, so as to avoid stress clashes (Taylor, 2009). Because stress in French is often said to be post-lexical and occurs at the end of intonational phrases, most speech French synthesizers rely on phrasing information only, often enriched with heuristics based on rhythmic constraints. It should however be noted that this only allows for the prediction of lexical and nuclear stresses but neglects emphatic stresses. The latter are not considered by most TTS because of their very complex prediction, which should be based on semantic and pragmatic information which can hardly be predicted from text (Boite et al., 2000). More recent studies have however tried to predict focus and emphasis information from text in English. Hirschberg (1993) proposes a pitch accent prediction based on how much new information a word contains, relying on features that correlate with these sophisticated semantics. A tree-based representation of the syntactic structure of the text to synthesize has also proved to be useful for such events prediction in Hirschberg and Rambow (2001). In Strom et al. (2006), emphasis is annotated based on capital letters or short quotations within a spoken utterance or short exclamations. Hovy et al. (2013) predict focus from a set of lexical and syntactic features, including the word POS, length, position and the sentence length.

As previously mentioned, speech synthesizers of TYPE 1 rely on phrasing and accentual information alone to generate prosody. This information is enriched by the contextual features provided by the linguistic analysis made by the front-end which include for instance the parts of speech of the neighboring words, syllable and word boundaries. Synthesizers of that type for French include Colotte and Beaufort (2005). In their unit-selection synthesizer, the units are selected based on linguistic features,

phrasing and accentual information. Phrasing is computed with an improved "chinks and chunks" algorithm and phonetic lexicons are used to infer accentual features. That type of prosody generation is also used for HMM-based speech synthesis in Lanchantin et al. (2010) and Picart (2013). This is the standard technique for HMM synthesizers if no specific attention is paid to prosody generation. Finally, Obin et al. (2010b) propose to include higher-level syntactical features to improve prosody of HMM-based French synthesis. Syntactic features integrated in the contextual information result from an automatic linguistic parsing of the sentence, based on a compact Tree Adjoining Grammar⁹. Perceptive tests showed a significant preference for this model against the baseline, which integrated basic phrasing and prominence features only.

Predicting prosodic labels (Step 2)

Tonal models as presented in Subsection 5.3.2 are usually exploited in this stage. A wealth of research has investigated the prediction of ToBI labels to generate prosody in English TTS, or in the synthesis of other languages to which ToBI was adapted. A prosodic label is generally assigned to each syllable. To predict labels from text, both deterministic and data-driven approaches have been developed. Most of these techniques rely on the previous prediction of phrasing and accentual information (*i.e.* Step 1). Rule-based approaches include Quene and Kager (1989) which automatically assigns ToBI-like labels to Dutch texts, based on criteria such as parts-of-speech and rhythmic constraints. This algorithm assigns however only three labels defining sentence accents and phrase boundaries. ToBI boundary tones are also predicted in Strom et al. (2006), with a very simple set of rules based on punctuation, POS of the sentence-initial word, and a flag indicating whether the sentence contains the word "or" immediately following a comma. This study distinguishes between different types of questions (Wh-questions, yes-no questions, and alternative questions) which are said to bear a different final boundary tone. To automatically predict ToBI labels (both pitch accents and boundary tones) for English from text, Ross and Ostendorf (1996) exploit CART and Markov models. They rely on features such as parts-of-speech, paragraph structure but also prosodic phrase structure as predicted in Step 1. In Syrdal et al. (2001), automatic prediction of ToBI labels from English texts, through machine-learning algorithms (as described in Sproat et al. (1992)) is also investigated to speed up manual annotation. They make use of various criteria including parts-of-speech, morphological information and mutual information scores based on a four-word window context.

For French, Véronis et al. (1998) proposed a probabilistic linguistic module to automatically generate INTSINT labels from text, based on part-of-speech tagging and punctuation. Piet Mertens also proposes a rule-based prediction of his labels from text based on prosodic phrasing (with the "chinks and chunks" algorithm as detailed in Subsection 5.3.3), the modality of the sentence (*e.g.* declarative or interrogative) and rhythmic constraints (Malfrère et al., 1998). These constraints help in avoiding too long sequences of unaccented syllables. A restriction of this

⁹This is computed by the Alpage Linguistic Processing Chain: <http://alpage.inria.fr/alpc.en.html>

study is that it does not predict initial stresses. In Mertens et al. (2001a), it is proposed to manually enrich the text with pragmatic intentions, by means of a set of pre-defined functional XML tags, which are exploited to derive tones, with heuristics. The "focus" tag, will result in a HL-accented word-final syllable while topic, for example, will trigger the prediction of an HH or H/H intonational boundary. This prediction is made possible by the very nature of Mertens' model which assigns specific functions to the tones (Mertens, 2008). In Mertens et al. (2001b), tones are predicted from a rich analysis of the syntactic structure of the text by the FIPS parser (Gaudinat and Wehrli, 1997), based on a set of hand-derived rules, the Mingus algorithm. Tones are defined according to the boundary level, the sentence modality and prosodic markers (such as focus and appendix). In the same line, Mertens (2008) derives the tonal structure from text based on a set of rules, relying on lexical stresses, the number of syllables, syntactic dependency relationships (based on the dependency grammar (Tesnière, 1959) enriched by the pronominal approach as defined by Blanche-Benveniste et al. (1984)) and syntactic constructions.

Speech synthesizers of TYPE 2 do not predict acoustic values from these prosodic labels but directly map training and synthesis samples/models based on these labels. Most HMM-based speech synthesizers aiming at producing natural prosody can be put in this category, as the HTS toolkit (Zen et al., 2007a) proposes to include ToBI labels as contextual information. In unit-selection synthesis, synthesizers of TYPE 2 include Wightman et al. (2000), which integrates a simplified version of ToBI ("ToBI Lite") and Strom et al. (2006), which includes ToBI boundary tones.

Predicting acoustic values (Step 2 bis. and Step 3)

Finally, synthesizers of TYPE 3 rely on the prediction of acoustic values, typically F0 and duration values.

Duration is usually predicted at the syllable or at the phone level (Taylor, 2009). As it is dictated by both phonetic and prosodic criteria, these two sources of information are usually exploited for the prediction. As for many other NLP tasks, methods range from pure heuristics to data-driven techniques. The most renown is Klatt's rules (Klatt, 1979) which were integrated in MITalk and compute the lengthening of the context-neutral intrinsic durations of phones based on several criteria including the phonetic context, stress, and phrase boundaries. Van Santen (1994) proposed a semi-trainable generalization of Klatt's rules called the "sums-of-products model". Data-driven methods include the use of CART (Riley, 1990) based on the segmental context, stress, the lexical position and the phrasing position. In Campbell (1992) the prediction of syllable durations is provided by neural networks, based on 6 contextual features, *i.e.* the number of phonemes, the nature of the syllabic peak, the position in the intonational phrase, the type of foot, stress and word class. For French, Obin et al. (2009) also propose a statistic multi-level prosodic model for duration prediction. The advantage is that it considers duration variations at different linguistic levels (syllable and interpausal group).

It should be noted that most duration prediction algorithms make use of stress

and phrasing information only, which were shown to highly influence phoneme duration. In that sense, they serve as Step 2. bis (see Figure 5.7) as they do not require a preceding prediction of prosodic labels. Exceptions to that rule include Malfrère et al. (1998) that include Mertens’ tone in the CART to predict phoneme durations. Mertens et al. (2001b) also propose a three-stage model of syllable duration prediction using a set of rules based on the syllabic position, the degree of accentuation, the number of phonemes, and, for nuclear-accented syllables, their tone.

The **prediction of F0 values** can be subdivided in two main techniques, *i.e.* the prediction from phrasing and accentuation (*i.e.* Step 2 bis) as proposed by acoustic models of prosody presented in Subsection 5.3.2 and the prediction from prosodic labels (*i.e.* Step 3), as obtained at the previous stage.

The use of acoustic models such as Fujisaki’s model or TILT facilitates the prediction of acoustic values as the parameters are given from the models themselves. The problem, however, lies in how to generate those parameters from text. Hirose et al. (2002) propose to train a model to predict the parameter values of phrase and accent commands for Japanese. This model (based on binary decision tree and multiple linear regression) relies on criteria such as the accent type, the position of the accent phrase in the sentence, and parts of speech. This gives fairly good results because of the low variability of Japanese prosody but would be applied to English or French with difficulty. Similarly, Dusterhoff et al. (1999) train a regression tree on phrasing and accentuation features to predict the TILT parameters.

Mapping from prosodic labels to acoustic values has given rise to a wealth of studies focusing on the prediction of F0 values (as shown below), as each tone level should be assigned a specific pitch value. F0 prediction from ToBI labels (Step 3) has been achieved with heuristics (Anderson et al., 1984; Jilka et al., 1999). Jilka et al. (1999), which can be seen as an extension of the APL¹⁰ algorithm (Anderson et al., 1984) assign a time position and a position in the pitch range to each tone, in a specific position in the phrase. It then makes use of interpolation to produce the pitch contours between the target points. Data-driven approaches have also been proposed, such as in Black and Hunt (1996) who use linear regression to assign 3 target F0 values to each labelled syllable. They exploit features like lexical stress, number of syllables between the current syllable and the end of the phrase, and identity of the previous labels. For French, INTSINT labels are mapped with F0 values in Véronis et al. (1998) based on a quadratic spline curve that goes through the set of target points, each target point being placed at 2/3 of the duration of the corresponding word. The frequencies F_T , F_M and F_B , corresponding to the INTSINT labels Top, Mid and Bottom are fixed and learned on the speaker. Finally, Mertens’ tones have also been used to predict F0 values in Malfrère et al. (1998) and Mertens et al. (2001b). The first relies on a dictionary of intonational patterns learned on corpus to predict F0 values for unit-selection synthesis. Each pattern contains a set of pitch targets and is defined according to contextual features (*e.g.* position in the sentence, last final accent, number of unaccented syllables). In Mingus (Mertens et al., 2001b), each tone is associated to pitch targets at different times in the syllable. The mapping between tones and frequency is defined according to

¹⁰Named after the last name of its authors, namely Anderson, Pierrehumbert and Liberman.

the speaker’s range and computed in semi-tones and depends on the declination slope.

The prediction of acoustic values was compulsory for concatenation synthesis, as prosodic modification of the waveform was required. It has subsequently been applied to unit-selection and HMM-based speech synthesis but becomes nowadays less popular than the two other types, especially for HMM-based speech synthesis for which acoustic values can be automatically learned from contextual features (phrasing/accident information or prosodic labels) as proposed in Yoshimura et al. (1999).

5.3.4 The question of global prosody features

Most of the aforementioned models, as presented in Lacheret-Dujour and Beaugendre (1999), are based on the sentence level. This restriction is pointed out by Simon (2004) who highlights the importance of prosody in its discourse context. Unlike read laboratory speech, which is prosodically monotonous, spontaneous speech displays prosodic characteristics that go beyond the sentence level. This is also pointed out by Taylor (2009) who insists on the importance of considering prosody in a wider context to model conversational, discourse, emotive or spontaneous communication. An insightful review of discourse and functional aspects of prosody is given by Hirschberg (2002).

In Obin et al. (2009), a distinction is made between 3 levels of prosodic variations:

- *Local variations*, associated with phonological properties, including accentuation.
- *Global variations*, associated with a speaker or speaking style (which are mean and variances of the prosodic parameters).
- *Intermediate variations*, associated with linguistic units that are longer than the syllable, such as intonational and accentual phrases. Modality (*e.g.* declarative or interrogative), for instance, can be seen as applying to this level ¹¹.

Our belief is that intermediate variations can be described at a local level. We have indeed shown that tonal models as presented in Section 5.3.2 propose to annotate such phenomena at a syllable level. For interrogative sentences in French, for instance, the prosodic marking is found in the accentual pattern of the last syllable. We will therefore consider only two levels of prosodic variation in the remainder of our thesis, *i.e.* local and global.

For French, the MOMEL algorithm as proposed by Hirst and Espesser (1993) also relies on the assumption that two different prosodic levels are co-existing: microprosody (which corresponds in their framework to short-term variations that are specific to segments) and macroprosody which is more global. This distinction is further detailed in De Looze (2010b). The importance of a global annotation of prosody was also highlighted in the development of the IVTS¹² annotation system

¹¹This level can be somehow compared to Morlec et al.’s global level (2001) which is related to prosodic contours.

¹²Intonational Variation Transcription System.

(Post et al., 2006), based on IViE¹³ (Grabe and Post, 2002) which dedicates a specific tier to the description of global prosody phenomena. In IVTS, the "global auditory phonetic tier" relates to phenomena that fall beyond the "Domaine d'association tonale" (ID), the ID being roughly defined as accentual phrases with the addition of the following syllable.

Global prosodic features include, among other, pitch range, speech rate, average intensity, accentual density and the proportion of rising syllables. This information can be derived from the summation of syllabic values. Such global features also apply to voice quality, some speech segments being, for instance, realized with a more creaky or breathy voice. This macro level can be somehow associated to what other authors have called "speaking styles" (as used in many speech synthesis studies like Abe (1997)), "speech styles" (Kern, 2010), "voice style" (Székely et al., 2011), "prosodies" (Odgen and Walker, 2001), "discourse ambiances" (Goldman, 2012) or even "phonogenres" or "phonostyles" (Goldman et al., 2009). While "phonostyle" usually refers to a specific communication situation (*e.g.* interview, read speech, political speech), the term "speaking style" is often seen as more generic, gathering also specific emotional states and attitudes. This term will be kept for the remainder of the thesis and is further defined in the next chapter.

Global changes play a crucial role in **expressive speech** (Léon, 1970). Expressive speech has been shown to be characterized by a higher pitch and intensity level (as shown on French emphatic speech by Fónagy (1979)). Léon (1970) also indicates that intonation modifications in expressive speech do not only appear at stress locations but are related to the entire utterance (*e.g.* register, intensity, duration). In their qualitative study of French emphatic speech Devillers and Vasilescu (2003) noticed that, among a large set of acoustic parameters, essentially the range of F0 and the maximal cross variations were relevant to that kind of speech. Murray and Arnott (1995, 2008) also proposed a global prosodic analysis of acted emotional speech and pointed out a correlation between various emotions (*i.e.* anger, happiness, sadness, fear and disgust) and global pitch, speed and intensity features. Intensity, for example, is higher in anger and happiness than in sadness. A wider range of pitch is also observed in these two first emotions. Fear is also shown to have a faster speech rate than sadness. A similar study was carried out by Hirose et al. (2000) who showed that pitch is the best discriminator of emotions in Japanese. Speech rate is also found to be increased in anger but decreased in sadness. Similar studies include Bänziger and Scherer (2005) who showed that F0 is mainly influenced by emotion arousal.

¹³Intonational Variation in English.

5.4 HMM-based expressive speech synthesis: The DSP

5.4.1 HMM-based speech synthesis, spectral and prosodic modeling

Most current research on speech synthesis concentrates on HMM-based speech synthesis which offers a greater variability compared to unit-selection synthesis while requiring less storage space. HMM-based speech synthesis relies on speech recognition techniques as described in the first part of this thesis. We provide here a brief illustrated summary of its global working, with a specific focus on how to modify prosody and voice quality as this constitutes a major objective of our study. For a more detailed description of HMM-based TTS, see Yoshimura et al. (1999); Masuko (2002); Tokuda et al. (2002b); Zen et al. (2007b, 2009) and Tokuda et al. (2013) for a more recent review.

The general framework of HMM-based speech synthesis is shown in Figure 5.9. As for all data-based synthesizers, it is mainly divided in two stages: training and synthesis.

The **training stage** is very similar to model training in speech recognition. The model parameters are estimated based on a Maximum Likelihood (ML) criterion as follows:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \{P(O|\omega, \lambda)\} \quad (5.2)$$

where λ is the set of model parameters, O is the training data (speech signal) and ω is the word sequence corresponding to observation O . This is computed by means of the Expectation Maximization (EM) algorithm (Dempster et al., 1977), as described in Yamagishi and Kobayashi (2007).

The features of the training data are extracted from a database of natural speech. They consist of *spectral features* (usually mel-frequency cepstral coefficients) and, unlike speech recognition, *excitation features* (usually log F0) along with their dynamics (delta and acceleration coefficients), as shown in Figure 5.10. They are modeled by a set of multi-stream context-dependent HMMs (Young et al., 2000). One stream is dedicated to spectral features which are modeled by means of multivariate Gaussian distributions. The second stream relates to excitation. Because of the very nature of pitch contours which are not continuous (*i.e.* for unvoiced regions), a similar distribution cannot be used. Excitation is therefore modeled with Multi-Space probability Distributions (MSD) (Tokuda et al., 1999). The integration of MSD in the HMM framework is referred to as MSD-HMM (Tokuda et al., 1999, 2002a). They include discrete or continuous distributions for the modeling of unvoiced or voiced regions of speech, respectively.

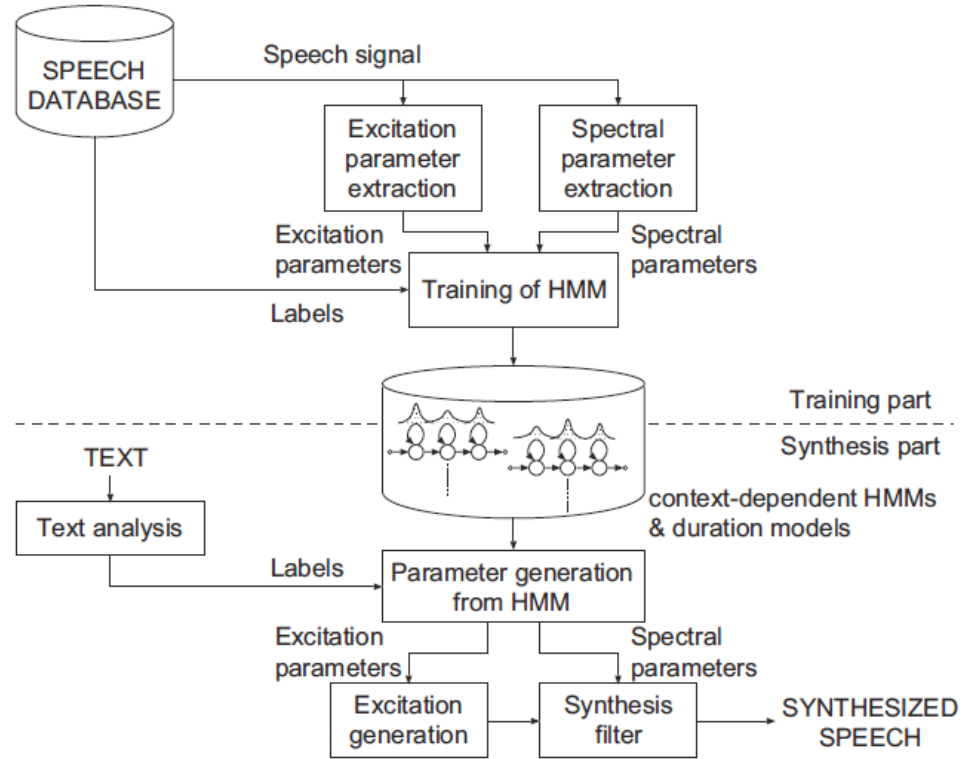


Figure 5.9 - Block diagram of HMM-based speech synthesis system (HTS) (Zen et al., 2007b).

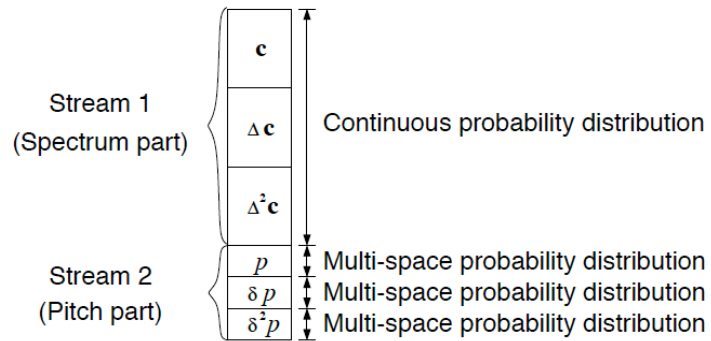


Figure 5.10 - Feature vector (Yoshimura et al., 1999).

Finally, each HMM is associated with a *state duration* distribution which is estimated at the last stage of the forward-backward algorithm by using state occupancy probabilities. They consist in multivariate Gaussian distributions, the dimensionality corresponding to the number of states in the model. This combination of HMM with explicit state duration models is called a Hidden Semi-Markov Model (HSMM) (Zen et al., 2007c). Spectrum, F0 and duration are thus modeled in a unified framework.

Each model corresponds to a phoneme and is context-dependent. Unlike speech recognition where only the phonetic context is usually considered by the models (*i.e.* by means of triphones or tied-state triphones as described in Chapter 2), the context considered is much wider in speech synthesis and includes various linguistic annotation levels, as produced by the front-end. For example, the contexts used in the HTS English synthesizer (Zen et al., 2007a) are:

- Phoneme:
 - current phoneme
 - preceding and succeeding two phonemes
 - position of current phoneme within current syllable
- Syllable:
 - numbers of phonemes in preceding, current and succeeding syllables
 - stress and accent of preceding, current and succeeding syllables
 - position of current syllable within current word and phrase
 - numbers of preceding and succeeding stressed syllables within current phrase
 - numbers of preceding and succeeding accented syllables within current phrase
 - number of syllables from previous stressed syllable
 - number of syllables to next stressed syllable
 - number of syllables from previous accented syllable
 - number of syllables to next accented syllable
 - vowel identity within current syllable
- Word:
 - guess at part of speech of preceding, current, and succeeding words
 - numbers of syllables within preceding, current, and succeeding words
 - position of current word within current phrase
 - number of preceding and succeeding content words within current phrase
 - number of words from previous content word
 - number of words to next content word

- Phrase:
 - numbers of syllables within preceding, current, and succeeding phrases
 - position of current phrase in major phrases (or utterances)
 - ToBI endtone of current phrase
- Utterance:
 - numbers of syllables, words, and phrases in utterance

Because of the large range of contextual features, and the resulting exponential number of combinations, a good coverage cannot be provided from a limited training set. To alleviate that issue, a decision tree context clustering technique is applied to the distributions of spectrum, duration and F0 (that technique has been extended to MSD-HMMs in Yoshimura (2002)). As the influence of the context may be different for those three types of features, three independent trees are built. An example of decision tree clustering for state duration models is shown in Figure 5.11.

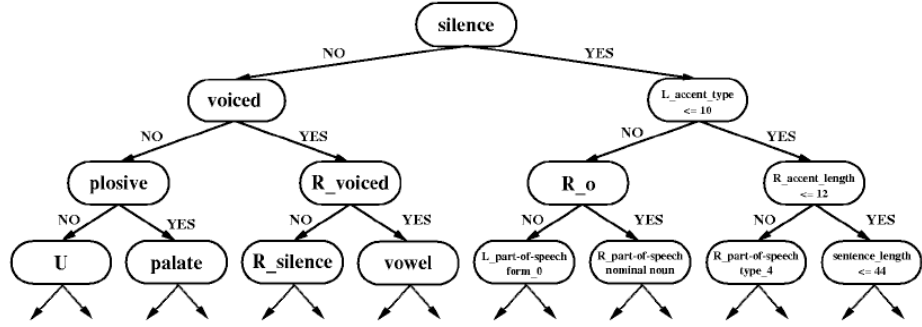


Figure 5.11 - Decision tree for state duration models (Yoshimura et al., 1998).

The **synthesis stage** is illustrated in Figure 5.12. Four main steps can be distinguished:

1. The text to be synthesized is transformed into context-dependent labels by the front-end component
2. The respective HMMs (obtained from the training stage) are concatenated for both spectrum and F0, and state durations are estimated to maximize the likelihood of state duration probability densities
3. The speech features (typically mel-cepstral coefficients and log F0) are generated by maximizing the output probabilities of the models as:

$$\hat{o} = \underset{o}{\operatorname{argmax}} \{P(o|\omega, \hat{\lambda})\} \quad (5.3)$$

where $\hat{o}$ are the generated speech features, ω is the word sequence to be synthesized and $\hat{\lambda}$ is the set of estimated parameters, learned during the training stage.

4. The speech signal is generated from excitation and spectral features with a speech synthesis filter, *e.g.* the Mel Log Spectrum Approximation (MLSA) filter (Imai et al., 1983). It should be noted that the generation of the signal from such a filter results in a somehow "buzzy" synthesized voice, which is often pointed out as a major drawback of HMM-based speech synthesis. Research has therefore been carried out to develop high-quality vocoders to alleviate that issue (*e.g.* The Mixed Excitation Linear Prediction (Yoshimura et al., 2001; Gonzalvo et al., 2007), the glottal-flow derivative model (Cabral et al., 2007, 2008) or the Speech Transformation and Representation Using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) (Kawahara et al., 1999; Kawahara and Morise, 2011)). In the framework of this thesis, we will mainly rely on the Deterministic plus Stochastic Model (DSM) of the residual signal, as presented in Drugman and Dutoit (2012).

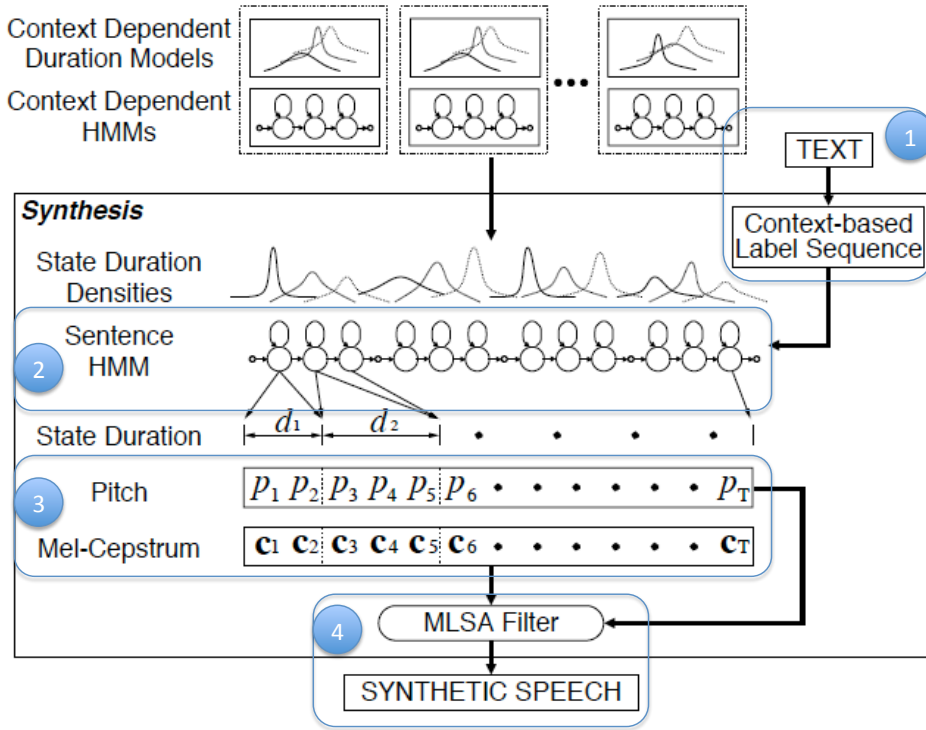


Figure 5.12 - *Synthesis part of the system (Yoshimura et al., 1999).*

A major advantage of HMM-based speech synthesis lies in the possibility to easily modify voice characteristics. **Adaptation techniques** were initially developed in

speech recognition to adapt the models to a specific speaker, thereby improving the agreement between the model and the production to recognize. Because HMM-based speech synthesis exploits similar paradigms, these adaptation techniques can be applied to adapt speaker-dependent models to a specific speaking style or emotion, using only a (possibly) small amount of speech adaptation data. We summarize here two major adaptation techniques, which will be exploited in the remainder of this thesis: Maximum A Posteriori (MAP) estimation (Gauvain and Lee, 1994; Tsurumi and Nakagawa, 1994; Masuko et al., 1997) and Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995; Tamura et al., 1998).

The **MAP** adaptation requires knowing what the parameters are likely to be as the training incorporates prior information of model parameters and new incoming training data. Let μ_i and Σ_i be the mean vector and covariance matrix of the i -th state output distribution, respectively. For a given training sample $O = O_1, \dots, O_N$, the new mean vector $\bar{\mu}_i$ and covariance matrix $\bar{\Sigma}_i$ are estimated as follows:

$$\bar{\mu}_i = \frac{\alpha\mu_i + \sum_{j=1}^N O_j}{\alpha + N} \quad (5.4)$$

$$\bar{\Sigma}_i = \frac{\sum_{j=1}^N O_j O_j' - (\alpha + N)\bar{\mu}_i \bar{\mu}_i' + \frac{\beta}{\alpha}\Sigma_i + \alpha\mu_i \mu_i'}{\beta + N} \quad (5.5)$$

where α and β are constants to control the influence of prior information. The major drawback of MAP adaptation is that every Gaussian is adapted individually, which causes issues if the data are sparse. Solutions have been developed to alleviate that issue, the most renown being the Vector Field Smoothing (VFS) algorithm (Takahashi and Sagayama, 1995) which (i) interpolates new parameters and (ii) smoothes the estimated parameters.

MLLR is known to perform better than MAP if the amount of adaptation data is limited, as it does not require estimating individual adaptations of each Gaussian distribution (Zen et al., 2009). A second advantage is that fewer parameters should be controlled, the only parameter being the number of regression matrices (which can be set to one for very small amounts of adaptation data). It relies on a set of linear transforms which are used to map an existing model to the target model. The state-output of the adapted model is a linear transformation of the original mean vector and covariance matrices as:

$$\bar{\mu}_i = \zeta_k \mu_i + \epsilon_k \quad (5.6)$$

$$\bar{\Sigma}_i = H_k \Sigma_i H_k' \quad (5.7)$$

where $\bar{\mu}_i$ and $\bar{\Sigma}_i$ correspond to the linearly transformed mean vector and covariance matrix of the i -th state output distribution and ζ_k, H_k and ϵ_k correspond

to the mean linear-transformation matrix, the covariance linear-transformation matrix, and the mean bias vector for the k -th regression class (Zen et al., 2009). For details about the computation of the regression matrices, see Tamura et al. (1998). The state-output distributions "are usually clustered by a regression-class tree, and transformation matrices and bias vectors are shared among state-output distributions clustered into the same regression class" (see Figure 5.13)(Zen et al., 2009). In case the same transforms are trained for ζ and H , it is referred to as Constrained Maximum-Likelihood Linear Regression (CMLLR).

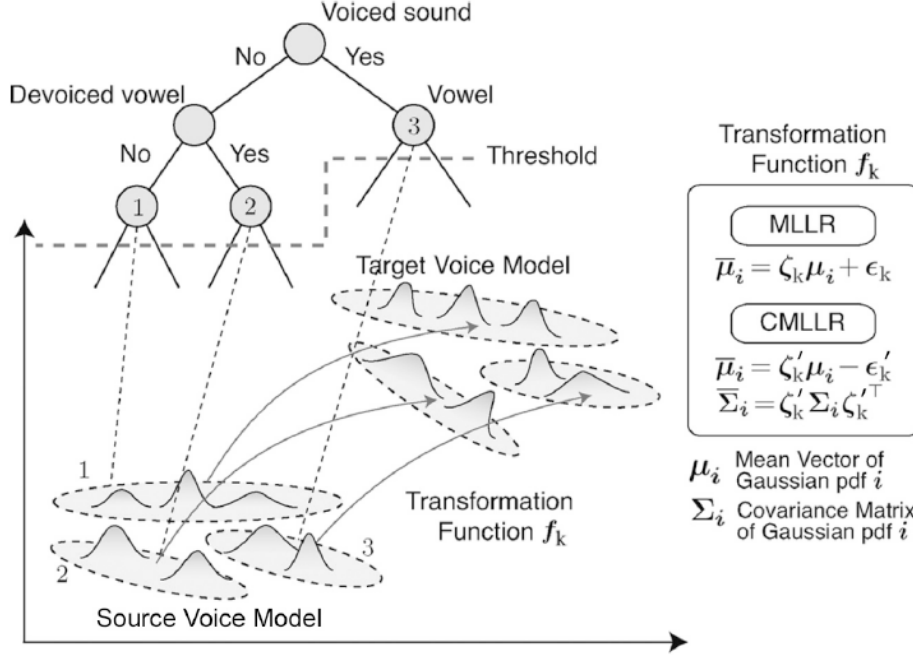


Figure 5.13 - *Constrained Maximum-Likelihood Linear Regression (CMLLR) and its related algorithms (Yamagishi et al., 2009b), as modified in Picart (2013).*

An interesting fact is that both techniques can be judiciously combined in a hybrid algorithm, implemented in HTS (see below)¹⁴. While MLLR proposes to apply a linear regression to the average model to target a new speaker or speaking style, MAP can be additionally applied to the model transformed by the linear regression to upgrade the distributions containing a sufficient amount of speech samples (see Figure 5.14). This method will be used in the framework of our research.

HTS (HMM-Based Speech Synthesis System) is an open-source software toolkit released as a patch code to HTK (Young et al., 2000), which provides a research

¹⁴The reader interested in technical details of this method is referred to Picart (2013).

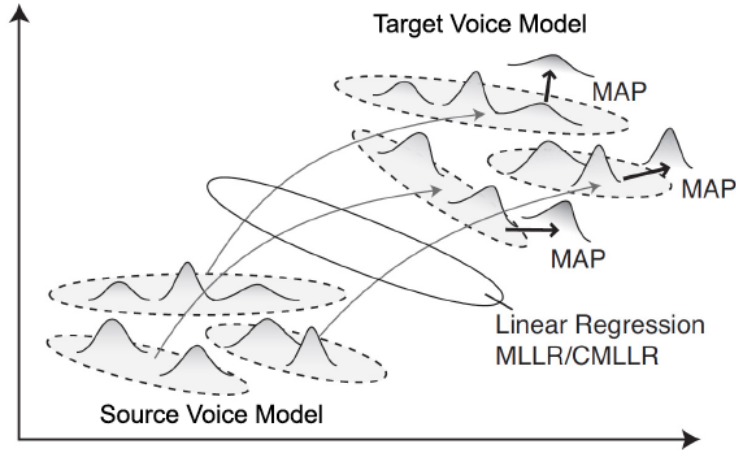


Figure 5.14 - Combined algorithm of the linear regression and MAP adaptation (Yamagishi et al., 2009a), as modified in Picart (2013).

and development platform for the speech synthesis community (Zen et al., 2007a). It implements all the necessary algorithms to build HMM-based synthetic voices and includes additional methods for speech adaptation, conversion, and interpolation. Specifically, HTS implements both CMLLR and MAP-VFS adaptation methods and makes it possible to combine them. For feature extraction, it relies on the Speech Signal Processing Toolkit (SPTK) (SPTK Working Group, 2009). HTS will be used in the remainder of this thesis. It should however be noted that it only consists in the DSP stage of the synthesis and does not provide any text analyzer (*i.e.* the front-end part of the synthesizer). For that matter, Zen et al. (2007a) refer to existing systems including the Festival Speech Synthesis System (Black et al., 2001), which is currently available in a variety of languages including English and Spanish. Festival was however not fully developed for French. Therefore, we explore in Section 5.5 available front-ends for French.

5.4.2 Expressive speech synthesis: State of the art

In this section, we provide an overview of proposed techniques to automatically generate expressive synthesized speech. We first briefly present the general methods used in formant, unit concatenation and unit selection synthesis. We then focus on studies exploiting HMM-based synthesis and point out at the main trends, their benefits and drawbacks.

The integration of expressiveness in speech synthesis has been a topic of research for decades. For an overview of the history of expressive speech synthesis, the interested reader may refer to Schröder (2001) and Burkhardt and Stegmann (2009).

Formant synthesis has the advantage of providing a high degree of flexibility

and was therefore seen as an interesting ground for experimentation. Studies include Montero et al. (1998) who proposed a set of rules to modify prosodic features such as speaking rate, F0 range and spectral tilt to produce expressive Spanish synthesis (*i.e.* sad, happy and angry speech). In the same line, the HAMLET synthesizer (Murray and Arnott, 1995) was implemented for English and provided a set of dedicated rules (modifying voice quality, pitch and duration) to model anger, sadness, happiness, fear, disgust and grief. Similar rules, aiming at modifying pitch, timing, voice quality and articulation were also implemented in the Affect Editor (Cahn, 1990), based on the DecTALK 3 formant-based synthesizer. The main drawback of formant synthesis lies however in its resulting low signal quality.

Integration of expressivity in **unit concatenation synthesis** usually relies on the modification of F0 and duration, sometimes through a "copy" synthesis, in which the acoustic values of the original emotion are transposed to the synthesized voice. To simulate emotional arousal, Emofilt (Burkhardt, 2005) proposes to define a set of parametrized rules, describing manipulation of pitch, duration, voice quality and articulatory features based on the MBROLA synthesis engine. It supports 35 languages. Schröder (2004) also proposed emotional prosodic rules for German speech synthesis to adapt fundamental frequency and tempo according to a dimensional approach of emotions, based on activation, evaluation and power. The main drawback is that modifications of the signal result in a degraded quality.

This is usually solved by **unit selection** techniques, which can exploit the richness of the database containing various units of a same phonetic content. Studies include the Japanese CHATAKO system (Iida et al., 2000) which relied on the recording of emotion-specific corpora (*i.e.* sad, angry and happy). They exploited a large database containing 400-500 sentences per emotion, uttered by non-professionals to avoid exaggerated emotions. The main issue lies here in the large required database which is very costly (Black, 2003). In fact, databases as used in unit-selection are usually very large (typically about 10 hours) and require much recording and annotation time. In the same trend, the IBM expressive synthesis system (Eide et al., 2004; Hamza et al., 2004) proposes to include recordings of various expressive styles (*e.g.* conveying bad news or good news) in the unit-selection database and to adapt the target cost to select the adequate unit at synthesis time. In addition, data from the various styles is also used to train style-specific prosodic models. Perceptive evaluation shows that this improves the overall quality of the synthesis.

It should be mentioned that, aside the aforementioned studies, a whole body of research has focused on voice conversion techniques to transform neutral speech into emotional speech (Kawanami et al., 1999; Türk et al., 2005; Türk and Schröder, 2008; Inanoglu and Young, 2007). These studies often focus on voice quality modifications (Kawanami et al., 1999; Türk et al., 2005; Türk and Schröder, 2008). Türk et al. (2005) propose an algorithm to automatically interpolate new databases for unit concatenation (based on MBROLA), with intermediate levels of vocal effort. For a comparison of voice conversion techniques for emotional speech generation, see Türk and Schröder (2008).

HMM-based speech synthesis provides an ideal framework for the development of expressive speech generation, being extremely flexible and of high quality. It has therefore become the favorite playground of most experts in expressive speech synthesis. For an insightful summary of the domain, the interested reader should refer to Nose and Kobayashi (2011).

Beside the basic training on a corpus tailored for a specific emotion called *style-dependent model* (Yamagishi et al., 2005), similarly to what has been done in unit selection, we propose here to classify the methods into three main trends:

- the use of adaptation techniques
- the modification of the tree-based clustering, *e.g.* through the exploitation of additional contextual features
- post-processing techniques directly modifying F0 and duration values

Expressive speech synthesis relying on *adaptation techniques* includes Yamagishi et al. (2004); Qin et al. (2006); Tachibana et al. (2006) and Eyben et al. (2012). A major advantage of adaptation techniques is that they only require a very small amount of expressive speech data, used to adapt an already trained (neutral) model. They allow for the adaptation of both voice quality (with an adaptation of the spectral coefficients) and prosody (adapting F0 and duration models).

In Yamagishi et al. (2004), MLLR techniques are exploited to produce joyful and sad speech from a neutral reading training corpus, with an adaptation corpus varying in size. They showed that using an emotional corpus of about 50 sentences yields quality that is judged as similar to speech from target models.

Qin et al. (2006) focused on the synthesis of happy, sad and angry speech in Chinese. Unlike Yamagishi et al. (2004), their seed model was trained on several emotions and not on neutral speech only, to provide an *average model*. It is here trained on sad, happy and neutral, to provide a good coverage of the emotional space and it is subsequently adapted to angry speech. They compared this technique with the adaptation of a neutral model and indicated that it provided more natural F0 curves.

Similarly, the use of an average model was exploited in Eyben et al. (2012) to adapt it to different speaking styles. This method was shown to achieve high quality rendering of the expressivity.

Next to these standard adaptation methods, it is worth mentioning that a specific type of adaptation was proposed in Miyanaga et al. (2004) to generate rough, joyful and sad speech. The method consists of three steps: (i) the training of context-dependent HMMs without context clustering for the respective styles independently, (ii) the computation of a shared decision tree context clustering to obtain a common tree structure for all styles and (iii) the modeling of mean vectors for each style computed by multiple regression with the style control vector. The interest here is that "the model retains multiple regression matrices and does not have mean vectors as the model parameters". This provides a more explicit way to modify expressivity,

the values of the style control vector being interpretable (here, 2 dimensions related to valence and the degree of politeness) and modifiable by the user.

Finally, statistical interpolation methods have also been applied to expressive HMM-based speech synthesis, to produce combined emotions in Tachibana et al. (2005).

Another approach consists in considering emotional information as a *contextual feature in the decision tree-based clustering* of the models. Yamagishi et al. (2004, 2005) referred to that technique as ‘style-mixed modeling’ and indicated that it yielded better results than a style-dependent modeling for the synthesis of neutral, rough, joyful and sad speech. Similar experiments were carried out by Tsuzuki et al. (2004). This method was compared to adaptation techniques in Eyben et al. (2012). They found no significant difference between both methods in terms of quality but pointed out that adaptation significantly outperformed the use of speaking style as contextual feature, in terms of expressiveness of the synthesized voice.

Finally, Hirose et al. (2004, 2005) proposed a *modification of F0 and duration* values based on Fujisaki’s acoustic model (Fujisaki and Kawai, 1982; Fujisaki, 1992). With a view to produce angry, joyful and sad speech in Japanese, they relied on a binary decision tree to predict the F0 model parameters and duration of command phrases and accent values. They motivate this choice by the fact that prosodic parameters are known to have a wider time span than segmental features. Conversely, the two first methods directly modify the phoneme models, therefore applying changes at the segmental level. They achieved good quality for anger but rather unsatisfactory results for both other emotions. It should be noted that the post-processing is here processed on a style-dependent model (*i.e.* one model trained for each emotion), which allows for a good modeling of global voice quality features of each emotion.

Outside these three main trends, other studies are worth mentioning such as the integration of articulatory features instead of spectral parameters in statistical speech synthesis (Black et al., 2012), which offers a meaningful representation allowing to generate stylistically varying speech. They achieved reasonable recognition of the four basic emotions (anger, sadness and happiness) for English speech synthesis.

In this thesis, we address three main criticisms to the aforementioned existing studies:

CRITICISM 1 The use of one corpus for each acted emotion poses two main problems:

- Emotions (*e.g.* happy, sad, angry) may be a simplistic reduction of the expressive naturalness lacking in speech synthesizers. We rather believe that synthesizers should include other types of expressivity such as various speaking styles or emphatic speech.
- Considering each emotion in isolation neglects global prosodic changes which occur in real speech, as we move from one emotion to another, with various degrees of intensity.

CRITICISM 2 Most techniques propose a global adaptation to each expressive speaking style without considering more local prosodic phenomena which are specific to expressive speech, such as emphatic stress.

CRITICISM 3 Most studies focus on modifications at the prosodic and voice quality levels only and clearly neglect potential phonetic modifications.

These criticisms are further described below.

Criticism 1. The use of one corpus per emotion

A first interesting observation is that most of the aforementioned expressive HMM-based synthesizers rely on acted emotional speech, typically one emotion per corpus (*e.g.* Yamagishi et al. (2004, 2005); Tsuzuki et al. (2004); Hirose et al. (2004); Miyanaga et al. (2004); Qin et al. (2006)). This approach was questioned by Campbell (2006) who indicated two main drawbacks:

1. Most corpora are based on the "big six" emotions of psychological research as illustrated by Ekman and his colleagues (Ekman, 1992). This is a clear reduction of the notion of expressivity, which should integrate "more subtle expressions of speaker states and relationships".
2. He insists on the lack of naturalness of the collected databases which focus on a single acted emotion. Conversely, real conversational speech is much more fluctuating and less stereotypic of specific emotions.

In his study, Campbell (2006) proposes to model expressivity in terms of various dimensions of the speaking style, related to the speaker (mood and interest) and the interlocutor (relationship and place). He implements this model in the framework of unit-selection speech synthesis by combining a neutral synthesizer (for content-based utterances) with the selection of longer units from a *large natural conversational corpus*. For the latter, however, no keyboard entry is allowed, which means that no text to synthesize can be explicitly defined by the user. Only functions and constraints will determine the lexical output. This constraints the set of possible speech output and can only provide high quality results with a very large database (here more than 5 years of one person's daily conversational speech). No formal evaluation of the resulting synthesis was carried out but the fact that large speech segments from the database are concatenated makes it very naturally-sounding.

The problem of dealing with emotions in isolation was also pointed out by Eyben et al. (2012). As a result, they relied on audiobooks, which avoids the use of corpora tailored for one specific emotion but contain a wide range of expressivity. The difference with human conversational speech still lies in the read acted aspect of the corpus but, similarly to conversational speech, there is no clear segmentation of the speech continuum into distinct stereotypical emotions. Eyben et al. propose an automatic clustering of the corpus to determine various speaking styles and emotions, based on a hierarchical K-means clustering. These fragments of data are then used to train various expressive models. The main drawback of this technique,

however, is that the human interpretation of the clusters is not always straightforward, which makes it harder to choose the right model for a new sentence to synthesize.

Interestingly, compared to emotion modeling, the synthesis of various speaking styles (*e.g.* TV speech, sports commentaries, political discourse) has arisen much less interest. As we mentioned, Eyben et al. (2012) tried to bridge that gap by identifying different speaking styles in audiobooks. Krstulović et al. (2007) also adapted a German HMM-Based synthesizer to expressive football announcements. They relied on a simple adaptation of a neutral model to synthesize neutral and expressive sports commentary voice. Similarly, Lorenzo-Trueba et al. (2013) exploited adaptation techniques to produce HMM-based speech synthesis of news broadcast, live sports commentaries, interviews and parliamentary speech. An interesting aspect of their work is that they propose a technique for speaking style *transplantation*, *i.e.* the transfer of "the nuances of particular speaking styles speech models into a different neutral speaker". In French, other exceptions include Obin et al. (2011) who generated HMM-based speech synthesis of mass, sports commentaries, political and journal speech. They integrated prosodic contextual features indicating boundaries and prosodic prominences. One model was trained for each speaking style. An interesting study was also carried out by Roekhaut et al. (2010) who modeled radio news, political and conversation speech, but in unit-selection synthesis. They mainly relied on post-processing of the signal, modifying speech rate and F0. It should also be noted that a whole body of research has focused on the synthesis of spontaneous speech (*e.g.* Sundaram and Narayanan (2003)), which can be seen as a specific speaking style, depending on the communicative situation (*e.g.* interview, conversation between friends). This is further investigated in Chapter 11.

Criticism 2. An overall adaptation

A second major criticism addressed by Maeno et al. (2011) to the aforementioned studies is that they only provide a *global* adaptation/modeling of each speaking style or emotion. No attention is paid to more local phenomena as presented in Section 5.3.1. While we believe that adaptation techniques adequately modify spectral characteristics, which can be seen as more global, they do not consider accentual phenomena typical of expressive speech. Such phenomena include changes in the stress distribution or specific realizations of emphatic stresses. An exception is found in Hirose et al. (2004, 2005) but focuses on the modeling of Japanese accentual patterns only, which are known to be less complex than English or French models of accentuation¹⁵. We have pointed out that some accentual phenomena (especially emphatic stress) play a major role in expressive speech. By globally modeling the

¹⁵Japanese can be distinguished from most Indo-European languages by its restricted prosodic variations. Beckman and Pierrehumbert (1986) showed that this restriction can be seen both in the realization of the emphatic stress and in its position. While the acoustic variability of this stress conveys, in English, the expression of nuances like doubt or impatience, Japanese only knows a falling realization. Beside, there is no total independence in Japanese between intonational phrase boundaries and emphatic stress which systematically falls at the beginning of the phrase. Conversely, this stress can fall on any highlighted syllable in English. A last interesting element is that duration and intensity play a much more subtle role in Japanese prosody than in English for instance. Black

emotion, no attention is paid to where and how these stresses should be modeled. This has given rise to a parallel body of research centered on emphasis modeling. Whereas many studies have tackled this issue in unit-selection synthesis (*e.g.* Strom et al. (2006); Fernandez and Ramabhadran (2007)) we will focus here on HMM-based studies only.

Badino et al. (2009) proposed to include an "emphatic" label, additionally to ToBI labels, in the contextual features of the decision tree. They however achieved poor results, with a preference of the testers for non-emphatic stress. The authors explained it by the fact that emphatic stresses were produced in an exaggerated way, perhaps due to the recording of the training corpus which was specifically tailored for emphatic stress pronunciation. In that sense, a more natural corpus might better fit the task. Such a corpus was exploited in Yu et al. (2010). Emphatic parts of the corpus were manually indicated and emphasis was produced at the word level. They compared four main techniques:

1. Simple adaptation (MLLR) on emphatic segments.
2. Integration of emphasis in context features of the decision tree.
3. Idem 2. but with a priority given to emphatic information. Here a first tree is trained based on emphatic information only and then extended with other features.
4. Factorized decision trees: a first decision tree is built with normal context questions only. The second is built with emphatic-related questions. Both trees are then combined, the leaf nodes corresponding to the intersection of nodes from the two decision trees.

They show that the third method provides the best results in terms of emphasis rendering but that method 4 outperforms all other methods in terms of naturalness of the synthesized voice. A similar experiment is proposed by Morizane et al. (2009) who propose two integrations of emphasis in Japanese synthetic speech, *i.e.* through the training of independent emphatic and non-emphatic models (as style-dependent models) or by integrating emphatic contextual information in the decision tree. While both techniques achieve a good rendering of emphatic stress, the latter synthesizes significantly more naturally-sounding speech. In that study, emphasis is considered at the phrase level, which seems to be effective for Japanese prosody modeling. Similar studies on Japanese include Maeno et al. (2011) who integrate emphasis as a context feature. Experiments indicates that it yields good reproduction of emphasis expression.

Interestingly, most of these studies focus on emphasis alone and do not integrate it into the broader framework of expressive speech synthesis. Research is usually done on corpora tailored for emphasis analysis (*e.g.* Badino et al. (2009); Morizane et al. (2009)) or on neutral read speech (*e.g.* Yu et al. (2010)). It is rather surprising

and Campbell (1995) had already faced these divergences which prevented a simple transposition of Japanese unit-selection synthesizers to English.

that such studies were not integrated in research for speech synthesis of emotions or expressive speaking styles, which are known to display a higher level of emphasis.

Criticism 3. Modifications at prosodic and voice quality levels only

While research is now targeting the generation of expressive, spontaneous and media-related speech, most current synthesizers with that objective focus on modifications at prosody and voice quality levels only, discarding potential phonetic modifications. This criticism also goes for studies focusing on the synthesis of hypo- and hyper-articulated speech (Picart et al., 2014). Only the attempt to generate spontaneous-like pronunciation has received attention (*e.g.* Prahallad et al. (2006); Werner and Hoffmann (2006); Lee et al. (2010)). Exceptions are Jande (2008) who models speaking-style specific pronunciation modifications for Swedish speech synthesis and Roekhaut et al. (2010) who investigate the possibility to model style-specific final schwa elision for French in unit-selection synthesis. This is rather surprising, as it is well acknowledged that phonetic variation is affected by expressivity and speaking style. A possible explanation may be that some studies have indicated that the integration of phonetic variants may degrade intelligibility (see Larreur and Sorin (1990)), which has always been regarded as more important than the naturalness of the synthesized message.

Those three criticisms are addressed in this second part of the thesis.

5.5 Available front-ends : The development of eLite-HTS

As mentioned in the introduction to this chapter, Text-to-speech (TTS) synthesis relies on two main modules, the NLP and the DSP. The NLP, also called "front-end", should provide a rich automatic annotation of the text and sound at different linguistic levels (*e.g.* phonetics, syllables, parts of speech). This usually relies on various NLP tools like part-of-speech (POS) taggers and grapheme-to-phoneme converters to process important contextual information. Based on these annotation layers, more complex information, like rhythmic group boundaries, can also be extracted and can offer a more complete description of suprasegmental phenomena. For HMM-based speech synthesis, the HTS toolkit (Zen et al., 2007a) provides a very complete implementation of the DSP. While it interfaces with existing NLPs for English, such as the Festival Speech Synthesis System (Black et al., 2001)¹⁶, no front-end is provided for French. To produce a full HMM-based expressive speech synthesis for French, the integration of a French annotation component in HTS is crucial and is explored in this section.

Few publicly available NLP tools have been proposed for French. An exception is the open source system LIA_PHON (Béchet, 2001), which offers grapheme-

¹⁶The Festival Speech Synthesis System was initially developed for unit-selection synthesis, and includes tools to produce such a complete linguistic analysis of the text.

to-phoneme conversion and grammatical analysis to provide a description file for MBROLA TTS (Dutoit et al., 1996). This tool, however, does not include prosodic and rhythmic information as required for HMM training with HTS or for the selection of units in TTS. The IrcamAlign tool (Veaux et al., 2008) proposed the generation of input HTS files for French, based on LIA_Phon analyzer, but this tool was not made available. Another existing tool to which we have access is eLite¹⁷ (Beaufort and Ruelle, 2006; Colotte and Beaufort, 2005). eLite includes its own full NLP component, providing both basic linguistic information (*e.g.* phonetics, POS) and more complex rhythmic and phrasing annotation (*e.g.* rhythmic groups). In the framework of our thesis, we further extended the tool to provide an automatic generation of HTS files¹⁸ from text and further developed it into a web service, eLite-HTS. This tool was developed in close collaboration with Sophie Roekhaut.

5.5.1 Architecture of eLite

This subsection briefly presents the different components of our NLP tools and details how they help to generate HTS description files. eLite NLP includes 3 modules (see Figure 5.15). A disambiguation module is made of preprocessing, morphological and syntactic analysis. A grapheme-to-phoneme conversion module is then responsible for the generation of the phonetic transcription. Finally, the prosodic module generates linguistically-driven phrasing and rhythmic information. The communication between the 3 modules is realized with a Multi-Layer Data Structure (MLDS) inspired by the structure proposed by the Festival Project. The MLDS is composed of 7 annotation layers, corresponding to different levels of segmentation of the text. *Sentences*, *tokens*, *units* and *words* are detected from the input text by the disambiguation model. *Phonemes* are generated by the grapheme-to-phoneme converter and, finally, *syllables* and *rhythmic groups* are produced by the prosodic module.

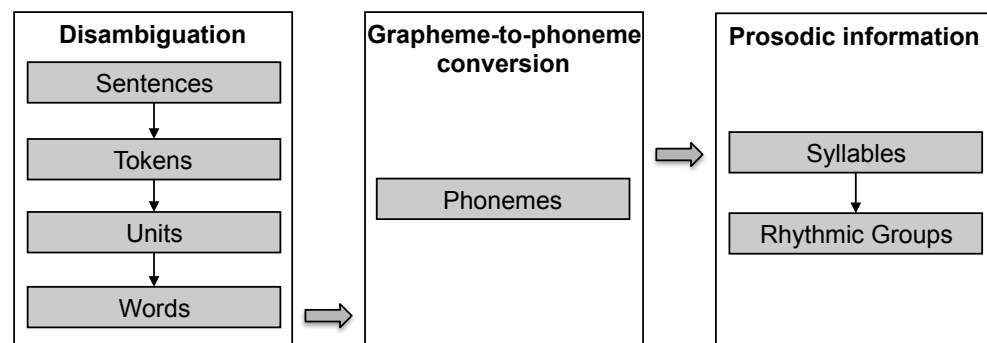


Figure 5.15 - Relations between the 3 modules of eLite in the MLDS.

¹⁷Enhanced LIngually-based TEExt-to-speech synthesizer.

¹⁸For a description of the structure of HTS label files, see Appendix A.

Disambiguation

First, in the disambiguation module, a preprocessing of the text is applied, based on a set of manually-tuned rewrite rules. It identifies sentence boundaries and detects some special tokens : URLs, phone numbers, dates, times, currencies and units of measurement. It produces the *sentence* and *token* units. Then, it carries out, sentence by sentence, a morphological analysis and a contextual disambiguation. The morphological analysis performs, at the *word* level, a lexicon look-up to determine the set of possible grammatical categories for each word, given its token. It then creates the *unit* level that groups together some *words* like compound nouns (*e.g.* "pomme de terre"¹⁹). A contextual disambiguation at the *unit* level selects the most likely grammatical category from the set of propositions, based on a statistical language model (Beaufort et al., 2002).

Grapheme-to-phoneme conversion

The grapheme-to-phoneme conversion module, *i.e.* the phonetizer, first produces the phonetic transcription of each *word* unit, with *phoneme* units, regardless its context. The phonetization is based on an ID3 decision tree learned from a phonetic dictionary (Pagel et al., 1998). Note that the dictionary is automatically accentuated. For non-clitic words, a primary stress is put on the last vowel while a secondary stress is put on the first. Previously extracted contextual information is then exploited to deal with phonetic variations at word boundaries. The most frequent phenomena that appear in French are the liaisons between two words and the deletion and insertion of schwa for the lubrication of speech.

Prosodic analysis

The prosodic module, or prosody generator, provides linguistic information that contributes in generating a more natural prosody in HMM-based or unit selection speech synthesis. *Syllables* are determined with a set of rules based on the possible grouping of *phonemes*, according to articulatory criteria. Each *syllable* is then associated with the stress status of its corresponding nucleus. Grammatical information is also used to detect the boundaries of *rhythmic groups*, which should correspond to breath groups ending in a boundary tone. The rhythmic groups are detected with a chinks and chunks algorithm (Lieberman and Church, 1992) as described in Section 5.3.3. Figure 5.16 gives an example of sentence in TextGrid format as output by eLite NLP.

Most components of eLite have been evaluated and further detailed in publications (see Pagel et al. (1998) for the phonetization module, Beaufort et al. (2002) for the syntactic analyzer and Beaufort (2008) for a detailed description of the other components).

¹⁹Potato.

²⁰The hummingbirds are singing on the branch.

PROSODIC AND PHONETIC IMPROVEMENTS FOR HMM-BASED
EXPRESSIVE TTS: A THEORETICAL AND PRACTICAL INTRODUCTION

Words	Les	oiseaux	mouches	chantent	sur	la	branche	.																
Phonemes	l	e	z	w	a	z	o	m	u	S	S	a~	t	s	y	R	l	a	b	R	a~	S	_	
Syllables	l	e	z	w	a	z	o	m	u	S	S	a~	t	s	y	R	l	a	b	R	a~	S	_	
Units	DET		NOUN				VERB		PREP		DET		NOUN		PUNCT									
Rythm groups	RG						RG		RG						PUNCT									
Sentences	DECLARATIVE																							

Figure 5.16 - Example of NLP analysis for sentence in TextGrid format "Les oiseaux-mouches chantent sur la branche²⁰".

eLite was further developed to automatically generate output files in the HTS format and is now available as a web service at : <http://cental.uclouvain.be/elitehts/v1/>. eLite-HTS generates input files for the training and synthesis stages of a French HMM-based synthesizer using the HTS toolkit. The web service is in a REST architectural style (Richardson and Ruby, 2008). A fundamental principle of a REST architecture service is to keep the implementation of the server and the client independent. REST Web services can use the HTTP protocol for the communication between the server and the client. The client of the service can be implemented in any language that allows for HTTP requests (*e.g.* php, C, perl, Matlab). In addition to the HTS file, our web service can also produce an output in TextGrid format for the Praat application (Boersma, 2001) dedicated to phonetic and phonological research. eLite-HTS has already been used by several laboratories at the international scale (Cernak and Honnet, 2015; Stan et al., 2016).

5.5.2 eLite generation of the HTS file

A global flowchart of the tool is shown in Figure 5.17. A text or TextGrid is first provided to the tool (1). The ability of eLite to accept TextGrid files is an invaluable advantage as it allows to directly provide the tool with a manually-checked phonetic transcription. All other annotation tiers will be produced and synchronized with this phonetic annotation without modifying it. It should be noted that, to our knowledge, most other front-ends do not offer that possibility. This means that, even if a manually-checked phonetic transcription of the training corpus (or of the new sentence to synthesize) is available, it cannot be directly exploited by original versions of other front-ends like Festival. These latter will automatically produce a new automatic phonetization of the text. This loss of information is very problematic when the real phonetization differs markedly from the automatic transcription, notably in the case of spontaneous corpora (see Chapter 9). This functionality is therefore essential for the remaining of this part of the thesis, which aims at the synthesis of sports commentaries, based on a spontaneous corpus which has been manually annotated.

If no phonetic transcription of the text is available, eLite will generate an automatic phonetic transcription (2), which can then be further aligned with existing alignment tools (3). It should be noted that eLite-HTS output files can be directly used as input to Train&Align described in Chapter 4. The aligned TextGrid can

then be provided as a new input to eLite which will automatically produce all other required annotation tiers (*e.g.* syllables, POS, rhythmic group) without modifying the alignment. At that stage, eLite also duplicates the syllable tier. This tier can be manually filled by the user (5) to insert prosodic information at the syllable level. eLite-HTS then allows for the generation of HTS files considering this new annotation level (6).

The minimal linguistic unit used in the HTS description files is the phoneme. For each phoneme, 53 characteristics are described, as detailed in Section 5.4. The other linguistic levels used by HTS are (from low- to high-level) : syllables, words, phrases and utterances. Each unit is described according to its context (left and right), its position in relation to the other linguistic units of higher level and the number of units of the lower level it contains. Some specific features are added to the different levels : stresses for syllables, POS for words and ToBI endtones for phrases (Wightman and Ostendorf, 1992). The internal structure of eLite MLDS provides most of the data required for the generation of HTS files (see Table 5.3).

Table 5.3 - *Correspondance between eLite and HTS layers.*

eLite		HTS
Sentences	→	Utterances
Units	→	Words
Phonemes	→	Phonemes
Syllables	→	Syllables
Rhythmic groups	→	Phrases

Only ToBI tones are not automatically generated by eLite-HTS. While this annotation system has been widely adopted by the English-speaking community, we have shown in Subsection 5.3.2 that several prosody annotation protocols are commonly used for French prosody. As previously mentioned, the web service offers the possibility to introduce prosody information, manually or with external scripts. Such an integration of prosodic information is exploited in Chapter 6 for the synthesis of sports commentaries.

Two types of output files can be produced, corresponding to the training or synthesis stages. Unlike synthesis description files, the training files output by eLite-HTS additionally contain the time boundaries of each phoneme in the sound files.

5.5.3 Specifications of the web service

The web service accepts two types of inputs: TextGrids and texts and can produce three types of output formats: TextGrids, HTS and DLS (eLite NLP format). Table 5.4 describes the main HTTP requests that can be addressed to the server.

PROSODIC AND PHONETIC IMPROVEMENTS FOR HMM-BASED EXPRESSIVE TTS: A THEORETICAL AND PRACTICAL INTRODUCTION

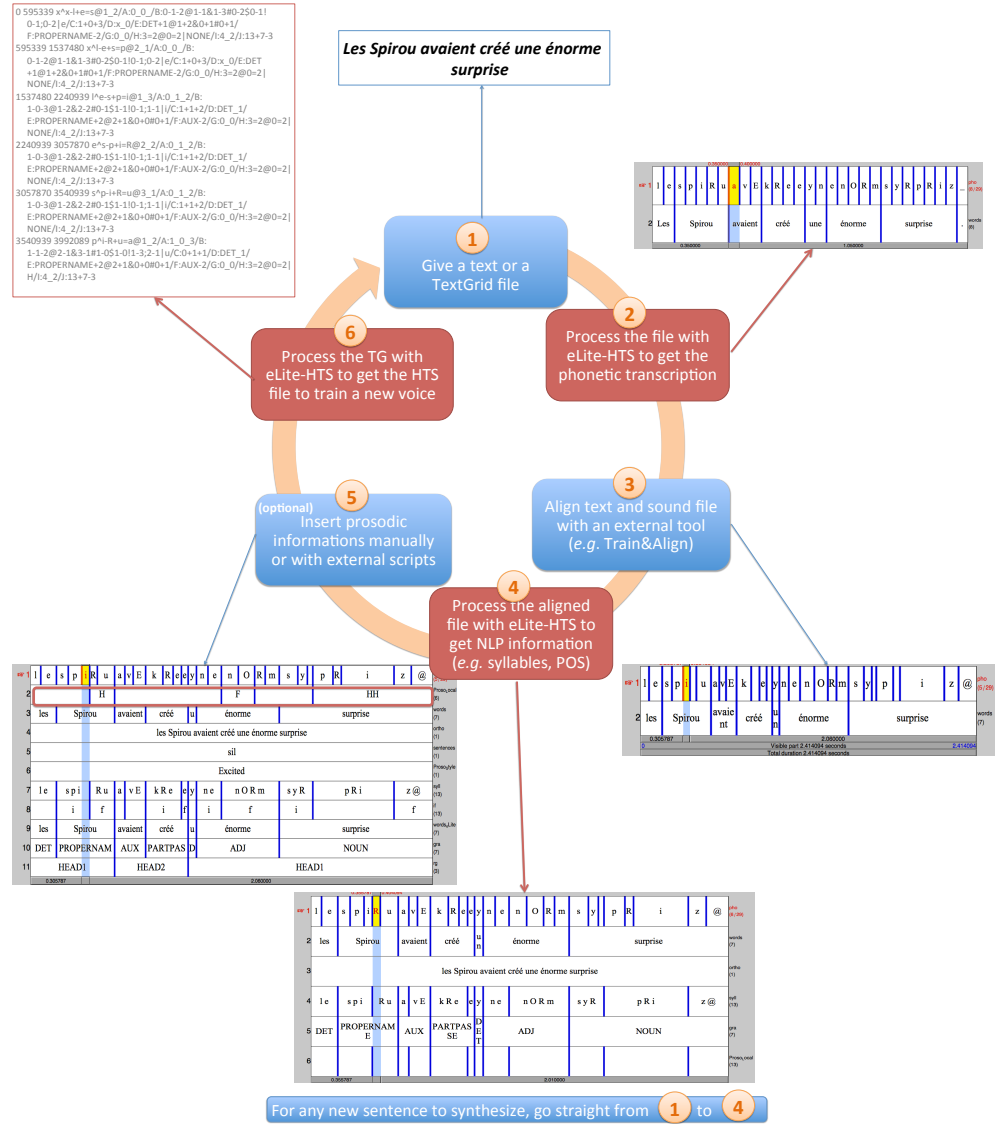


Figure 5.17 - Flowchart of eLite-HTS.

5.5.4 Future work

Future work include the complete connection of the web service with Train&Align (see Chapter 4), for the automatic alignment of sound files before training, and with HTS modules to fully generate French HMM-based speech synthesis. Later versions should also allow for more personalization of the different modules of the NLP. Administrative requests are currently done to investigate the possibility to provide an open source version of eLite-HTS.

Table 5.4 - *Main requests of the web service.*

Request	Resource	Description
POST	/textgrids	record a new resource on the server and it returns the location of the resource (id)
GET	/texts /textgrids:id or /texts:id	get the resource recorded on the server
GET	/textgrids:id/hts or /texts:id/its	get the HTS data corresponding to the resource
GET	/textgrids:id/textgrid_hts or texts:id/textgrid_hts	get the NLP data in TextGrid format corresponding to the resource
GET	/textgrids:id/dls or /texts:id/dls	get the NLP data in dls format (eLite) corresponding to the resource

5.6 A concrete application of expressive speech synthesis: The Sportic project

This second part of the thesis falls within the framework of the Walloon Region project WIST 3.0 Sportic. The objective of Sportic is the acquisition and the rendering of sports events. Our main interest lies in the synthesis of sports commentaries, which fits into the field of expressive speech synthesis. We provide in this section a brief description of the project.

5.6.1 Outline of the project

Sportic considers the interpretation of sports scenes and aims at producing an audio-visual representation of the recorded event. It focuses on basketball games but should be generalizable to other sports. The outline of the project is shown in Figure 5.18. From the automatic acquisition of game scenes, by a set of cameras placed around the field, Sportic proposes an automatic framework to (i) annotate the actions, (ii) provide a professional coverage of the game and (iii) generate personalized audio-visual summaries. This project is innovative given that no existing alternative is known to provide such a complete framework to generate sports commentaries, from image to speech ²¹.

This application meets the needs of various publics:

- **Sports clubs and coaches** should be interested in the scene annotation as it

²¹Noteworthy previous research in this domain include the Goalgetter project (Klabbers et al., 1996; Theune and Klabbers, 1998) which provided an automatic generation of spoken soccer reports in Dutch from tabular data derived from the Teletext pages, and the SOCCER project (André et al., 1988) which offered a simultaneous interpretation of soccer image sequences to provide their natural language description.

would make it possible to automatically retrieve all actions (of a specific type) from a specific player. This should help the coach in the analysis of the players' behavior. Focusing on a specific player should also allow for the automatic generation of an audiovisual business card.

- **Televisual broadcasters** would benefit from a low-cost coverage of sports games. In fact, most broadcasters buy the rights for an entire season but cannot afford to cover all games because of human costs. An automatic production of games images could extend the offer of covered events.
- **Journalists** could take advantage of an automatic annotation of sports actions as it would facilitate their archiving and fasten documentary research.
- **The general public** could enjoy the personalized summaries. In fact, many viewers have neither the time nor the desire to watch the whole game. They usually want to concentrate on interesting actions only, possibly with a specific focus on a particular player or type of action.

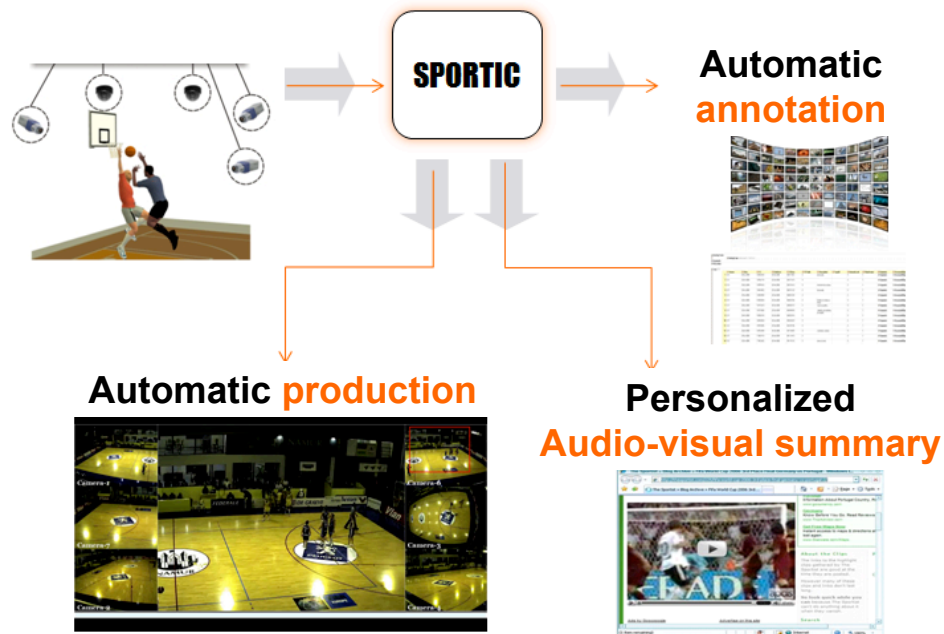


Figure 5.18 - Outline of the Sportic project²².

To achieve this aim, Sportic relies on three scientific domains:

- **Intelligent video analysis** should allow for the image production from a large set of cameras (*i.e.* automatic camera control, interpolation between images from different cameras, and image editing), players and ball detection and tracking, image selection from multiple views of a single event, and scene analysis and interpretation (*i.e.* determine the type of action from a recorded scene).

- **Automatic text generation** aims to generate text from non-linguistic data (*i.e.* scenes description as furnished by the video analysis module). The text generation module should produce automatic summaries based on two types of constraints: text should be time-aligned with the scenes (to provide a resulting synchronized speech synthesis) and adapted to the user's specifications (*e.g.* duration, focus on specific actions or players). It should rely on the preceding generation of language resources including lexical databases (lexical field of basketball with synonyms, antonyms, and so on), syntactic patterns (set of similar syntactic constructions fulfilling a same function) and discursive models (*e.g.* structuring of summaries, paragraphs). From a set of selected scenes, it should produce textual descriptions, joined together with temporal connectors. These descriptions should enrich the video content.
- **Expressive speech synthesis** aims at the generation of voice that is characteristic of sports commentaries to propose an automatic reading of the scene descriptions.

While the low-cost coverage of sports games and the automatic annotation of the actions result from visual intelligence only, the automatic generation of personalized game summaries relies on all three domains. We will focus on this specific application in this part of the thesis and will concentrate on the speech synthesis stage (see Figure 5.19). The recording of scene events, and their resulting analysis, is used to generate textual descriptions which will be synthesized with a voice characteristic of sports commentaries. This project is an interesting application of speech synthesis that serves the needs of our research.

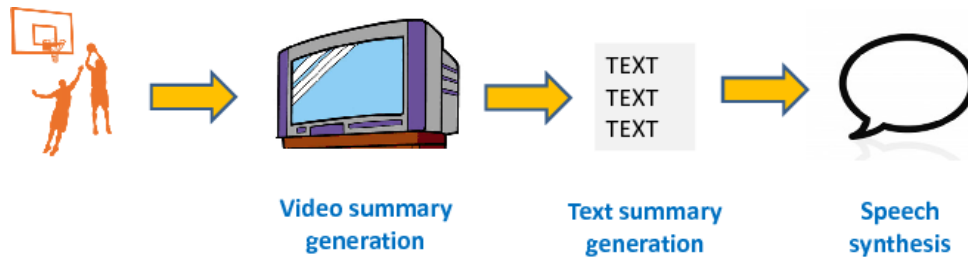


Figure 5.19 - Flowchart of the audiovisual game summary application.

5.6.2 Concept-to-speech synthesis

We focus here on the "speech synthesis" component of Sportic, which will be our main topic of study. The objective of this part is to be able to reflect the intensity of the game through the synthesized voice. As will be shown in the next chapter, this objective poses challenges at various levels, including voice quality, prosody and

²²This illustration originates from the Sportic project description files, as submitted to the Walloon Region.

phonetics.

Beyond traditional "text-to-speech", which receives a text to pronounce and only partially perceive its meaning, "concept-to-speech" has been developing and generates itself the text from structured data (*e.g.* (Young and Fallside, 1979), the SOLE concept-to-speech system (Hitzeman et al., 1998, 1999), Paul Taylor's work (2000; 2009) and Shimei Pan's PhD at Columbia University (2002)). CTS has access to the meaning of the text, along with complementary information such as the affective state to imitate. In that sense, the concatenation of both language generation and speech synthesis blocks of Sportic constitutes a specific case of CTS.

The text generating system being developed by a distinct team, it was not available at the beginning of our study. We therefore addressed the synthesis problem as a simple "text-to-speech" application, which could be able to synthesize any given text. The advantage of the inclusion of this synthesis, in the long term, in a "concept-to-speech" application can still be seen as a considerable advantage as prosody information would be more easily predicted with the help of semantic and/or pragmatic information provided by the speech generation system, as further described in the next Chapter (see also Hitzeman et al. (1998); Pan (2002)). For the reasons previously mentioned in the introduction, the synthesizer is implemented here by means of Hidden Markov Models, which offer greater flexibility than the other synthesis techniques.

To provide a quality speech synthesizer, our study consists of two main steps :

1. An analysis of real basketball commentaries should allow us to highlight the main characteristics of this speaking style. Special emphasis is put on segmental features (*i.e.* voice quality), supra-segmental features (*i.e.* prosody as described in 5.3), and pronunciation (in terms of phonetic modifications like elision or liaisons as described in 5.2). An automatic and/or manual annotation of the commentaries is therefore required, at least in terms of prosody and phonetic realization.
2. These characteristics should then be integrated in HMM-based speech synthesis. As will be shown in the next two chapters, most modifications are made in the annotation of the training and synthesis sentences. Protocols of annotation are developed and their integration in speech synthesis is assessed through subjective tests.

Both the analysis and the synthesis of sports commentaries require the use of a corpus recorded in high quality conditions. Such a corpus was recorded to meet the needs of our study and is described in the next subsection.

5.6.3 Corpus design

While corpora of basketball commentaries are available online²³, their low acoustic quality and the general background noise hinders their use for a precise prosodic analysis. This is even more problematic for the training of a new synthetic voice. For these reasons, a new corpus was recorded for the purposes of our study: the *Sportic* corpus.

The *Sportic* corpus is made of the commentaries of two basketball matches by Pierre Robert, a professional French commentator from the RTBF, a Belgian TV channel. The speaker watched the game and commented on it without any prompt. Both matches star the Spirou Belgian team with very tight final scores, which induces a high level of excitement. The total corpus duration is of 162 minutes with a high proportion of silences. The corpus was recorded in sound-proof conditions with speech synthesis in view. The room was equipped with a screen displaying the game and with an AKG C3000B microphone, as illustrated in Figure 5.20²⁴. As previously mentioned, the issue with sports commentaries corpora is usually the high level of background noise which precludes their precise acoustic analysis (Trouvain, 2011). Conversely, our corpus exhibits the advantage of being spontaneous and of high acoustic quality, being therefore suited for speech synthesis.

A whole process of corpus annotation was carried out, based on automatic tools and manual checks by students in linguistics. The corpus was first manually orthographically transcribed and its phonetization was automatically produced by eLite-HTS (as presented in the previous section). The phonetic transcription was manually corrected in terms of modifications such as elisions or liaisons. It should however be noted that glottal stops were not annotated²⁵. We believed that their low rate of appearance in the corpus would result in them being averaged out as noise in the corresponding vocalic models. Their annotation would have been essential in the framework of unit selection synthesis in which, as previously mentioned, a single annotation error may ruin the naturalness of the resulting synthesized speech. One should however be aware that their absence from the annotation may still have an impact on the training of some HMMs, especially in specific prosodic positions where glottal stops tend to be often realized like syllables with an emphatic stress.

The phonetic transcription was then automatically aligned with the sound with the Train&Align tool presented in Chapter 4. This alignment took advantage of the bootstrap option to reach alignment rates higher than 80 % with a 20 ms tolerance threshold. The eLite NLP system, as described in Subsection 5.17 produced other required annotation tiers (*e.g.* syllables, parts of speech, rhythmic groups). Sentence boundaries form another important annotation level. Such a segmentation of a spon-

²³Such corpora include basketball game commentaries from "Basket1" magazine, which are made available on the RTBF website: <http://www.rtbf.be/video/recherche/tag?pattern=basket+1>

²⁴Outside this room, a Motu 8pre was capturing audio from the microphone, with a sampling rate of 44.1 kHz, using XLR connections on one side, and was sending the results to a computer connected through Firewire. For more information about the various components, see Picart (2013)

²⁵Glottal stops are optional phenomena in French, their realization depending on diverse factors including speaking rate and degree of articulation. They follow rather strict distribution rules and are inevitable in some contexts, *e.g.* "le un".



Figure 5.20 - *Recording conditions of the Sportic corpus.*

taneous speech corpus is rather complex as we do not have access to punctuation. The corpus was therefore manually annotated to define segments corresponding to both a prosodic and a semantic completeness.

5.6.4 Outline of the next chapters

Part II presents our methodology and results to synthesize sports commentaries, in the framework of the Sportic project. For that matter, it aims at answering to the 3 major criticisms we addressed to existing expressive speech synthesis studies in Section 5.4.2. Here is a very brief description of the solution we propose:

- CRITICISM 1** To avoid the use of acted emotional corpora (typically one corpus per basic emotion), we rely on a corpus of sports commentaries (see Subsection 5.6.3). It has the advantage of containing various types of expressivity (*e.g.* excitement and frustration) and a high degree of fluctuation, which is much more similar to everyday conversational speech. This corpus is also spontaneous and rather natural (*i.e.* not "acted" as most corpora used by other studies).
- CRITICISM 2** In Chapter 6, we propose a two-tier prosody annotation to deal with both local (*i.e.* accentual) and global (*i.e.* speaking style) prosody phenomena.

Unlike most studies on emphatic speech synthesis, the advantage of our corpus is that it displays "natural" emphasis (which was not elicited and therefore not exaggerated as in Badino et al. (2009) but which is still highly perceptible). This is in contrast with studies dealing with emphasis on neutral read speech (*e.g.* (Yu et al., 2010)). A set of local and global prosodic tags, specific to sports commentaries, is designed and applied to the Sportic corpus. This chapter also assesses the quality of the resulting synthesis when integrating these two prosodic tiers. The prosodic annotation is further investigated in Chapters 7 and 8.

CRITICISM 3 Chapter 9, focuses on phonetic aspects. After a comprehensive phonetic analysis of the corpus, it assesses the role played by phonetic variations in synthesized speech.

Chapter 6

Modeling Prosody for the Synthesis of Sports Commentaries

Contents

6.1	Introduction	142
6.2	Proposed prosody annotation protocol	145
6.2.1	Description of the two-layer prosody annotation	145
6.2.2	Inter-annotator rate	149
6.3	Acoustic analysis of the prosodic labels	149
6.3.1	Analysis of the local annotation	150
6.3.2	Analysis of the global annotation	152
6.4	Application to speech synthesis	154
6.4.1	Methodology	155
6.4.2	Integration of the local labels	155
6.4.3	Integration of the global labels	157
6.4.4	Integration of both local and global labels	162
6.4.5	Discussion of the evaluation framework	163
6.5	Conclusion and perspectives	164

Abstract

This chapter proposes a new prosody annotation protocol specific to live sports commentaries. Two levels of annotation are defined with HMM-based speech synthesis in view. Local labels are assigned to all syllables and refer to accentual phenomena. Global labels classify sequences of words into five distinct sub-genres, or speaking styles, defined in terms of valence and arousal. The objective of the study is to provide a set of labels complying with two constraints: they should be (i) related to a specific function and (ii) characterized by a distinct acoustic realization. The consideration of these constraints should allow for an automatic prediction of the labels both from the text or from the speech signal. Reasonable inter-annotator scores are achieved for both annotation levels. A prosodic analysis of all labels also shows that they can usually be distinguished by specific acoustic realizations. The injection of this new annotation protocol within HMM-based is also investigated. The integration of global labels (*i.e.* speaking styles) is carried out either using speaker-dependent training or adaptation methods. Local labels are integrated as contextual features in the decision tree clustering of the models. A comprehensive study allows evaluating the effects achieved by each prosody annotation layer on the generated speech. The evaluation process is based on three subjective criteria: intelligibility, expressivity and segmental quality. Our experiments indicate that: (i) for the integration of global labels, adaptation techniques outperform speaking style-dependent models both in terms of intelligibility and segmental quality; (ii) the integration of local labels results in an enhanced expressivity, while it provides slightly higher intelligibility and segmental quality performance; (iii) combining the two levels of annotation (local and global) leads to the best results.

This chapter is based upon the following publications:

- Sandrine Brognaux, Benjamin Picart & Thomas Drugman, A New Prosody Annotation Protocol for Live Sports Commentaries, *Interspeech*, Lyon (France), 2013.
- Benjamin Picart, Sandrine Brognaux & Thomas Drugman, HMM-based Speech Synthesis of Live Sports Commentaries: Integration of a Two-Layer Prosody Annotation, *8th ISCA Speech Synthesis Workshop (SSW8)*, Barcelona (Spain), 2013.
- Sandrine Brognaux, Benjamin Picart, & Thomas Drugman, La synthèse de commentaires sportifs: intégration d’une annotation prosodique à deux niveaux dans un synthétiseur HMM, *Nouveaux Cahiers de la Linguistique Française*, Vol. 31, 63-76, 2014.

Many thanks to Dr. M. Avanzi and Prof. A. C. Simon for their insightful advice.
The synthesis part of this chapter was realized in collaboration with Dr. B. Picart.

6.1 Introduction

As shown in the previous chapter, expressive speech synthesis based on Hidden Markov Models (HMMs) has been the focus of many studies in the last ten years. Unlike unit-selection based synthesis, HMM-based synthesis offers a rich playground in terms of controllability of the generated speech. However, state-of-the-art synthesizers are often criticized for their lack of naturalness. This is notably triggered by the inability of existing systems to deal with expressivity, which is either not modeled at all or only in a very limited way (Campbell, 2006). As previously mentioned, current research on expressive speech synthesis presents a certain number of drawbacks. We briefly summarize here two main flaws of existing studies which will be tackled in this chapter.

First, expressivity is often exclusively generated via adaptation or training on corpora with the targeted expressivity (*e.g.* Yamagishi et al. (2005); Qin et al. (2006)). Generally, training or feature adaptation is achieved globally, with no consideration of local phenomena specific to expressive speech, like accentuation. However, it is widely acknowledged that the accentual structure of a sentence and the realization of emphatic stress play a crucial role in the expressive function. Fónagy (1979) notably emphasized the greater accentual density in emphatic speech. A few isolated studies (*e.g.* Hirose et al. (2004)) have tried to integrate some expressive accentual information in speech synthesis. However, they were applied to Japanese, a language with more restricted accentual patterns than French or English (Beckman and Pierrehumbert, 1986).

Beside the omission of local phenomena, very little attention has been paid to macro-prosodic changes as described in Subsection 5.3.4. Indeed, most current studies rely on acted corpora of each emotion. These corpora are very constant regarding expressivity, being stereotypical with respect to the considered emotion. However, various expressivity types follow each other in real human speech. As stated by Braunschweiler et al. (2010), "a coherent speech corpus includes prosodic effects that go beyond the sentence level". These global prosodic changes should be modeled to improve the quality of the generated expressive speech (Eyben et al., 2012).

The generation of an expressive prosodic realization is of utmost importance when synthesizing sports commentaries. Their high degree of expressivity has been shown to fulfill various functions (*e.g.* expressing excitement, frustration, catching attention) which should be reproduced in the synthetic voice (Kern, 2010; Trouvain, 2011). The speech synthesis of sports commentaries has been little investigated. Krstulović et al. (2007) propose to adapt a neutral German model, based on a small set of sports commentary sentences. A similar study is proposed in Lorenzo-Trueba et al. (2013) who further exploit the adapted models for a voice transplantation application. A *style-dependent* HMM-based model is also trained by Obin et al. (2011) on a corpus of sports commentaries. It is complemented with the training of a style-specific prosodic model, which predicts a sequence of prosodic labels for the sentence to synthesize. Subjective tests show that the speaking style is correctly recognized when compared to synthesizers of mass, political and journal speech. Finally, the Goalgetter project (Klabbers et al., 1996; Theune and Klabbers, 1998) provides a

speech synthesizer of Dutch soccer commentaries, based on the concatenation of pre-recorded phrases, available in a variety of intonational patterns.

Several studies have focused on the prosodic analysis of sports commentaries (*i.e.* basketball, football and rugby (Audrit et al., 2012), horse races (Trouvain and Barry, 2000), soccer (Obin et al., 2010a) and football (Krstulović et al., 2007; Kern, 2010; Trouvain, 2011)). This phonogenre has been shown to display a very specific prosodic realization that highly differs from other genres (Obin et al., 2010a). It was shown to be importantly characterized by prosodic variations at the local but also at the global or macro level, sports commentaries containing multiple distinctive speech styles (Kern, 2010; Trouvain, 2011) (see Figure 6.1). These speech styles will be referred to as ‘sub-genres’ or ‘speaking styles’ in the remainder of the thesis and relate to what others have called ‘prosodies’ (Odgen and Walker, 2001) or ‘discourse ambiances’ (Goldman, 2012)¹.

These sub-genres can be seen as gathering macro-discursive units associated with specific communicative actions. While the *elaboration* relates to a somehow neutral speaking style, dramatic speech is characterized by a high arousal degree, which rises during the *building up suspense* and reaches a climax at the *presentation of a highlight*. These sub-genres were shown to display a specific function and to be characterized by a rather stable acoustic realizations. Audrit et al. (2012) show, for example, that highlight events like shots or confrontations are usually realized with a significantly higher F0. The analysis of horse race commentaries (Trouvain and Barry, 2000) indicates similar results for high emotional phases like the end of the race. Krstulović et al. (2007) also point out a high vocal effort, high pitch and level range and an increased speech rate for excited parts of football commentaries.

The analysis of sequences happening just after a goal in football commentaries further indicates that the prosodic realization depends whether the goal is for or against the supported team (Trouvain, 2011). Sports commentaries being highly ‘listener-oriented’, these two distinct acoustic realizations help the listener decode the message more quickly. Unlike the classification of Kern (2010), the distinction is based on valence rather than on the arousal level. On the whole, most studies tend to suggest that a prosody annotation of sports commentaries requires, beside local accentual information, a more global annotation level assigning a specific sub-genre to the speech segments.

Several prosodic models have been developed for speech synthesis purposes and could be used for a local annotation of expressive speech (see Sections 5.3.2 and 5.3.3). We rely here on tonal models and propose a simplified version of Piet Mertens’

¹The term ‘sub-genre’ can be here interpreted as a sub-part of a *phonogenre*, defined as the model or image that people have in mind of speech typical of a specific communicative situation (Goldman et al., 2009). In the framework of our study, we consider sports commentaries as a specific *phonogenre* that we further subdivide in *sub-genres*. The use of the term *speaking style* in that respect is also justified by its much more common use in speech synthesis. Its acceptance is however generally wider and more generic. It has been seen to refer to emotions, like angry or sad speech (Tachibana et al., 2005; Yamagishi et al., 2005) and communicative situations like political speeches of sports commentaries (Obin et al., 2011). Both ‘speaking style’ and ‘sub-genre’ will be used to refer to the subdivision of the sports commentary phonogenre in the framework of our study. The reader interested in terminology issues should refer to Goldman et al. (2009).

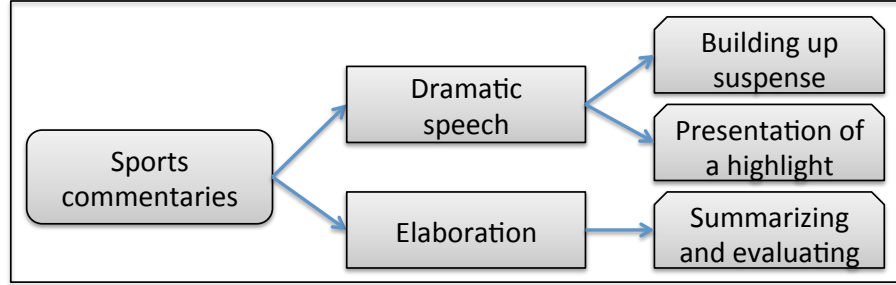


Figure 6.1 - *Speech styles in sports commentaries (Kern, 2010).*

annotation framework which is further described in Section 6.2. The global annotation in sub-genres is based on the classification of Kern (2010) and Trouvain (2011).

This chapter first sets out to present a prosody annotation protocol specific to sports commentaries (basketball in particular) with HMM-based speech synthesis in view. Our study relies on the *Sportic* corpus as presented in Subsection 5.6.3. This prosodic annotation relies on two annotation levels. A local annotation is associated to the syllable level and aims at annotating accentual events. This is motivated by the fact that the syllable is considered to be a central unit for the characterization of many aspects of prosody (Mertens, 2014). A global annotation classifies groups of words into specific sub-genres.

The use of this annotation for speech synthesis implies several constraints. The local and global labels should be associated with a specific expressive function. Assuming that a semantic analysis of the text is available (*i.e.* in the framework of concept-to-speech synthesis), it would then be easy to predict labels from the text. Secondly, to avoid averaging very different acoustic realizations, each label should be characterized by specific acoustic features.

In a second stage, the integration of this annotation protocol in speech synthesis is investigated. To allow for their integration in the speech synthesis system, **local labels** are considered as contextual information, similarly to parts of speech or positions of the syllable in the words. This is the standard solution proposed by HTS (Zen et al., 2009), which suggests to integrate ToBI labels at that level. Local labels are therefore considered in the questions to build the decision tree which computes the context-dependent clustering of the models. Regarding the integration of the **global labels** in the speech synthesizer, three possibilities are investigated. First, *style-dependent* models are trained on each sub-corpus corresponding to each sub-genre. This application is rather straightforward. However, it may produce low signal quality if the training corpora are too small. The two other investigated possibilities make use of adaptation techniques. A generic sports model is learned either on the most neutral sub-genre only or on all mixed sub-genres (*i.e.* to produce an average model as in Qin et al. (2006) and Eyben et al. (2012)). This model is then adapted to each sub-genre with adaptation algorithms.

The chapter is organized as follows. Section 6.2 details the proposed annotation protocol. An acoustic analysis of both annotation levels (global and local) is then provided in Section 6.3. In a second stage, the integration of the proposed annotation protocol within a HMM-based speech synthesis system is evaluated and discussed in Section 6.4. Finally, Section 6.5 concludes the chapter and proposes some perspectives.

6.2 Proposed prosody annotation protocol

6.2.1 Description of the two-layer prosody annotation

A two-tier prosody annotation is proposed.

A **local tier**, linked to the syllable level, refers to accentual phenomena. We have shown in Subsection 5.3.3 that three types of synthesizers can be distinguished according to their level of prosodic prediction, *i.e.* acoustic values (TYPE 1), phrasing and accentuation information (TYPE 2) or prosodic labels (TYPE 3). We have pointed out that the prediction of F0 and duration values presents a major drawback, *i.e.* the fact that they cannot be easily predicted from text. Besides, HMM-based synthesis is particularly suited for synthesis of TYPE 2 or 3 as it allows for integrating prosodic contextual information in the decision tree features exploited to cluster the models. Unlike unit-selection, data sparsity is not an issue in this synthesis technique, as the clustering allows for a generalization of the models to unseen events.

Including phrasing and accentual information only (*i.e.* TYPE 2), as predicted by the NLP, reduces, in our view, the quality of the prosodic annotation. A typical example is that a major boundary could be continuative or conclusive. The pitch of the corresponding final syllable highly differs in both cases and would not be indicated in the phrasing annotation. Another example relates to emphatic stresses, which are very frequent in sports commentaries, but are not predicted as accentual information provided by NLPs like eLite-HTS (presented in Section 5.5). In our case, we could however take advantage of the inclusion of Sportic in a concept-to-speech application to access this information. For these reasons, the generation of local prosodic information will rely on the prediction of prosodic labels, *i.e.* in line with techniques of TYPE 3.

For that matter, several linguistic prosodic models have been presented in Subsection 5.3.2. The advantage of *linguistic models* is that they propose an intermediate representation, between the acoustics of the signal and linguistic functions, which facilitates their prediction from text. HMM-based speech synthesis is more fitted to *tonal models* than holistic representations. This is due to the fact that speech is modeled at the phoneme level and that each phoneme should be assigned a specific label. In the case of tonal models, the label of a syllable can be easily transposed to the corresponding phonemes.

This justifies our choice to investigate existing tonal models, with a specific attention for ToBI and Mertens' tones in particular. Several studies have criticized

the ToBI system (*e.g.* Martin (2001); Wightman (2002)). Initially developed for English, its straightforward application to French is rather complex (Martin, 2001). Besides, it contains a large set of labels (*e.g.* 5 different labels for pitch accents) which makes their prediction from text rather complex, especially if tones should be predicted along with the break-index levels. It should also be noted that some redundancy can be found in the annotation. The presence of a break index 4, for instance, is redundant with the occurrence of a boundary tone label (Wightman, 2002). This complexity can be observed in inter-annotation experiments, which indicate high inter-rater agreement for the prediction of the location of phrase accent or boundary tones, but rather low agreement on the choice of the right label for each of these events (see Pitrelli et al. (1994); Syrdal and McGorgy (2000) and Grice et al. (1996) for G_ToBI²). Pitrelli et al. indicate that, while the disagreement rate is rather high in that case, confusions are mainly related to relatively fine distinctions between similar labels. For that reason, many reduced variants of the ToBI system have been proposed (*e.g.* Wightman and Ostendorf (1994); Wightman and Rose (1999); Greenberg et al. (2002); Ananthakrishnan and Narayanan (2008); Jeon and Liu (2012)). We propose such a reduced version in the framework of our study, inspired by Mertens’ prosodic annotation system.

Mertens’ tones are very similar to ToBI tones in many aspects, mainly in the fact that they propose a unified framework to annotate pitch, accentuation and phrasing at the syllable-level. The main advantage of the system is that it was originally designed for French and therefore takes into consideration specificities of French prosody. Similarly to ToBI, however, it includes a very large number of labels, which makes their prediction from text rather complex. Two main issues can be distinguished in that sense:

- Several labels fulfill a similar function. Mertens (2008) points out, for example, that /LL, LL and \LL all indicate a minor continuation while HL and HL- are two manifestations of a new information or a specific implication of the speaker. The prediction of the right label from text is therefore made complex.
- Some labels are related to very specific functions which can hardly be predicted from text by existing NLP systems. An example is the LH label which can be interpreted as inviting a reaction from the interlocutor.

We propose here a reduction of the set of labels to provide a limited number of tags which can be easily predicted from text (*i.e.* with the limited semantic information provided by the concept-to-speech system). Obviously, this reduction results in a simplification of existing acoustic realizations of the syllables. Our hypothesis is that some of these acoustic subtleties could be automatically learned by the models, based on linguistic contextual information exploited by model clustering.

Our annotation protocol results from a **functional** approach of accentuation, as proposed by Di Cristo (1999, 2000). It assigns a label to each syllable, as in Mertens (1987). Similarly to ToBI and Mertens, it aims at providing a unified framework to

²G_ToBI is an adaptation of the ToBI original system to German

annotate phrasing, accentuation and intonation. The set of labels is listed in Table 6.1. In terms of *accentuation*, five labels are related to non-emphatic stresses (Di Cristo, 2000) and one to emphatic stresses.

The first five labels are assigned to the end of prosodic phrases. Two levels of *phrasing* are determined, similarly to boundary tones in ToBI. To facilitate the automatic annotation of the labels (from the text or from the acoustics), these two levels are distinguished according to the presence or absence of a subsequent silence. Unlike H and L syllables, HH and LL syllables are directly followed by a silent pause. In terms of *intonation*, we define two pitch levels (H and L), similarly to ToBI. This integrates an **acoustic** dimension in the model, even if the two pitch levels are supposed to be related to two different functions of phrasing, *i.e.* continuative or conclusive breaks. It should be noted that the model makes no distinction between rising vs. high syllables and falling vs. low syllables. This choice was dictated by the functional objective of our approach which facilitates the prediction of labels from text. In that sense, both rising and high tones could be seen as indicating continuative contours (vs. conclusive contours for low or falling tones). Our acoustic analysis of the labels and the resulting speech synthesis will be evaluated to determine if such a reduction triggers major issues. A specific tag E is assigned to the final boundary of player names enumerations, which are very common in sports commentaries and may display a specific acoustic realization.

A focus stress (F) relates to emphatic stresses. No specification is given regarding its position, which means that it can be assigned to initial, middle or final syllables. The use of a single emphatic label constitutes a clear reduction of the set of possible acoustic realizations of that stress into a single averaged manifestation. This is further investigated in Chapter 7. Here again, our assumption is that some of the acoustic variations will be automatically modeled based on linguistic contextual information, by the automatic clustering of the models.

In addition, an hesitation label (He), and a creaky label (C) allow avoiding the degradation of the models at training time. Indeed, hesitations are realized with long durations while creaky syllables are characterized by a very low pitch (Drugman et al., 2012). If these syllables are not singled out, their prosodic features may influence the synthesized prosody.

All remaining syllables are assigned a NA symbol.

Table 6.1 - *List of local labels.*

Stresses		Unstressed	Other
Not emphatic	Emphatic		
H <i>Rising/high minor phrase boundary</i>	F <i>Focus</i>	NA <i>Unstressed</i>	He <i>Hesitation</i>
HH <i>Rising/high major phrase boundary</i>			C <i>Creaky</i>
L <i>Falling/low minor phrase boundary</i>			
LL <i>Falling/low major phrase boundary</i>			
E <i>Player name enumeration boundary</i>			

It should be noted that a parallel can be established between our annotation framework and that proposed in Obin et al. (2011). Likewise, they distinguish

between two prosodic boundary levels, based on the presence of a subsequent pause. They also indicate what they call "prosodic prominences" by a single label, similarly to our F label. It could be seen as a further reduction of our model in which syllable height is discarded.

The global tier is inspired by Kern (2010) and Trouvain (2011) and is assigned to groups of words, unlike the local annotation that assigns a symbol to each syllable. It basically classifies the speech segments into sub-genres (or speaking styles), based on a dimensional analysis of emotions (Mehrabian and Russel, 1974; Russell, 1980). Valence and arousal levels drove us to define five sub-genres (Figure 6.2). This representation can be seen as similar to the activation-evaluation space proposed in Cowie et al. (2001). Activation is interpreted as "the strength of the person's disposition to take some action rather than none", and is opposed to passivity (Cowie et al., 2001). It can be seen to be related to the excitation level (Liscombe et al., 2003). Valence describes the positive or negative aspect of the emotion³. It should be noted that other studies including Schröder et al. (2001) introduce a third dimension related to power which regards the dominant/submitted position of the speaker. This dimension is not considered in our annotation as it can be seen as rather stable throughout the corpus.

Noteworthy is the fact that the corpus was initially annotated with an additional tag, *MinorInfo*. This tag was associated with discursive parentheses which usually occur when nothing interesting is happening in the game. Such parentheses generally include facts about the players' biographies or general contextual information. The acoustic analysis of the corresponding segments showed that this sub-genre was not acoustically specific and could rather be relabeled *Neutral*, which was done for the whole corpus. While these sub-genres are assumed to display specific global **prosodic** realizations (*e.g.* F0 range, speaking rate) they should also be characterized by specific **voice quality** realizations (*e.g.* more tense or relaxed voice).

The corpus was manually annotated with the two-tier annotation. Only one stress-related label could be assigned to each syllable, possibly associated with He or C. If an emphatic stress fell on an accentual phrase boundary, the F label overwrote the non-emphatic stress. The annotation results from two or three listenings of each sequence of 4-5 words and was submitted to a second check by the same annotator. This was performed with Praat (Boersma, 2001) with acoustic cues displayed (F0, intensity and formants). For local labels, it allowed avoiding the annotation of stresses with no specific acoustic realization, which would degrade the quality of the synthesis. Indeed, it has been shown in Goldman et al. (2010) that human annotators tend to overdetect prominences in specific grammatical locations where a stress is expected, or underdetect them on clitic words like determiners. We therefore carefully tried to avoid such issues, the objective of our annotation protocol being to assess the acoustic

³It should be noted that the term 'neutral' for the sub-genre with a null valence and low excitation level may have not been ideally chosen as it does not correspond to the definition of neutral as proposed in the introduction of this thesis (*i.e.* neutral speech being seen as corresponding to read non-emotional laboratory speech as usually used in speech synthesis). It should be mentioned that the 'neutral' sub-genre as annotated here is still very specific to sports commentaries and can therefore be considered as rather expressive in that sense.

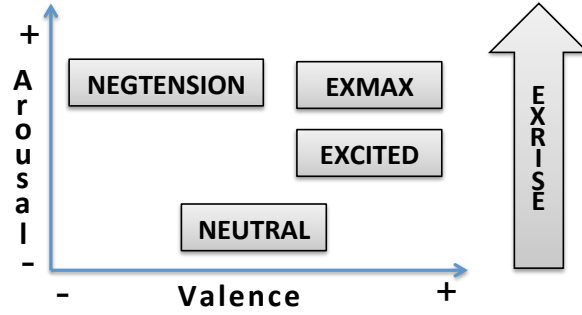


Figure 6.2 - The global labels on a dimensional scale.

realization and not to offer a perceptive representation of the corpus. In that sense, our model clearly differs from **perceptive** approaches as proposed by Mertens (1987) and d’Alessandro and Mertens (1995).

6.2.2 Inter-annotator rate

Twenty percents of the corpus was annotated by a second expert, the initial annotation being hidden. The local tier reaches a Cohen’s kappa score (Cohen, 1960) of 0.66, with an observed agreement of 80.22 %, for the 7 local labels (see confusion matrix in Table 6.2). This is comparable to the inter-annotator rate achieved by ToBI (Silverman et al., 1992; Pitrelli et al., 1994), when relying on the absence vs. presence of phrase accents, pitch accents or phrase boundaries only. For the global tier, a lower kappa score was achieved, *i.e.* 0.38, with an observed agreement of 54.33 %. As for the local annotation, this score was computed at the syllable level. It should however be noted that the confusion matrix shows logical confusions between the sub-genres (see Table 6.3). It is not surprising that *Neutral* and *Excited* (as well as *ExMax* and *Excited*) tend to be interchanged, the distinction between *Neutral*, *Excited*, and *ExMax* proposing a discretization of the excitation/arousal continuum.

It should be noted that this annotation of a portion of the corpus by a second expert was only used for evaluation purposes of the annotation protocol. That second annotation was not further exploited in the remainder of this thesis and disagreement cases were therefore not resolved. This results from the fact that only part of the corpus was re-annotated and that a correction of that part (*e.g.* based on discussions between the two annotators) would have led to a non-uniform annotation of the corpus.

6.3 Acoustic analysis of the prosodic labels

As mentioned in the previous section, our annotation framework is based on a functional approach to prosody, to allow an easy prediction of the labels from text.

Table 6.2 - *Inter-annotator confusion matrix for the local annotation of the corpus, based on the annotation of 20 % of the corpus by a second annotator.*

	NA	L	LL	E	H	HH	F
NA	2345	29	22	4	207	5	46
L	3	117	1	1	42	2	2
LL	5	1	245	0	1	33	0
E	0	0	4	112	3	5	0
H	30	53	0	1	197	0	7
HH	6	1	37	0	0	111	1
F	60	4	15	2	20	20	114

Table 6.3 - *Inter-annotator confusion matrix for the global annotation of the corpus in sub-genres, based on the annotation of 20 % of the corpus by a second annotator. Agreement is computed at the syllable level.*

	Neutral	Excited	ExMax	NegTension	ExRise
Neutral	1047	147	0	77	17
Excited	869	499	144	53	28
ExMax	14	20	152	24	0
NegTension	122	68	17	247	3
ExRise	97	124	58	9	305

In this section, we investigate to what extent our labels are also characterized by distinct acoustic realizations. A good acoustic discrimination between the prosodic labels should improve their generation at synthesis. In fact, if syllables with an identical prosodic label highly differ in terms of acoustics, the HMM might produce an averaged realization which would not fit the intended function. This would occur only if other linguistic contextual information (*e.g.* parts of speech, position of the syllable in the word) do not discriminate between the different acoustic realizations of a same label.

6.3.1 Analysis of the local annotation

Table 6.4 displays the average acoustic values for four discriminant prosodic features. Duration measures are extracted with Prosogram(Mertens, 2004a). The total perceptual loudness (Peeters, 2004) and the fundamental frequency (using the SRH (Drugman and Alwan, 2011) algorithm) are also analyzed. The fundamental frequency is further transformed in semi-tones (ST) which provide a more accurate representation of perceived pitch. Dynamic values are computed as the difference between the value of the last and first frames of 10 ms of the syllable.

It should first be noted that the vocalic nucleus duration (*i.e.* the duration of the vowel) helps distinguishing between intonational boundaries followed (LL and HH) and not followed (L and H) by a silence. The duration of focused nuclei (F) remains

Table 6.4 - *Average acoustic values of the local labels in the corpus.*

	Vocalic nucleus duration (sec)	Mean F0 (ST)	Dyn F0 (ST)	Dyn loudness (dB)
H	0.07	56.90	1.03	-2.2
HH	0.14	56.87	0.16	-26.2
E	0.11	53.45	2.02	-16.5
L	0.10	56.85	-1.33	-2.7
LL	0.12	55.08	-2.49	-26.0
F	0.07	58.18	4.75	5.8
He	0.18	53.70	-1.23	-3.3
C	0.03	51.74	-0.58	13.6
NA	0.05	55.76	0.81	9.2

fairly low. On the other hand, the mean pitch for F is higher than that of all other syllables. This confirms the widespread assumption stating that emphatic stresses in French are more realized in terms of pitch than lengthening, unlike boundary stresses (Lacheret-Dujour and Beaugendre, 1999). The dynamic pitch distinguishes between H and L pitch levels, but it should be noted that the average amplitude of the slope is rather low which indicates that most corresponding syllables are not characterized by large intrasyllabic pitch variations.

Emphatic stresses in our corpus show a markedly rising pitch. Finally, the loudness dynamics draw the distinction between emphatic and non-emphatic stresses, the latter displaying a falling loudness pattern because of their location at the end of an accentual or an intonational phrase. Like HH, E is characterized by a rising pitch, vocalic nucleus lengthening and falling loudness. It is, however, realized with a clearly lower pitch.

A detailed study of the emphatic stresses in our corpus highlighted several interesting points. Unlike non-emphatic stresses, emphatic stresses tend to fall on numbers (in 27 % of the cases against 9 % for non-emphatic stresses). This finding is clearly specific to sports commentaries. Numbers often refer to scores and play a significant role in the expressive function. It was also found that emphatic stresses tend to fall on the first syllable of the word (as stated by Séguinot (1976)) unlike non-emphatic stresses. They are also often preceded by a silence.

A final aspect of our study focused on the comparison between the local labels and annotations provided by automatic prominence detection tools. Prosoprom (Goldman et al., 2007) automatically annotates each syllable as prominent or not. The annotation relies on rules learned on a French annotated corpus and considers various prosodic features (*e.g.* F0, duration, silences). 42.1 % of our stressed syllables (*i.e.* H, HH, L, LL, E and F) are regarded as prominent by Prosoprom, against 10.9 % for the unstressed syllables. F-labelled syllables achieve the highest percentage (60.6 %). Prosoprom was recently developed into a gradual prominence annotation tool, Prom-Grad (Goldman et al., 2012), which defines five prominence levels, ranging from 0 to 4. Table 6.5 shows the average prominence value (and its standard deviation) assigned

to each label of our corpus. The two boundary levels related to non-emphatic stresses are clearly reflected by the automatic annotation.

It should be noted that PromGrad defines the prominence level of a syllable based on a complex value combining various prosodic parameters (*i.e.* mean pitch, duration, pitch movement and duration of the following pause). This means that, if a stress is realized by few distinct features, it is less likely to be assigned a high prominence level. It is here clearly the case for emphatic stresses that are less characterized by syllable lengthening and are usually not followed by a pause. An algorithm considering the presence of a preceding pause might alleviate this issue.

While large variability is found in the predicted levels, it should be noted that some labels are clearly distinguished; HH, for instance, is assigned a prominence of level 4 in 69 % of the cases and NA receives a 0 prominence level for 85 % of its occurrences. This relation between our annotation and PromGrad (Goldman et al., 2012) implies that it could be possible, to a certain extent, to automatically predict the local labels from their acoustic realization. This is further investigated in Chapter 8.

Table 6.5 - *Average prominence levels (and their standard deviation) assigned by Promgrad (Goldman et al., 2012) to our labels in the corpus.*

H	HH	E	L	LL	F	He	C	NA
0.9 $\pm$ 1.1	3.2 $\pm$ 1.4	2.7 $\pm$ 1.7	0.7 $\pm$ 1.0	2.8 $\pm$ 1.6	1.5 $\pm$ 1.5	1.4 $\pm$ 1.5	0.3 $\pm$ 1.5	0.2 $\pm$ 0.8

6.3.2 Analysis of the global annotation

An analysis of the prosodic realization of each sub-genre is shown in Table 6.6. The ratio of rises corresponds to the percentage of rising syllables, as detected by Prosogram (Mertens, 2004a). The articulation rate is computed as the number of syllables divided by the articulation duration. Finally, the ratio of silences is computed as the total duration of silences in each sub-genre divided by the total duration of this sub-genre. It should be noted that initial and final silences are not considered as they could be arbitrarily assigned to the previous or following sub-genre.

Table 6.6 - *Average acoustic values of the global labels in the corpus.*

	Mean F0 (ST)	Mean loudness (dB)	Ratio rise (%)	Articulation rate (syll/sec)	Ratio sil (%)
Neutral	55.93	29.7	2.6	5.6	46
Excited	56.93	33.4	2.3	6	32.8
ExMax	58.83	41.1	1.5	5.2	18
NegTension	56.11	32.6	2.7	5.6	32.7
ExRise	56.11	29.4	15.5	5.5	29.5

A first insightful observation concerns the higher pitch value of the *ExMax*

sub-genre. This finding is in line with other studies pointing out higher F0 values during and after shots (Trouvain, 2011; Audrit et al., 2012), which are, in our corpus, often assigned an *ExMax* label. As expected, *Excited* is realized with a pitch frequency lying between *Neutral* and *ExMax*. Similar observations can be made for the loudness, with higher values for *ExMax* and intermediate values for *Excited*. Noteworthy, however, is that *NegTension* is characterized by a relatively low pitch, despite its high arousal level on the dimensional scale. This confirms the assumption of Trouvain (2011) who states that disappointment should lead to a depressed pitch value. The ratio rise clearly draws a distinction between *ExRise* and the other sub-genres, *ExRise* displaying a continuous rising of the pitch. Unlike what might be expected, the articulation rate is rather stable across all styles. This phenomenon has already been noticed in other studies on sports commentaries (Trouvain and Barry, 2000; Kern, 2010). However, we observe a clear reduction of the ratio of silences, according to the arousal level, which indicates a rise in the speaking rate, when including silences.

We also analyzed the probability of transition from one sub-genre to another (see Table 6.7). As expected, *ExRise* usually leads to *ExMax* (0.38) or *NegTension* (0.38), depending on the success or failure of the action and on the involved team. The end of *ExMax* and *NegTension* usually indicates the end of the action and is followed by a *Neutral* tag. This transition matrix, along with the average duration of each sub-genre segment (Table 6.8), may be considered when automatically selecting global tags at synthesis time (see Chapter 8).

Table 6.7 - *Transition probability matrix for global prosody annotation computed on the corpus.*

		To				
		Neutral	Excited	ExMax	NegTension	ExRise
From	Neutral	0	0.35	0.06	0.16	0.43
	Excited	0.46	0	0.15	0.14	0.25
	ExMax	0.43	0.32	0	0.16	0.09
	NegTension	0.66	0.22	0.03	0	0.10
	ExRise	0.12	0.12	0.38	0.38	0

Table 6.8 - *Average duration of each sub-genre in the corpus.*

	Neutral	Excited	ExMax	NegTension	ExRise
Average duration (sec)	13.4	4.1	7.1	5.1	3.9

Finally, the evolution of the acoustic features was studied. The objective was to investigate whether a clear change was noticed at the intersection between two sub-genres, as this would facilitate an automatic annotation of new corpora based on the acoustics only. ProsoDyn (Goldman, 2012) produces such a representation

of the prosodic dynamism of a speech segment. The average value of the prosodic features is computed on a default sliding window of 19 syllables. Its application on a representative speech file of the corpus is shown in Figure 6.3, with the segmentation in sub-genres at the bottom. The evolution of the pitch, in semi-tones, is shown to clearly follow the arousal level.

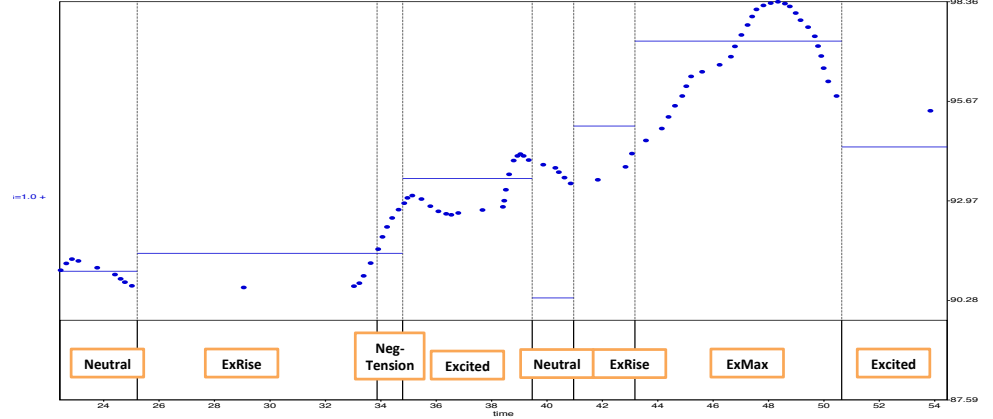


Figure 6.3 - *Dynamic representation of the pitch on a small part of the corpus with ProsoDyn (Goldman, 2012) and its relationship with sub-genre labels.*

Further studies should investigate the interconnections between both levels of annotation. It is indeed De Looze’s assumption that local events depend on the macro prosodic level (De Looze, 2010b). Subsection 6.3.1 showed that F syllables display a higher pitch. However, this might be due to the fact that F syllables are more frequent in *ExMax*, in which the average F_0 is higher. The different realizations of the stresses should therefore be studied within the distinct sub-genres. For more generic values of the local labels, the acoustic values could, for instance, be normalized by the average values of the sub-genre. If there is an interconnection between both levels, this would make up for this two-level annotation, as it would show that each tier is dependent on the other. Beside their realization, the location and frequency of the local labels in each sub-genre should also be further investigated. This would help in predicting the local tags directly from text only, when the global tag of the sentence is known.

6.4 Application to speech synthesis

In this section, we aim at integrating efficiently the local and global annotations into an HMM-based speech synthesizer.

The goal of this study is two-fold:

- Quantifying the possible improvements brought by each annotation layer on various aspects of speech synthesis;
- Comparing different training methods regarding the integration of the global labels.

To that end, our study relies on the *Sportic* corpus, presented in Section 5.6.3, manually-annotated in terms of local and global prosody according to the aforementioned protocol.

6.4.1 Methodology

In order to assess the validity of our local and global labels definition, several HMM-based speech synthesizers were built, relying on the implementation of the HTS toolkit (version 2.1) (Zen et al., 2007a). The speech waveforms were sampled at 16 kHz and windowed by a 25 ms Hamming window with a 5 ms shift. The observation vector consisted in the static, delta and delta-delta of the log frequency and of Mel Generalized Cepstral (MGC) coefficients (with $\alpha = 0.42$, $\gamma = 0$ and order of MGC analysis = 24), traditionally used in parametric synthesis. We used 5-state left-to-right HMMs. The context features of the models were determined based on eLite-HTS as presented in Subsection 5.5. As excitation modeling, we applied the Deterministic plus Stochastic Model (DSM) of the residual signal proposed in Drugman and Dutoit (2012), as it was shown to significantly improve the naturalness of the delivered speech.

For each synthesizer, 90 % of the corresponding database was used for the training (called the *training set*), leaving around 10 % for the synthesis (called the *synthesis set*). The training and synthesis label files corresponded to the manually-checked annotation of the corpus.

The influence of the integration of the local and global labels is first analyzed and quantified independently in Subsections 6.4.2 and 6.4.3, respectively. Using the conclusions drawn from these latter evaluations, the combination of both local and global labels is investigated in Subsection 6.4.4.

6.4.2 Integration of the local labels

This section is devoted to the integration of the single local annotation layer into HMM-based speech synthesis.

Method

As pointed out in the previous chapter, HMM-based speech synthesis allows for a straightforward integration of prosodic information directly in the contextual information exploited by the decision trees to cluster the acoustic models. This technique was shown to yield high quality prosody when integrating ToBI labels in Tokuda et al. (2002b) and is investigated in this section with our reduced set of local prosodic labels.

A first synthesizer is trained on the entire training set of the *Sportic* database, regardless of the speaking styles. This is our baseline system, called *Base*. The only contextual information provided during the training and synthesis stages is that automatically generated by eLite-HTS. It should be noted that this information is based on the manually-checked phonetic transcription of the corpus, which is made possible by the structure of eLite-HTS (see Section 5.5). The same training procedure is applied to the second synthesizer, called *Loc*. It makes use of the same contextual information, but complemented in this case with the specific information from the local prosody annotation level (Table 6.1).

Evaluation protocol

A first Mean Opinion Score (MOS) test (Grancharov and Kleijn, 2008) is conducted in order to quantify the impact of the local annotation layer in comparison with the baseline system. For this evaluation, testers were asked to listen to two versions of the same sentence synthesized by (i) the baseline system (*Base*) and (ii) the system integrating local labels (*Loc*). The two versions were randomly shuffled. Each sentence was scored according to three criteria: intelligibility⁴, expressivity and segmental quality. Listeners were given three continuous MOS scales for each version (one for each criterion) ranging from 1 (meaning “poor”) to 5 (meaning “excellent”). These scales were extended one point further on both sides (ranging therefore from 0 to 6) in order to prevent border effects.

It should be noted that assessing intelligibility in such an evaluation framework is rather unusual. Most intelligibility assessments commonly rely on dedicated tests, based on Semantically-Unpredictable Sentences (SUS), in which the listener is asked to transcribe the synthesized sentences. In our case, seeing the high number of subjective tests in our study, this would have made the evaluation more complex. Another criticism might be that HMM-based synthesis has reached a fairly high quality level in which intelligibility is not an issue anymore. However, sports commentaries are an exception in that sense because of the high density of player and team names. This clearly hinders the understanding of the message, especially when prosodic boundaries are erroneous. In that case, it is sometimes hard to identify names in the speech flow. This explains why we included intelligibility in our criteria. One should be careful when interpreting those results as they may have been biased by the unusual evaluation framework.

The test consists of 10 randomized pairwise comparisons. Sentences were randomly chosen amongst the synthesis set of the database. These sentences were not divided into speaking styles, which means that a sentence may correspond to a sequence of various speaking styles. 10 native French speakers, mainly naive listeners, participated in this evaluation. They were carefully explained what was meant by the criteria of ‘intelligibility’, ‘expressivity’ and ‘segmental quality’.

⁴Testers were given no prior information about player names and teams to imitate real listening conditions.

Results

MOS scores are computed for all evaluations in this chapter to provide comparable results in a coherent evaluation framework. The actual MOS scores are, however, less informative in this first evaluation. Besides, averaging MOS values results in losing individual preferences. Therefore, our analysis relies on the preference percentages which are computed as the listener’s relative preference for a synthesis method compared to another. Figure 6.4 shows the preference scores for the three criteria. The light grey segment corresponds to the proportion of cases in which both methods are assigned the same MOS score. It can be observed that *Loc* is preferred for the rendering of the expressivity. This means that local labels were properly learned during the training of *Loc* and that specific accentual Probability Density Functions (PDFs) were properly estimated. At synthesis time, the model was thus able to predict more precise accentual realizations. Interestingly, it is also shown to improve the segmental quality, while achieving an intelligibility level that is similar to the baseline (*i.e.* *Base*).

An example of synthesized sentences is shown in Figure 6.5. It shows that the integration of local prosodic information allows producing, for a focused syllable, a higher and more dynamic pitch contour, along with a lengthening of the syllable. Conversely, since no prosodic annotation was provided for the *Base* training, all acoustic realizations that should have corresponded to distinct local labels were merged into more global PDFs.

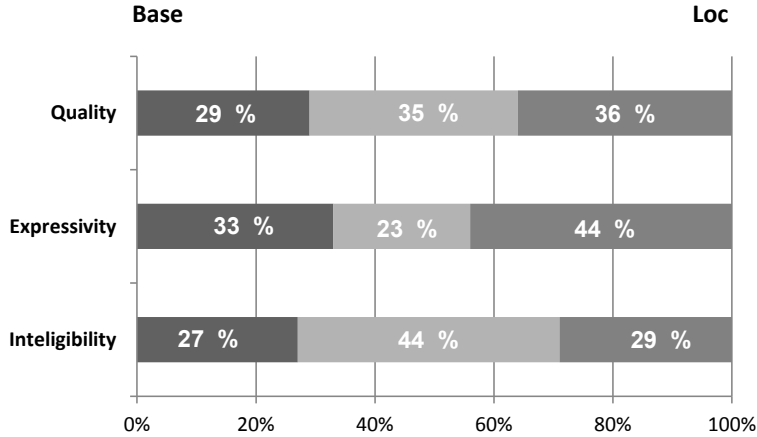


Figure 6.4 - Preference scores for each criterion and each synthesis method, *i.e.* with (*Loc*) and without (*Base*) the integration of the local labels.

6.4.3 Integration of the global labels

This section studies the integration of the single global annotation layer into HMM-based speech synthesis. It should be noted that the speaking styles defined

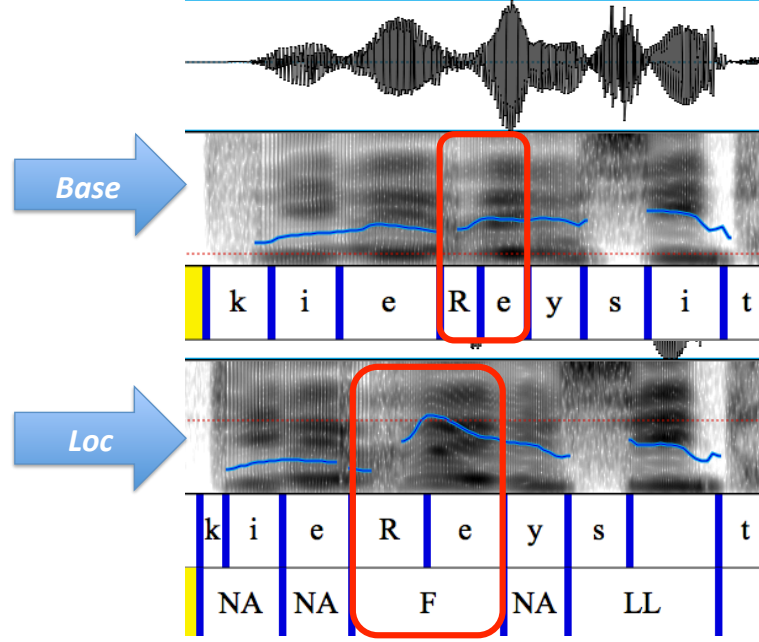


Figure 6.5 - *Synthesis (Syn1 and Syn2) of a sentence with focus.*

by the global annotation layer are not uniformly distributed throughout the corpus. The total duration of each speaking style composing the database is shown in Table 6.9.

As shown in Subsection 5.4.2, one way to perform expressive HMM-based speech synthesis is to train a model, called a *style-dependent model* (Yamagishi et al., 2005), using a database containing specific data (*e.g.* data corresponding to a particular speaker or emotion). Another way to build the models is to make use of adaptation techniques, which allow changing the voice characteristics and prosodic features of a source speaker into those of a target speaker (Yamagishi et al., 2009b). These latter adapt the source HMM-based model with a limited amount of target speech data. The resulting model is called *adapted model*. The same concept holds for speaking style adaptation (Yamagishi et al., 2004). This technique allows providing high quality speech synthesis using a limited amount of adaptation data (Yamagishi, 2006).

Another common technique lies in the integration of the speaking styles as contextual information exploited by the decision trees (Tsuzuki et al., 2004; Yamagishi et al., 2005; Eyben et al., 2012). This would imply a similar processing of both local and global labels in our case. In this study, we rely only on the two first approaches (*i.e.* style-dependent modeling and adaptation) as this third method was shown to achieve poorer results in terms of expressivity in Eyben et al. (2012). We believe that adaptation techniques are particularly suited in the framework of the synthesis

of global sub-genres or speaking style, thereby modeling the global prosody and voice quality feature of that specific style.

Table 6.9 - *Total duration (in sec.) of the various sub-genres in our database, long silences (>1 sec.) being excluded.*

Excited	ExMax	ExRise	NegTension	Neutral
1032	475	485	582	2955

Method

Three distinct methods are investigated regarding the integration of the global labels in speech synthesis. The first method consists in training speaking style-dependent models on exclusive subsets of the whole corpus, specific to the global label they correspond to (see Table 6.9). They will be referred to as *style-dependent models* in the remainder of the thesis. At the end of this step, 5 different style-dependent models, called *Glob1*, are obtained (*i.e.* Excited, ExMax, ExRise, NegTension and Neutral).

The two other methods exploit adaptation techniques. The second method relies on the fact that the Neutral style has the highest amount of speech data amongst the different speaking styles. It can also be considered as a central point in the space of emotions (Mehrabian and Russel, 1974). Assuming that this amount of speech data is sufficient to obtain a strong Neutral *style-dependent model*, voice adaptation techniques (Yamagishi et al., 2009b) can be applied to train more reliably the remaining models, for which less speech data are available. The *Glob1* Neutral *style-dependent model* was then adapted using the Constrained Maximum Likelihood Linear Regression (CMLLR) transform (Digalakis et al., 1995; Gales, 1998) in the framework of Hidden Semi Markov Model (HSMM) (Ferguson, 1980) with the adaptation sets of the four remaining speaking styles. It produces respectively Excited, ExMax, ExRise and NegTension HMM-based synthesizers. The linearly-transformed models were further optimized using MAP adaptation (Yamagishi et al., 2009b), providing the 4 adapted models called *Glob2*.

A potential drawback of the second method is that the speech data used to train the Neutral *style-dependent model* may not be large enough. To alleviate this issue, Yamagishi and Kobayashi (2007) proposed to adapt a so-called *average-voice model* to a particular target speaker. The average-voice model is computed once and for all over a database containing many different speakers. This technique proved to be efficient when little speech data is available. Such average voice models were also proposed in the framework of emotion or speaking style synthesis in Qin et al. (2006) and Eyben et al. (2012) and yielded high quality results. In this section, we propose to train an average-voice model on the entire database, regardless of the speaking styles. This model is then adapted following the same procedure as for *Glob2*. We

finally obtain 5 different adapted sub-synthesizers called *Glob3* (*i.e.* Excited, ExMax, ExRise, NegTension and Neutral).

Evaluation protocol

A second MOS test is conducted in order to determine which of the three methods is the most suited for the integration of global labels representing the speaking styles. For this evaluation, participants were asked to listen to three versions of the same sentence, synthesized by the models corresponding to the three aforementioned methods. Each sentence was scored according to two criteria: its intelligibility, and its segmental quality. The same experimental protocol as in Section 6.4.2 was applied. However, unlike Section 6.4.2, the expressivity was not assessed here. Indeed our first informal experiments showed that *Glob* models exhibit some intelligibility and segmental quality issues. These had to be addressed before focusing on a good rendering of the expressivity.

The test consists of 15 randomized triplets. Sentences were randomly chosen amongst the synthesis set of the database. Unlike Section 6.4.2, each sentence only contains one speaking style.

12 native French-speaking testers, mainly naive listeners, participated in this evaluation. During the test, they could listen to the triplet of sentences as many times as wanted in the order they preferred. They were nonetheless advised to first listen to the three sentences in a row so as to estimate approximately their relative position. However, they were not allowed to come back to previous sentences after validating their decisions.

Results

For each synthesis technique and each speaking style, Figures 6.6 and 6.7 display respectively the averaged intelligibility and segmental quality MOS scores, together with their 95 % confidence intervals (CI). It clearly turns out that *Glob3*, *i.e.* the adapted average models, provides the highest results, both in terms of intelligibility and segmental quality of the generated speech. The *style-dependent models*, conversely, achieve the lowest scores in most speaking styles. As a reminder, a score of 3 or 4 on the MOS scale means respectively “Fair” or “Good”.

This preference for the adapted average models can be explained by the fact that they are computed using all the training sets for each speaking style, thus providing a robust model which is then adapted to each speaking style. It can however be noted that Neutral, Excited and NegTension voices are better rendered than ExMax and ExRise ones. This is probably due to the fact that Neutral, Excited and NegTension have more speech data, leading to a better average-voice model adaptation compared to ExMax and ExRise.

It should also be noted that all the synthesizers achieved the same performance for the Neutral speaking style. This can be understood by the fact that this latter style

is the only one having a comfortable amount of speech data for a reliable estimation of the model, independently of the training method.

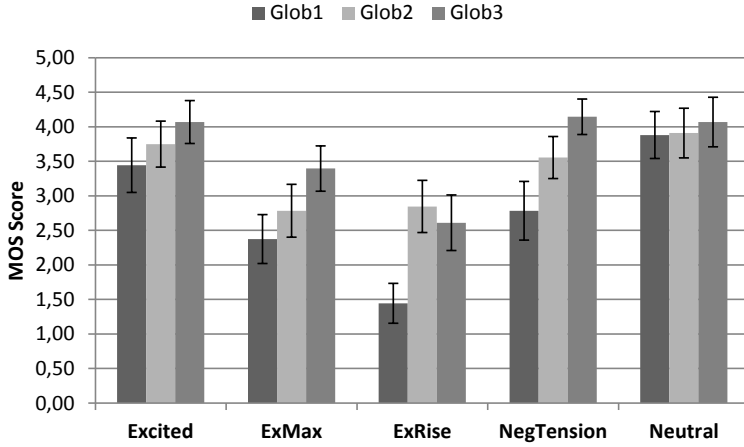


Figure 6.6 - Averaged *intelligibility* MOS scores for each synthesis method and each speaking style, with their 95 % confidence interval.

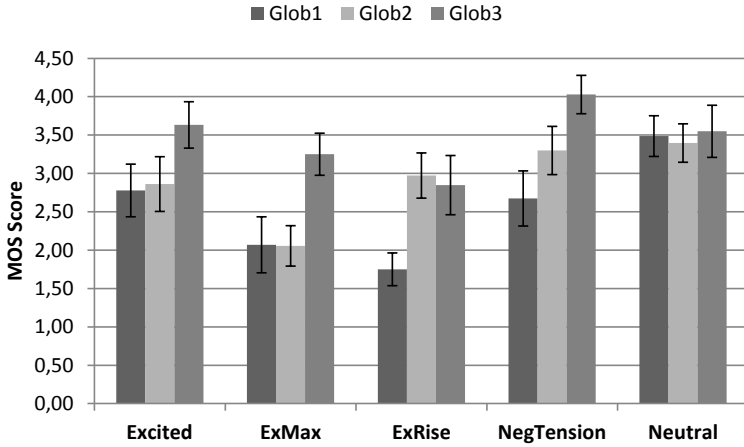


Figure 6.7 - Averaged *segmental quality* MOS scores for each synthesis method and each speaking style, with their 95 % confidence interval.

Preference scores corroborate these findings. Regarding intelligibility, *Glob3* is preferred in 83.9 % and 66.7 % of the cases compared to *Glob1* and *Glob2*, respectively. It is respectively chosen 83.1 % and 73.9 % of the time for its segmental quality.

This shows that, as the amount of speech data is unevenly distributed amongst the different speaking styles, adapting a robustly trained average-voice model with an efficient technique such as CMLLR+MAP allows generating various speaking styles

of reasonable intelligibility and segmental quality.

Comparison with the baseline

Similarly to the integration of local labels (Section 6.4.2), the integration of global labels was compared to the baseline through a MOS test. For this comparison, the best integration technique, *i.e.* the adaptation of the average model, was used. The baseline is the same as in Section 6.4.2, which means that it disregards both prosody annotation layers. 20 native French speakers, mainly naive listeners, participated in this evaluation. The averaged MOS scores obtained for each of the three criteria are shown in Figure 6.8. No statistically significant differences were observed between both methods. This integration achieves indeed comparable or even slightly lower scores in terms of intelligibility, expressivity and segmental quality.

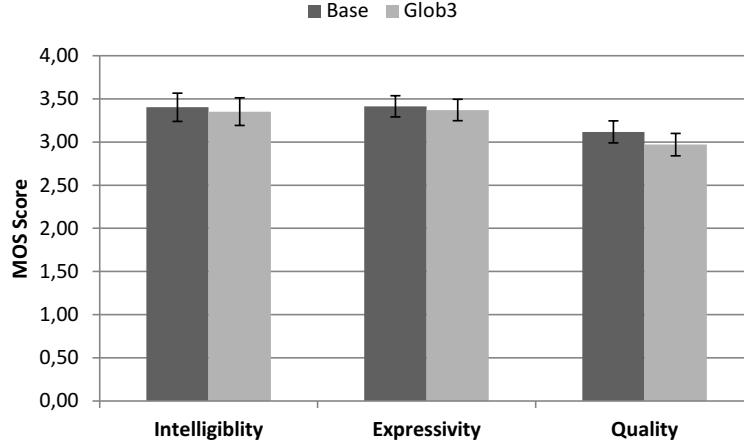


Figure 6.8 - Averaged quality MOS scores for each criterion and each synthesis method, *i.e.* with (Glob3) and without (Base) the integration of the global labels, together with their 95 % confidence interval.

6.4.4 Integration of both local and global labels

This section is devoted to the integration of the two-layer annotation (both local and global labels) into HMM-based synthesis.

Method

We showed in Section 6.4.2 that the integration of local labels results in an enhanced expressivity, while it provides slightly higher intelligibility and segmental quality performance. Regarding the integration of global labels, the use of adaptation techniques, from an average model, was shown to provide the best results (see Section 6.4.3). However, the integration of this annotation level seemed to achieve no improvement regarding expressivity in comparison with a baseline model.

We investigate, in this last section, whether the integration of both local and global labels achieves higher scores. These combined models are referred to as *Loc+Glob3*.

Evaluation protocol

A third MOS test is conducted. For this evaluation, listeners were asked to listen to four versions of the same sentence synthesized by the following models (randomly shuffled): (i) the baseline model (*Base*); (ii) the model integrating local labels (*Loc*); (iii) the model adapted from the average-voice model (*Glob3*); (iv) the model adapted from the average-voice model integrating local labels (*Loc+Glob3*). Here again, intelligibility, expressivity and segmental quality are evaluated. The same experimental protocol as in Section 6.4.2 was applied.

The test consists of 10 randomized quadruplets. Sentences were randomly chosen amongst the synthesis sets of each speaking style of the database. Similarly to Section 6.4.3, each sentence only contains one speaking style. 20 native French-speaking people, mainly naive listeners, participated in this evaluation.

Results

Here again, the results are presented in terms of preferences. Table 6.10 shows the preference scores for the four methods. It should be noted that the scores obtained by two reverse pairs are not summing to 100 %. This is due to the fact that cases when both methods composing the considered pair are found to be equivalent are not mentioned in the table. Regarding intelligibility, Table 6.10 shows for example that *Base* is preferred to *Loc* in 31 % of the cases, while *Loc* is preferred to *Base* in 35.5 % of the cases. The remaining percentage, *i.e.* 33.5 %, corresponds then to the cases where both *Base* and *Loc* are equivalently preferred.

As in Section 6.4.2, it is observed that the integration of the local labels carries out an improvement in the rendering of the expressivity and provides comparable or slightly better intelligibility and segmental quality of the generated speech. On the other hand, the integration of the global labels alone (using the adapted average models) does not improve any of the analyzed criteria compared to the baseline, which corroborates the results obtained in Section 6.4.3. Regarding the integration of both prosody annotation levels, an insightful observation is that *Loc+Glob3* is preferred in 40 % of the cases in terms of expressivity against *Loc*, which is assigned 35 % of the preferences. The segmental quality degrades, however, from *Loc* to *Loc+Glob3* as they are respectively preferred 39.5 % and 31.5 % of the time. Nonetheless, both methods achieve similar intelligibility performance.

6.4.5 Discussion of the evaluation framework

It should be noted that the evaluation carried out in the framework of our study differs greatly from most other evaluations of expressive speech synthesizers. In fact, listeners are usually asked to make a forced choice between various emotions. While

such evaluations assess the degree of discrimination of the models across the different emotions, it provides "no information about the quality of the stimulus in terms of naturalness of believability" (Schröder, 2001). The evaluation protocol developed here also better fitted our purpose as it allowed for the comparison between the integration of the two annotation levels. Indeed, the integration of local prosody information only does not allow for a discrimination between the different speaking styles or sub-genres. It should however be noted that an evaluation based on the identification of the various speaking styles is proposed in Chapter 8, focusing on the integration of both local and global annotation tiers.

A second element that is worth pointing out is that, unlike many other studies, test sentences used in our evaluation are extracted from the *Sportic* test corpus and are not, in that sense, semantically neutral carrier sentences. The unnaturalness of the use of neutral sentences to evaluate expressive speech synthesis has been pointed out by Schröder (2004). While such sentences are usually required to avoid an influence of the content in the framework of forced-choice tests, our specific evaluation protocol relaxes this constraint. We believe that this provides a much more believable testing framework (assessing the quality of a sports commentary voice by synthesizing semantically-neutral sentences would, in our opinion, make no sense).

Table 6.10 - *Integration of both Global and Local Labels - Preference scores (in [%]) for each method and each criterion.*

		<i>Base</i>	<i>Loc</i>	<i>Glob3</i>	<i>Loc+Glob3</i>
Intelligibility	<i>Base</i>	0	31	33	29.5
	<i>Loc</i>	35.5	0	37.5	29.5
	<i>Glob3</i>	29	29	0	26.5
	<i>Loc+Glob3</i>	40	27.5	38	0
Expressivity	<i>Base</i>	0	37.5	40	32.5
	<i>Loc</i>	45.5	0	46	35
	<i>Glob3</i>	35.5	35	0	29.5
	<i>Loc+Glob3</i>	51	41	47.5	0
Quality	<i>Base</i>	0	39.5	40	44
	<i>Loc</i>	38.5	0	46.5	39.5
	<i>Glob3</i>	27.5	33.5	0	35
	<i>Loc+Glob3</i>	36	31.5	45.5	0

6.5 Conclusion and perspectives

This chapter proposed a two-layer prosody annotation specific to live sports commentaries with HMM-based speech synthesis in view. The local annotation relates to accentual phenomena while the global layer classifies the speech segments into distinct speaking styles. The interest was to propose labels which were characterized by both a distinct expressive function and a rather stable acoustic realization. This double

constraint should facilitate the automatic prediction of the labels both from the text or from the speech signal (see Chapter 8). The manual annotation of a 2 hour-long corpus of basketball commentaries showed that our different labels are, as expected, related to specific acoustic realizations. Good inter-annotator rates were also reached.

The integration of this prosodic annotation in HMM-based speech synthesis was further investigated. This study was divided into three parts.

First, the improvement carried out by local labels was quantified by comparing:

- a baseline model including no explicit prosodic contextual information
- a model integrating local labels as contextual information, similarly to ToBI tones

Subjective tests revealed that, compared to the baseline, the integration of local labels results in an enhanced expressivity, while providing slightly higher intelligibility and segmental quality scores.

Secondly, the integration of global labels (*i.e.* speaking styles) was evaluated. Three methods were investigated: (i) a speaking style-dependent training and the adaptation of (ii) the neutral model or (iii) the average-voice model to each speaking style. It was shown that adaptation techniques, and the adaptation from an average-voice model in particular, outperform style-dependent models both in terms of intelligibility and segmental quality. However, the comparison with the baseline, *i.e.* the models disregarding global labels, showed that, unlike what was expected, the integration of global labels did not enhance expressivity and slightly degraded the segmental quality.

A last experiment allowed evaluating the effects achieved by the combination of both prosody annotation layers on the generated speech. Interestingly, the complete integration of the two-layer annotation, compared to the model integrating local labels only, led to an even better rendering of expressivity, while achieving similar intelligibility scores. However, it slightly degrades the segmental quality. Future work should thus focus on the improvement of speaking style adaptation techniques in order to increase the segmental quality of the generated speech. An additional comment is that *ExRise* is a very specific label which is more dynamic than the other labels. It might therefore benefit from a specific processing. It would, for example, be interesting to determine whether its replacement by an interpolation between a *Neutral* and an *ExMax* model achieves a good representation of the expressivity (see Tachibana et al. (2005)).

Chapter 7

Improving the local prosodic annotation: One or several emphatic stresses ?

Contents

7.1	Introduction	169
7.2	Statistical analysis of emphatic stresses	170
7.2.1	Extraction of acoustic features	170
7.2.2	Dimensionality reduction	171
7.2.3	Clustering	172
7.2.4	Correlation between clusters and linguistic contexts	174
7.3	Speech synthesis: A perceptual study	176
7.3.1	Evaluation protocol	176
7.3.2	Results	176
7.4	Conclusion	177

Abstract

Emphatic stress is known to fulfill essential functions in expressive speech. Its integration in speech synthesis usually relies on a prosodic annotation of the training corpus. In the previous chapter, we have proposed to assign a single "F" label to all emphatic stresses. A possible alternative consists in the definition of several labels according to the acoustic realization of the respective syllables. While it would be more complex to predict those various labels for a new text to synthesize, it might allow for a better rendering of the stress in the synthesized speech. This chapter examines whether the use of more than one emphatic label improves the perceived expressivity of the synthesized speech. It relies on our manually-annotated expressive corpus of sports commentaries. Statistical acoustic analyses show that four distinct realizations of emphatic stresses can be distinguished. However, subjective evaluation of the resulting synthesis indicate that the integration of this distinction in HMM-based speech synthesis does not lead to significant improvement of the expressivity. This implies that the different acoustic realizations of the stress may not be required to be explicitly annotated in the training corpus.

This chapter is based upon the following publication:

- Sandrine Brognaux, Thomas Drugman & Marco Saerens, Synthesizing sports commentaries: One or several emphatic stresses ?, *Speech Prosody*, Dublin (Ireland), 2014, 270-274.

Many thanks to Dr. B. Picart with whom we collaborated for the training of the synthetic voices.

7.1 Introduction

Albeit rather infrequent in neutral speech, emphasis plays a crucial role in the prosody of expressive speech (Fernandez and Ramabhadran, 2007; Strom et al., 2006). This phenomenon relates to highly prominent syllables which are sometimes seen as (particularly prominent) ‘pitch accents’ (Hirschberg, 1990; Badino et al., 2009). Emphatic stresses are known to fulfill various functions like contrasting or highlighting elements and contribute to the liveliness of the message (Fernandez and Ramabhadran, 2007). Their generation in expressive speech is therefore essential. It is even more the case when synthesizing sports commentaries which have been shown to display a high rate of emphatic stresses, falling on specific positions like numbers in scores, as indicated in the previous chapter.

As described in Section 5.4.2, several attempts to integrate emphasis have been proposed, both in unit-selection (Raux and Black, 2003; Fernandez and Ramabhadran, 2007; Strom et al., 2006, 2007) and HMM-based speech synthesis (Badino et al., 2009; Yu et al., 2010). They usually rely on a prosodic annotation of the corpus in terms of emphatic stresses. The acoustic characteristics of the corresponding syllables are then learned to be reproduced at synthesis stage. While some annotations present various labels associated with different acoustic realizations of the emphatic stress (as in ToBI), most studies only use a single label for emphasis (Yu et al., 2010; Hovy et al., 2013). As previously mentioned, ToBI achieves high inter-annotator rates in terms of presence vs. absence of phrase and pitch accents but much poorer results regarding the choice of the right pitch accent (Wightman, 2002; Taylor, 2009). For that reason, some studies have annotated their corpus with a simplified version of ToBI, dropping many of the pitch-accent distinctions (*e.g.* Wightman and Ostendorf (1994); Wightman and Rose (1999)).

The obvious advantage of using a single label for emphatic stresses is that it makes it easier to predict it for a new sentence to synthesize. Most studies investigating the automatic prediction of emphasis from text have, for that matter, considered only one emphatic label (Strom et al., 2006; Hirschberg, 1990; Hovy et al., 2013). Conversely, predicting several emphatic labels from text requires a correlation between the labels and specific distinct functions, which is rarely the case.

However, emphatic stress is often regarded as a gathering of different kinds of stresses with various functions, positions, and, importantly, different acoustic realizations (as proposed in ToBI). Badino et al. (2009), for example, distinguish between contrastive focus and other emphatic stresses by defining it as a particularly strong pitch accent. While some studies have mentioned the potential existence of different *levels* of emphasis (Fernandez and Ramabhadran, 2007; Goldman et al., 2012), we rather believe that different *kinds* of emphasis may co-exist, with no specific order relation between them. If one label is associated with each type of acoustic realization, it facilitates the training of more acoustically-consistent models, more inclined to generate suitable emphatic stresses in speech synthesis.

The question that remains is whether the use of a single emphatic label in HMM-based speech synthesis still allows for an appropriate rendering of the various acoustic realizations. In other words, it should be assessed whether the acoustically-different emphatic stresses are learned by the models, based on the linguistic context, or if the distinction requires being made explicit by annotating with distinct emphatic labels.

This chapter investigates the benefit of annotating emphatic stresses for expressive HMM-based speech synthesis with several labels instead of one. For that matter, it relies on the *Sportic* corpus presented in Section 5.6.3. The corpus is spontaneous while containing a natural variety of prosody, unlike studies relying on artificially-produced emphatic stresses by actors (see Strom et al. (2006)). Besides, we have shown in the previous chapter that sports commentaries are characterized by a high density of emphatic stresses with strong acoustic correlates, which makes them much more suitable for the study of emphasis than rather neutral read speech as used in Yu et al. (2010). Emphatic stresses were manually annotated in the corpus and were statistically analyzed in order to define several sets of labels, corresponding to stresses with distinct acoustic realizations. The manual annotation having been realized on a functional basis, our study partly answers the criticisms formulated in Hirst (2005), *i.e.* the fact that prosodic function and form tend to be merged in prosody annotation.

The objective is here to distinguish between various forms of a single emphatic function and to assess the resulting improvement reached in the expressivity of the synthesized speech.

Our study first sets out to statistically analyze the acoustic realization of the emphatic stresses in the corpus in Section 7.2. The integration of different sets of emphatic labels in HMM-based synthesis is then investigated in Section 7.3 through a subjective evaluation. Finally, Section 7.4 concludes the chapter.

7.2 Statistical analysis of emphatic stresses

The statistical acoustic analysis of the emphatic stresses in the corpus consists of four steps. First, a set of acoustic features is extracted for each emphasized syllable (Subsection 7.2.1). Dimensionality reduction techniques are then used to reduce the set of features and delete potential redundancy (Subsection 7.2.2). The reduced feature set is then used to cluster the emphasized syllables, as an attempt to find the more suitable number of distinct emphatic stresses (Subsection 7.2.3). These new sets of stresses are then investigated for potential correlations with specific linguistic contexts (Subsection 7.2.4). For further information about the exploited statistical methods, see Johnson and Wichern (2002) and Izenman (2008).

7.2.1 Extraction of acoustic features

For each emphasized syllable, 65 acoustic values are extracted. The features consist of prosodic measurements: F0 extracted with SRH (Drugman and Alwan, 2011) (*e.g.* mean, max, standard deviation), energy (*e.g.* mean, max, standard deviation) and duration (both of syllable and nucleus). A prominence value is added by PromGrad

(Goldman et al., 2012) which assigns a prominence score from 0 to 4 to each syllable, on an acoustic basis. Two additional features indicate the presence of a preceding or following silence and its duration. Finally, contextual information, *i.e.* comparisons with the acoustic values of the two previous and the next syllable, are also computed as they were shown to be efficient for prominence detection in French (Goldman et al., 2007, 2012; Mertens and Simon, 2013). These latter measurements are only extracted if both syllables are not separated by a silence, as it is known that silences tend to be associated with a resetting of the prosodic parameters, especially regarding F0 (Vaissière, 1983).

It should be noted that duration values are normalized with respect to the average and standard deviation of the duration of the corresponding phonemes in the corpus. This choice relies on the fact that the nature of the phoneme clearly affects its duration (Rositzke, 1939; Di Cristo, 1985). Missing values (for contextual information) are replaced by the average value of the feature. Finally, all variables are normalized into standard scores.

7.2.2 Dimensionality reduction

The second stage of our analysis aims at reducing the number of features. For that matter, a principal component analysis (PCA) is carried out on the data. The scree plot (see Figure 7.1) shows the contribution of the components to the global variance. Since there is no universal technique for selecting the natural number of dimensions, we relied on two popular rules of thumb: (i) keeping the dimensionality accounting for 70 % of total variance and (ii) removing the dimensions for which the contribution to the global variance remains "stationary". Based on these considerations, we chose to keep ten and four dimensions.

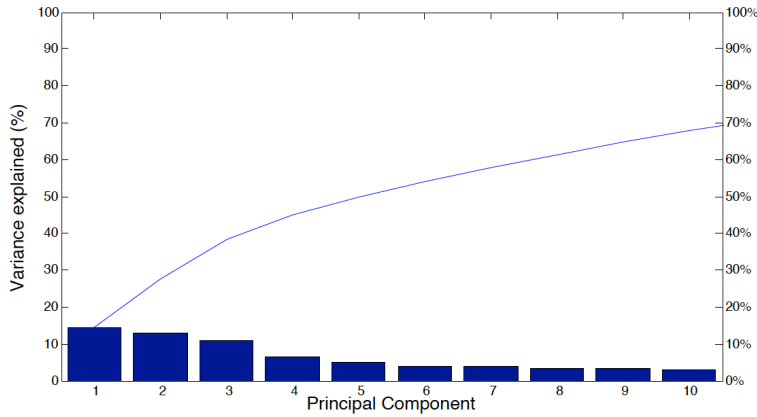


Figure 7.1 - Scree plot displaying the proportions of global variance (in %) carried out by the ten first components of the PCA.

It is interesting to examine the weights assigned to the different variables for each

IMPROVING THE LOCAL PROSODIC ANNOTATION: ONE OR SEVERAL EMPHATIC STRESSES ?

component. Interestingly, the first component is clearly related to energy, all 20 higher weights being assigned to energy values. The second component is linked to F0 with the 18 highest weights corresponding to F0-based features. Finally, duration-based measurements have most of the heaviest weights in the third PCA component. The fourth dimension is a mix of different types of variables (F0 and energy in particular). The three higher weights for the first three components are assigned respectively to mean energy, mean F0 and syllable duration.

7.2.3 Clustering

A clustering is now carried out on the reduced data obtained from the PCA. The main objective is to define different sets of emphatic stresses characterized by distinct acoustic values.

We first apply a Ward hierarchical clustering (Ward, 1963). The advantage of this clustering technique is that it visually shows the gathering of the various clusters, which helps in determining the "natural" number of clusters. Figure 7.2 shows the dendrogram obtained on the first four PCA components of our data. The algorithm assigns a unique color to each group of nodes where the linkage is less than a specific fixed threshold. This dendrogram clearly shows 4 distinct groups of syllables. Interestingly, when applied to the first 10 PCA components, the dendrogram also points at four distinct clusters.

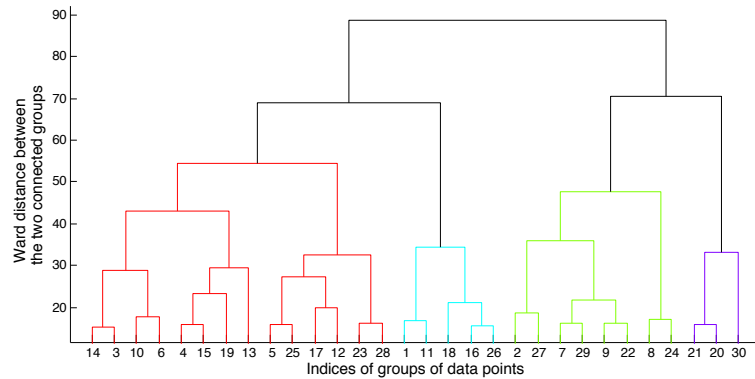


Figure 7.2 - Dendrogram computed from Ward hierarchical clustering applied to the first four PCA components of our data (the leaves in the plot correspond to more than one data point).

A second advantage of first applying a dendrogram algorithm is that the centroids of the generated clusters can then be exploited for the initialization of a K-means clustering, for further optimizing the within-cluster variance (Lebart et al., 1995). This is done in this second stage of our analysis. To assess the quality of the clustering obtained when using various numbers of clusters, we also compute K-means

clusterings with 2 to 10 clusters. For that matter, initialization points are selected randomly and the algorithm is run 50 times, the best clustering being kept for analysis.

The silhouette value provides an evaluation of the clustering quality, the higher the value, the better the distinction between the various clusters (Rousseeuw, 1987). Figure 7.3 interestingly shows that, even when launching a K-means with random initialization on the 4 PCA components of our data, the algorithm achieves the best clustering quality with four clusters, which confirms what was shown by the dendrogram. It should be noted that a similar curve, with a peak for 4 clusters, is also observed when launching the K-means on 10 PCA components instead of four. Another interesting finding is that the silhouette value reached with 4 clusters and random initialization is identical to that obtained when initializing the K-means on the centroids from the dendrogram, which indicates a similarity between both clusterings. It should be highlighted, however, that the obtained silhouette values are rather low, indicating a rather uncertain distinction between the clusters.

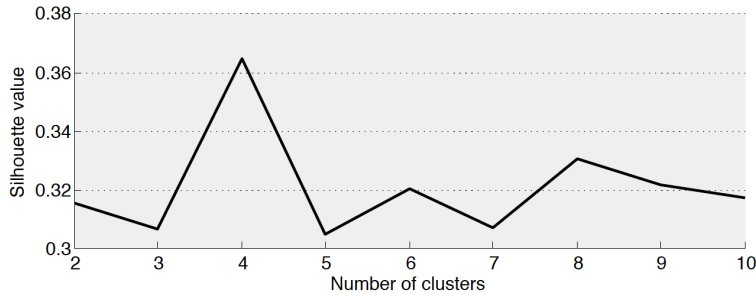


Figure 7.3 - Value of the silhouette of the K-means on four PCA components according to the number of clusters, with random initialization.

The rand-index (Rand, 1971) allows us to compare two clusterings. Its value ranges between 0 and 1, 1 corresponding to two identical clusterings. Paired comparisons were performed for four K-means clusterings in 4 clusters:

- with initializations on centroids from dendrogram on 4 PCA components
- with initializations on centroids from dendrogram on 10 PCA components
- with random initialization on 4 PCA components
- with random initialization on 10 PCA components

All rand-index values reach a level above 0.93 which indicates a certain stability, all clusterings converging towards similar solutions.

The first of the four clusterings is used in the remainder of this study (see Figure 7.4). A prosodic analysis of the syllables contained in the four clusters indicates that each cluster can be associated to a specific realization in terms of the three

IMPROVING THE LOCAL PROSODIC ANNOTATION: ONE OR SEVERAL EMPHATIC STRESSES ?

main prosodic features (*i.e.* energy, F0 and duration), as shown in Table 7.1. To investigate whether the increase in the number of clusters is in line with an increase in the naturalness of the expressivity, the version with 10 clusters, as obtained with random initialization on the 4 PCA components, is also assessed in the perceptual evaluation. This clustering provides a vectorial quantification of the acoustic space, each region being assigned to a different cluster. It should be noted, however, that it is obviously more complex to predict such a high number of tags from a text to synthesize.

Table 7.1 - *Acoustic characteristics of the clusters, compared to average acoustic values of emphatic stresses.*

Cluster	Energy	F0	Duration
Cluster 1	-	+	-
Cluster 2	-	-	-
Cluster 3	+	+	-
Cluster 4	+	+	+

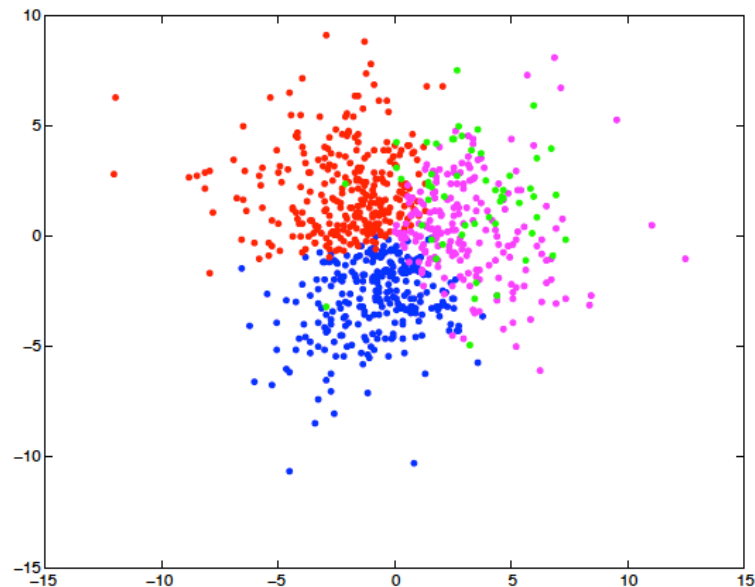


Figure 7.4 - *K-Means clustering on our data projected on the two first components of the PCA and initialized by the centroids of the dendrogram.*

7.2.4 Correlation between clusters and linguistic contexts

Potential correlations between the four defined clusters and linguistic information (such as syllable position or structure) are investigated. Provided that HMM-based

speech synthesizers rely on such contextual criteria to cluster the models (as detailed in Section 5.4), correlations would indicate a possible automatic distinction between the various acoustic realizations. This would imply that the stresses belonging to different clusters would not need to be explicitly distinguished in the annotation.

To investigate that question, we analyzed 13 linguistic variables used as contextual information for synthesis: position of the syllable or word in the word or rhythmic group (RG)¹, amount of syllables in word and RG, amount of (content) words in RG, structure of the syllable, nature of the nucleus and part of speech of the word. Seeing the high amount of samples, Chi-square tests tend to be significant for most variables. Cramer’s V (Cramér, 1946) allows interpreting chi-squares for high effectives. Table 7.2 shows that only weak associations (*i.e.* $V < 0.2$ (Rea and Parker, 1992)) can be seen between the acoustic clusters and contextual linguistic information.

Table 7.2 - *Correlation, computed with Cramer’s V , between the four clusters and linguistic contextual information (first five variables).*

Variable	Cramer’s V
Syllable position in word	0.1379
Nature of the vowel	0.1165
Word position in rhythmic group (forward)	0.1164
Syllable position in word (forward)	0.1141
Syllable position in word (backward)	0.1137

Interestingly, the highest value (*i.e.* 0.14) is assigned to ‘syllable position in word’, mainly informative about whether a syllable is initial or final. The omnipresence of both syllable and word position in the ranking drove us to investigate whether some acoustic differences may be due to final syllables at the end of the RG. In that case, they might coincide with what is commonly referred to as boundary tone (Silverman et al., 1992), which could influence their realization. Table 7.3 shows that the acoustic values of those syllables are significantly higher compared to the other emphatic stresses (for the three acoustic features, $p < 0.001$ for a bilateral ranksum test performed on 82 final emphatic syllables and 721 other emphatic syllables). This indicates that the distinction between the clusters can partly be explained by linguistic contextual information. Associations are however rather weak, suggesting that other factors probably play a role in the acoustic realization of emphatic stresses².

¹Positions are computed both from the beginning and from the end of the corresponding unit.

²As previously mentioned, glottal stops were not annotated in the manual phonetization of the corpus. They could however play a role in distinguishing between different acoustic realizations of emphasized syllables. Further analyses in that direction should be carried out to potentially confirm that hypothesis.

Table 7.3 - *Acoustic realization of final emphatic stresses (end of word at the end of RG) and other emphatic stresses, together with their 95 % confidence intervals.*

Emphatic stress	Mean F0 (Hz)	Syllable Dur (z-score)	Mean Energy (dB)
Final	262.9 ± 7.4	1.9 ± 0.6	49.4 ± 1.8
Other	242.9 ± 3	0.84 ± 0.1	44.2 ± 0.4

7.3 Speech synthesis: A perceptual study

7.3.1 Evaluation protocol

In order to assess the quality of the expressivity produced when integrating various types of emphatic stresses, several HMM-based speech synthesizers (Zen et al., 2009) were built, relying on the implementation of the HTS toolkit (version 2.1) (Zen et al., 2007a). For each synthesizer, 90 % of the corresponding database was used for the training (called the *training set*), leaving around 10 % for the synthesis (called the *synthesis set*). Model configuration, filter parametrization and excitation modeling were identical to those exploited in the evaluation protocol of Chapter 6. Emphatic annotation was used as contextual information, in the same way as linguistic information.

Three models are compared: the baseline model (*Baseline*), using only one emphatic stress, and the models with 4 (*4 Stresses*) and 10 (*10 Stresses*) emphatic stresses, as obtained by annotating the emphatic syllables with the 4 and 10 clusters defined in the previous section. Test sentences were automatically selected from the synthesis set, as being shorter than 5 seconds and displaying at least two emphatic stresses in their annotation. It is indeed much easier to compare short sentences in which more than one difference appears. The test consisted in 18 pairs of sentences, 6 from each comparison, randomly selected from 63 pairs (21 for each comparison).

30 native French-speaking testers, mainly naive listeners, participated in the evaluation. During the test, they could listen to the pair of sentences as many times as wanted. For each comparison, they were first asked whether they heard any difference between both versions of the sentence. If so, they were asked to compare them in terms of naturalness of the expressivity³. The scale ranged from -3 (much less natural) to +3 (much more natural). A score of 0 was given if both versions were found to be different but with equivalent naturalness of the expressivity.

7.3.2 Results

A first interesting finding is that the testers did not hear any difference between both versions for around 20 % of the pairs. This percentage is even higher (*i.e.* 28 %)

³Compared to the previous chapter, MOS tests are replaced by preference tests, to better evaluate whether a model tends to be preferred to the other. The overall quality of the synthesis may however be still interpreted as MOS evaluations of the baseline were provided in Chapter 6.

for models with 4 and 10 clusters which tend to display rather similar intonational patterns. Figure 7.5 shows the preference percentages for the remaining pairs. Middle sections correspond to pairs which were considered as similarly natural in terms of expressivity, but with a perceived difference between both syntheses. We can observe that the model with 4 emphatic stresses slightly outperforms both other models. This might be explained by the fact that it more accurately synthesizes the various acoustic realizations of the stress. In the 10-cluster model, we notice a degradation which may be due to the reduced number of occurrences for each stress, which is partly alleviated with the 4-cluster model. However, the preference for 4 clusters rather than one single stress (*i.e.* the baseline) is quite weak and is not statistically significant ($p = 0.21$ with a unilateral signed rank test comparing the average percentage of preferences on the 30 testers).

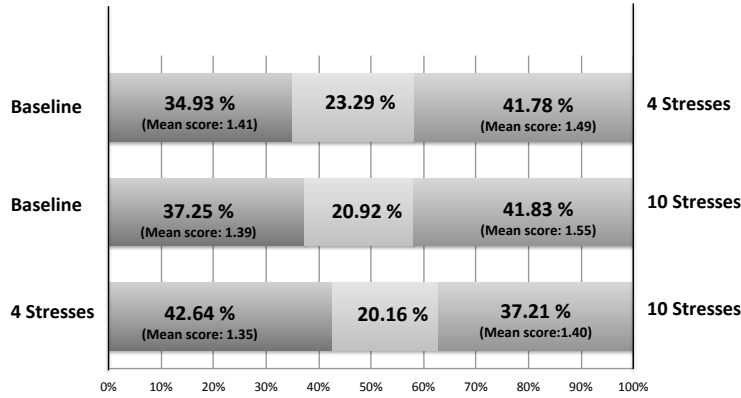


Figure 7.5 - Percentage of preferences for each model (Baseline with 1 emphatic label, and proposed models with 4 and 10 emphatic labels) in the three comparison pairs.

Figure 7.5 also shows, in parenthesis, the mean score obtained by the three models when preferred in the comparison. These scores are barely higher than 1 because testers mostly assigned a score of ‘1’, reflecting only a ‘slight’ preference.

7.4 Conclusion

This chapter focused on the modeling of emphatic stresses. Emphatic stresses can be assigned a single label or be divided into distinct labels according to their acoustic realization. While the prediction of a single label from text is easier, the use of different tags might allow for the generation of more suitable stresses. The question was then whether the use of several emphatic labels effectively improved the naturalness of the expressivity.

We addressed this latter question by investigating HMM-based speech synthesis

IMPROVING THE LOCAL PROSODIC ANNOTATION: ONE OR SEVERAL EMPHATIC STRESSES ?

using one or several emphatic labels. Statistical acoustic analyses allowed determining 4 and 10 distinct emphatic labels based on a set of extracted acoustic features at the syllable level. The definition of 4 clusters was shown to achieve the best clustering quality. The model with 10 labels was computed to propose a more accurate vectorial quantification of the acoustic space. Both models were compared to a baseline model using a single emphatic label.

Perceptual tests showed that the model with four emphatic stresses was slightly preferred over both other models. However, the differences are not significant. While participants did not perceive any difference in 20 % of cases, more than 20 % of the remaining pairs were scored as ‘similar’ regarding the naturalness of expressivity. For pairs for which a preference was given, the score was usually low, denoting a weak degree of preference. These results tend to indicate that it might not be required to explicitly annotate different kinds of emphatic stresses in the training corpus, when using HMM-based speech synthesis. This is a clear advantage for the annotation of the text to synthesize and justifies the use a single "F" label in our local prosodic annotation. Further studies should investigate whether this finding is generalizable to other languages or speaking styles.

Chapter 8

Automatic prediction of the expressive prosody annotation

Contents

8.1	Introduction	181
8.2	Automatic prediction of local labels from text and sound	183
8.3	Automatic prediction of global labels from text and sound	186
8.3.1	Which minimal unit should be chosen ?	186
8.3.2	Extraction of acoustic features	189
8.3.3	Automatic prediction of macro-prosodic labels: Method .	190
8.3.4	Integrating the automatic annotation in speech synthesis: A subjective evaluation	193
8.4	Automatic prediction of prosodic labels from text only: Perspectives	196
8.5	Conclusion	197

Abstract

In the previous chapters, we have proposed a prosody annotation protocol for live sports commentaries. Experiments have indicated that its integration in a HMM-based speech synthesis system yielded improvement in the expressivity of the synthesized voice. However, manually annotating large amounts of data as regards this protocol, as required for the training corpus, is costly and therefore prohibitive. The labels should also be predicted for any new sentence to synthesize. This chapter investigates the possibility to automatically predict the prosodic labels from textual and/or acoustics characteristics. Both local and global levels of prosodic annotation have been developed from a functional perspective, which facilitates their prediction from text, especially in the framework of a concept-to-speech application. In terms of prediction from acoustics (and text), *i.e.* for the annotation of the training corpus, we first show that a SVM approach exploiting acoustic features of the syllables yields reasonable accuracy for the local prosody annotation. In a second stage, the chapter mainly focuses on the prediction of global prosodic labels from acoustics with two objectives in mind: (i) a possible extension of the training corpus to achieve higher coverage of the various sub-genres and (ii) a potential exploitation of the automatic labeling to correct outliers in the manual annotation. Our study focuses on the three excitation levels: Neutral, Excited and ExMax. We propose two models based on Gaussian mixtures and logistic regression to automatically annotate the corpus and correct manual annotation errors. We point out, through subjective tests, that the synthesis trained on the corrected annotation allows for a better discrimination between the three arousal levels, the first approach achieving the highest results. This indicates that automatic labeling predictors may be exploited to refine the manual annotation and thereby contribute to the modeling of expressivity by the synthesizer.

We would like to thank Prof. M. Saerens for his insightful advice and Dr. T. François for his collaboration in the statistical processing of the data.

8.1 Introduction

In Chapter 6, we proposed a two-tier annotation protocol for sports commentaries. The integration of that annotation into a HMM-based speech synthesis system further indicated that such annotation yielded good quality results and that it significantly improved the expressivity of the synthesized message. The annotation needs to be provided to the synthesizer both for the training and for the synthesis. The need for an automatic prosody annotator is twofold:

- Manually annotating prosody, from text and audio or text only, is a time-consuming task which can be tedious and costly (Campione and Véronis, 2001). This is stressed by Syrdal et al. (2001) who indicated that manually labeling speech with ToBI labels, even when performed by experienced labelers, typically takes from 100 to 200 more time than the real time of the signal.
- Another problematic aspect of manual annotation lies in its rather low level of consistency, especially if several annotators collaborate on the same corpus (see Pickering et al. (1996)). Interestingly, Goldman et al. (2010) pointed out that human annotation is highly influenced by linguistic information. They show, for example, that annotators tend to over-detect final prominence on lexical words, and that the opposite tendency occurs for grammatical words. This indicates that human annotation relies on other factors than just acoustics ones and is dictated by an "auditory illusion" based on linguistic factors (Goldman et al., 2010). While such an annotation may be of interest for studies on prominence perception, such discrepancies between acoustics and annotation should clearly be avoided in the annotation of speech synthesis corpora, as it may bring noise in the acoustic realization of the various labels. In that sense, an automatic annotator based on acoustics features alone should provide a more consistent annotation.

As all other aspects of modern data-driven speech synthesizers can be derived automatically (*e.g.* by the front-end component), the interest for automatic prosodic labeling of speech corpora is high and has given rise to a wealth of research. Studies include Wang and Hirschberg (1992); Wightman and Ostendorf (1994); Syrdal et al. (2001); Chen et al. (2004a); Braunschweiler (2006); Sridhar et al. (2008); Ananthakrishnan and Narayanan (2008); Wagner (2009); Rosenberg (2010) and Jeon and Liu (2012) for (reduced versions of) ToBI, Campione et al. (2000) for INTSINT and Bartkova et al. (2012); Mertens (2013a, 2014) for pitch levels and movements in French.

As pointed out in Chapter 6, the annotation protocol we proposed is based on a functional approach. Each label is supposed to correspond to a very specific function in the communication process. In that sense, prediction from text at *synthesis* should be rather straightforward, provided that enough semantic information is available. This should not be problematic in the framework of our project, as the speech synthesizer is supposed to be included in a wider concept-to-speech application. In this chapter, we briefly outline a few basic steps which could be taken to produce an automatic prosodic annotation from text, and which are based on automatic

phrasing and accentuation labeling, as presented in Section 5.3.3.

The major focus of this chapter consists, however, in the prediction of the annotation framework from sound and text, *i.e.* the automatic annotation of the *training* corpus. In Chapter 6, an acoustic analysis of the labels highlighted the fact that the different prosodic tags were characterized by distinct acoustic realizations. While this allows for specific acoustic generation at synthesis, this should also facilitate the automatic prosodic annotation of the training corpus, based on its acoustic realization.

This meets two specific objectives:

1. As previously mentioned, manually annotating a large database, required for speech synthesis purposes, is costly. Automatic labeling could take advantage of the existing manual annotation to train a model with a view to enrich the database with more speech data. As shown in Chapter 6, some sub-genres achieve lower results when integrated in speech synthesis because of their smaller relative size in the corpus. That issue could be alleviated by integrating more annotated data.
2. We have noticed in Chapter 6 that the sub-genre or macro-prosodic annotation only improves to a small extent the resulting speech synthesis, and only in combination with local prosodic features. This was a surprising finding as adaptation to each sub-genre should enhance their synthesis. Our belief is that the manual annotation is highly subjective and strongly influenced by the semantic content of the corresponding speech segments. A successful basket may have been considered as highly excited while a section describing a player biography was classified as *Neutral*, even if the segment did not display characteristic acoustic features of the corresponding sub-genre. In that sense, an automatic annotation of the corpus, based on acoustic features, could lead to an improvement of the annotation. It could take advantage of the existing annotation to generalize global tendencies and automatically correct outliers. This approach is in line with our previous study (Brognaux et al., 2012a) in which we proposed a post-processing technique to reduce the amount of prosody labeling errors. It was shown that a predictor trained on (partially erroneous) manual annotation could be exploited to check existing labels and suggest modifications. Evaluations indicated that the proposed technique significantly improved the annotation quality.

With that second objective in mind, this chapter especially focuses on the automatic prediction of the global labels from text and acoustics.

The chapter is organized as follows. We first investigate the possibility to automatically predict the prosodic annotation at *training* time, *i.e.* making use of both textual and acoustic characteristics. In Section 8.2, local labels are predicted by training support vector machines (SVM). The automatic annotation is evaluated in comparison with the manual annotation, which was shown to achieve high inter-annotator rates in Chapter 6. Then, the automatic prediction of global prosodic

labels is investigated in Section 8.3. The main focus is put on the prediction of three levels of excitation. The automatic annotation is then exploited as a correction of the manual labeling and integrated in speech synthesis. It is evaluated through subjective tests of the resulting speech synthesis. Finally, Section 8.4 briefly outlines possible avenues to automatically predict the labels from text only, at synthesis stage. Section 8.5 concludes the chapter and discusses further perspectives.

8.2 Automatic prediction of local labels from text and sound

We have observed, in Chapter 6, that the local prosodic labels are characterized by rather specific acoustic features. Besides, the definition of these tags also implies specific positions. For example, in French, boundary stresses (*i.e.* H , HH , L , LL and E) usually fall on the last syllable of the word while emphatic stresses (F) tend to be assigned to initial syllables (Lacheret-Dujour and Beaugendre, 1999). We propose in this section a preliminary modeling of the local prosody annotation to indicate whether its prediction from text and acoustic achieves reasonable annotation quality.

In order to automatically predict those tags, two types of approaches could be investigated:

- Fully-automatic approaches, *e.g.* decision trees, neural networks, support vector machines or conditional random fields (CRF). The models should be trained with insightful acoustic and linguistic features for each syllable. The set of parameters could be enriched by more complex features like phrasing information (*e.g.* based on Chunks & Chunks Liberman and Church (1992) or improved Chunks & Chunks Beaufort (2008)).
- Existing prominence detection algorithms like Prosoprom (Goldman et al., 2007), Promgrad (Goldman et al., 2012) or Promise (Christodoulides and Avanzi, 2014). These algorithms were shown, in Section 6.3, to help distinguishing between our local tags. The prominence value provided by Promgrad could then be converted into a specific tag according to other linguistic constraints like the position of the syllable in the word. These conversions could be heuristically defined. It should be noted that these two approaches can be combined as the predictor of Christodoulides and Avanzi (2014), for example, is based on CRF. Conversely, Prosoprom and Promgrad are rule-based methods (with parameters trained on corpora).

This section briefly investigates the first solution, *i.e.* the use of classification methods. For that matter, acoustic and linguistic information is extracted at the syllable level. Acoustic features are identical to those extracted for the analysis of emphatic stresses in Chapter 7 and include raw prosodic features (pitch, duration, pauses and energy), along with more complex prominence values. Contextual

information is also exploited and includes the two previous syllables and the next one, provided they are not separated from the current syllable by a silence. Duration values are normalized with respect to the average and mean deviation of the duration of the respective phonemes. Missing values are replaced by average values and all variables are normalized into standard scores. Most acoustic features were automatically extracted with Prosogram (Mertens, 2004a).

Additional linguistic information (at the syllable level) is extracted with eLite-HTS (presented in Chapter 5) and includes:

- structure of the syllable
- position of current syllable/word/phrase in current word/phrase/utterance (forward and backward)
- vowel of the current syllable
- part of speech of the previous/current/next word
- lemma
- primary or secondary stress as predicted by eLite
- number of phonemes in the previous/current/next syllable
- number of syllables in the previous/current/next word/phrase/utterance
- number of words in the previous/current/next phrase
- number of phrases in current utterance
- number of content words before/after current word in current phrase
- number of words from previous content word to current word
- number of words from current word to next content word

In terms of classification algorithms, we opted for SVMs as they were shown to achieve reasonable classification results for classification of ToBI pitch accents and phrase ending tones in AuToBI (Rosenberg, 2010). To avoid bias due to the higher frequency of certain labels (especially NA, *i.e.* unstressed syllables, which make up for 63 % of the syllables), the SVMs were trained on a balanced subset of the corpus, which contained about 500 occurrences of each local prosodic label (7 tags). This balanced dataset was obtained with a stratified random sampling of the whole data. Creaky and hesitation syllables were not considered in this prediction task as they should be excluded from the corpus in a preliminary stage, based on other dedicated criteria (see Martin (2012) and Kane et al. (2013) for creak detection). All acoustic and linguistic features were exploited except for the parts of speech, the structure of the syllable, its stress status and the vowel of the current syllable, which should be parametrized in further research. The experiment relied on a 5-fold cross-validation.

Table 8.1 shows the average prediction rates and their standard deviations in relation with different values for the c -parameter of the SVM model¹. The value of the c -parameter in SVM allows defining the trade-off between the flexibility of the model and the number of accepted wrongly-classified cases on the training data. Lower costs define a large margin, allowing constraints to be easily ignored. Our experiments show that a cost of 1 provides the best results, with a correct classification reaching more than 70 %. This is largely above chance, as 7 labels are predicted on a balanced test corpus (expected random score is therefore 14 %). It should however be noted that this percentage is only indicative and does not reflect the real performance of the classifier. 70 % may indeed be considered as the most optimistic results that we could achieve with an "optimal" value for the c -parameter. Ideally, an in-depth study should rely on a nested cross-validation, considering the fact that the optimal c -parameter will not be known for a new sub-corpus to annotate. It should however be noted that the performances are seen to be rather stable across the various values of the parameter and that performance above 65 % may be expected.

This result is rather similar to the results obtained by AuToBI for the classification of pitch accents (71.6 % for only 6 labels) and exceeds results for AuToBI classification of phrase endings (54.95 % for the 5 phrase final tones). Further studies should investigate which types of features are the best predictor for local labels. For that matter, features could be grouped in broader categories, such as duration, energy, pitch, pause, prominence and linguistic. To include additional categorical features, part of speech values could be transformed into binary values, based on their lexical or grammatical nature. This could also be done for the vowel of the current syllable, which could be reduced to a schwa/non schwa opposition.

Table 8.1 - *Correct prediction rate in % (average and standard deviation) using different values of c -parameter with support vector machines.*

Cost	Average correct prediction rate (%)	Standard deviation of the correct prediction rate (%)
100	58.22	2.07
10	63.78	0.98
5	67.96	1.44
1	70.39	1.95
0.5	70.34	1.99
0.1	69.93	1.73
0.01	69.72	1.81
0.001	68.26	2.30

Further research could also investigate sub-genre specific local prosody prediction, as the position and density of stresses may change according to the sub-genre (*e.g.* more F stresses may be found in ExMax than in Neutral speech). Other strategies for automatic local prosody annotation of the training corpus are further explored in

¹The SVM uses a linear kernel, and the parameters were L2-normalized. To compute multi-class classification, we used a one-against-all strategy.

Chapter 10 and are applied to other types of corpora.

8.3 Automatic prediction of global labels from text and sound

This section investigates how acoustic features may be exploited to automatically predict macro-prosodic tags, *i.e.* the set of sub-genre labels defined in Chapter 6. We aim here at determining whether the homogeneity of human annotations can be improved through an automatic reclassification system.

As previously mentioned, this study focuses on the prediction of arousal/activation, *i.e.* the level of excitation. Our automatic annotator is therefore dedicated to the annotation of the Neutral, Excited and ExMax labels. This choice is motivated by the fact that those three labels were shown in Chapter 6 to display specific acoustic characteristics. The discriminative power of acoustics on NegTension was pointed out as being lower. Finally, ExRise is also a very specific sub-genre, due to its dynamic nature, which should be treated separately. Research on emotion recognition has indicated that acoustic features such as pitch, timing, energy and voice quality of the speech signal are highly correlated with the activation level, which allows for high accuracies of automatic classification between high-activation and low-activation emotions (Tato et al., 2002; Liscombe et al., 2003), see El Ayadi et al. (2011) for a review. Schröder et al. (2001) indicate that activation is the emotional dimension that is most correlated with specific acoustic realizations. They show, for example, that activated speech displays higher F0 mean and range, longer and faster F0 rises and falls and increased intensity. By capturing long-term dependencies between the acoustic observations derived from hierarchical functionals of prosodic, spectral and voice quality features, Wöllmer et al. (2008) highlight that it is possible to predict the arousal level with an accuracy similar to human performance.

In the case of this preliminary study, we therefore relied on arousal prediction only to determine whether the homogeneity of human annotations can be improved through an automatic reclassification system. This can be compared to Eyben et al. (2012) and Székely et al. (2011) who propose to automatically cluster various speaking styles within audiobooks. The main difference lies in the fact that their clustering is not based on a manual annotation and that the interpretation of the resulting speaking style remains rather opaque.

8.3.1 Which minimal unit should be chosen ?

Unlike local labels, global prosody annotation should not be predicted at the syllable level. A sub-genre, or macro-prosodic unit, is rather assigned to a sequence of words. The main challenge is therefore to determine where a relevant sub-genre begins and where it ends. Interestingly, the question of determining the boundaries of emotional segments has aroused very little interest, mainly because many studies on emotion recognition are based on acted speech. As pointed out in Batliner

et al. (2010), the beginning and the end are defined trivially for such data, as usually one sentence is pronounced with a prompted emotion (*e.g.* Amir et al. (2001)). Even in more realistic data, such as Schuller et al. (2003), classification is often related to pre-segmented chunks, based on the manual annotation of emotions.

The question of the unit of analysis has been investigated in Schuller et al. (2011), who propose a state of the art of emotion recognition in speech. They point out that the speaking turn is often considered in conversational speech, but highlight the fact that turns can sometimes be rather long and contain several shorter emotional episodes. They describe two strategies to cope with this problem, *i.e.* the use of ‘technical units’ which are time slices or simple proportions of longer units and the use of ‘meaningful units’ which are linguistically and semantically well-defined. This second approach relies on units like syllables, words or phrases. In Batliner et al. (2010), they propose to rely on the word level, considered as the smallest meaningful emotional unit, or *eneme*. They also investigate the use of larger units like syntactic chunks (based on high syntactic boundaries or lower syntactic boundaries followed by a long pause) and eneme chunks, obtained by clustering adjacent enemes belonging to the same class. The latter type of units relies on a manual annotation of the corpus and is hardly automatizable. However, performance with eneme chunks is shown to outperform the use of syntactic chunks. They indicate that the best compromise between automation and performance seems to rely on the use of word units.

Automatic clustering of various speaking styles in audiobooks has also pointed out the need for defining minimal units. In Eyben et al. (2012), no preliminary experiments are carried out to define minimal units and they propose to work at the sentence level, which they consider to be the longest possible chunk in that respect. It should however be noted that sentence segmentation is more straightforward in the case of audiobooks as it can rely on the punctuation of the text, while we have no punctuation information in spontaneous speech. In a similar study, Székely et al. (2011) point out that significant reading style changes may occur within a single sentence. This dictates their choice for interpausal units, which avoided abrupt changes of voice style within a speech segment.

A preliminary study was carried out to address that question, based on the manual prosodic annotation of the *Sportic* corpus.

We discuss here the possibility to choose either interpausal units (IPU) or rhythmic groups (RG). Such middle-sized chunks should provide a good compromise being short enough to ensure emotion stability, but of sufficient duration to compute global prosodic measures, such as articulation rate or accentual density, which have been shown to be correlated with the activation level (Murray and Arnott, 1993; Schröder et al., 2001). The first relies on the (manually checked) phonetization of the corpus in which silences are indicated. The segmentation in RG is provided by eLite-HTS which proposes a phrasing algorithm based on an improved version of Liberman’s chunks & chunks (Liberman and Church, 1992; Beaufort, 2008). We observe the percentage of sub-genre boundaries which correspond to the boundaries

of one or the other type of groups in our corpus. We also investigate the possibility of choosing other segmentation levels, *i.e.* words, accentual phrases and sentences. Accentual phrases are here considered as sequences of words ending in a boundary tone (*i.e.* H, L, LL, HH or E). We also define a specific level of accentual phrase (which could be considered as closer to intonational phrases) which only considers groups of words ending in higher level boundaries, *i.e.* LL and HH tones. It should be remembered that sentence boundaries were manually determined. Results are shown in Figure 8.1. Bars in red (*i.e.* accentual phrases and sentences) correspond to information which can only be obtained with the help of a manual annotation. Ideally, our study should be based on automatically provided units such as RGs or IPUs.

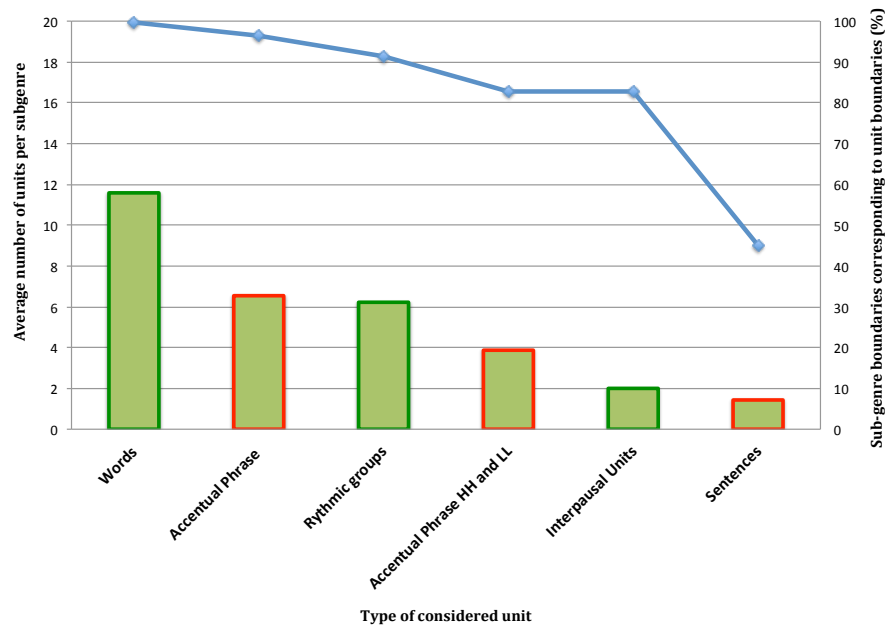


Figure 8.1 - *Correlation between linguistic segmentations and prosodic macro changes.*

This study shows that 91.45 % of the boundaries of sub-genre segments in our corpus also correspond to a RG boundary while they only correspond, in 82.81 % of the cases, to IPU boundaries. On average, a sub-genre segment contains 2 IPUs and 6.22 RGs. It should be noted here that silences are excluded from the ‘chinks and chunks’ algorithm and constitute a rhythmic group on their own, which explains this high amount of rhythmic groups (the same goes for sentences and accentual phrases). To provide further insight into the length of macro-prosodic units in our corpus,

we also extended this study to words and sentences. 99.43 % of the boundaries correspond to word boundaries. Macro-prosodic groups are, on average, made of 11.58 words and 17.86 syllables. On the whole, this study tends to imply that RG constitute the best automatic unit to predict macro prosodic changes.

It should be noted, however, that other types of units could be considered, using available tools. This segmentation of the corpus could rely on automatic tools like ADoReVA and ADoTeVA (De Looze, 2010b,a), two PRAAT plugins that automatically detect changes in speech rate or pitch. ProsoDyn (Goldman, 2012) could also be exploited as it computes global acoustic values on a sliding window of about 20 syllables. Significant changes in these values could assess a change in sub-genre and therefore delimit minimal macro-prosodic units. This could constitute a starting point for further studies.

8.3.2 Extraction of acoustic features

In a first stage, a set of global acoustic features is extracted at the rhythmic group level. In total, 3999 RG are considered.

Acoustic features include²:

- **Pitch** : mean, median, mode, min, max, inter-quartile range, variance
- **Energy** : mean, median, mode, min, max, inter-quartile range, variance
- **Duration** : mean phone duration, mean syllable duration, mean vocalic nucleus duration, mean normalized syllable duration and their variance
- **Pauses** : length of the preceding and following silence. If the RG is not preceded and/or followed by a silence, this parameter is set to 0
- **Complex values** : percentage of prominent syllables (as detected by Proso-prom (Goldman et al., 2007)), average difference of pitch between a syllable and the preceding syllable (*delta pitch*), in absolute value (as a measure of pitch dynamism)
- **Speaking rate** : number of syllables per second (silences do not need to be considered/excluded here as RG never include silences)

Duration values are normalized by the average duration of the corresponding phoneme in the corpus and its standard deviation (for mean phone and nucleus duration). The mean syllable duration is the sum of the normalized durations of its corresponding phonemes. The normalized version of the mean syllable duration is equal to the mean syllable duration divided by the number of phonemes in the syllable.

This accounts for a total of 27 features. All values are continuous. Duration values tend to be log-normal and the pause parameters have a particular distribution

²Further studies could explore the integration of voice quality features, as proposed by Székely et al. (2011) and Eyben et al. (2012).

as they are either equal to 0, when there is no silence, or to a value greater than 0 that corresponds to the silence duration. It should be noted that only the *delta pitch* parameter can have an undefined value if the RG is made of only one syllable. Our feature vector additionally includes the macro-prosodic label (which should be predicted), the name of the file and the beginning and the end of the RG. This should facilitate the relabeling of the corpus after prediction.

No linguistic feature was integrated in our classifier. The motivation behind this choice comes from the fact that the objective is to automatically distinguish between segments of speech that are *acoustically* different to better train the synthesizer. Linguistic contexts which may have (erroneously) influenced the manual annotators should therefore be discarded.

Preprocessing of the data included:

1. **Missing values:** All missing values (delta pitch parameters) were replaced by the mean of the variable.
2. **Standardization:** All values were normalized.
3. **Dimensionality reduction:** The features were reduced with a principal component analysis (PCA), in order to reduce redundancy. As shown in Figure 8.2, results indicate that more than 70 % of the variance is explained by the 6 first components. The analysis of the scree plot further highlights that the contribution of the next dimensions to the variance remained almost stationary. Based on these considerations, we chose to keep those 6 first components only.

8.3.3 Automatic prediction of macro-prosodic labels: Method

Based on the reduced set of aforementioned acoustic features, a prediction model is trained.

In that framework, two approaches are considered:

1. A Bayesian approach based on a simple Gaussian mixture
2. A logistic regression

In the **Bayesian approach**, a priori probabilities for each level of excitation e of the initial data are computed as:

$$P(e) = \frac{nb(e)}{\sum_{e \in E} nb(e)} \quad (8.1)$$

where $e \in E = \{Neutral, Excited, \text{ and } ExMax\}$ and $nb(e)$ is the number of occurrences of level e in the corpus

Then, for each feature vector v , the marginal probability is computed as:

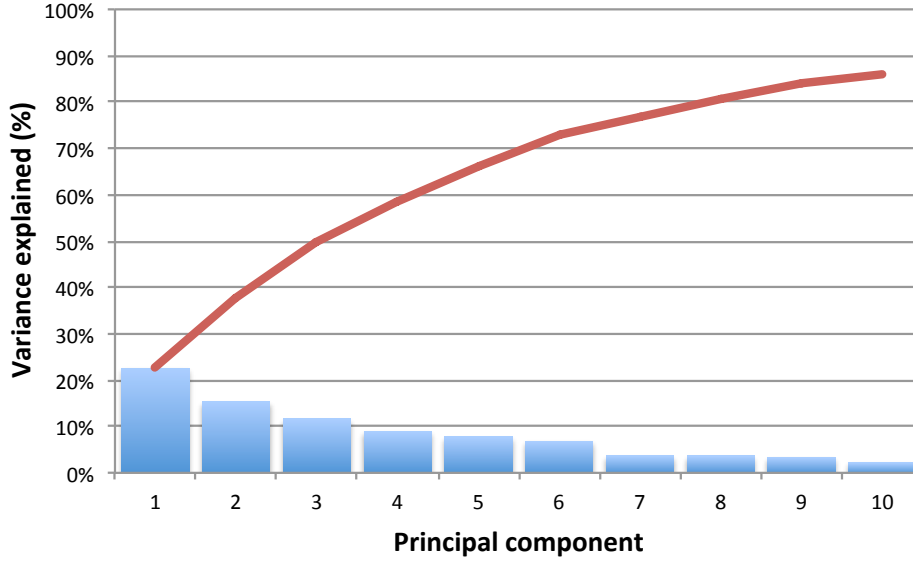


Figure 8.2 - Scree plot displaying the proportions of global variance carried out by the ten first components of the PCA.

$$P(v) = \sum_{\forall e \in E} P(v | e) \cdot P(e) \quad (8.2)$$

where $P(v | e) = f(v, \mu_e, \Sigma_e) = \frac{1}{\sqrt{|\Sigma_e|(2\pi)^d}} e^{-\frac{1}{2}(v-\mu_e)\Sigma_e^{-1}(v-\mu_e)}$

where d is the dimensionality of v (here 6), μ_e is 1-by-6 vector, containing average values of the 6 components for level of excitation e . Σ_e is the corresponding 6-by-6 covariance matrix. This allows considering the mean and variance of each class. It should however be noted that the variance-covariance matrices of all three classes are rather similar and should not greatly influence the classification task.

In the prediction stage, an excitation level is re-assigned to each rhythmic group according to its maximal a posteriori probability:

$$e_{new}(v) = \arg \max_e P(e | v) = \frac{P(v | e) \cdot P(e)}{P(v)} \quad (8.3)$$

Because we are interested in the argmax, $P(v)$ may be neglected.

A reclassification of all segments, based on the automatic prediction, indicated that 64.35 % of the initial tags were kept (see Table 8.2).

Table 8.2 - *Confusion matrix between manual annotation and automatic reclassification, with Gaussian mixture models.*

		Automatic annotation		
		Neutral	Excited	ExMax
Manual annotation	Neutral	2018	338	101
	Excited	627	315	118
	ExMax	129	112	239

The confusion matrix shows a large reclassification of Excited and ExMax labels to Neutral (and Excited), which may be due to an erroneous manual labeling of these segments based on misleading semantic information.

For the **logistic regression** approach, we trained ordinal logistic regression models (cf. Agresti (2002)) to highlight the natural ordering across the response categories. This considers the linear ranking between all three categories, which relate to levels of arousal. Based on that model, a level of excitation is predicted for each vector.

Results of the prediction are shown in Table 8.3. Interestingly, the proportion of original tags kept by the predictor is very close to that of the first method (64.50 %).

Table 8.3 - *Confusion matrix between manual annotation and automatic reclassification, with logistic regression.*

		Automatic annotation		
		Neutral	Excited	ExMax
Manual annotation	Neutral	2253	191	13
	Excited	813	229	18
	ExMax	184	200	96

Noteworthy is that both algorithms tend to favor the Neutral style and significantly reduce the number of ExMax labels. This is even more the case for the regression-based classifier (see Table 8.4).

Table 8.4 - *Number of labels for each annotation.*

	Neutral	Excited	ExMax
Manual annotation	2457	1060	480
Automatic annotation 1 : Bayes	2774	765	458
Automatic annotation 2 : Regression	3250	620	127

A Krippendorff alpha (Hayes and Krippendorff, 2007) was applied to the data. It computes an inter-rater agreement specific to ordinal data, in the sense that it penalizes less divergences between contiguous classes than a nominal approach. We

obtained an alpha of 0.4016 between the manual labels and those predicted by the model based on Bayes and an alpha of 0.3030 between the manual labels and those yielded by the logistic regression. This seems to indicate that the Bayes model makes finer changes.

As the results provided by both methods tend to go in the same direction, we also compared the classes they predict for each rhythmic group (RG). The comparison indicated that both predictors assign an identical label in 70.48 % of the cases. We also computed the percentage of large differences between both classifications, *i.e.* Neutral classified as ExMax or the opposite³. This only accounts for 0.98% of the cases.

8.3.4 Integrating the automatic annotation in speech synthesis: A subjective evaluation

Method

In order to assess the quality of the automatic re-annotation of the corpus in sub-genres, several HMM-based speech synthesizers were built, relying on the implementation of the HTS toolkit (version 2.1) (Zen et al., 2007a). For each synthesizer, 90 % of the corresponding database was used for the training (called the *training set*), leaving around 10 % for the synthesis (called the *synthesis set*). Model configuration, filter parametrization and excitation modeling were identical to those exploited in the evaluation protocol of Chapter 6. The training and synthesis label files corresponded to the manually-checked annotation of the corpus. As we observed in Chapter 6 that the integration of local labels contributed to the expressivity of the synthesized voice, they were included in our HMM-models. The building of speaking style-dependent HMMs (*i.e.* one for each excitation level) relied on adaptation (MLLR + MAP) of an average-voice model trained on the whole corpus, as this technique was shown to yield the best results in Section 6.4. Three synthesizers were built, respectively based on the manual annotation (*Baseline*), the automatic annotation with Bayes (*Auto1*) and the automatic annotation with logistic regression (*Auto2*).

The synthesis set is made of rhythmic groups (RG) or sequences of rhythmic groups which are assigned a similar label in the manual annotation and are not divided by a silence. They are referred to as "sentences" in the remainder of this section. Only sentences of a few seconds, containing more than one word or player name were selected for the test.

Two perceptive test were designed to answer to the following research questions:

1. Does one of the three models better synthesize the naturalness of expressivity ?
2. Are the different levels of excitation better discriminated by one of the three models ?

³This can be seen as a counterpart of the adjacent accuracy (Heilman et al., 2008), defined as the proportion of prediction with maximum one level of error.

The first test consisted in 18 pairwise comparisons, *i.e.* 6 for each level of excitation. Sentences (*i.e.* sequences of RG) were randomly chosen amongst the synthesis set of the database. The listeners were asked to indicate, for each pair, which version seemed more natural in the framework of sports commentaries. The aim was to compare the naturalness of the expressivity for both syntheses, and indicate which one sounded more like the way a real sports commentator would speak. The scale ranged from -3 (Synthesis A is much less natural than synthesis B) to 3 (synthesis A is much more natural than synthesis B). If both versions sounded equally natural (or if they both sounded mediocre) they could rate them as "equivalent". For this first test, sentences were synthesized with their original level of excitation, *i.e.* that of the manual annotation.

The second test also consisted in 18 pairwise comparisons, including 6 comparisons for each model, *i.e.* 2 for each combination of sub-genres: Neutral/Excited, Excited/ExMax and Neutral/ExMax. The two synthesized sentences, in each pair, were synthesized with the same synthesizer but a different level of excitation. Sentences were randomly chosen amongst the synthesis set of the database. For each pair, the listener was asked in which version the commentator sounded more excited. The scale ranged from -3 (Synthesis A sounds much less excited than synthesis B) to 3 (synthesis A sounds much more excited than synthesis B). For similar level of excitation, the "equivalent" label could be chosen.

20 native French speakers, mainly naive listeners, participated in the evaluation. To avoid intelligibility issues (namely due to player and teams names, as pointed out in Section 6.4), the transcription of the speech files was displayed under the audio players.

Results

Preferences for the first test are shown in Figure 8.3. All preferences shown are very slight. Significance measures were computed, with a unilateral signed rank sum test comparing the average percentage of preferences on the 20 testers. No preference was shown to be significant ($p = 0.10$, $p = 0.06$ and $p = 0.36$ for pairs Auto1/Base, Auto2/Base and Auto1/Auto2 respectively). This indicates that using the corrected annotation does not enhance the naturalness of the synthesized speech.

The second perceptive test offered insightful complementary results. Figure 8.4 indicates the percentage of cases in which each sub-genre synthesis (*i.e.* Neutral, Excited and ExMax) is considered as the most excited, in the pairwise comparison. In an ideal setting, Neutral should be assigned 0, Excited 50 % (*i.e.* more excited than neutral but less excited than ExMax) and ExMax 100 %. The results clearly indicate that the Auto1 model offers the best discrimination between the three levels of excitation. The baseline suffers from a lack of discrimination between neutral and excited syntheses ($p = 0.09$ with a unilateral signed rank sum test comparing the average percentage of cases in which Neutral and Excited were



Figure 8.3 - Subjective preferences for each pair of models (Baseline with manual annotation, Auto 1 with the automatically-corrected annotation based on the Gaussian mixture model and Auto 2 with the automatically-corrected annotation based on logistic regression), in terms of naturalness of the synthesized voice.

considered as the most excited, in the comparison of that pair). Conversely, this discrimination is shown to be significant both in Auto1 and Auto2 (with $p < 0.0001$ for both models). This can also be observed from the average degree of excitation assigned to Excited samples compared to Neutral samples, which achieves 0.2 for the Baseline, against 0.98 and 1.56 for Auto1 and Auto2, respectively. This indicates that excited samples, in the two latter models, are seen as slightly more excited to more excited, compared to neutral samples. While the discrimination between Excited and ExMax sentences is significant in all three models ($p < 0.01$), the average relative excitation degree between the two arousal degrees indicates a better distinction in the baseline (0.93) and Auto1 (1.03) compared to Auto2 (0.55).

These observations indicate that, while Auto1 synthesis is similar to both other models in terms of naturalness, it clearly allows for a better discrimination between the three excitation levels. This indicates that the automatic correction of the global annotation helps in synthesizing more adequately the various levels of arousal.

Further studies could investigate the integration of contextual information in the models, *i.e.* by considering the sequential aspect of the units to better predict the excitation level. As shown in Chapter 6, Neutral segments tend to be more followed by Excited speech than ExMax speech. Beside, the probability of assigning a same label to subsequent segments is higher than excitation changes between each rhythmic group. For that matter, the Bayesian approach could be integrated in a HMM framework, in which the emission probabilities would be extracted from the model.

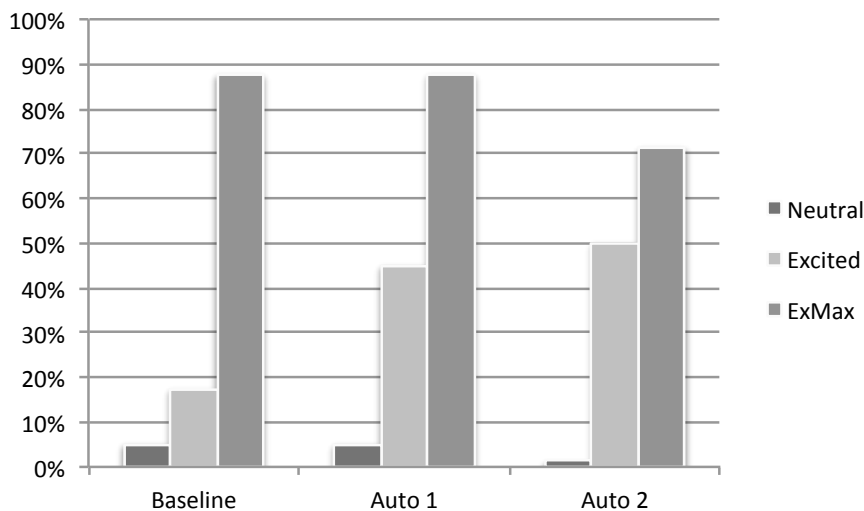


Figure 8.4 - *Percentage of cases in which each sub-genre is perceived as the most excited, for the three models (Baseline with manual annotation, Auto 1 with the automatically-corrected annotation based on the Gaussian mixture model and Auto 2 with the automatically-corrected annotation based on logistic regression).*

However, while we admit that this may allow for a better agreement with the manual annotation (which unconsciously relies on these constraints), we believe that integrating non-acoustic predictors would, on the contrary, likely degrade the quality of the resulting synthesizers.

8.4 Automatic prediction of prosodic labels from text only: Perspectives

This section offers perspectives on how to automatically predict local or global prosodic labels from text only, when synthesizing a new sentence. It takes advantage of the major characteristic of our annotation protocol, *i.e.* the fact that each label corresponds to a specific function. In the Sportic project, we can also rely on semantic information which is provided by the Natural Language Generation (NLG) component, responsible for the automatic generation of commentary sentences from specific events detected by the camcorders located around the sports ground (baskets, faults, and so on).

As previously mentioned, algorithms providing an automatic segmentation of the text in prosodic phrases, like the ‘chinks and chunks’ algorithm (Lieberman and Church, 1992), could be further exploited to predict boundary labels. The main challenge is then to automatically assign *F* labels as they highly depend on semantics. For

that matter, the text patterns used by Sportic NLG component could be manually enriched with focus information. A sentence about a score, and happening just after a basket, could, for instance, be assigned a *F* label to the initial syllable of the number corresponding to the scoring team. An example of pattern provided by the NLG is:

"Boom! Le sniper <PLAYER> est de sortie, 3 points pour <TEAM>".⁴

which could be enriched by **local** tags as follows:

"Boom!	Le sniper	<PLAYER>	est	de	sortie,	3	points	pour	<TEAM>"
F		H			HH	F			LL

The prediction of **global** tags should also benefit from the NLG component. In order to rightly associate the various sub-genres with specific events (*e.g.* scoring shot, missed shot, fault, time out), it might be useful to further annotate our corpus with the various types of events (the comprehensive list being provided by the *Tele-communication* team of the project).

Several strategies should then be investigated:

- A heuristic assignation of the most probable sub-genre to each event. This could be considered as a baseline model.
- An additional consideration of the context to account for sub-genre transition probabilities presented in Subsection 6.3.1. This could be integrated in the framework of HMM-models.
- Finally, while the two first strategies hypothesize that the sequence of words related to one event (*e.g.* scoring shot, missed shot, fault) is assigned only one tag, it could be assumed that two tags could follow each other in a same event. For example, a winning shot could begin with an *Excited* sub-genre and end with *ExMax*, potentially connected with an *ExRise*. In order to find the position of the boundaries between the sub-genres, we might consider other information sources like rhythmic groups, as proposed in the previous section.

Here again, the evaluation should be performed on the resulting speech synthesis.

8.5 Conclusion

In the previous chapters, we have presented a new prosody annotation protocol for sports commentaries. Perceptive evaluations have shown that it yields improvement of the naturalness and expressivity of the synthesized message. Such results relied on the use of a manual prosodic annotation, both of the training corpus and of the

⁴"Boom! Sniper <PLAYER> is there, 3 points for <TEAM>" (our translation)

test sentences. Such a manual annotation constituting a tedious and costly task, we presented in this chapter avenues on how global and local prosodic tags may be automatically predicted from acoustics and/or text.

At *synthesis stage*, prosodic labels should be predicted from text only. In that respect, we can take advantage of two aspects of our study: (i) prosodic labels were defined on a functional approach which facilitates their mapping with text and (ii) our study is included in a broader concept-to-speech synthesis which grants access to semantic information on the sentences to synthesize.

For the *training* corpus, prosodic labels should be predicted from text and acoustics. Preliminary results indicated that a reasonable prediction rate can be achieved for the prediction of local labels, exploiting SVMs. This chapter then mainly focused on the prediction of global tags (here, the levels of excitation only, *i.e.* Neutral, Excited and ExMax) from acoustics, with two objectives in mind. First, an automatic prediction of the prosodic sub-genre would facilitate an extension of the corpus and allow for a better coverage of the various sub-genres. Secondly, automatically predicted labels could serve for the correction of manual annotation errors, which may have been caused by an influence of the semantic content. In that framework, two predictors were developed, based on a Gaussian mixture model and logistic regression. Synthesis with the manual and the automatic annotations were then assessed through perceptive tests. The evaluation pointed at two main findings: (i) no significant difference was found across the 3 models in terms of naturalness of the synthesized message but (ii) models exploiting the automatic annotation achieved a much better discrimination between the various levels of excitation. The annotation based on the Gaussian mixture model, in particular, yielded the best results in terms of excitation rendering. This indicates that the automatic annotation allows for an automatic detection of annotation outliers and thereby improves the quality of the annotation and that of the resulting synthesis.

Chapter 9

Impact of phonetic variations on the synthesis of sports commentaries

Contents

9.1	Introduction and state of the art	201
9.2	Corpus design	204
9.3	Methodology for the phonetic analysis of the corpus . .	206
9.4	Global phonetic analysis of the corpus	208
9.4.1	Overall proportion of phonetic changes	208
9.4.2	Schwa elision	209
9.4.3	Final schwa pronunciation	209
9.4.4	Final liquid elision	211
9.4.5	Liaison	212
9.4.6	Summary	213
9.5	Phonetic variations : Focus on sports commentaries . .	213
9.6	Correlations between phonetic and rhythmic features .	214
9.7	Impact of phonetic variation on synthesized speech . . .	217
9.7.1	Experimental protocol	217
9.7.2	Evaluation and discussion	217
9.8	Conclusion and perspectives	219

Abstract

While speech synthesis research is now focusing on the generation of various speaking styles or emotions, very few studies have considered the possibility of including phonetic variations according to the communicative situation of the targeted speech (*e.g.* sports commentaries, TV news). This chapter proposes a phonetic analysis of large French corpora to assess the influence exerted by three situational ‘traits’: read/spontaneous, media/non-media and expressive/non-expressive. It shows that some variations, like elisions, tend to be more frequent in spontaneous and non-media speech, unlike liaisons which appear more often in read and media speech. Interestingly, no phonetic variation draws a clearcut distinction between expressive and non-expressive speech. In a second stage, a prosodic analysis indicates that the phonetic variations are not directly correlated with the rhythmic features of their corresponding situational ‘trait’. Finally, we also set out to analyze whether these alternative pronunciations contribute to the plausibility of the message and should therefore be considered in synthesis. To this end, subjective tests are performed on synthesized French sports commentaries. They aim at comparing HMM-based speech synthesis with genuine pronunciation and with neutral NLP-produced phonetization. Results show that the integration of the phonetic variation significantly improves the perceived naturalness of the generated speech. They also highlight the relative importance of the various types of variations and show that schwa elisions, in particular, play a crucial role in that respect.

This chapter is based upon the following publications:

- Brognaux Sandrine & Thomas Drugman, Phonetic variations: Impact of the communicative situation, *Speech Prosody*, Dublin (Ireland), 2014, 428-432.
- Brognaux Sandrine, Benjamin Picart & Thomas Drugman, Speech synthesis in various communicative situations: Impact of pronunciation variations, *Inter-speech*, Singapour, 2014. [ISCA Best Student Paper Award]
- Brognaux Sandrine & Thomas Drugman, Variations phonétiques: Impact de la situation de communication, *Nouveaux Cahiers de la Linguistique Française*, Vol. 31, pp. 223-236, 2014.

Authors are grateful to J.-P. Goldman for his insightful advice. Many thanks to Dr. B. Picart with whom we collaborated for the training of the synthetic voices.

9.1 Introduction and state of the art

As mentioned in the previous chapters, Text-To-Speech (TTS) synthesis has reached in the last few decades a fairly good level of quality and intelligibility. However, the neutral carefully read speech it produces is often criticized, as it usually differs from natural human conversations (Prahallad et al., 2006; Lee et al., 2010; Werner and Hoffmann, 2006). The need for speech synthesis to better adapt to the context of communication has been recently underlined (*e.g.* the LISTA project (Cooke et al., 2014)) and the interest has now shifted to the production of speech that suits different speaking styles and emotions. In that respect, the three previous chapters focused on prosodic and voice quality modifications of synthetic speech to produce naturally-sounding sports commentaries. Interestingly, most research on the synthesis of various speaking styles and emotions have considered modifications at these two levels only (Miyanaga et al., 2004; Yamagishi et al., 2005; Tsuzuki et al., 2004; Obin et al., 2011). Surprisingly little attention has been paid by current studies to potential phonetic modifications of the sentence to synthesize.

This concern is particularly important in French, where words are characterized by a high amount of phonetic variation. As described in Section 5.2, schwa ([ə]) elisions and liaisons are the most frequent phonetic modifications. The first consists in the optional pronunciation of a schwa vowel in the middle or at the end of a word (*e.g.* *petite* pronounced [ptit]). The second relates to latent consonant at the end of a word which can be pronounced when followed by a vowel or a mute h (*e.g.* *ils vont au cinéma* pronounced [ilvõtɔsinema]). Many linguistic studies have analyzed the modalities of appearance of these phenomena (see Fougeron and Frauenfelder (2001) and Fougeron et al. (2001) for liaisons, Hansen (1991) and Candea (2002) for final schwa pronunciation, Bürki et al. (2011) and Hambye (2005) for elisions). They show that the realization of the different variants can be explained by many factors: morpho-syntax (Boula de Mareüil et al., 2003), speech rate (Lacheret-Dujour, 1991; Bürki et al., 2011; Fougeron et al., 2001), word frequency (Adda-Decker and Lamel, 1999; Fougeron and Frauenfelder, 2001; Fougeron et al., 2001), word probability (Jurafsky et al., 2001b), degree of articulation (Picart et al., 2010), origin of the speaker (Hambye, 2005; Martinet, 1971), age of the speaker (Hambye, 2005), and so on.

As mentioned in Section 5.2, surprisingly few linguistic analyses however, have investigated the influence exerted on phonetic variations by the communicative situation (CS) (*e.g.* TV news, political speech, text reading, sports commentaries), also sometimes referred to as the "phonoggenre" (Goldman et al., 2009; Pršir et al., 2013)¹. Yet, the potential interaction between both linguistic levels is commonly ac-

¹Interestingly, a workshop on phonetics and phonology of speaking styles was organized by ESCA in 1991. Most published papers, however, focused on accentual patterns, speech rate, segmental duration, and articulation features and few studies provided an analysis of phonetic changes (*i.e.* phoneme insertion, deletion or substitution) according to the communicative situation. An exception is Hansen (1991) who analysed the pronunciation of schwas in French read speech, interviews and free conversations. In that work, she mainly focused on the distinction between read and spontaneous speech.

knowledge (Simon et al., 2009; Bürki et al., 2011) and is known to be very influential for prosody (Simon et al., 2009; Goldman et al., 2009; Roekhaut et al., 2010; Pršir et al., 2013). Only the phonetic differences between read and spontaneous speech have aroused some interest (Hansen, 1991; Adda-Decker et al., 1999; Fougeron and Frauenfelder, 2001; Lucci, 1983). Regarding the impact of speaking style, Van Bael et al. (2004) analyzed the phonetic realization in read speech, public lectures and telephone dialogues in Dutch. However, their study mainly draws conclusions on the differences between spontaneous and read speech. Jande (2008) also proposed to include style specific criteria (*e.g.* the degree of spontaneity or the degree of formality) to automatically predict phonetic variations in Swedish, based on speech from radio news, radio interviews and elicited monologues. An exception for French in Adda-Decker and Snoeren (2011) who analyzes temporal speech reductions in broadcast news, telephone and face-to-face conversations in the framework of speech recognition.

While modeling phonetic variation has attracted interest in speech recognition (Ostendorf et al., 1996; Strik and Cucchiaroni, 1999; Riley et al., 1999; Wester, 2003), it is also noticeable that few studies have addressed the integration of these variants in speech synthesis. Most speech synthesizers integrate basic phonetic variations. However, they are trained to produce a phonetic transcription corresponding to standard read speech. For optional variations, the most likely variant is generally produced, independently of the communicative situation. While research is now targeting the generation of expressive (Yamagishi et al., 2005; Tsuzuki et al., 2004) and media-related speech (*e.g.* sports commentaries in Krstulović et al. (2007)), the need for a broad study of the influence of these situational ‘traits’ (as further defined) on phonetics is striking.

Among existing studies, only the attempt to generate spontaneous-like speech has received attention (Prahallad et al., 2006; Lee et al., 2010; Werner and Hoffmann, 2006; Bennett and Black, 2005). Phonetic modeling of some spontaneous speech variation has been proposed in Bennett and Black (2005) and Prahallad et al. (2006) but its impact on speech synthesis has not been evaluated. In Lee et al. (2010), a state-based transformation of speech synthesis with Hidden Markov Models (HMM) is presented to convert neutral synthesized speech signal into spontaneous speech with no modification at the language processing level. Finally, Werner and Hoffmann (2006) integrate phonetic variation in the synthesis of German travel information, albeit with a corpus not well suited to spontaneous speech synthesis. They reported, from subjective evaluations, an improvement of the naturalness; however, no consideration was made of the relative importance of the different types of phonetic variation (be they elisions, insertions, or others). Jande (2003a,b) also investigated the integration of knowledge-based rules in diphone concatenation synthesis to integrate reduction phonetic phenomena according to the speaking rate in Swedish. Perceptive tests indicated that the reduced version is perceived as significantly more natural in fast synthesized speech. As mentioned in Section 5.2, the modeling of speaker specific variants has also aroused some interest (Miller et al., 1997; Miller, 1998b,a; Kim et al., 2004; Bennett and Black, 2005). Phonetic changes in synthesis according to regional varieties have also been proposed in

Fitt (1997) and Briony and Isard (1997). With the exception of the distinction between read and spontaneous speech, phonetic variations have rarely been applied to the synthesis of expressive speech in different CS, with the exception of Roekhaut et al. (2010) who propose to modify the pronunciation of final schwas only, in unit-selection synthesis of radio news, political addresses and conversational speech.

Our study first proposes an analysis of the influence exerted by the *communicative situation* on the phonetic realization. Because they cannot be easily ranked on a single scale, the various communicative situations are defined according to three binary ‘traits’, defined in the next section: media/non-media, expressive/non-expressive and read/spontaneous. The main objective of the study is to offer an insightful description of the phonetic features of each ‘trait’ to outline what should be considered when synthesizing a certain communicative situation. One of the questions raised by this first part of our study is whether these phonetic variations contribute to the plausibility of the message and should therefore be considered when synthesizing speech with a specific communicative purpose. That question is addressed in the second part of the chapter.

Our analysis has the advantage of relying on a very large corpus in French of about 300 minutes from 32 speakers and 10 communicative situations (*e.g.* sports commentaries, TV news, political speech). The study of the phonetic realization is based on a strategy making use of natural language processing (NLP) techniques. Rhythmic features are also considered as it has been shown to be one of the prosodic correlates of phonetic variations (Lacheret-Dujour, 1991; Fougeron et al., 2001). The potential correlation between phonetic and rhythmic features is evaluated. In a second stage, specific attention is paid to live sports commentaries which are further used for synthesis experiments. A subjective evaluation aims at comparing speech synthesized with a neutral read pronunciation and with the real pronunciation of the speaker. The relative importance of the different types of phonetic modifications is also evaluated.

The chapter is organized as follows. Section 9.2 presents the corpus and its annotation. The methodology exploited to carry out our phonetic analysis is detailed in Section 9.3. The main phonetic analysis of the corpus is described and discussed in Section 9.4. Possible correlations between phonetic variation and rhythmic features is further investigated in Section 9.6. The perceptive role of these variations in speech synthesis is evaluated through perceptive tests in Section 9.7, which presents a discussion of the results. Finally, Section 9.8 concludes the chapter and discusses further works.

9.2 Corpus design

Our corpus is an extended version of C-PROM (Avanzi et al., 2010) including additional sub-corpora (sports commentaries and corpora designed for TTS purposes)². Special attention is paid to sports commentaries (*i.e.* a sub-portion of the Sportic corpus, presented in Subsection 5.6.3, and other sports commentaries from Audrit et al. (2012)) which have also been added to the corpus. The corpus consists of 315 minutes from 32 French-speaking speakers (French, Belgian and Swiss) and ten sub-corpora covering various CS (interview, political speech, and others).

Because they cannot be easily ranked on a single scale, the various CS are described in this study according to three binary ‘traits’ which characterize the situation of communication: media/non-media, expressive/non-expressive and read/non-read. This classification of the sub-corpora according to situational traits is inspired by Goldman et al. (2009). Non-read speech is here considered as spontaneous speech. These ‘traits’ are referred to as ‘situational traits’ in the remainder of this chapter³. These are obviously simplistic reductions of the situational complexity of the sub-corpora which inevitably influences the speech features. It however offers a basic general framework to study the influence of these traits which have been extensively used in speech synthesis (especially for the distinctions read/spontaneous and expressive/non expressive). It should be noted that these traits may be indirectly related to the notion of proximity vs. distance, defined by Koch and Oesterreicher (2001), which is regarded as a crucial criterion influencing the speaker’s communicative behavior. Media, non expressive and read speech can be seen, in that framework, as characteristics of a more distant speech. Making the distinction between those three dimensions however allows for studying independently the respective role played by each dimension.

Each situation contains 2 to 7 speakers. The number of speakers per trait ranges from 13 to 17 with an average duration of about 2 hours of speech per trait. Expressive is here defined as an audible emotional/expressive implication of the speaker (*e.g.* excitement, anger, happiness), be it acted or not. Conversely, non-expressive is seen as equivalent to neutral, in accordance with the definition we provided in the introduction. Expressive speech therefore includes a wide range of types of expressivity. Emotion valence, for example, can be positive (*e.g.* happy) or negative (*e.g.* sad). This could lead to averaged effects in our analysis, and hinder the interpretation of the role played by the various aspects of expressivity. A summary of the different sub-corpora is shown in Table 9.1.

²The reader interested in more details about the corpus is referred to Avanzi et al. (2010)

³These situational traits are partly inspired from the situational characteristics which are pointed out as useful in Biber and Conrad (2009) to study situational variation. Media is both related to the *place of communication*, which is not shared, and the public *setting*. Expressive can be seen as related to the *communication purpose* which goes further than a simple information transfer by expressing emotions, emphasis or attitudes. Finally, spontaneous, or non-read relates to the *production circumstances*, indicating whether the message was prepared, planned in advance, or whether "[t]he speaker is producing language at the same time that he is thinking about what he wants to say".

Table 9.1 - *Distribution of the sub-corpora according to the three studied situational traits (TTS corpora were recorded for speech synthesis purposes). NC = not classified because ambiguous regarding that trait. For Read, "+" is read, "-" is spontaneous. For Media, "+" is media, "-" is non-media. For Expressive, "+" is expressive, "-" is non-expressive, i.e. neutral.*

Communicative situation	Read	Media	Expressive
Sports commentaries	-	+	+
Conference	NC	NC	-
Political discourse	+	+	NC
Interview	-	+	+
Itinerary explanations	-	-	-
TV news	+	+	NC
Expressive speech TTS	+	NC	+
Neutral speech TTS	+	NC	-
Neutral reading	+	-	-
Narration	-	-	+

As these traits are continuums, some corpora were not classified; if their nature regarding a dimension was ambiguous, they were not included in the corresponding subset of the corpus. In fact only corpora belonging to extremes of the scales (*e.g.* completely read or spontaneous) were classified as the three dimensions are transformed into binary criteria.

The continuum between read and spontaneous speech, as we define it, goes for instance through ‘prepared’, which could be assigned to conferences. Conferences were therefore not classified in that trait. Another example relates to text-to-speech corpora which are not broadcast as such, but could be used for public announcements. They were consequently not classified for the media trait. Conferences being public and possibly recorded depending on the type of conference, they were not classified for the media dimension either. Regarding the expressive trait, political discourse and TV news were excluded as they might contain some emphatic speech for important facts, while being rather unexpressive for other. Finally, for interviews, only the parts of the interviewee were kept.

In Section 9.5, special attention is paid to two subparts of the corpus : neutral read speech recorded for speech synthesis purposes (READ) and sports commentaries (SPORTS). READ is made of three sub-corpora from different speakers for a total duration of 124 minutes. SPORTS consists of 5 sub-corpora with a total duration of 54 minutes, each corresponding to a different speaker commenting on a basketball, football or rugby match.

For the subjective evaluation (Section 9.7), an extended version of a basketball commentary from SPORTS (*i.e.* ‘Sportic’, as described in Section 5.6.3) was used to train an HMM-based speech synthesizer. This corpus is 162 minutes long, silences

included. It has the advantage of being spontaneous and of high acoustic quality, and therefore is suited to speech synthesis.

The phonetization of all speech files was done automatically and further corrected manually. The automatic phonetization is performed by eLite-HTS which proposes a two-pass approach (The phoneme set used by eLite-HTS is shown in Appendix B). The first stage is based on an ID3 decision tree (Pagel et al., 1998) learned from the Brulex dictionary (Content et al., 1990). It should be noted that Brulex is a French phonetic dictionary which integrates phonetic modifications like elisions and (de)voicing. It provides, in that sense, a transcription closer to phonetics than phonology. Regarding phonological variants, Brulex adopted a specific strategy which consists in selecting the form which is considered as the most frequent in isolated presentation. For the schwa variation, the adopted representation always includes the schwa. In a second stage, contextual linguistic information is exploited to deal with phonetic variations at word boundaries (mainly liaisons and deletion/insertion of schwa for speech fluency). Liaisons are generated based on a set of rules considering POS, phonetic criteria and rhythmic group boundaries.

It should be noted that the manual correction was mainly based on phone elisions and insertions. This reasons behind this choice are twofold:

- The phonetic annotation of the corpus was carried out by a large number of experts and, while the annotation of phone elisions and insertions should be rather straightforward, the annotation of phone substitution (*e.g.* depending on the voicing state of the consonants) relies on a continuum of phonetic realizations and may be more subjective when several annotators are collaborating.
- The phonetic annotation of the CPROM part of our database was realized outside the framework of our thesis. The objective was first and foremost to improve the automatic alignment and, in that sense, indicating phone insertions and deletions was crucial. For that reason, most minor substitution phenomena were not annotated and manually checking the entire corpus would have been a very tedious and costly task.

Regarding phone substitutions, only large differences were annotated, in a more phonological perspective. The entire corpus was then automatically phonetically aligned with EasyAlign (Goldman, 2011) and Train&Align (presented in Chapter 4).

9.3 Methodology for the phonetic analysis of the corpus

For the phonetic analysis, we developed a specific methodology based on NLP techniques, similar in some aspects to what was proposed in Van Bael et al. (2004). For each sub-corpus, the orthographic transcription was used to produce its automatic phonetization with eLite-HTS, as described in Section 5.5. This produced a ‘standard’ phonetization of the text, corresponding to neutral read speech. This

transcription was then automatically aligned with the genuine phonetic transcription as pronounced by the speaker (which was manually checked).

This alignment relies on a slightly modified version of Levenshtein’s edit distance (Levenshtein, 1966). Several adaptations were made:

- Each phoneme is represented by only one character,
- Some phonemes substitutions are not penalized (*e.g.* $\emptyset \rightarrow \text{œ}$, $\text{e} \rightarrow \text{ɛ}$) as they might result from a subjective perception of the annotator,
- Insertions and deletions of silences are not penalized.

To retrieve the alignment, the matrix obtained by the algorithm is backtracked. Finally, all modifications are stored according to their type (insertion, deletion or substitution). Here is an example of alignment of the sentence “Parce que ça, je pense pas que c’était prévu” with schwa and liquid deletions :

p	a	R	s	k	@	s	a	Z	@	p	4	s	p	a	k	@	s	e	t	E	p	R	e	v	y
p	a	/	s	k	@	s	a	Z	/	p	4	s	p	a	k	/	s	e	t	E	p	R	e	v	y

To avoid potential phonetization errors of the NLP, all sound files containing numbers (written as ciphers) were deleted. This solves issues due to the pronunciation variants for numbers from ‘70’ to ‘90’ in French French, Belgian French and Swiss French. Proper names being very frequent in sports commentaries, their deletion would have resulted in a highly reduced sub-corpus. For that reason, we decided to consider only modifications of phonemes not belonging to proper names or to syllables just before and after a proper name.

A first analysis of the alignment highlighted recurrent errors made by the algorithm when two modifications occurred within a small phonetic context. This led to some refinements:

- The deletion and insertion of schwas being more frequent, they were favored and assigned a reduced cost.
- The cost was also reduced for substitution of [i] by [j], which is rather frequent.

This avoided alignments such as:

b	E	l	Z	_	s	/	E	t	y	n	<i>instead of</i>	b	E	l	Z	/	_	s	E	t	y	n
b	E	l	Z	/	@	_	s	t	y	n		b	E	l	Z	@	_	s	/	t	y	n

The main advantage of using NLP-produced phonetization is that it allows for an easy comparison of the pronunciation of the corpus with a so-called ‘standard’ pronunciation. The latter already considers most mandatory phonetic variations such as liaisons or elisions dictated by the linguistic context. It provides a more precise analysis than a comparison with a phonetized dictionary (as proposed in Van Bael et al. (2004)) while being fully automatic. The alignment of our corpus with NLP-produced phonetization revealed 4 main types of phonetic modifications (as described in Section 5.2) which will be further discussed in the next sections: schwa elision, final schwa pronunciation, final liquid elision and liaison.

Throughout this study, the statistical significance of the results is calculated via unilateral t-tests or Wilcoxon tests depending on the normality of the variable distribution. Correlations are evaluated using Spearman’s coefficient.

9.4 Global phonetic analysis of the corpus

In this section, we first consider the overall proportion of phonetic variations for each situational trait (9.4.1). We then focus on the analysis of four phonetic variations that were qualitatively assessed to be the most frequent in our corpus: schwa elision (9.4.2), final schwa pronunciation (9.4.3), final liquid elision (9.4.4) and liaison (9.4.5)⁴.

9.4.1 Overall proportion of phonetic changes

Phonetic variations are analyzed by comparing the NLP-produced standard phonetization with the real pronunciation by the speaker. The overall proportion of phonetic changes is computed as the total amount of modifications (deletions, insertions or substitutions) divided by the maximal number of characters, *i.e.* the number of characters of the longest of both strings.

Table 9.2 shows significant differences in the amount of phonetic changes for the read dimension ($p = 3.7e - 06$). This indicates that spontaneous speech differs rather strongly from what produces a generic NLP. Conversely, read speech corpora exploited for speech synthesis display, on average, only 1.33 % of phonetic changes⁵. This finding indicates that, while NLP phonetizers are suited for read speech, they may require some modifications when synthesizing spontaneous messages. Non-media and

⁴It should be noted that all these phonetic modification phenomena relate to the insertion or deletion of one or several phonemes. As previously mentioned in Section 5.2, phone substitution phenomena, such as the modification of the voicing state of consonants, are also known to be frequent in French and are often associated with elision and insertion phenomena. We may, for instance, observe consonant devoicing resulting from schwa elision, such as in *je pense* pronounced [ʃpãs] and not [ʒpãs] following the schwa elision in the pronoun "je". These phone substitution phenomena were however not analyzed in the framework of our study because the phonetic annotation protocol did not allow for it (for reasons detailed above). Further studies should however been carried out in that direction.

⁵One should however be aware that this very low percentage of phonetic differences between manual and automatic phonetization is partly due to the nature of the dictionary exploited by our NLP. As already mentioned, that dictionary provides a rather precise phonetic transcription which already includes some phonetic variations.

expressive speech are also shown to display a higher rate of phonetic variation but the differences are not significant (with $p = 0.086$ and $p = 0.10$, respectively). It should be noted that expressive corpora may be characterized by many other factors like the emotion valence (*i.e.* happy or sad) which could influence their realization. The next sections investigate typical phonetic phenomena for each trait.

Table 9.2 - *Summary of the phonetic changes (in %), computed as the average percentage of differences between the manual and automatic phonetization, for the three situational trait dimensions. For Read, "+" is read, "-" is spontaneous. For Media, "+" is media, "-" is non-media. For Expressive, "+" is expressive, "-" is non-expressive, i.e. neutral. The corresponding sub-corpora are shown in Table 9.1.*

Situational trait	All changes		Elision of [l] in 'il'		Elision of liquid in obstruent+liquid		Liaisons	
	+	-	+	-	+	-	+	-
Read	1.79	4.54	8.33	76.56	3.65	50.34	59.33	42.15
Media	3.25	4.20	51.87	96.67	18.53	49.99	54.52	44.77
Expressive	3.85	2.98	52.53	34.72	33.54	21.74	44.53	50.27

9.4.2 Schwa elision

Schwa elision is one of the most intricate phonetic variations in French. It relates to schwas which can be pronounced or not at the middle or the end of a word. Our analysis excludes final schwas which may rather be related to liaisons and are further investigated in the next subsection. The percentage of elided schwa is here computed as the number of schwa deletions, inside words, divided by the total amount of schwas inside words. Figure 9.1 shows the significant role played by the distinctions spontaneous/read and media/non-media (with respectively $p = 0.005$ and $p = 1.7e - 04$). It shows that more schwas are elided in spontaneous speech, corroborating the results of earlier studies (Hansen, 1991; Adda-Decker et al., 1999; Fougeron and Frauenfelder, 2001; Van Bael et al., 2004; Hambye, 2005). An interesting finding is that more schwas are also elided in non-media speech. This may be explained by the fact that media speech tends to belong to a higher level of language which has been said to be correlated with lower elision rates (Warnant, 1996). As in Bürki et al. (2011), we observe rather high inter-speaker variability.

9.4.3 Final schwa pronunciation

Final schwa pronunciation also called 'supporting schwa', refers to the insertion of a final schwa in a word ending or not in -e (*e.g.* *match* pronounced [matʃə]) (Hansen, 1991, 1994; Carton, 1999; Hansen and Hansen, 2003). To describe that phenomenon, Hansen (1991) talks of a "schwa-like echo that can be observed in a oratorical style, where final consonants are particularly clearly articulated". Candea (2002) shows that its frequency has increased in the last decades and that it can occur independently of

IMPACT OF PHONETIC VARIATIONS ON THE SYNTHESIS OF SPORTS COMMENTARIES

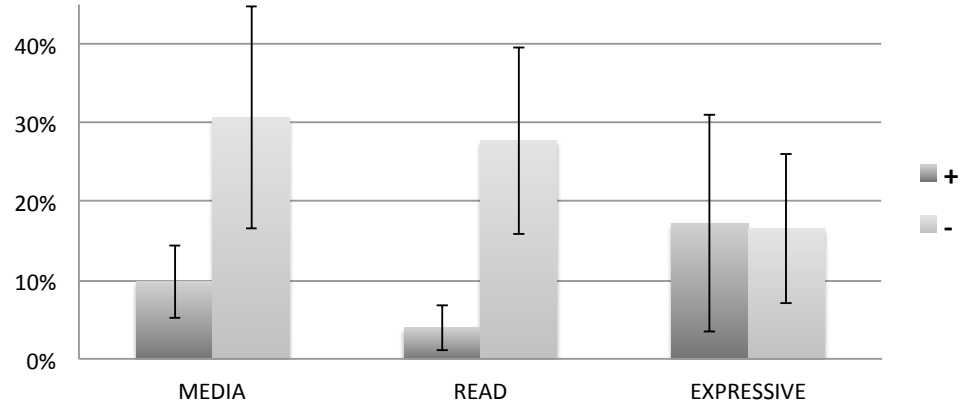


Figure 9.1 - Percentage of elided schwas in middle-word position, together with their 95 % confidence intervals. For Read, "+" is read, "-" is spontaneous. For Media, "+" is media, "-" is non-media. For Expressive, "+" is expressive, "-" is non-expressive, i.e. neutral. The corresponding sub-corpora are shown in Table 9.1.

the phonetic or rhythmic context. While it was previously seen as a sign of informality, the sociolinguistic aspect is now fading out.

Our analysis focused only final schwa pronunciation in words not ending in final -e which will be called "epenthetic schwas" in the remainder of this chapter. This phenomenon has sometimes also been referred to as "parasitic schwas" (Hansen and Hansen, 2003)⁶. Interestingly, while this phenomenon is commonly considered as typical of Parisian French (Candea, 2002; Hansen and Hansen, 2003), we found no significant influence of the speaker's origin on the realization of epenthetic schwa.

While global phonetic changes and elisions are shown to display similar behaviors in read and media speech, the study of epenthetic schwas highlights characteristics of media speech. Interestingly, this variation is significantly more frequent in media compared to non-media speech, as shown in Figure 9.2 ($p = 0.013$). This goes in line with Roekhaut et al. (2010) who pointed out a higher rate of ending schwa pronunciations (all words considered) in radio news and political speech compared to conversational speech. This rate is also significantly higher in spontaneous and expressive speech ($p = 0.019$ and $p = 0.008$), most likely due to their high frequency in sports commentaries. High inter-speaker variability is however observed. While they are often studied on Parisian French, no difference was witnessed in our corpus between French, Belgian and Swiss speakers.

⁶It should be noted that the transcription of such epenthetic schwas is rather complex as they can be easily confused with hesitation marks "euh" (Candea, 2002). Carton (1999) highlights that issue and indicates that both phenomena can be considered as belonging to a same continuum of events. Our annotation is mainly based on Candea (2002) who considers epenthetic schwas as less "intense" and shorter than hesitation schwas. Duration was shown to be an insightful criterion to distinguish between both phenomena.

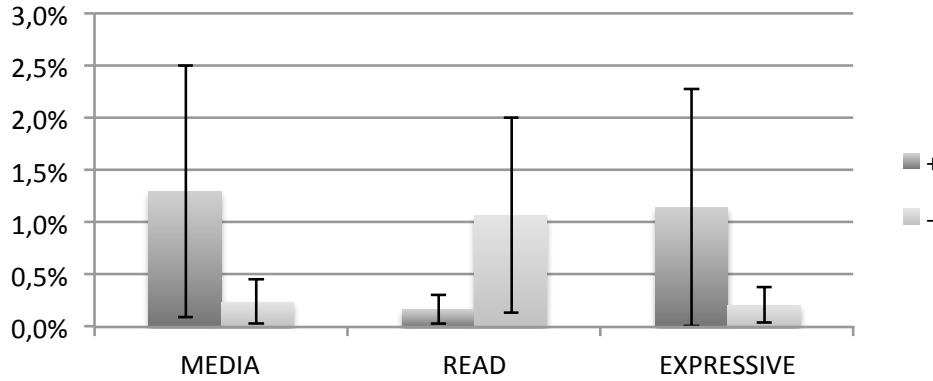


Figure 9.2 - Percentage of words, not ending in -e, pronounced with a final schwa, together with their 95 % confidence intervals. For Read, "+" is read, "-" is spontaneous. For Media, "+" is media, "-" is non-media. For Expressive, "+" is expressive, "-" is non-expressive, i.e. neutral. The corresponding sub-corpora are shown in Table 9.1.

In a second stage, a statistical analysis of final schwas was carried out on all words (ending or not in -e) except for monosyllabic function words (e.g. *le*, *que*, *se*), for which the non-pronunciation of the final schwa could be studied in terms of elision. Here again, final schwa insertion is significantly more frequent ($p=0.0028$) in media compared to non-media speech. Such significant results differ from Hansen (1991) who found no straightforward influence of the speaking style on this phonotactic variable. This may be due to their use of a reduced corpus, containing read speech, conversations and interviews only and their focus on the spontaneous/read speech dimension alone. Besides no significant differences are observed for the other two situational traits. Figure 9.3 shows the percentage of final schwa pronunciation across the various CS. It shows that higher rates are observed in media corpora (i.e. sports commentaries, TV news, political speech and interviews).

9.4.4 Final liquid elision

When analyzing the alignment of both predicted and real phonetizations, we observed that the 'il' pronoun (meaning 'he' or 'it') is often pronounced [i], with elision of the final 'l'. This phenomenon occurring quasi-exclusively before consonants, we analyzed its appearance in that specific context (see Table 9.2). Only sub-corpora containing at least 3 occurrences of the pronoun in that context were kept. Both media and read traits are significant, with higher elision rates in spontaneous and non-media speech (respectively $p = 6.1529e - 05$ and $p = 0.002$) which goes in line with results obtained for the elision of schwa in Section 9.4.2.

IMPACT OF PHONETIC VARIATIONS ON THE SYNTHESIS OF SPORTS COMMENTARIES

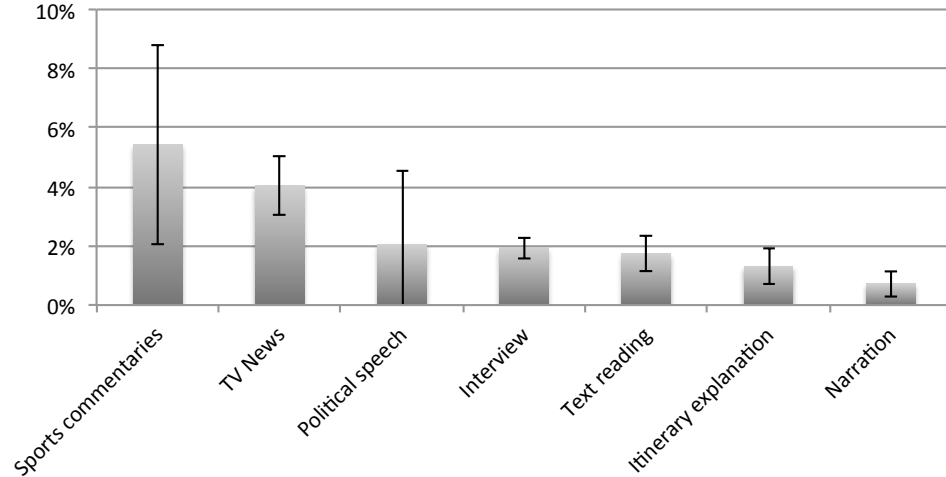


Figure 9.3 - *Percentage of words (excluding monosyllabic function words) pronounced with final schwa, for each CS, with their 95 % confidence intervals.*

Another type of consonant elision regards the elision of the word-final liquid when preceded by an obstruent (*e.g.* ‘peut-être’ pronounced [pøtɛt]) (de Reuse, 1987). A first qualitative analysis shows that this phonetic variation highly depends on the phonetic context. The liquid is nearly always pronounced when followed by a vowel. Conversely, it may be dropped when followed by a consonant. Our analysis was carried out in this latter phonetic context only, on corpora containing at least 4 occurrences of such phonetic context. Table 9.2 shows that, here again, spontaneous and non-media corpora display significantly more elisions of the liquid than read and media speech (with respectively $p = 1.1e - 05$ and $p = 0.04$).

9.4.5 Liaison

We defined potential liaison contexts as words ending in a French liaison consonant /t, n, z, R, p/ and followed by a vowel, as in Fougeron and Frauenfelder (2001); Fougeron et al. (2001); and Boula de Mareüil et al. (2003). It should be noted that these potential liaisons do not only refer to the so-called ‘optional liaisons’, all liaisons being considered, be they mandatory, facultative or prohibited. Table 9.2 shows that read speech displays a significantly higher rate of liaisons ($p = 4.6396e - 05$) which confirms findings in Fougeron and Frauenfelder (2001); Fougeron et al. (2001); Lucci (1983). Interestingly, media speech also shows more liaisons, even if the difference is not significant. This might be due to the fact that media and read speech tend to be more formal, ‘sustained’ speech being more inclined to high liaison rates (Argod-Dutard, 1996; Warnant, 1996).

9.4.6 Summary

In summary, the statistical analysis of the corpus points out that:

- Read speech is characterized by significantly:

- Less phonetic variations in general
- Less schwa elisions
- Less epenthetic schwas
- Less final liquid elision
- More liaisons

compared to spontaneous speech.

- Media speech is characterized by significantly:

- Less phonetic variations in general
- Less schwa elisions
- More epenthetic schwas, and final schwa pronunciation in general
- Less final liquid elision

compared to non-media speech.

- No significant phonetic difference is found between expressive and non-expressive speech.

9.5 Phonetic variations : Focus on sports commentaries

This second stage of our analysis focuses on differences between sports commentaries and neutral read speech recorded for TTS purposes. The motivation behind this comparison is that our ultimate objective is to produce naturally-sounding sports commentaries while generic synthesizers tend to produce neutral read speech. We should therefore provide greater insight into the major discrepancies between those two speech styles.

We expect sports commentaries to display phonetic variation related to spontaneous, media (expressive) speech. Neutral read speech should be, on the other hand, more similar to what is produced by a standard NLP phonetizer. The aim of our analysis is to highlight the most significant variations observed in sports commentaries to evaluate their impact on speech synthesis in Section 9.7. The significance of the differences between both corpora is computed on all samples with the tests for equality of two proportions.

Table 9.3 shows that the percentage of phone variations in sports commentaries is significantly higher than in neutral read speech ($p < 0.0001$). Regarding schwa

elision, our analysis only focuses on non-final schwa (except for monosyllabic words) as final schwa may instead be linked to liaisons and is studied separately. As displayed in Table 9.3, significantly more schwa elisions are found in sports commentaries ($p < 1e - 08$). The analysis of the corresponding occurrences shows that they primarily appear in monosyllabic words; more than 20 % of the monosyllabic words ending in -e are shortened (see Figure 9.4). We observe large inter-speaker variations in the elision of schwa in the middle of a word. Our analysis also showed that mainly function words tend to undergo elision (about 21 % against 9 % for content words). These two observations are probably correlated as most function words are monosyllabic in French.

Sports commentaries, as expected, also exhibit a significantly higher rate of final schwa pronunciation for all words and of epenthetic schwas, *i.e.* schwa insertions at the end of words not ending in -e ($p < 1e - 08$ and $p < 1e - 08$, respectively). Further investigations indicated that the insertion of a final schwa usually occurs before a silence, as studied in Hansen (1991). Epenthetic schwas were observed to appear usually after a consonant, most probably for articulation purposes. The elision of liquids, both in the ‘il’ pronoun and after an obstruent are also significantly more frequent in sports commentaries ($p = 1e - 08$ and $p = 1.288e - 07$, respectively), compared to our READ corpus in which they are almost never elided. Finally, as expected in spontaneous speech, sports commentaries display significantly fewer liaisons than TTS neutral read speech ($p = 5.119e - 06$). High inter-speaker variability is however observed for that measure.

Table 9.3 - *Phonetic changes, computed as the average percentage of differences between the manual and automatic phonetization, for sports commentaries (SPORTS) and neutral read speech for TTS (READ).*

	All changes	[ə] elision	Final [ə] pronunciation	Epenthetic [ə]	[l] elision in ‘il’	Elision of liquid after obstruent	Liaison
READ	1.33 %	5.92 %	2.20 %	0.13 %	0 %	5.68 %	55.90 %
SPORTS	3.69 %	17.33 %	5.43 %	2.98 %	41.48 %	12.34 %	47.21 %

9.6 Correlations between phonetic and rhythmic features

Rhythm, and speaking rate in particular, has been shown to be correlated with schwa elisions, more elisions appearing in fast speech (Lacheret-Dujour, 1991). Similarly, Jande (2003a) indicates that phonetic reductions are also more frequent in fast speech in Swedish. Jande (2008) also pointed out that the realization of Swedish phonetic variants was influenced by multiple prosodic factors including prominence and mean vowel duration. This section analyzes various rhythmic features to assess their correspondence with the situational traits and evaluate their correlation with the aforementioned phonetic variations. It focuses on three rhythmic measures: the

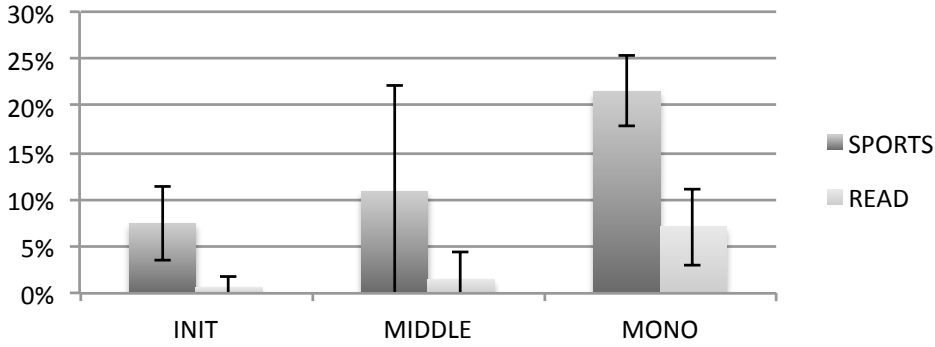


Figure 9.4 - Percentage of elided schwas (monosyllabic, initial, or in middle-word position) in sports commentaries (SPORTS) and neutral read speech for TTS (READ), with their 95 % confidence intervals.

articulation rate, the mean duration of interpausal units (IPU) and the proportion of prominent syllables.

The analysis of the *articulation rate*⁷ highlights significant differences for the read trait only, with lower rates in spontaneous speech ($p = 0.019$). This can be explained by the presence of lengthened syllables, due to hesitations. As in Simon et al. (2009), we also observe a significantly lower percentage of articulation for spontaneous speech ($p = 0.049$). Unlike existing studies (*e.g.* Lacheret-Dujour (1991)), however, no significant correlation is found between speaking rate, or articulation rate and schwa elision ($|Rho| < 0.10$ and $|Rho| < 0.05$, respectively), or any other phonetic variation. We have shown, on the contrary, that elisions are more frequent in spontaneous speech which displays a lower articulation rate. This difference might be due to the fact that most existing studies focus on one specific task (*e.g.* text reading) in which only speaking rate is modified to observe the frequency of elisions. In our corpus, too many factors are influencing the speaking rate, *e.g.* hesitations, communicative situation and speaker idiosyncrasies.

The mean duration of IPU turns out to be significantly longer in spontaneous speech compared to read speech ($p = 0.04$). This seems to indicate that spontaneous speech displays less but longer silences. A possible explanation is that short pauses may be rarely silent in spontaneous speech and rather tend to be filled by disfluency markers like hesitations. Finally, the *percentage of prominent syllables* is assessed with Prosoprom (Goldman et al., 2007), an automatic algorithm for detecting prominent syllables on an acoustic basis. Only the media dimension seems to be influential for that measure, media speech containing more prominent syllables (see Figure 9.5). This distinction also stands out when looking at initial stresses only,

⁷We focus here on the articulation rate, *i.e.* the speaking rate (number of syllables per second) with silences excluded, the sub-corpora displaying different silence densities.

IMPACT OF PHONETIC VARIATIONS ON THE SYNTHESIS OF SPORTS COMMENTARIES

media displaying 19.1 % of prominent initial syllables against 16.4 % in non-media speech. This confirms findings in Goldman et al. (2009). However, the inter-speaker variability is rather high which makes this difference not significant ($p = 0.07$).

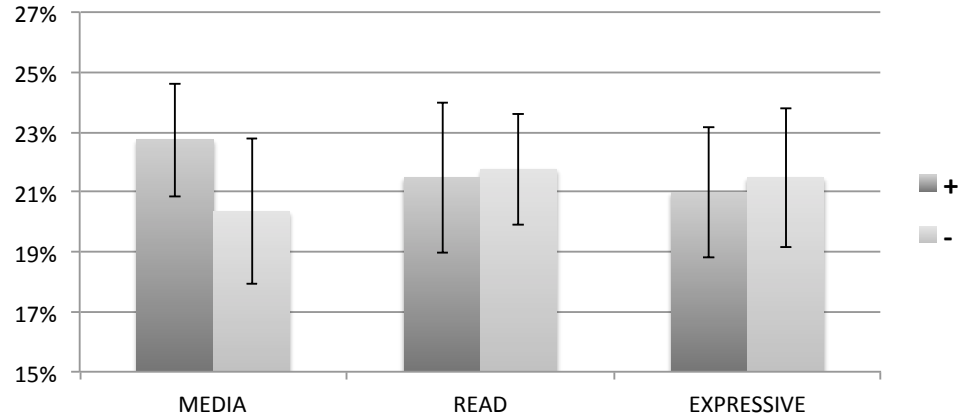


Figure 9.5 - Percentage of prominent syllables, together with their 95% confidence intervals. For Read, "+" is read, "-" is spontaneous. For Media, "+" is media, "-" is non-media. For Expressive, "+" is expressive, "-" is non-expressive, i.e. neutral. The corresponding sub-corpora are shown in Table 9.1.

IPU duration is moderately correlated with schwa elisions and 'l' elisions of the 'il' pronoun (respectively $Rho = 0.59$ and $Rho = 0.43$). This may imply that more elisions are realized when longer sections of speech are uttered without pauses. This, however, does not hold for liquid elisions after obstruents. Interestingly, no correlation can be found between the percentage of prominent syllables and phonetic variations.

On the whole, correlations between rhythmic and phonetic variations are rather weak. This seems to indicate that phonetic variation is more directly dependent on the situational trait itself than on the rhythmic features of the trait. It should be noted that, as for phonetic variation, the expressive trait is not characterized by any specific rhythmic feature.

Further studies could investigate the link between phonetic changes and specific labels in the local or global prosody annotation proposed in Chapter 6. We might expect, for example, that some phenomena often occur in a sub-genre with high arousal level or on a stressed syllable. This could indicate whether these three levels are closely intertwined.

9.7 Impact of phonetic variation on synthesized speech

9.7.1 Experimental protocol

To evaluate the naturalness of the generated speech with and without specific phonetic variation, an HMM-based speech synthesizer was built, relying on the implementation of the HTS toolkit (version 2.1) (Zen et al., 2007a). For training, 90% of the Sportic corpus (presented in Section 5.6.3) was used, leaving 10% for synthesis. The training process relies on manually-corrected phonetic transcriptions, which is an advantage of eLite-HTS (see Section 5.5). Model configuration, filter parametrization and excitation modeling were identical to those exploited in the evaluation protocol of Chapter 6. Local prosody annotation was used as contextual information in the same way as linguistic information (For more details about our prosody annotation protocol and its integration in TTS, see Chapter 6).

Two methods of phonetization were compared: use of a canonical NLP-produced phonetization (*NLP*), or use of the actually realized pronunciation of the sentence in the corpus (*Real*). This comparison was analyzed through two subjective tests. For the first test, 30 sentences were manually selected from the synthesis set as displaying three or four modifications compared to the NLP-produced phonetization. Around 20% of the sentences of the corpus contain 3 or more modifications. Occasional errors made by the NLP phonetizer were manually corrected in order not to influence the listener. 22 native French-speaking participants, mainly naive, took the test. For 15 randomly selected sentences, two versions were compared, with the real and the NLP phonetization. Listeners were asked to score the naturalness of the message as corresponding to sports commentaries. The scale ranged from -3 (much less natural) to +3 (much more natural). A score of 0 was to be chosen if both versions were seen as equivalent.

The purpose of the second test was to assess the relative role played by 4 types of phonetic variation: schwa elision, final schwa insertion, liquid elision and liaison. 60 sentences (or chunks of sentences) were manually selected from the synthesis test, *i.e.*, 15 for each type of phonetic variation. Each sentence originally contained only one phonetic modification of the corresponding type. Another 21 native French participants, mainly naive, took the test. They were asked to compare 16 randomly selected pairs of sentences, 4 for each type of phonetic variation. The other evaluation conditions were the same as in the first test. For the interpretation of the results, statistical significance is computed with Wilcoxon signed rank tests comparing the average percentage of preferences of the 22 (or 21) testers.

9.7.2 Evaluation and discussion

Results of the first test are displayed in Figure 9.6, which shows a significant degree of preference for the genuine/real phonetization, as produced by the human speaker and containing phonetic variation ($p = 0.012$). Synthesis with real phonetization is

IMPACT OF PHONETIC VARIATIONS ON THE SYNTHESIS OF SPORTS COMMENTARIES

preferred in 45.15% of the cases against 31.21% for the NLP-produced phonetization, while the two were found to be equivalent in 23.64% of cases. Figure 9.7 highlights the significantly higher percentage of preference for the real pronunciation in sentences in which at least two schwa elisions appear ($p = 1.86e - 04$). In this case, real phonetization is preferred in 66.67% of the cases against 19.44% for NLP-produced phonetization (average preference score of 1.6 when the real phonetization is preferred). Conversely, sentences with more than two final schwa insertions were assigned lower scores. This indicates that schwa elisions may be required, as opposed to final schwa insertion. This is further investigated in our second subjective test.

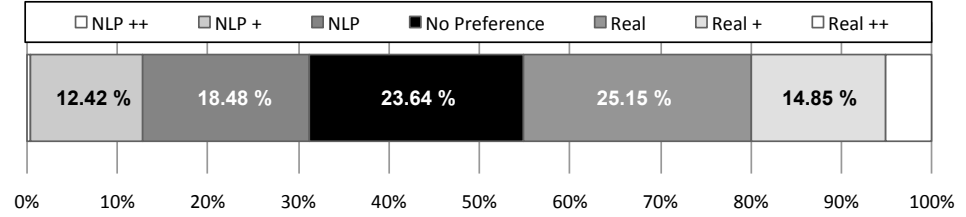


Figure 9.6 - Preferences (%) for NLP-produced and real phonetization for the first test. "+" refers to stronger preference scores.

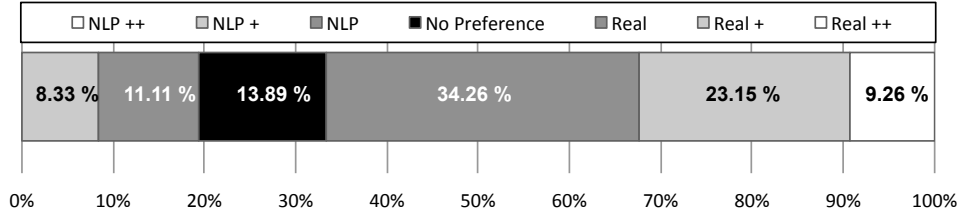


Figure 9.7 - Preferences (%) for NLP-produced and real phonetization for the first test in sentences with at least two [ə] elisions. "+" refers to stronger preference scores.

Overall results of the second test show substantial differences in ratings according to the type of phonetic variation (cf. Figure 9.8). It should first be pointed out that the overall preference for the real phonetization is slightly lower than in the first experiment, which may be explained by the fact that both versions of the test sentences only differ by a single variation. This may imply that the use of such short chunks of speech, out of context and independent of other variations, somehow influences the results and thus may not be representative of real spontaneous speech. This should prevent us from drawing hasty conclusions.

For liaisons and final schwa, the NLP-produced phonetization tends to be preferred, *i.e.* a realized liaison and no insertion of final schwa. For final schwa insertion, the real phonetization is however slightly preferred in the case of epenthetic schwa, *i.e.* insertion of final schwa in words not ending in -e. This phenomenon

was indeed shown to be highly specific to sports commentaries in the previous section. A more detailed analysis of the prosody patterns of such insertions might allow for a better realization of that type of variation (see the description of the complex contours accompanying epenthetic schwas in Hansen and Hansen (2003)). Such research could rely on Hansen (1991) which points out that the speaking style influences the pitch contour realization of this phenomenon. For the elision of liquids, the real phonetization tends to be preferred, *i.e.* with elision of the final liquid after an obstruent or in the ‘il’ pronoun. Further investigation shows that this preference rate is higher (46.67% vs. 33.33% for NLP-produced phonetization) when considering only final sequences with obstruent and liquid; the variation in the ‘il’ pronoun usually results in a ‘no preference’ rate. This difference, however, is not statically significant ($p = 0.06$). Finally, as noticed in the first experiments, the real phonetization is rated significantly more natural than the NLP-produced pronunciation for the elision of schwa ($p = 0.006$).

It should be noted that the perceptive tests focused on the evaluation of the naturalness of the expressivity only, thereby neglecting other aspects like intelligibility. This resulted from a common assumption that "a synthesis perceived as more natural is also easier to understand and less demanding to listen to" (Jande, 2003a). This indicates the interconnection between both evaluation criteria and implies that the lower rates, obtained for example for epenthetic schwas pronunciations, may be due to a degradation of the intelligibility⁸. This should be confirmed by further studies.

9.8 Conclusion and perspectives

While speech synthesis of various speaking styles and emotions is now the focus of much research, very few studies consider potential phonetic variations according to the communicative situation (*e.g.* expressive or spontaneous speech). This chapter offered a phonetic analysis of a large corpus in French to assess the influence played by three situational traits: read/spontaneous, media/non-media and expressive/non-expressive. It first showed that spontaneous speech exhibits a significantly higher percentage of phonetic variations compared to standard read speech as produced by a conventional NLP for TTS. Regarding the various phenomena, we showed that spontaneous speech is characterized by a higher elision rate and less liaisons, which confirms results of earlier studies. Our study also indicates that media speech usually follows the same phonetic tendencies as read speech, which may be due to the overall higher level of language compared to spontaneous and non-media speech. However, it displays much more epenthetic schwas, which seems to be particularly characteristic of sports commentaries. Interestingly, the expressive trait is not associated with any specific phonetic feature. The diversity of the corpora in that trait (*e.g.* emotions with different valences) should be further investigated to evaluate the role played by the different aspects of expressivity.

⁸Such deterioration of intelligibility due to phonetic changes was found in Larreur and Sorin (1990) with schwa elision.

IMPACT OF PHONETIC VARIATIONS ON THE SYNTHESIS OF SPORTS COMMENTARIES

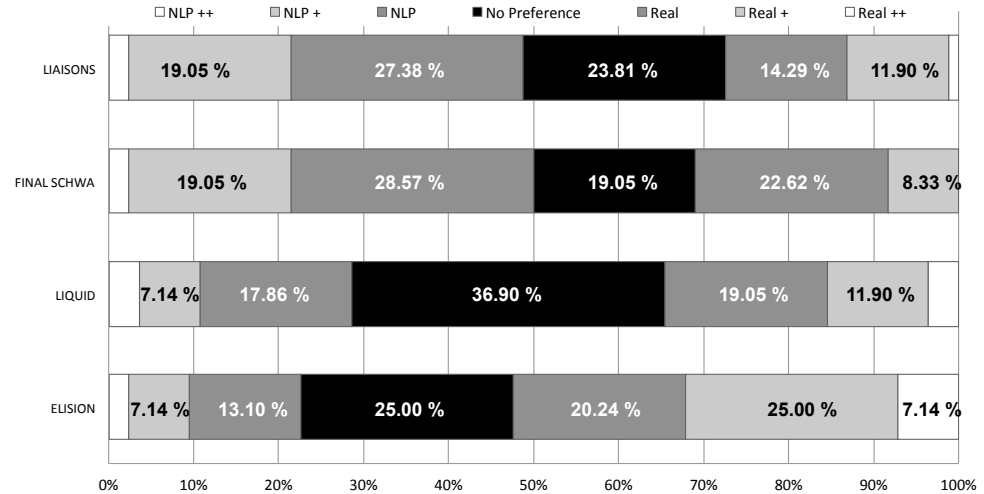


Figure 9.8 - Preferences (%) for NLP-produced and real phonetization for the second test, according to the type of variation (liaison, final schwa pronunciation, liquid elision and schwa elision). "+" refers to stronger preference scores.

Rhythm was analyzed in a second stage and showed higher speaking rates and shorter interpausal units for read speech. Higher proportions of prominences were observed for media speech. Here again, no specific rhythmic feature was associated with the expressive trait. Finally, low levels of correlation were found between the rhythmic parameters and the phonetic variations, except for a moderate correlation between the duration of the interpausal unit and some types of elisions. This implies that phonetic variations may depend primarily on the situational trait (read/spontaneous and media/non-media) and not on the rhythmic features of that trait.

The third stage of our analysis aimed at assessing whether the phonetic variation that appears in various communicative situations (CS) contributes to the plausibility of the message and should, in that respect, be integrated in speech synthesis. It focused on sports commentaries, which were shown to be characterized by higher rates of schwa and liquid elisions, more final schwa pronunciations and fewer liaisons. Subjective evaluations with HMM-based synthesized extracts of sports commentaries indicate that, on the whole, integrating phonetic variants significantly improves the naturalness of the synthesized message. While the insertion of final schwa or the reduction of the number of liaisons does not enhance the naturalness, integrating schwa elision and, to a lesser extent, final liquid elision after obstruents turns out to improve the overall naturalness. This indicates that the synthesis of French sports commentaries should benefit from such phonetic modifications. One should however be aware that, ideally, not only these phone elision phenomena should be predicted

but also the phonetic contextual changes resulting from it, such as the devoicing of surrounding consonants.

Further studies should be carried out to explore whether or not these conclusions also hold for other speaking styles and languages. This is investigated in Chapter 11. Perspectives obviously include the subsequent integration of phonetic modification rules to transform the phonetic transcription of the sentences to synthesize according to the speaking style. In that sense, the use of a more phonological transcription as input (as opposed to the more phonetic transcription proposed by Brulex, and subsequently by eLite-HTS) would constitute a better basis for the design of pronunciation modification rules. This constitutes the major perspective of our research and is further detailed in Chapter 12.

Part III

Prosodic and phonetic improvements of HMM-based speech synthesis: Further developments

This third part of the thesis aims at investigating the possible generalization of the techniques and observations made in Part II. Regarding prosodic modifications, experiments on the speech synthesis of sports commentaries have indicated that the integration of a local prosody annotation, in particular, achieved high improvement in the naturalness and expressivity of the synthesized voice. In terms of phonetic realization, we have pointed out that a wide range of phonetic variations may be assigned to spontaneous and media speech. We have further highlighted that the integration of these phonetic changes in the speech synthesis of sports commentaries significantly improved the naturalness of the resulting speech. Through perceptive tests, we have shown that schwa elision and liquid deletion, in particular, played a crucial role in that respect. Such observations are further investigated in this last part of the thesis with the use of other corpora containing different speaking styles and languages.

Chapter 10 focuses on prosodic modifications. It sets out to investigate the integration of a local prosody annotation in the speech synthesis of audiobooks in French. In a first stage, we present a method to automatically predict an autosegmental annotation of a French corpus, requiring no manually-annotated data. In a second stage, we explore the integration of such annotation in HMM-based expressive speech synthesis. To that end, our automatic annotation tool is applied to a corpus of audiobooks on which a synthesizer is trained. The granularity of the annotation (*i.e.* the number of prosodic labels) is also questioned and evaluated through preceptive tests of the resulting synthesis.

In **Chapter 11**, we focus on the integration of phonetic variation in the framework of spontaneous speech synthesis, as this speaking style was shown to induce high phonetic variability in the second part of the thesis. The chapter builds upon the observations made in Chapter 9 to provide further insight on how phonetic modifications should be dealt with when generating spontaneous speech. The chapter provides preliminary results of two extended studies we carried out on that topic and which are still currently in progress.

First, an in-depth study of the impact of spontaneous speech on phonetic changes is carried out, through the statistical analysis of a large corpus in French. This study presents the advantage of relying on a corpus of more than 13 hours containing various speakers and a wide range of linguistic contexts. Each speaker is recorded in both spontaneous and read speech which facilitates the comparison between both speaking styles. This analysis also relies on a rich annotation of the corpus which was manually checked. It gives insight into the way phonotactic, lexical, prosodic, sociolinguistic and stylistic variables interact when modeling the behavior of optional phonetic facility phenomena in French. This constitutes, in our view, a crucial preliminary step to the definition of phonetic modification rules in French, which should consider the role played by a wide range of contextual factors and the way in which they interact.

The second study presented in this chapter focuses on the synthesis of spontaneous speech in English. It sets out to question the widespread consistency assumption according to which errors made in the phonetization of the training corpus should not affect the resulting synthesized voice if the same errors are also produced at synthesis. Our study investigates the validity of that hypothesis

in the case of synthesis training on spontaneous speech, which was shown in Chapter 9 to display a high amount of phonetic variation. To that aim, it relies on a parallel corpus of sentences read and pronounced spontaneously by the same speaker. This allows for a precise analysis of phonetic changes due to the speaking style only, by neutralizing all other factors potentially impacting the pronunciation. The evaluation relies on perceptive tests of the resulting HMM-based speech synthesis.

It should be noted that this third part of the thesis benefits from collaborations with researchers from other universities that we have met during the PhD, namely Prof. Piet Mertens (KULeuven, Belgium), Dr. Mathieu Avanzi (Cambridge University, UK) and Rasmus Dall and his colleagues (Center of Speech Technology Research (CSTR), Edinburgh, UK). Beside the obvious added value of knowledge and competence transfer arising from interactions with other experts of connected domains, it also granted us access to other valuable resources. In the framework of prosodic modeling, our research benefited from Polytonia (Mertens, 2013a,c, 2014), an automatic tonal annotator for French speech corpora. Regarding French phonetic analysis, the CPROM-PFC corpus (Avanzi, 2014) was also put at our disposal. It contains both spontaneous and read speech and has been manually checked in terms of phonetic realization. Finally, our research on spontaneous speech synthesis could rely on an English parallel corpus (read-spontaneous) recorded at CSTR (Dall et al., 2014).

Chapter 10

Automatic annotation of an autosegmental description of French prosody

Contents

10.1 Introduction	229
10.2 Developing an automatic prosody annotator	232
10.2.1 Development corpus	234
10.2.2 Annotation protocol	234
10.2.3 Choice for an automatic prominence detector	239
10.2.4 An alternative option: Training an AuToBI classifier on a French annotated corpus	241
10.3 Integration of the local annotation in speech synthesis	242
10.3.1 Description of the corpus	242
10.3.2 Method	243
10.3.3 Evaluation and discussion	245
10.4 Reducing the set of labels	247
10.4.1 Annotation protocol	247
10.4.2 Perceptive evaluation	250
10.5 Conclusion	251

Abstract

This chapter builds upon the results of Chapter 6 which indicated that the integration of a local prosody annotation significantly improved the naturalness and expressivity of the synthesis of live sports commentaries. In that framework, we present a method to automatically provide an autosegmental annotation of French prosody from any speech corpus and its orthographic transcription, in view of speech synthesis. To our knowledge, no such generic automatic annotator has been proposed in French, despite the wide range of research dealing with that topic in English. To that aim, we investigate the possibility to combine three sources of information as provided by prominence detectors, tonal annotators and NLPs (providing phonetization, syllabification and rhythmic group segmentation of speech). Based on phonological intonation models for French, we set out to implement a set of rules to merge these annotation layers into a unified annotation protocol, called ToPPos-Fr. A major advantage of that method is that it does not require the use of a manually-annotated corpus in terms of prosody specifications. Qualitative analyses of the resulting annotation on a development corpus show that it provides a good-quality description of the accentual patterns. In a second stage, the integration of this annotation in HMM-based speech synthesis of audiobooks is assessed, through a set of preliminary perception tests. The results indicate that it leads to significant improvement of the naturalness of the synthesized voice, which is shown to display prosodic contours that resemble those of the original speech segments. Finally, we investigate the possibility to reduce the set of labels with the intention to facilitate their prediction from text. The resulting speech synthesis is however not shown to provide significant improvement over the baseline.

A note on collaborative work:

This study was carried out in strong collaboration with **Prof. Piet Mertens**. Through fruitful discussions on the topic, we designed together a set of rules to provide an automatic annotation of local prosodic phenomena in French. He then designed a Praat script to automate the process. In a second stage, I tested different prominence detectors which resulted in the training of new Promise models by **George Christodoulides**. I trained the synthesis and implemented the listening tests. The reduced set of labels was defined through discussions with **Prof. Mertens**. Based on that, I designed a script to automate the labels rewriting, trained the synthesizer and ran perception tests. I also carried out the analysis of the results.

We are very grateful to **Prof. Mertens** for his invaluable advice and the long and fruitful discussions we have had on this topic. We also want to thank **Prof. Andrew Rosenberg**, who proposed an adaptation of AuToBI for French, and **Prof. Julia Hirschberg** who guided us and provided us with insightful advice throughout this study.

10.1 Introduction

We have shown in Part II that the integration of a local prosody annotation, considering accentual, intonational and phrasing aspects of the corpus, significantly improves the naturalness and expressivity of HMM-based speech synthesis of sports commentaries. In Chapter 8 we have however pointed out that manually annotating large amounts of speech data (as required for the training of speech synthesizers) with prosodic information is a tedious and costly task (Campione and Véronis, 2001; Jeon and Liu, 2012). This is pointed out by Geoffrois (1995) who indicated that manually-annotating one hour of spontaneous speech in French with Mertens’ tones (as presented in Subsection 5.3.2) required more than 500 hours. We have therefore proposed avenues to automatically annotate local labels from acoustics and have indicated that SVMs relying on syllable-based acoustic criteria achieved reasonable rates in that respect. The models were however trained and tested on different parts of the same corpus which obviously facilitated the task of the annotator. The application of such an automatic annotator to other corpora with different speaking styles is therefore questionable.

Ideally, an automatic prosody annotator should provide generic models trained on a large database containing various speakers and speaking styles. The development of such an automatic annotator is investigated in this chapter, in view of expressive HMM-based speech synthesis. The resulting prosodic annotation is evaluated through perception tests of the generated speech synthesis. As local prosodic annotation highly depends on the language, we develop a language-dependent method specifically designed for French corpora.

In the linguistic research community, the automatic autosegmental annotation of prosody¹ of speech corpora from acoustics (and additional linguistic features) has aroused great interest. For an insightful recent review of English studies on that topic, see Jeon and Liu (2012).

A wide range of studies has focused on the detection of prominence and phrase boundaries (possibly complemented by an indication of the break index tier level). Wightman and Ostendorf (1994) present an automatic labeler based on decision trees and Markov sequences. A similar annotation framework is proposed in Chen et al. (2004a), in which both acoustic and syntactic criteria are exploited by means of Artificial Neural Networks (ANN) and Gaussian Mixture Models (GMM), respectively. They achieve annotation rates similar to the agreement rates between different human labelers. In the same line, Sridhar et al. (2008) and Ananthakrishnan and Narayanan (2008) also detect prominences and prosodic phrase boundaries based on a maximum entropy framework and a comparison of various classification methods (including Linear Discriminant, GMM and neural networks), respectively. Sridhar et al. (2008) additionally provide an insightful state-of-the-art of existing studies on automatic prominence and boundary detection. Jeon and Liu (2012) complement these studies by proposing a co-training framework in which much less hand-labelled data is re-

¹Considered here as an annotation of both phrasing and prominence, enriched by a description of pitch level and movement.

quired as unlabelled data can be exploited to improve the model. Recent work has also investigated the possibility to train phrase boundary and pitch detectors on one language and apply them to an unseen language (in Soto et al. (2013) and Rosenberg et al. (2012), respectively). Phrase boundary detection was performed with AdaBoost (Freund and Schapire, 1997) and resulted in reasonable detection rates, mainly relying on the presence of a subsequent silence. Pitch accent detection was performed with a logistic regression classifier but achieved poor accuracy in a cross-language setting.

Wagner (2009) concentrates on pitch accents only and proposes a pitch accent type classification, based on a reduced set of acoustic features. She compares several machine learning techniques including neural networks, classification trees and discriminant function analysis. Her study indicates that neural networks achieve the best rates for the detection of prominent syllables while classification trees outperform all other models in terms of pitch accent type classification. Conversely, Wang and Hirschberg (1992) focus on phrase boundary detection only. They indicate that the use of decision trees based on both acoustic and linguistic features achieve good boundary detection, with success rates just over 90 %.

Other studies have focused on the annotation of both pitch accent and boundary tone types. Providing a rule-based automatic annotation of ToBI labels based on acoustic and linguistic criteria, the Prosodizer (Braunschweiler, 2006) achieves reasonable rates on German and English. Syrdal et al. (2001) also propose to predict default labels for pitch and accent events to speed up the manual annotation of American English ToBI. More recently, AuToBI² (Rosenberg, 2010) was also developed as an automatic English ToBI annotator based on several machine learning techniques (including SVMs and logistic regression classifiers). It relies on specific acoustic criteria to achieve four distinct tasks: pitch accent detection, pitch accent classification, phrase detection and phrase ending classification. While the AuToBI website proposes a French version of the annotator, it was trained on a corpus annotated in terms of prominences only and therefore annotates the corpus with a single *P* label.

Because autosegmental annotation of prosody relies on a linguistic description of the language, it is very much language-dependent. Therefore, annotators designed for a language can hardly be transposed to another. To our knowledge, no generic automatic annotator of a detailed autosegmental annotation of prosody has been proposed for French. Prosody annotators have however been developed and can globally be divided in two groups: prominence detectors, and pitch level annotators. They concentrate, respectively, on accentuation (or more precisely prominence) or intonation only but do not combine both sources of information.

In relation to **accentuation**, research has mainly focused on automatic prominence detection (Goldman et al., 2007; Avanzi et al., 2008, 2011; Veaux et al., 2008; Goldman et al., 2012; Christodoulides and Avanzi, 2014). These studies usually annotate with a single prominence label, with the exception of PromGrad (Goldman et al., 2012) which distinguishes between 4 levels of prominence and Avanzi et al. (2011) who propose two prominence levels. Automatic **intonation** annotation includes pitch stylization algorithms (*e.g.* Prosogram (Mertens, 2004a) or MoMEL (Hirst and Di Cristo, 1998; Hirst, 2007)), or further categorization of the pitch continuum through a sym-

²Available at: <http://eniac.cs.qc.cuny.edu/andrew/autobi/>

bolic annotation (indicating pitch target, possibly associated with pitch movements), as in Campione et al. (2000); Bartkova et al. (2012) and Mertens (2013a)³. Bartkova et al. (2012) additionally encodes a symbolic representation of duration and energy, in two separate annotation tiers.

Outside these two main trends, Mertens and Simon (2013) proposed a rule-based boundary tone annotator for French. It defines four different boundary levels (*i.e.* major, intermediate, minor and none) and relies on acoustic criteria including pitch height, syllabic lengthening and pause duration. A major drawback of the algorithm is that it often mis-detects boundary tones on emphatic-stressed syllables. The authors indicate that access to lexical information could alleviate this issue. They also mention the high sensitivity of the system to speaking style variation.

Another insightful study is presented in Campione and Véronis (2001) who propose a semi-automatic annotator of French prosody. The annotation relates however to broad prosodic movements, as it automatically detects major prosodic boundaries to which it assigns major prosodic movements (with different sizes of rise and fall), in a more holistic approach of pitch contours (as presented in Section 5.3.2). Unlike these annotators, the labels are not assigned at the word or syllable level and provide an annotation more suited to syntactic or pragmatic studies.

Finally, it should be noted that Geoffrois (1995) proposed an automatic annotator of Mertens' tones 1987 in the framework of a French speech recognizer, based on acoustic information. The annotator relies on hidden Markov models in a speech recognition fashion, trained on a manually-annotated corpus. That annotator was however not evaluated on another corpus or in the framework of speech synthesis.

Our study sets out to develop an automatic autosegmental prosody annotator for French, at the syllable-level, and to integrate it in the synthesis of expressive speech. It relies on an expressive speech corpus which differs in many aspects from sports commentaries which were exploited in the second part of the thesis, *i.e.* an audiobook corpus.

Audiobooks have been shown to be of great value for speech synthesis training and have been extensively used in the last decade (Zhao et al., 2006; Prahallad et al., 2007; Székely et al., 2011; Hinterleitner et al., 2011; Eyben et al., 2012; Charfuelan and Steiner, 2013; Doukhan, 2013). They are valuable for several reasons, as large corpora of a single speaker with high quality recording and containing expressive speech. Such corpora can be further divided into multiple sub-genres corresponding to different emotions, characters or attitudes (see Székely et al. (2011); Eyben et al. (2012)). Besides, speech synthesis of audiobooks is of considerable interest for commercial use, thereby answering a growing demand for automatic reading of narrative texts.

Many studies have focused on the automatic annotation and alignment of audiobooks, which presents major challenges mainly due to the length of the speech files and the noisy orthographic transcription, since words are often changed or discarded by the reader (Prahallad et al., 2007; Braunschweiler et al., 2010). Since the phonetic annotation of such corpora and their alignment with the correspond-

³As these automatic annotators rely on a pure description of acoustic features only, they are supposed to be language-independent.

ing sound are costly, few aligned audiobooks corpora are made available⁴. This chapter is based on a French book read by a professional actor which was developed in the framework of Nicolas Obin’s thesis (2011) and annotated for that purpose.

The current study meets two main objectives:

- Proposing an automatic annotator to generate, from speech and text, a local prosody annotation of French corpora, based on an autosegmental approach of prosody. This annotator is developed with speech synthesis in mind and assessed through subjective evaluation of the resulting synthesized voice. This is in line with Mertens (2004b) and Rillard (2014) who indicated that speech synthesis constitutes a privileged tool to simulate, evaluate and refine prosodic models.
- Investigating the degree of granularity needed for the annotation, *i.e.* the number of required prosodic labels. As described in Chapter 6, this number should correspond to an optimal balance between acoustic coverage and functional conciseness, to facilitate the prediction of the labels from text. In a first stage, we integrate in the synthesizer a large number of prosodic labels, which can be compared to the set of tags proposed by ToBI (Silverman et al., 1992) or Mertens (1987). In a second stage, a reduction of the set of labels is proposed and evaluated in speech synthesis. This reduced set is very similar to that presented in Chapter 6 and allows, to a certain extent, an evaluation of our annotation protocol on another type of corpus. This second part of the study is in line with experiments in Chapter 7 which investigated the level of granularity required to accurately describe emphatic stresses in expressive speech synthesis.

The chapter is organized as follows. In Section 10.2, we present our automatic annotation protocol and the background research it relies on. In that framework, a development corpus is exploited to provide a first qualitative analysis of the resulting annotation and potentially to improve the developed algorithm. In Section 10.3, our automatic annotation is applied to a large corpus of audiobooks and integrated in an HMM-based speech synthesizer. The evaluation of the resulting synthesized voice is assessed through perception tests. In a second stage, we investigate in Section 10.4 the possibility of reducing our set of labels to facilitate their prediction from text. The resulting annotation is then integrated in the synthesizer to be evaluated with a second set of perception tests. Finally, Section 10.5 concludes the chapter.

10.2 Developing an automatic prosody annotator

To provide an automatic prosody annotator of French corpora, based on an autosegmental approach of prosody, two main avenues could be explored:

- Exploiting existing accentual and intonational annotators as presented in the introduction of the chapter and combine them to annotate stress, pitch events and prosodic phrasing in a unified framework.

⁴Librivox (www.librivox.org) proposes a series of freely available audiobooks. However, they are usually read by amateurs, sometimes in poor recording conditions. They are obviously not annotated nor aligned with the text.

- Exploiting an existing prosody annotator based on machine learning methods (*e.g.* AuTOBI) to train new models based on a large annotated French corpus.

This chapter mainly focuses on the first option. The second option was also investigated but provided preliminary low-quality results. This is briefly outlined in Subsection 10.2.4.

The autosegmental annotation proposed in our study is based on the combination of three sources of information, all obtained at the syllable-level:

- The prominence level (binary value)
- The pitch level and movement
- Phonetic and prosodic information:
 - The position of the syllable in the word
 - The position of the word in the rhythmic group (RG)
 - The vocalic nucleus (binary value indicating the presence of a [ə])

The interest of these three sources of information is that they can all be automatically provided, namely with an automatic prominence detector, a tonal annotator and an NLP module, respectively. We describe in this section our development protocol which aims at merging these 3 levels of prosodic information into a single annotation of Tones in Prosodic Positions in French, called ToPPos-Fr in the remainder of this chapter.

The choice for a prominence detector is postponed to Section 10.2.3 as our development protocol relies, in a first stage, on a manual annotation of prominences. Conversely, the annotation of pitch levels and movements is carried out automatically. The motivation behind this choice is that a pitch level annotator is more reliable as it is based on pitch acoustic criteria which are quantifiable and easily extracted from the corpus. Prominence detection, on the contrary, tends to be less reliable as the notion of prominence is much harder to define objectively⁵. Besides, the quality of the prominence detection is highly dependent on the quality of the manual annotation on which the detector was trained. Prominence detection may also be very sensitive to the characteristics of the corpus (*e.g.* spontaneous or expressive speech).

In the same way, the quality of RG boundaries as predicted by NLPs (usually through a (modified version of a) chunks & chunks algorithm) may be questionable as it highly depends on the quality of the parts-of-speech annotation (which may be low on spontaneous speech). This explains why our development protocol is based on a manual annotation both in terms of prominence and parts-of-speech. In a second stage, existing algorithms and tools are investigated to automate the process.

⁵This may also be due to the fact that the exact parameters related to the phonatory power have not yet been uncovered.

10.2.1 Development corpus

Our development protocol is based on the CPROM corpus (Avanzi et al., 2010)⁶, as described in Section 9.2. CPROM is a 70 minute-long speech corpus, uttered by 28 speakers (16 males and 12 females) in 7 different communicative situations (*i.e.* read speech, political speech, news broadcast, conference, radio interviews, map task and life story). The development of our annotation protocol on such a varied corpus should allow for its generalization to various speaking styles, be they expressive and/or spontaneous. The corpus is annotated in terms of phonemes, syllables and words, all aligned with the speech signal. The phonetic annotation was manually checked.

The main interest of the corpus lies in its manual annotation of prominences, based on the judgements of two linguistic experts. This annotation layer was shown to reach high inter-agreement rates in Avanzi et al. (2010) and can be considered as of high quality. Besides, all divergences were discussed in a joint session during which both experts agreed upon a final common annotation. While CPROM proposes two levels of prominences (p and P), both levels were merged into a single prominence label in the framework of our study, as suggested in Avanzi et al. (2010).

Our study also benefits from an enrichment of the annotation of CPROM realized in Goldman et al. (2010). In the framework of that study, the eLite NLP tool (as presented in Section 5.5) was applied to the corpus, to provide other annotation layers including parts of speech (POS) and RG. Interestingly, the parts-of-speech annotation was manually checked and the resulting RG boundaries were recomputed on the manual POS annotation. Both manual prominence and RG tiers are exploited to develop our annotation protocol. Our annotation also relies on the syllabification and phonetic transcription of the corpus, along with their alignment with the speech signal.

10.2.2 Annotation protocol

As described in the previous section, CPROM provides an annotation in terms of phonemes, syllables, RG and prominence, which accounts for 2 of the 3 aforementioned prosodic information sources. To fulfill our requirements, the corpus still needs to be annotated with pitch levels and movements.

For this, Mertens (2013a,c, 2014) recently proposed an automatic labeling of pitch height and movement at the syllable-level, called Polytonia. Pitch is extracted at the syllable nucleus level⁷, stylized (based on d’Alessandro and Mertens (1995) and Mertens et al. (1997)) and further translated into a reduced set of pitch labels, based on heuristics. Pitch height is both relative to the speaker and to the context, similarly to Mertens’ tones as presented in Subsection 5.3.2. That context includes up to three preceding syllables. Polytonia proposes five pitch levels: B (bottom), L (low), M (medium), H (high) and T (top). These levels are combined with pitch movements,

⁶Available at: <https://sites.google.com/site/corpusprom/telechargement>

⁷Characterized as the "central part of the voiced area of a syllable rhyme (vowel and coda, as determined from the phonetic alignment)" (Mertens, 2013a).

characterized by their direction and pitch range: *R* (large rise), *r* (small rise), *F* (large fall), *f* (small fall) and *_* (flat/level). Compound movements are also possible and use a sequence of symbols, *e.g.* *RF* (rise-fall), *_R* (level-rise), *R_* (rise-level).

Unlike the majority of automatic prosody annotators (*e.g.* Wang and Hirschberg (1992); Wightman and Ostendorf (1994); Braunschweiler (2006); Ananthakrishnan and Narayanan (2008); Rosenberg (2010); Jeon and Liu (2012)), it relies on acoustic information only and is independent from phonological theories. In that sense, each syllable is assigned a label, independently from its stressed/unstressed status. The algorithm is speaker-independent and does not require any training data. A preliminary evaluation of the algorithm on French and Dutch in Mertens (2014) indicated that it reached encouraging results, with an F-measure between 0.946 and 0.815 for pitch levels low, mid and high and between 0.708 to 1 for pitch movements. Such high F-measure values were obtained through a manual correction of automatically-annotated labels. The Polytonia output is used as one of the inputs for our prosody annotator. It was therefore applied to the CPROM corpus.

In collaboration with Prof. Mertens, we developed a set of mapping rules to merge all three levels of prosodic information (*i.e.* prominence, pitch height/movement and linguistic annotation) into a single ToPPos-Fr annotation.

Theoretical motivation behind the definition of these rules is detailed in Chapter 5. It mainly relies on a few major characteristics of French prosody that we summarize here:

- French distinguishes between two types of stresses: final and initial stresses.
- Final stresses usually fall on the last syllable of the word and indicate prosodic phrasing boundaries. In that sense, they always coincide with prosodic boundaries (unlike English), except for words ending in [ə]⁸, for which the final stress falls on the penultimate syllable. Initial stresses tend to fall on the initial syllable of the word.
- Final stresses may have a static or dynamic realization (rising or falling) while initial stresses are usually static.
- The penultimate before a final stress may contribute to the final contour (Mertens, 2008).

Based on this, ToPPos-Fr proposes a protocol to annotate accentual contours in French⁹. It basically identifies three structural positions: final stress, penultimate and initial stress. Each event is assigned a specific label, indicating its accentual

⁸This rule does not apply to a few exceptional cases, including *quoique* (when used alone, without subordinate clause), *parce que* and *sur ce* (where *ce* is a demonstrative pronoun).

⁹In that sense, it does not consider specific prosodic contours which are not directly related to stressed syllables, such as the appendix (*i.e.* a sequence of unstressed syllables with special distributional properties displaying a flat pitch contour (see Mertens (2013b))). These specific contours would however be complex to predict from text at synthesis as they require an in-depth syntactic analysis of the text, which is not provided by most automatic NLP tools.

AUTOMATIC ANNOTATION OF AN AUTOSEGMENTAL DESCRIPTION OF FRENCH PROSODY

status along with its pitch level and movement. A final stress is assigned to word-final prominent syllables, provided that they do not contain a [ə] vowel. In that sense, final stresses indicate the right boundary of prosodic groups. If an (optional) [ə] appears after the final stress within the same word it constitutes the last syllable of the intonation group, whereas it is not itself in final stress position. This is in line with the consideration of final [ə] in the annotation provided in Geoffrois (1995). Word-initial prominent syllables which are not themselves in final stress position are labelled as carrying an initial stress. Finally, the syllable preceding a final stress is assigned a penultimate label. Additional rules account for specific cases. For example, if the last syllable of a stress group is followed by a pause (or by a syllable containing a [ə] followed by a pause) and reaches the bottom pitch level, it is labelled as a final stress, regardless of its prominence status.

It should be noted that, unlike ToBI, ToPPos-Fr proposes a single tonal annotation and does not include a break index tier. This can be justified by the fact that, in French, the hierarchy of prosodic grouping may be directly derived from the peak pitch level of the final stresses, as detailed in Mertens (1987, 1990) and Mertens et al. (2001b). In that sense, the number of boundary levels depends upon the number of distinct pitch labels.

The algorithm implementing these rules identifies positions in the prosodic structure from the position of the syllable in the RG and in the word. A description of the algorithm is shown below:

First step**1. If syllable is prominent**

- (a) If final syllable of the word
 - i. If monosyllabic word not at the end of a RG
 - Assign initial stress, written Xi , where X is replaced by the pitch level annotated by Polytonia (see mappings below)
(*e.g.* Il est très ^{H_i} beau ^{B})
 - ii. Else, if monosyllabic word at the end of a RG
 - Assign final stress, written X
(*e.g.* La ligne du tram ^{H})
 - iii. Else (*i.e.* pluri-syllabic words)
 - Assign final stress
(*e.g.* La jeune maman ^{H})
- (b) Else, if penultimate syllable of the word
 - i. If last syllable contains a [ə]
 - Assign final stress
(*e.g.* Le touris ^{H} me)
- (c) Else, if first syllable of the word
 - Assign initial stress
(*e.g.* La ca ^{H_i} thédrale)

2. Else

- (a) If last syllable of the word
 - i. If before pause or noise, and Polytonia B pitch level
 - Assign B final stress (*e.g.* Elle est partie ^{B})
- (b) If penultimate syllable of the word
 - i. If last syllable contains a [ə], penultimate has Polytonia B pitch level and word followed by a pause or noise
 - Assign B final stress (*e.g.* Il l'a repri ^{B} se)

Second step**1. If penultimate syllable of the word**

- (a) If last syllable is assigned final stress
 - Assign penultimate label (*e.g.* C'est un beau bâti ^{H_p} ment ^{B})

AUTOMATIC ANNOTATION OF AN AUTOSEGMENTAL DESCRIPTION OF FRENCH PROSODY

It should be noted that all labels are assigned to prominent syllables, except for:

- Penultimate syllables before final stresses which are known to participate in the final contour, especially for expressive contours (Mertens, 2008). It should be noted that these labels are assigned in a second stage, as their annotation relies on the previous detection of final stresses.
- Final syllables with an extra low tone (here *B*), as they "can be short and weak because their stressed status is already indicated by tone level" (Mertens, 1991).

In our annotation framework, the Polytonia labels, which are relatively narrow, are reinterpreted into fewer, broader, more abstract categories. This implies several reductions. First, complex pitch movements are simplified with simple ones, as complex movements are known to be extremely rare in French (Mertens, 1987). Small pitch movements are also assimilated to flat pitch. Then, a reduction of the resulting labels is carried out based on the mappings detailed in Table 10.1. Motivations behind these reductions mainly relate to the fact that some labels are particularly rare or refer to functionally equivalent phenomenon. The integration of a very large set of labels in synthesis is also inconvenient as it would multiply the number of corresponding models and makes the prediction of the labels at synthesis more complex. The remaining labels in Table 10.1 are combined with the information about position in prosodic structure to give rise to a unique set of symbols.

Table 10.1 - *Mapping of Polytonia labels to our autosegmental annotation.*

Final stress							
H	HR	HF	Hr⇒HR	Hf⇒H	HF,B⇒HF	HR,T⇒T	H,T⇒T
M	MR	MF	Mr⇒H	Mf⇒M	MF,B⇒MF	MR,T⇒T	
L	LR	LF⇒B	Lr⇒LR	Lf⇒B	LF,B⇒B	Lf,B⇒B	
B	BR⇒LR		Br⇒B				
T		TF⇒HF		Tf⇒T			
Initial stress							
H ⇒ Hi	M ⇒ Mi	Mr ⇒ Mi	MR⇒Mi	L⇒Li	T⇒Hi		
Penultimate (unstressed position before final accent)							
H ⇒ Hp	M ⇒ Mp	L ⇒ Lp					

This amounts to a total of 17 ToPPos-Fr labels:

- **Final:** H HR HF M MR MF L LR B T
- **Initial:** Hi Mi Li
- **Penultimate:** Hp Mp Lp
- **Unstressed:** NA

Based on these rules and labels rewriting, a Praat script was developed to automatically map prominence, pitch and linguistic information into a ToPPos-Fr annotation tier. C-PROM was automatically annotated with that protocol. Syllables lacking Polytonia pitch information (due to diverse reasons including speaker overlaps, hesitations and extraction issues), were assigned an *X* label to allow for their discarding at training stage. All syllables with no label were assigned a *NA* (unstressed) label.

A first qualitative analysis of the resulting prosody annotation indicates a readable although relatively rich prosody transcription. Interestingly, the amount of error generated by the reduction of complex pitch movements to simple ones is remarkably low.

10.2.3 Choice for an automatic prominence detector

The aforementioned annotation protocol was developed on the basis of the manual prominence annotation. As previously mentioned, such manual intervention is costly and may be inconsistent if several annotators collaborate on a single corpus. To automate our protocol, it should be replaced by the output of an automatic prominence detection tool.

Several automatic prominence detection methods/tools have been proposed in the literature. They can be divided into three main groups:

- Statistical tools trained on annotated corpora (*e.g.* the method developed by Wightman and Ostendorf (1992), by Ananthakrishnan and Narayanan (2008), IrcamCorpusTool (Veaux et al., 2008), Promise (Christodoulides and Avanzi, 2014));
- Methods relying on rules trained on annotated corpora (*e.g.* the method developed by Silipo and Greenberg (1999), Prosoprom (Goldman et al., 2007), Analor (Avanzi et al., 2008), the method presented in Avanzi et al. (2011), PromGrad (Goldman et al., 2012));
- Methods relying on heuristics only, without the need for an annotated corpus (as proposed by Tamburini (2003)).

While most tools offer a binary classification of speech segments (usually syllables), as prominent or not, others (like PromGrad (Goldman et al., 2012)) propose a gradual prominence annotation. Among these tools, several apply to French (Veaux et al., 2008; Goldman et al., 2012, 2007; Avanzi et al., 2008; Christodoulides and Avanzi, 2014).

Promise (Christodoulides and Avanzi, 2014) is the more recent prominence detection tool for French. It offers a binary classification of the syllables as prominent or not. It was trained on a very large database containing both spontaneous and read speech and relies on conditional random fields (CRF), and random forest classification (RF). Unlike most other existing tools, the use of CRF allows defining the annotation

as a sequence labeling problem: contextual features including previously-detected prominences are considered for the annotation. This tool was shown to achieve high quality annotation, reaching an F-measure higher than 80 %.

As a qualitative evaluation of our ToPPos-Fr annotation on CPROM pointed out high quality results, we set out to select an automatic detector providing good prominence annotation rates compared to the CPROM manual annotation. In that framework, we decided to focus on 4 of the aforementioned available detectors in French, *i.e.* Analor, ProsoProm, PromGrad and Promise. Regrettably, automatic annotation rates of CPROM cannot be evaluated with Prosoprom and Analor as both tools were trained on CPROM, which would bias the evaluation of the annotation. Therefore, our comparison mainly focused on Promise and PromGrad. Here again, both levels of manual prominence annotation of CPROM (*i.e.* p and P) were merged into a single label. b syllables (*i.e.* syllables reaching an infra-low pitch level) were also recoded as prominent. All syllables annotated z (*i.e.* hesitations) were discarded for the evaluation.

For PromGrad, levels of prominence 1, 2, 3 and 4 were considered as "prominent" against "non prominent" for level 0. This was shown to achieve the highest F-measure.

Two Promise models were exploited in the framework of our study:

- We first used the original Promise model as described in Christodoulides and Avanzi (2014). That model was trained on the CPROM-PFC corpus (Avanzi, 2014), which constitutes a subpart of the PFC corpus (Durand et al., 2009). Despite its name, it does not overlap with the CPROM corpus but was also annotated in terms of manual prominences. It is referred to as **Promise-PFC** in the remainder of the chapter.
- Qualitative analysis of the CPROM-PFC and CPROM manual prominence annotations indicated that the CPROM annotation relied on a much more restrictive definition of prominence, compared to CPROM-PFC which contained a higher proportion of syllables annotated as prominent. To provide a better agreement with the degree of prominence annotation of CPROM, a second model was trained. First, a Promise model was trained on CPROM. This model was then exploited to automatically annotate the CPROM-PFC corpus (this should standardize the degree of prominence annotation in both corpora). In a second stage, a new Promise model was trained on the resulting automatic annotation of CPROM-PFC. It is called **Promise-CROSS** in the remainder of this chapter. This model introduces a bias as it was indirectly trained on the test corpus. It should however provide an annotation of prominences closer to the degree of prominence considered in CPROM.

Both Promise models were trained on acoustic features only (to avoid redundancy with the rules implemented in our Praat script). Table 10.2 shows the annotation rates obtained by the three annotators, when compared to the manual CPROM annotation. Promise-CROSS is shown to achieve the best results with high precision

rates. It should however be remembered that these results are partly biased by the indirect training of the model on the test corpus. The models may therefore generalize poorly to other corpora. For that reason, both Promise models are integrated in the expressive speech synthesizer and compared in Section 10.3.

Table 10.2 - *Prominence annotation rates.*

Model	Precision	Recall	F-Measure
Promise-PFC	63.30	83.53	72.02
PromGrad	70.27	80.98	75.25
Promise-CROSS	95.81	84.35	89.72

10.2.4 An alternative option: Training an AuToBI classifier on a French annotated corpus

As detailed in the introduction to this section, a second possible option to offer an automatic autosegmental annotation of French lies in the training of a new AuToBI model on a French annotated corpus. This option was investigated in the framework of our study but achieved poorer preliminary results and was therefore discarded. We provide here a brief summary of that study.

Training a new AuToBI model requires the use of a large corpus in French, annotated with autosegmental prosodic labels, and ideally based on the collaboration of multiple human annotators. To our knowledge, no such corpus is available. While ToBI has recently been adapted to French in Delais-Roussarie et al. (2015), the study only relies on a few manually-annotated examples. An entire corpus was annotated with Mertens' tones in Mertens (1987) but the annotation was carried out by a single annotator and is not readily available in electronic format. An exception may be the LATIS corpus (Lamel et al., 1995) which was manually-annotated in Geoffrois (1995) with Merten's tones. To our knowledge, that annotated corpus is however not made available.

Instead, we investigated the possibility to exploit an automatic annotation of a large multi-speaker corpus. To provide a larger corpus than that used in the previous section, we complemented CPROM with the Rhapsodie corpus (Lacheret et al., 2014a,b), which also provides a manual annotation of prominences (as "S"trong or "W"eak) as well as an annotation in phonemes, syllables and words. The corpus includes read and spontaneous speech, produced by a large number of speakers. Two hours of the corpus were selected for our experiment. Polytonia was further applied to the corpus and then exploited to compute an automatic autosegmental annotation, in a similar fashion to the protocol described above¹⁰. The annotation

¹⁰Some differences however existed compared to that protocol. The Polytonia annotation, in particular, was converted into a transcription more similar to ToBI. For conciseness, we do not detail that protocol here as the experiment achieved poor results. The results may however be influenced

was based on strong prominences only.

In collaboration with Prof. Andrew Rosenberg, we trained a new AuTOBI model including some adaptations of the training procedure to better fit our French annotation system. Such changes include the consideration of a larger acoustic context, better processing of penultimate syllables and the addition of prominence features in the detection of boundary tones.

Preliminary results, for final stresses only reach a F-measure of about 28 %¹¹ compared to our automatic annotation. A possible cause is the unbalanced nature of the corpus in which unstressed syllables are largely predominant. The acoustic features used by the annotator should also be adapted to better fit the French prosodic realization. A first qualitative analysis of the corpus also indicates that the manual prominence annotation of the Rhapsodie corpus may not be very reliable. This should be investigated in further studies.

10.3 Integration of the local annotation in speech synthesis

10.3.1 Description of the corpus

Audiobooks are considered for speech synthesis training in the framework of this study. The exploited corpus is a 8-hour corpus which corresponds to the reading of the first volume of "À la recherche du temps perdu"(Proust, 1913) by the French author Marcel Proust (Obin, 2011). The text was read by a professional actor, André Dussolier, with a rather expressive voice. The text was "interpreted by the actor according to his understanding of the narrative structure". The corpus displays high prosodic and syntactic complexity.

Each audio file is associated with its textual transcription and its aligned phonetic transcription. The alignment was realized with IrcamAlign (Lanchantin et al., 2008) which makes use of French models learned on a multi-speaker database (BREF80). Regarding the phonetic transcription, the automatic phonetization was automatically enriched with phonetic variants, the best phonetic sequence being selected at alignment time. The models proposed by IrcamAlign were adapted to André Dussolier on the first hundred files of the corpus, which were manually aligned. The alignment is therefore automatic, but of good quality. The corpus was previously used for unit-selection synthesis and offered good-quality results (Obin, 2011).

For the purposes of our study, the corpus was cleaned¹² and processed by eLite-HTS to provide an annotation in syllables, POS, and RG (based on a Chinks

by the conversion we proposed.

¹¹It should however be noted that the F-measure hides some interesting aspects of the performance. Specifically, the recall is fairly low (about 15-18 %), while the precision is quite high, *i.e.* 70-80 %. This means that few boundaries are detected but, when they are, they tend to coincide with real boundaries. This may be due to high detection rates when a final syllable is followed by a pause. This gap between precision and recall remains surprising and should be further investigated.

¹²The text transcription was first synchronized with the phoneme transcription. Most problems occurred in the surroundings of punctuation marks, hyphens and apostrophes. Ciphers written in

& Chunks algorithm (Lieberman and Church, 1992)). The ToPPos-Fr annotation protocol was then applied to the corpus, using two sources of prominence information, as provided by the Promise-PFC and Promise-CROSS models.

A quantitative analysis of the two resulting annotations indicated the number of occurrences of each labels and is shown in Table 10.3. As expected, the annotation based on Promise-CROSS contains many more NA syllables, as the prominence detector is known to be more restrictive. This reduction should not be considered as an issue, as it may allow for an annotation of prominences only when they are highly acoustically marked. Interestingly, initial prominences are rarely annotated with that model. Noteworthy is also the high amount of X labels, mainly due to detection issues of Polytonia. As detailed later, most detection issues are due to the whispered and creaky segments of the corpus.

Table 10.3 - *Number of occurrences of each ToPPos-Fr annotation label in our corpus, with two prominence detectors (Promise-PFC and Promise-CROSS).*

Label	ToPPos-Fr-PFC	ToPPos-Fr-CROSS
NA	60778	106267
X	20920	8550
H	6175	1947
L	5549	1290
M	4836	876
HF	1089	2253
B	1191	731
LR	868	609
T	862	269
MF	585	256
HR	323	252
MR	246	175
Lp	12539	4338
Mp	3358	1136
Hp	2458	1051
Li	3216	70
Hi	1876	119
Mi	1408	56

10.3.2 Method

In order to assess the validity of our automatic prosody annotator, three HMM-based speech synthesizers were built, relying on the implementation of the publicly available HTS toolkit (version 2.1) (Zen et al., 2007a). The speech waveforms were sampled at 16 kHz. Model configuration, filter parametrization and excitation modeling were identical to those presented in the evaluation protocol of Chapter 6. Three

Roman letters also had to be modified, along with some abbreviations.

models are compared: the baseline model (*Baseline*) integrates no explicit prosody annotation. The two other models, *ToPPos-Fr-CROSS* and *ToPPos-Fr-PFC* integrate the ToPPos-Fr automatic annotation, resulting from the Promise-CROSS and Promise-PFC models, respectively. This local prosody annotation is integrated as contextual information in the tree clustering process, similarly to the integration of local prosodic labels in Chapter 6.

For each synthesizer, 90 % of the corresponding database was used for the training (called the *training set*), leaving around 10 % for the synthesis (called the *synthesis set*), randomly shuffled. ToPPos-Fr annotation of the test sentences relied on their automatic annotation from the speech files. Test sentences were automatically selected from the synthesis set, as complying with the following criteria:

- Containing minimum 3 prosodic labels other than *NA* and *X*
- Containing maximum 2 *X* labels
- With a total duration between 3 and 8 seconds

These criteria were driven by the fact that enough prosodic differences should appear between the baseline and the two other models and that short sentences are easier to compare. This resulted in the extraction of 33 test sentences. As *X* labels allow for a discarding of the respective syllables at training, they were replaced by an *NA* label in the test sentences.

Two perception tests were designed.

The first test consists of 15 pairs of sentences, 5 for each comparison, randomly selected from the 33 test sentences. The listeners were asked to indicate, for each pair, which version seemed more natural in the framework of audiobook reading. The scale ranged from -3 (Synthesis A is much more natural than synthesis B) to 3 (synthesis A is much less natural than synthesis B). If both versions sounded equally natural (or if they both sounded mediocre) they could rate them as "equivalent".

In Mertens (2004b) it is also pointed out that evaluations of annotation systems may be best carried out through comparisons between the synthetic voice and the original voice. This is assessed through our second experiment, which contains 9 pairwise comparisons. For each sentence, the listeners are first required to listen to the original version of the sentence. In a second stage, they are asked to listen to two synthesized versions and should indicate which version is closer to the original in terms of intonation¹³. The scale ranged from -3 (Synthesis A is much closer to the original version than synthesis B) to 3 (synthesis B is much closer to the original version than synthesis A). This type of experiment is similar to that carried out in Eyben et al. (2012) to evaluate expressive speech synthesis.

Fifteen native French speakers participated in the evaluation, mainly naive listeners.

¹³The term *intonation* is used here as it can be seen as fairly generic to non prosody experts.

10.3.3 Evaluation and discussion

Results of the first test are displayed in Figure 10.1. An in-depth analysis of the results highlighted that the preference rates are highly dependent on the sentences. The percentages of preference are therefore computed at the sentence level. It shows that the ToPPos-Fr-CROSS annotation is clearly preferred compared to the baseline (58.9 % against 30.56 %). Interestingly, however, the baseline tends to be preferred to ToPPos-Fr-PFC. Statistical significance was computed through unilateral signed ranksum tests comparing the average percentage of preferences on the 33 sentences. They indicated that ToPPos-Fr-CROSS is significantly preferred to the baseline ($p = 0.03$) but that the preference difference between the baseline and ToPPos-Fr-PFC is not significant ($p = 0.16$). The difference between both ToPPos-Fr annotations is also not significant ($p = 0.24$).

The results indicate that ToPPos-Fr-CROSS achieves the highest rates in terms of naturalness of the synthesized voice. It should be noted that this annotation relies on a stricter approach to prominence detection, which results in a higher proportion of syllables annotated as unstressed, compared to ToPPos-Fr-PFC. In that sense, it could be assumed that syllables annotated with a non *NA* label are more marked in terms of accentuation, which improves their training, thereby providing a more adequate generation at synthesis.

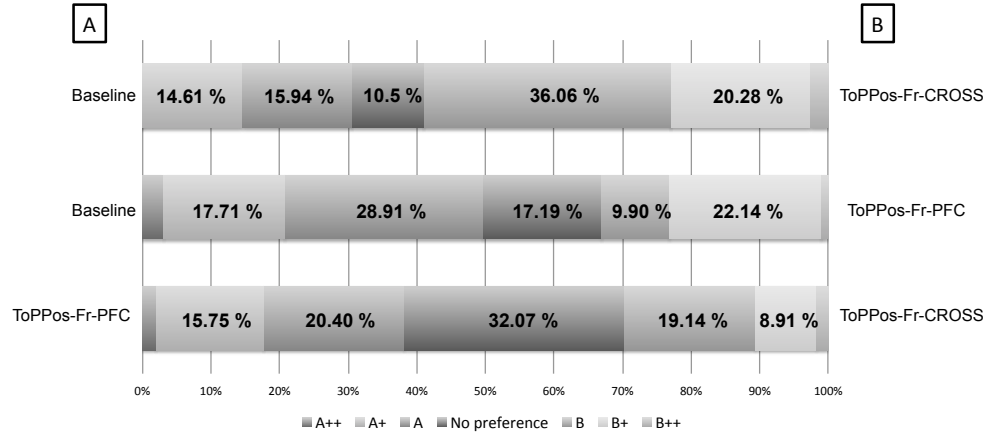


Figure 10.1 - Preference rates, at the sentence level, between the syntheses enriched by automatic prosody information (ToPPos-Fr-CROSS and ToPPos-Fr-PFC) and the baseline (including no explicit prosody annotation), in terms of naturalness of the expressivity of the synthesized voice. "+" refers to stronger preference scores.

Figure 10.2 indicates the results of the second test, in which each of the three synthesizers is compared to the original version of the sentence. It can be observed that both ToPPos-Fr synthesizers are pointed out as producing a speech intonation closer to the original voice, compared to the baseline (with 57.99 % against 22.22 %

AUTOMATIC ANNOTATION OF AN AUTOSEGMENTAL DESCRIPTION OF FRENCH PROSODY

for the comparison ToPPos-Fr-CROSS vs. Baseline and 54.51 % against 38.89 % for the comparison ToPPos-Fr-PFC vs. Baseline). Here again, unilateral signed ranksum tests comparing the average percentage of preferences on the test sentences were computed. They indicate that ToPPos-Fr-CROSS is significantly preferred over the baseline ($p = 0.01$). Differences between ToPPos-Fr-PFC and the baseline and between both ToPPos-Fr synthesizers are however shown to be non-significant (with $p = 0.20$ and $p = 0.27$, respectively).

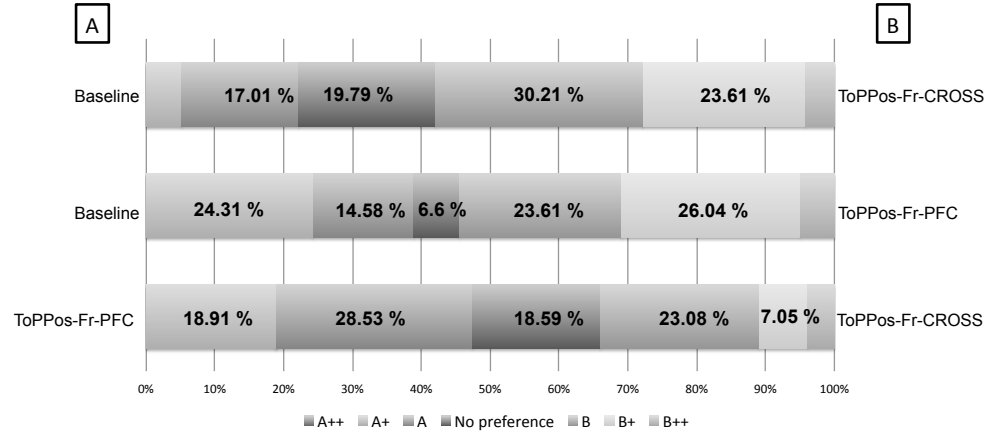


Figure 10.2 - Preference rates, at the sentence level, between the syntheses enriched by automatic prosody information (ToPPos-Fr-CROSS and ToPPos-Fr-PFC) and the baseline (including no explicit prosody annotation), in terms of resemblance to the original sentences. "+" refers to stronger preference scores.

A qualitative evaluation investigating the divergences between the original voice and the ToPPos-Fr synthesized voices points at various issues :

- The use of our corpus is problematic as:
 1. it contains a high proportion of creaky and whispered segments, which prevents an easy F0 extraction. This results in a large number of unannotated syllables, which are assigned a *X* label (16 % and 7 % of the syllables in ToPPos-Fr-PFC and ToPPos-Fr-CROSS, respectively). That may have degraded the training of the distinct prosodic models.
 2. the characteristics of the expressive reading of the text include very complex intonational patterns which may sound rather unnatural, especially when generated by a synthetic voice.

This indicates that experiments on other corpora may provide better results.

- Sentences with *X* labels may be badly synthesized as the *X* is replaced by a *NA* for the test samples. Ideally, tests should focus on comparisons with no *X* in the prosodic transcription but this would highly reduce the set of possible test

sentences, seeing the high amount of such labels, as mentioned in the previous point. Another possible solution would be to manually annotate the respective syllables. A typical example of resulting issues is that *NA* labels are scarcely ever followed by a silence in the ToPPos-Fr model (as silences tend to be preceded by a final stress). If such a label is predicted at synthesis (to replace a *X* label), a model corresponding to that prosodic context may not be found in the database, which could explain a preference for the baseline in that case.

- The realization of *HF* stresses sometimes leads to a saturation of the signal, which hinders its quality.

These issues should be investigated in further studies. In-depth analysis of the test sentences could also indicate whether some labels are better generated than others.

10.4 Reducing the set of labels

In this section, we investigate the possible reduction of the set of labels proposed in ToPPos-Fr, which will be referred to as ToPPos-Fr-Reduced. The annotation used in the previous sections contains 17 distinct labels (not considering the *X* label). As mentioned in Chapter 6, the exploitation of such a large set of tags poses a major issue for its integration in text synthesis, namely the fact that these labels may not be easily predicted from text.

Ideally, each tag should be (i) associated with a specific function that (ii) should be easily predictable from text. We propose in this section to reduce the set of labels according to the two aforementioned constraints. In that sense, the resulting reduced annotation can be seen as more phonological and less acoustic than the original ToPPos-Fr annotation. Obviously, such a reduction implies the merging of acoustically different syllables, which may degrade the resulting synthesis. The integration of ToPPos-Fr-Reduced in a speech synthesizer is therefore assessed through a second set of perception tests.

It should be noted that the current research bears some similarity to the study presented in Chapter 7 in which we investigated the use of a single emphatic label against that of various labels associated with distinct acoustic realizations. One should recall that it was observed that the use of a single label did not impair the resulting synthesis.

10.4.1 Annotation protocol

The reduction of the set of labels relies on a functional approach. In an ideal framework, it should be based on a detailed analysis of the occurrences of each label in our corpus aiming at defining their major prosodic functions. Such analyses are time-consuming, hence costly. We can however build on existing studies carried out by Prof. Mertens, which were based on a similar annotation protocol (Mertens, 2008).

Final stresses are known to delimit prosodic phrases in French. Mertens (1987, 1990, 2006) indicates that the pitch height and movement of the final stresses further determine their respective degree of boundary. In Mertens (2008) four boundary levels are distinguished (see Table 10.4). To each level correspond several labels in his system. Based on their similarities with ToPPos-Fr labels, we propose to rely on that classification to rewrite our set of final stress labels into four main tags.

The motivation behind this reduction is that it should facilitate the prediction of the labels from text (*i.e.* with punctuation and pauses) and its automatic analysis, relying on a simple chunks & chunks algorithm or more in-depth syntactic analyses as proposed in Mertens et al. (2001b). Ideally, such predictors should integrate rules defining the specific contexts in which a prosodic boundary is mandatory (see Mertens (2008)). It should also be noted that superior continuations (Sup) could be predicted at the end of certain types of interrogative sentences (*e.g.* in interrogations without interrogative word, with or without inversion, such as *Il travaille déjà* ^{Sup} ?¹⁴ and *Travaille-t-il* ^{Sup} ?¹⁵ (Mertens, 2004b).

It is noteworthy that some final stresses do not only fulfill a delimitative function but may also additionally inform the listener of the way the speaker presents the information to the listeners. This explains the inclusion of an additional label in our reduced annotation set: *Focus*, characterized by a high pitch fall. It should however be distinguished from the *F* label of the prosody annotation protocol presented in the second part of the thesis, which indicated emphatic stresses and in that sense mainly related to initial stresses. It should be pointed out that the prediction of the *Focus* label from text is rather complex. It could however be predicted from punctuation (*i.e.* before an exclamation mark at the end of a sentence) or based on a detailed syntactic analysis of the sentence, which would identify cleft constructions¹⁶, as in *C'est précisément là*^{Focus} *que des problèmes d'emploi surgissent*¹⁷ (Mertens, 2004b).

Initial stresses are usually observed with a high pitch level (H or M) which indicates the beginning of an informational topic in the speech flow (Mertens, 2004b). Low initial stresses have been said not to appear in French (Delais-Roussarie et al., 2015) but were found in rare cases in Mertens (1987). Initial stresses can be seen

¹⁴He already works ? (our translation).

¹⁵Does he work ? (our translation).

¹⁶Prince (1978) proposes to distinguish between two types of cleft constructions, summarized in Mertens (2012). In the first, called "stressed focus IT-cleft", the left part is strongly stressed and presents new information, usually with a contrastive value, while the right part contains known or old information. The second type of cleft constructions, called "informative presupposition IT-cleft" (Prince, 1978), is much more frequent in corpora of spontaneous (unprepared) speech. The left part contains a brief anaphoric focus, while the second part is "normally" accentuated, *i.e.* following the general rules for syntactic structures corresponding to a single dependency network (rather than complex constructions, such as cleft, pseudo-cleft, left and right dislocation, and so on). In Mertens (2008) it is indicated that cleft constructions of Type 1 display a focus on their first part and an appendix contour on the second. To allow for that prediction of focus, one should however rely on an automatic distinction between these two types of cleft constructions. This could rely on the presence of an anaphoric word in the first part of the construction but it is highlighted in Mertens (2012) that this is not an essential characteristic of the second type of cleft constructions. It should also consider the particular case of interrogative cleft construction in which a major continuation is observed in both parts of the construction.

¹⁷It's precisely there, that job issues arise (our translation).

Table 10.4 - For each function assigned to final stresses, the tones in Mertens' annotation scheme (2008), the corresponding labels in ToPPos-Fr, and our reduced label.

Function	Mertens (2008) tone	ToPPos-Fr label	ToPPos-Fr-Reduced label
Boundaries			
Final boundary	B-	B	Fin
Superior continuation	H/H	T, MR, HR	Sup
Major continuation	HH	H, LR	Maj
Minor continuation	/BB, BB, \BB	M, L	Min
Other			
Focus	HB, HB-	HF, MF	Focus

as fulfilling a single function: highlighting a specific word or constituent. Based on these considerations, we propose to rewrite the ToPPos-Fr initial low stress (*Li*) as *NA* and to merge *Hi* and *Mi* into a single *AI* label, similarly to their representation in the prosody annotation protocol presented in Chapter 6. As detailed in Chapter 5, a wide range of research has focused on the prediction of emphatic stresses from text and could be used for that purpose. In Chapter 7 we have also pointed out that the use of a single emphatic label highly facilitated its prediction from text.

Finally, a last specific prosodic realization which ought to be considered is the **high penultimate**. It can be distinguished from the aforementioned phenomena in the sense that it does not apply to a stressed syllable. The high penultimate is however directly related to accentual phenomena as it precedes a final stressed syllable, thereby participating in the final contour. It is known to fulfill an expressive function. Because the function of low and medium-high penultimate is respectively unmarked and weakly marked, we propose to rewrite both *Lp* and *Mp* as *NA* in our reduced annotation protocol.

However, rewriting *Hp* as *NA* would probably bring noise in the *NA* models. The *Hp* label is therefore kept in our ToPPos-Fr-Reduced annotation. It should however be noted that its prediction from text may be complex. In Mertens (2008), it is said to be found in contours of type *Hp B* and *Hp H*. The first creates "an assertion/evidence effect" while the second brings a "connotation of shared knowledge". As these functions may only be predicted through an in-depth semantic analysis of the sentence, existing front-ends would not be able to predict them¹⁸. Their separate processing at training remains however useful.

This amounts to a total of 9 labels: *NA*, *X*, *Fin*, *Sup*, *Maj*, *Min*, *Focus*, *AI* and *Hp*. This annotation will be referred to as ToPPos-Fr-Reduced in the remainder of this chapter.

¹⁸Such a prediction could rely on indices including punctuation marks, like exclamation marks and suspension points.

10.4.2 Perceptive evaluation

Our corpus was automatically relabeled with the ToPPos-Fr-Reduced annotation set. As perception tests in Section 10.3 indicated a higher performance of the ToPPos-Fr-CROSS annotation, the automatic relabelling relied on that initial annotation. A new synthesizer was then built, integrating the reduced annotation. The methodological protocol is similar to that presented in Section 10.3. Here again, two tests are carried out to assess (i) the naturalness of the synthesized speech and (ii) its similarity with the original version of the sentence. 20 French speakers, mainly naive listeners, participated in the evaluation.

Figure 10.3 shows the results for the first perception test. Here again, percentages are computed at the sentence level. As experiments in the previous sections indicated that test sentences containing *X* labels tended to bias the results, they were discarded from this evaluation. Results still indicate a preference of the ToPPos-Fr-CROSS synthesis over the baseline. We also observe that ToPPos-Fr-Reduced slightly outperforms the baseline, but is much less preferred than ToPPos-Fr-CROSS, which was expected as the acoustic granularity of the labels has been reduced.

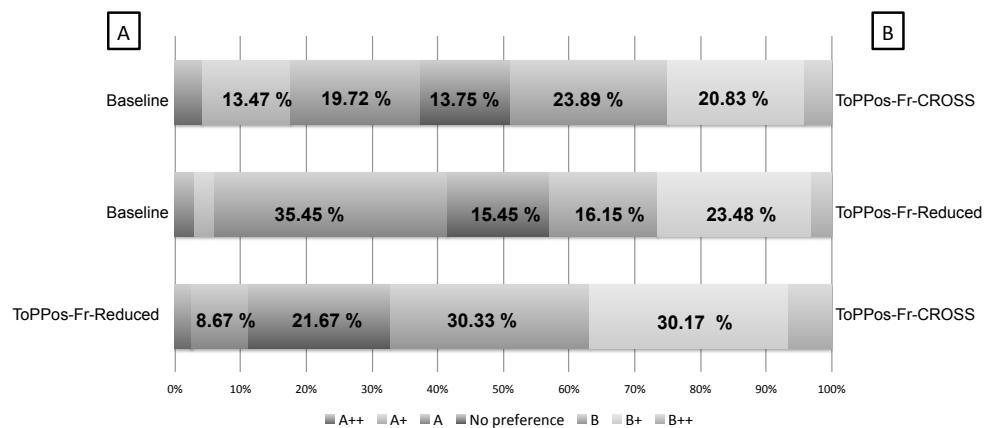


Figure 10.3 - *Preference rates, at the sentence level, between the syntheses enriched by automatic prosody information (ToPPos-Fr and ToPPos-Fr-Reduced) and the baseline (including no explicit prosody annotation), in terms of naturalness of the expressivity of the synthesized voice. "+" refers to stronger preference scores.*

Similar observations can be made from the results of the second part of the test (see Figure 10.4). It should however be noted that none of these preferences were shown to be significant. This difference with the results of the tests carried out in the previous section could be partly explained by the fact that the size of the sample is reduced as we rely on test sentences without *X* only. This divides by three the number of sentences which accounts for a total of 12, which is too small to reliably compute statistical significance.

It should also be noted that, beside the high variance across the different test sentences, the preference rates for a specific sentence are not always consistent. This can be explained by the fact that test sentences contain a minimum of 3 ToPPos-Fr labels (other than *NA*), which should account for prosody modifications at these points. It happens that modifications brought by one of the labels improve the synthesized speech while another degrades it. The evaluation task may therefore not be trivial. This should be further analyzed to assess the quality of the models corresponding to each ToPPos-Fr label.

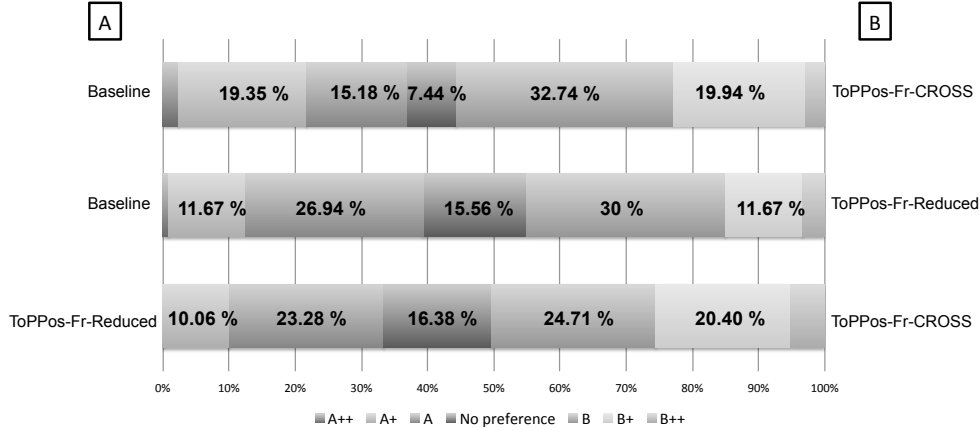


Figure 10.4 - Preference rates, at the sentence level, between the syntheses enriched by automatic prosody information (*ToPPos-Fr* and *ToPPos-Fr-Reduced*) and the baseline (including no explicit prosody annotation), in terms of resemblance to the original sentences. "+" refers to stronger preference scores.

10.5 Conclusion

In this chapter, we have proposed an approach to automatically produce an autosegmental prosody labeling of French corpora (called *ToPPos-Fr*), in view of speech synthesis. To our knowledge, no such method has been developed for French, despite the large number of existing studies exploring that topic in English. We have indicated that such annotation can rely on the combination of the output of existing systems, *i.e.* prominence detectors, tonal annotators and NLPs providing a phonetization, syllabification and rhythmic group segmentation of the corpus. Based on phonological descriptions of the French language, we have developed a set of rules to map these inputs into a single prosody annotation, which was shown to offer a high-quality representation of the prosody of the corresponding speech.

In a second stage, we have investigated the integration of *ToPPos-Fr* in the HMM-based speech synthesis of audiobooks. Perception experiments indicated that

AUTOMATIC ANNOTATION OF AN AUTOSEGMENTAL DESCRIPTION OF FRENCH PROSODY

this integration results in significant improvement of the naturalness of the synthesized voice, which provides intonation patterns closer to the original realizations of the test sentences.

The reduction of the set of labels was investigated in a third stage, aiming at facilitating their prediction from text. In that respect, we designed a set of labels corresponding to distinct functions, motivated by existing analyses of French tonal aspects. The integration of the labels in synthesis showed slight improvement compared to the baseline but no statistical significance was found in that respect.

While we firmly believe that ToPPos-Fr provides an insightful description of the prosody of the corpus, perception tests of the resulting synthesis were not as satisfactory as expected. A closer look at the corpus exploited for synthesis indicated major drawbacks of that testing material:

- Its characteristic voice quality features, which included large segments of creaky and breathy voice, prevent from high quality F0 extraction by the tonal annotator.
- The complex prosody patterns realized by the actor may sound unnatural when produced by a synthetic voice. This can be witnessed in the results of the perception tests, which indicate higher preference rates in terms of resemblance to the original voice than in terms of naturalness of the synthesized voice.
- The rather solemn tone of the speaker does not induce highly expressive prosody patterns, which could then be appreciated in the synthesized voice.

Consequently, we hypothesize that tests on other expressive corpora, such as sports commentaries analyzed in the second part of the thesis, may provide better results. This should be investigated in further studies.

Chapter 11

Modeling spontaneous speech: The integration of phonetic variants in speech synthesis

Contents

11.1 Introduction	257
11.2 A multifactorial analysis of phonotactic phenomena in French: Impact of the speaking style	259
11.2.1 Introduction	259
11.2.2 Overview of the factors affecting French pronunciation . .	261
11.2.3 Material	263
11.2.4 Overview of the impact of speaking style on phonotactic variation	264
11.2.5 An in-depth multifactorial analysis of post-obstruent liquid elision	268
11.2.6 Conclusion	272
11.3 Spontaneous speech synthesis in English: Investigating the impact of the phonetization of the training corpus .	273
11.3.1 Introduction	273
11.3.2 Material: A parallel corpus	275
11.3.3 Phonetic analysis of the corpus	276
11.3.4 Integrating pronunciation variation in forced alignment .	278
11.3.5 Synthesis evaluation	282
11.3.6 Conclusion	284

Abstract

This chapter builds upon the observations made in Chapter 9 which indicated that the integration of phonetic modifications at synthesis significantly improved the naturalness of the synthesized speech. The current study focuses on spontaneous speech synthesis since it was shown that spontaneous speech, in particular, was characterized by a high amount of phonetic variation. It summarizes the preliminary results of two ongoing studies that we have carried out in collaboration with other experts in the field.

The first part of the chapter deals with the integration of phonetic variants in the synthesis of spontaneous speech in French. To that aim, it sets out to investigate which factors should be considered when designing phonetic rewrite rules. In our view, the drawback of most existing studies dealing with the modalities of appearance of phonetic variation phenomena in French is that they focus on the impact exerted by a single or only a few variables in a corpus tailored for that task. Conversely, we here provide an in-depth statistical study investigating *simultaneously* the role played by a wide range of linguistic and extralinguistic factors. Preliminary results confirm the observations made in Chapter 9 by highlighting the higher rate of schwa and liquid deletion and of liaison non-realization in spontaneous speech. This first analysis is carried out on a large corpus balanced in terms of origin, age, socio-educational level and sex of the speakers to neutralize the potential effect of these extralinguistic features. In a second stage, a detailed analysis of word-final post-obstruent liquid deletion is performed and underlines the role played by three linguistic factors, namely the nature of the following segment, the number of syllables in the word and the articulation rate. This indicates that such factors should be considered when designing post-phonetic rewrite rules with spontaneous speech synthesis in mind.

The second part of the chapter deals with spontaneous speech synthesis in English. It investigates the widespread consistency assumption according to which errors made in the phonetization of the training corpus should not affect the resulting synthesized voice as long as the same errors are also made at synthesis time. Our hypothesis is that this may hold if only minor differences appear between the automatic phonetization of the corpus and the real pronunciation. In the case of spontaneous speech, however, the discrepancy may be bigger, thereby degrading the training of the models. To explore that question, the current study relies on a parallel corpus containing identical sentences read and pronounced in a spontaneous context by a same speaker. A preliminary linguistic analysis of the corpus confirms observations made in Chapter 9 which indicate that the real pronunciation of the spontaneous sentences highly differs from their canonical phonetization as provided by standard front-ends. To improve that auto-

matic phonetization, a lattice-based forced alignment system allowing for pronunciation variation is implemented. It is shown to achieve improvement in phoneme identity accuracy on the training corpus. Based on this, HMM-based synthesis is trained and evaluated through side-by-side MOS tests assessing the naturalness of the synthesized voice. Results indicate that, while the assumption of consistency holds in the case of read speech, the improvement of the phonetization of the training corpus is shown to significantly improve the synthesis of spontaneous speech, despite breaking the consistency assumption.

This chapter is based upon the following publications:

- Sandrine Brognaux & Mathieu Avanzi, Dialectal Variation of French Schwa in the light of Speaker-related Factors, *Phonetics and Phonology in Europe (PAPE 2015)*, Cambridge, 2015.
- Mathieu Avanzi & Sandrine Brognaux, A Multifactorial Analysis of French Liaison, *Phonetics and Phonology in Europe (PAPE 2015)*, Cambridge, 2015.
- Sandrine Brognaux & Mathieu Avanzi, Sociophonetics of phonotactic phenomena in French, *International Conference on Phonetic Sciences (ICPhS)*, Glasgow, 2015.
- Mathieu Avanzi & Sandrine Brognaux, A Multifactorial Analysis of Word-Final Post Obstruent Liquid Behavior in French, *Journal of Phonetics*, (Under review).
- Rasmus Dall, Sandrine Brognaux, Korin Richmond, Cassia Valentini-Botinhao, Gustav Eje Henter, Julia Hirschberg, Junichi Yamagishi & Simon King, Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis, *International Conference on Audio, Speech and Signal Processing (ICASSP2016)*, (Under review).

A note on collaborative work:

Both parts of this chapter relate to studies carried out in collaboration with other researchers.

The first part of the chapter summarizes studies carried out in collaboration with **Dr. Mathieu Avanzi** (Cambridge University). He provided us with the CPROM-PFC manually-annotated corpus. For the needs of our experiments, I designed scripts to automatically extract all phonetic modification phenomena. This relied on existing techniques as developed in Chapter 9 along with new extraction methods, based on phonetic dictionary lookups and frequency computations. I also developed scripts to automatically annotate the corpus in terms of accentual phrases. This resulted in a table containing all tokens of the corpus, associated with 128 descriptors, regarding their phonetic transcription and that of surrounding words, but also syntactic, lexical and prosodic characteristics. This table provided a rich playground for an in-depth analysis of the phonetic realization of the corpus. **Dr. Avanzi** then carried out the statistical analysis of the results.

The second part of the chapter was realized in collaboration with **Rasmus Dall** and his colleagues from the Center for Speech Technology Research (CSTR) in Edinburgh. **Rasmus Dall** provided us with the parallel corpus. We then both manually annotated it, to create a Gold standard transcription in terms of phonemic realization. All divergences were discussed during several joint sessions. **Cassia Valentini-Botinhao** provided a third manual annotation of the corpus. **Rasmus Dall** and I both listed the main types of phonetic variations we had come across to include them in the aligner. **Korin Richmond** implemented the system for creating lattices with pronunciation variants and gave advice throughout the study, especially regarding which manual rules to implement. **Junichi Yamagishi** created and provided the large female average voice model used for model initialization. **Rasmus Dall** trained the synthesizers and designed the perception tests. In that framework, **Gustav Eje Henter** (CSTR) assisted with the setup of the listening test and provided analysis scripts for the statistical tests.

We would also like to thank **Simon King** (CSTR), **Rob Clark** (CSTR/Google) and **Julia Hirschberg** (Columbia University in the city of New York) who provided us with insightful advice throughout the study.

11.1 Introduction

In Chapter 9, we have observed that including pronunciation variants in the HMM-based synthesis of sports commentaries significantly improved the naturalness of the resulting voice. Through the statistical analysis of a large multi-style corpus, we have pointed out that these variations mainly appeared in spontaneous speech. The current chapter therefore focuses on phonetic modifications in the framework of spontaneous speech synthesis, as we believe that this topic rightly fits into the research on expressive speech synthesis.

Indeed, we should consider that "natural" expressivity, as we aim to produce, typically goes through spontaneous speech. Spontaneous speech, also referred to as conversational speech, may be seen as a privileged channel of expressivity. In fact, Campbell (2005) indicated that the majority of sentences uttered in everyday conversational communication are not pronounced for their propositional content but rather express interpersonal relationships and affective information. In that sense, spontaneous speech is rarely completely neutral. We therefore believe that the integration of phonetic variants in expressive speech synthesis should first be tackled through spontaneous speech synthesis. Preliminary work on that topic is provided in this chapter.

Spontaneous speech synthesis has been the topic of a wide range of studies in the last decade (Sundaram and Narayanan, 2003; Bennett and Black, 2005; Carlson et al., 2006; Prahallad et al., 2006; Werner and Hoffmann, 2006; Campbell, 2007; Lee et al., 2010; Andersson et al., 2010b,a; Adell et al., 2010; Koriyama et al., 2011; Andersson et al., 2012; Dall et al., 2014). The interest has in fact shifted from intelligibility, which has reached a high level in most current speech synthesizers, to naturalness which is often pointed out as lacking in existing systems. The general assumption according to which spontaneous speech would be perceived as more natural than read speech is based on the fact that it is the style to which people are most used to, as nearly all natural communication situations imply spontaneous speech (Laan, 1997). This hypothesis was only tested recently in Dall et al. (2014). They pointed out that sentences pronounced in a spontaneous context were assessed as significantly more natural than the exactly same sentences read by the same speakers.

The interest has therefore shifted from "reading" synthesizers to "talking" synthesizers (Adell et al., 2010). The commercial applications of spontaneous speech synthesizers are numerous and range from speech-to-speech translation (which should reproduce the speaking style) to conversational agents and believable video-game characters.

A large number of studies focusing on that topic (*e.g.* Sundaram and Narayanan (2003); Adell et al. (2010); Andersson et al. (2010b, 2012); Carlson et al. (2006); Campbell (2007)) have tackled the issue of producing non verbal speech sounds (*e.g.* laughs, grunts, hesitations filled pauses) which have been referred to as "wrappers" and "fillers" in Campbell (2007). Andersson et al. (2010b) compared speech generated by two HMM-based synthesizers, trained on spontaneous speech and read speech respectively. They found that the spontaneous synthesizer was rated as significantly

more natural. The introduction of fillers was however not shown to improve naturalness in their study, but this may be due to the resulting lack of naturalness of the propositional content of the test sentences. Similar findings were highlighted in Adell et al. (2010) who found no significant improvement in naturalness by integrating filled pauses in unit selection synthesis.

Other studies, such as Koriyama et al. (2011), focused on generating the prosodic variability of spontaneous speech, notably through the integration of more detailed contextual features indicating phenomena such as phone prolongation, utterance style (*e.g.* laughing and whisper) and disfluencies. Perception tests of the resulting HMM-based synthesis in Japanese indicated that it was perceived as significantly more natural.

Surprisingly few studies have dealt with the integration of pronunciation variants in spontaneous speech synthesis, unlike the high attention paid to that aspect in speech recognition (*e.g.* Ostendorf et al. (1996); Riley et al. (1999); Wester (2003) for a review, see Strik and Cucchiari (1999); Jurafsky and Martin (2009)). A few exceptions to that rule have been mentioned in Chapter 9 and are summarized below.

Both Prahallad et al. (2006) and Lee et al. (2010) proposed to modify the phonetic models to better fit pronunciation variation in HMM-based conversational speech synthesis. The first relies on the assumption that a phone is never totally deleted or substituted but is only partially modified. In that sense, they propose to modify the usual linear model configuration to allow for different transitions and skips between the states of each phonetic model. The models were however not trained on spontaneous speech. The performance was assessed in terms of phone recognition but was not evaluated in the resulting synthesis. Lee et al. (2010) mainly focus on phoneme substitutions (*e.g.* nasalisation, assimilation, lengthening). The modifications between read and spontaneous speech are modeled through transformations of the phonemic models. In these two studies, changes are made in the Digital Signal Processor (DSP) only, and no modification is brought to the front-end component, which still produces a standard canonical speech phonetization.

Conversely, Bennett and Black (2005) propose to train a CART model to modify the phonetic transcription of a few frequent grammatical words like "for", "to" and "the". The impact on the resulting synthesis is however not assessed. Werner and Hoffmann (2006) also set out to integrate phonetic variants, based on a large phonetic dictionary. The choice is dictated by both word probability and sequential constraints. The resulting synthesis is shown to be perceived as more natural but a degradation in the intelligibility level is observed. Finally, Jande (2003a,b) proposed a set of phonetic rewrite rules to produce reduced forms in Swedish according to speech rate. Through perception tests of the resulting speech synthesis, he shows that reduced forms are preferred at high speaking rates.

This chapter builds upon the results obtained in Chapter 9 which indicated that:

1. Spontaneous speech is characterized by a high amount of pronunciation variants.
2. The integration of these variants in synthesized speech (trained on a spontaneous corpus manually checked in terms of phonetic pronunciation) significantly

improves the naturalness of the voice.

These observations drove us to investigate two major topics in the current chapter:

1. To integrate such phonetic modifications in the synthesized voice, phonetic modification rules should be learned. For this, the impact played by linguistic and extralinguistic factors should be assessed to determine which contextual features ought to be considered when predicting phonetic variants.
2. In perception tests of Chapter 9, the best results were achieved with manually-checked phonetic transcription for both training and synthesis. It is, however, unclear whether this was due to the consistency between phonetization at both stages or simply due to a more natural pronunciation during synthesis. In other words, our study did not investigate the role played by the correct phonetization of the training corpus.

Based on this, the chapter is divided in two main sections.

Section 11.2 sets out to address the first topic by providing an analysis of a large corpus in French, containing both spontaneous and read speech. This section was realized in collaboration with Dr. Avanzi. The impact of the speaking style (here spontaneous vs. read speech) is first assessed to investigate whether the observations made in Chapter 9 may be generalized. We additionally provide an in-depth analysis of liquid elision which was shown to greatly impact the naturalness of the synthesized speech in our previous tests. This is done by considering simultaneously a wide range of linguistic and extralinguistic factors.

Section 11.3 deals with the second topic by investigating the impact exerted by the quality of the phonetic transcription of the training corpus in the framework of spontaneous speech synthesis in English. In that sense, it questions the widespread consistency assumption which states that an imprecise phonetization of the training corpus should not impact synthesis, as long as identical phonetization errors are made at the synthesis stage. This assumption is put into question through perceptive evaluations of the resulting synthesized speech. This section was realized in collaboration with experts from the Center for Speech Technology Research (CSTR) in Edinburgh. It relies on a parallel corpus containing identical sentences in read and spontaneous speech.

11.2 A multifactorial analysis of phonotactic phenomena in French: Impact of the speaking style

11.2.1 Introduction

Several studies have proposed phonetic rewrite rules to modify the canonical phonetic transcription provided by existing synthesis front-ends. Such studies have mainly focused on the generation of speaker's specific variants (Miller et al., 1997; Miller, 1998b,a), variations due to speaking rate (Jande, 2003a,b) or to speech style, such as

Van Bael et al. (2004) who model style-specific phonetic variations in read speech, public lectures and telephone dialogues and Jande (2008) who include phonetic variants with respect to the degree of formality and the degree of spontaneity. These modifications have been implemented through neural networks (Miller et al., 1997; Miller, 1998b,a), decision trees (Jande, 2008; Bennett and Black, 2005), trained rewrite-rules (Van Bael et al., 2004) or hand-driven rules (Jande, 2003a,b). Whatever the technique, the prediction of phonetic modifications requires relying on linguistic and extra-linguistic contextual features. Miller considers a set of features including the distance to the word, phrase, clause and sentence boundaries. She also relies on a contextual window of nine phones. In Bennett and Black (2005), contextual parts of speech, identity of the word and stress information are also used. Jande (2003a,b) also considers word stress and accent features, while the prediction in Van Bael et al. (2004) is only based on the consideration of a five-phone contextual window. Conversely, Jande (2008) benefits from a very rich annotation of the corpus, at various linguistic levels (*i.e.* discourse, utterance, phrase, word, syllable and phoneme with their corresponding characteristics).

The choice for the adequate set of contextual features is usually dictated by in-depth phonetic analyses of the language. Hand-driven rules in Jande (2003a,b), for instance, are based on earlier work on reduction rules for Swedish. In French, as detailed in Chapter 9, a large number of linguistic and extralinguistic factors have been shown to influence pronunciation. In our opinion, the drawback of existing studies dealing with phonotactic phenomena variation in French is that they rarely consider all these variables simultaneously. This criticism can also be addressed to the analysis we have proposed in Chapter 9 in which we focused on three situational traits but neglected the impact exerted by other linguistic and extralinguistic factors. Besides, most studies usually rely on manually-annotated low-size corpora, tailored for a specific analysis, which prevents from analyzing possible interactions between the various factors. This precludes the possibility of grasping whether these phenomena are influenced by the same factors in the system of a given speaker.

In this context, this section aims at shedding new light on phonotactic phenomena in French. It presents the preliminary results of ongoing studies we have been carrying out on that topic. We first set out to test simultaneously the role of 4 sociolinguistic factors (speaking style, age, socioeconomic status and origin of the speaker) on the behavior of various phonotactic phenomena. This should indicate whether the observations made in Chapter 9 may be generalized. As liquid deletion was shown to play a crucial role in Chapter 9, a more in-depth analysis of that phenomenon is also carried out, considering a wide range of linguistic and extra-linguistic contextual factors. It has the advantage of relying on a very large corpus of more than 13 hours, balanced in terms of speaking style, region of origin, age and sex of the speakers. Such analysis should provide greater insight into the contextual features that ought to be considered when modeling pronunciation modifications in French, especially in the case of spontaneous speech synthesis.

This section is organized as follows. Subsection 11.2.2 presents an overview of the

factors that have been shown to impact phonetic realization in French. In Subsection 11.2.3 we present the large multi-speaker multi-style corpus that will be used in the framework of our study. Subsection 11.2.4 presents the preliminary results of a study of the impact of speaking style (*i.e.* spontaneous vs. read speech) on the phonetic variation. Subsection 11.2.5 provides a more in-depth analysis of Word-Final Post-Obstruent Liquid (WFPOL) deletion in the light of linguistic and extralinguistic factors¹. Finally, Subsection 11.2.6 concludes the section.

11.2.2 Overview of the factors affecting French pronunciation

In line with Chapter 9, the current study deals with 5 phonotactic phenomena in French:

- Schwa deletion in monosyllabic grammatical words, such as in *j(e) pense*
- Schwa deletion at the initial syllable of polysyllabic words, such as in *il lui a d(e)mandé*
- Optional liaison, *i.e.* the phenomenon whereby a latent final consonant in a word (Word-1) may or may not be pronounced as the onset of a following vowel-initial word (Word-2), such as in *ils vont-(t)-au cinéma*²
- The deletion of /l/ and /R/ in Word-Final Post-Obstruent Liquid (WFPOL), such as in *pénib(le)*
- The deletion of the final /l/ in the singular personal clitic subject pronoun ‘il’, such as in *i(l) va*

The modalities of appearance of these phenomena have been challenging issues for decades in the French phonological field, because the reasons that motivate their behavior (realization or deletion) are multiple and related to different levels of the linguistic analysis, as well as to extra-linguistic factors.

Linguistic factors relate to various levels of the linguistic analysis, such as segmental, phonotactic, lexical, morphological, grammatical and prosodic variables. Recently, Bürki et al. (2011) focused on the role played by such predictors in the presence vs. absence of the schwa vowel in a French corpus of radio-broadcast news. Her analysis included additional features such as speech rate and word predictability to account for a total of 17 predictors. Five of these predictors (namely the position of the syllable in the word and of the word in the utterance, the speech rate, the respect of the sonority principle and the number of consonants in the resulting consonant sequence when computed within the word) were shown to be sufficient to account for schwa alternation. In this study, however, the effect of extralinguistic variables related to the speaker, such as age, geographical origin and socio-educational level were not considered due to the nature of the corpus.

¹This is a summary of the results presented in Avanzi and Brognaux (2015).

²Since Delattre (1947), liaison is usually described as obligatory, facultative or prohibited. We focus here mainly on facultative liaisons.

The realization of liaison has also been shown to vary according to a certain number of linguistic cues (Encrevé, 1988; Morin and Kaye, 1982; Fougeron and Frauenfelder, 2001) among which: segmental variables, the morphological and grammatical status of the two connected words, their frequency and syntactic and prosodic variables.

Finally, the realization of WFPOL was shown to be influenced by the nature of the liquid and of the following environment (Pooley, 1996; Villeneuve, 2010), the grammatical category of the word and its lexical frequency (Pooley, 1996; Armstrong, 2001) and the accentuable status of the syllable hosting the WFPOL (Pooley, 1996; Putska, 2011). The interested reader is referred to Avanzi and Brognaux (2015) in which we provide a detailed overview of studies investigating WFPOL realization, in the light of linguistic and/or extralinguistic factors.

Studies of **extralinguistic factors** usually rely on the analysis of a single dimension, be it the age of the speaker, its origin or socioeducational level. Such factors have been shown to play a crucial role in all the aforementioned phonotactic phenomena. They are claimed to behave similarly regarding the read vs. spontaneous (referred to as speaking style in the remainder of this chapter) dimension, and speakers' *socio-economic status*. Formal speaking situations (such as reading) would favor a more controlled speech, and, consequently, a higher rate of liaisons and of realized schwas and liquids in the contexts mentioned above (for studies on the effect of speaking style on schwa, see Lucci (1983); Adda-Decker et al. (1999); Léon (1992); Hansen (2000), for liaison see Durand and Lyche (2008)). Conversely, speakers with a lower socio-economic status and spontaneous speech would trigger fewer schwas (Léon, 1992), less liaisons (Encrevé, 1988) and allow in a greater extent liquids deletion (Laks, 1977; Armstrong, 1996; Boughton, 2014).

As for the region of origin of the speaker, studies indicate that it would play a role in the behavior of only schwa (Hambye, 2005; Racine and Andreassen, 2012) and liquid deletion (Hambye and Simon, 2012; Putska, 2011; Boughton, 2013).

Finally, the impact of the age of the speaker was notably pointed out by Racine and Andreassen (2012) regarding schwa deletion and Villeneuve (2010) for WFPOL deletion.

Interestingly, little is known about how these sociolinguistic factors interact.

In Chapter 9 we have analyzed the realization of the aforementioned phonotactic phenomena according to three situational traits: spontaneous/read, media/non media and expressive/non expressive. In the framework of that study, we have however neglected the impact exerted by other dimensions of the corpus such as the origin of the speaker (CPRM contains speech from Belgium, France and Switzerland), the age of the speaker or his socio-educational level. In Section 11.2.4, we set out to analyze the impact of the speaking style on a very large corpus in French balanced in terms of age, origin, socioeducational level and sex of the speakers³. This aims at exploring whether the observations made in Chapter 9 may be generalized.

In Chapter 9, perception tests also indicated that both schwa elisions and

³The impact of these other factors has also been studied and presented in Brognaux and Avanzi (2015b) but falls beyond the scope of our thesis.

Word-Final Post-Obstruent Liquid (WFPOL) deletion played a significant role in the naturalness of the synthesized voice. Our phonetic analysis however discarded purely linguistic influences, such as the phonetic context or the syntactic construction in which the word is found, relying on the assumption that such contexts would be balanced in the corpus and would therefore not influence the interpretation of our results. Acknowledging the precise role played by linguistic factors is essential for the design of post-phonetic rewrite rules. To that aim, Section 11.2.5 focuses on post-obstruent liquid deletion, which was the topic of much less studies than schwa elision. As previously mentioned many factors have been pointed out as impacting the WFPOL deletion. However, the presented results are often contradictory, as thoroughly detailed in Avanzi and Brognaux (2015). We believe that this is mainly due to the fact that most studies do not take simultaneously into account the impact of the different variables, which often interact.

To address that criticism, we examine in the current study the role played by speaking style (read vs. spontaneous), by considering it simultaneously with the linguistic factors which have been shown to explain most of the variance in Bürki et al. (2011). Our study complements state-of-the-art analyses dealing with sociolinguistic aspects of the variable only, by providing insight into the way phonotactic, lexical, prosodic, sociolinguistic and stylistic variables interact when modeling the behavior of this optional pronunciation facility phenomenon in French.

Our study relies on a large French corpus, containing, for each speaker both spontaneous and read speech, described in Section 11.2.3. Conversely, the corpus studied in Chapter 9 did not include both speaking styles for each speaker which may have induced biases due to idiosyncratic features of the speakers in each style. The next sections largely apply the methodology developed in Chapter 9 to provide a more detailed analysis of the factors influencing the realization of phonotactic phenomena in French.

11.2.3 Material

Our corpus is made up of a subpart of the material recorded within the Phonologie du Français Contemporain framework (Durand et al., 2009). The current version of the corpus, called CPROM-PFC (Avanzi, 2014) includes 15 regional varieties of French, recorded in 3 different countries of Europe: 5 varieties spoken in the Northern part of metropolitan France (Béthune, Brécey, Lyon, Paris and Ogéviller); 5 varieties spoken in Switzerland (Fribourg, Geneva, Martigny, Neuchâtel and Nyon) and 5 varieties spoken in Belgium (Brussels, Gembloux, Liège, Marche-en-Famenne and Tournai). For each of the 15 locales, 4 female and 4 male speakers were selected; they were born and raised in the city in which they were recorded. As for the age of the speakers, it ranges from 19 to 82 years, and is controlled for each of the 15 groups. Each of the 120 speakers was recorded in a reading text task (the text is 398 words-long) and in semi-directed sociolinguistic interviews, in which the informant had minimal interaction with the interviewer. A segment of 3 minutes of spontaneous speech was selected for each speaker. In

total, the corpus is more than 13 hours long, and includes approximately 123k tokens.

The entire corpus was orthographically transcribed and automatically aligned within Praat (Boersma, 2001) with the EasyAlign script (Goldman, 2011), which provides a 3-layer annotation in phones, syllables and words (see Goslin et al. (1999) for the implementation of the syllabification rules). The phonetic transcription and its alignment were manually verified and corrected when necessary by inspecting both spectrogram and waveforms (*e.g.* boundary adjustments, segments deletion or addition especially in the case of schwas, liaisons and liquid consonants). The orthographic transcription was then annotated in part-of-speech (POS) tags using the DisMo software (Christodoulides et al., 2014) and was manually corrected by hand. A dedicated tier ("delivery") was manually added, in order to identify overlapping speech and short non-audible or unusable segments (*e.g.* due to the presence of external noises in the recording). Participants were broadly categorized as belonging to the working class (WC) or a higher class (HC) according to whether their occupation was manual or non-manual (Chambers and Trudgill, 1998). We are conscious that the factors that may determine socioeconomic class are numerous, but we followed Boughton (2014) and Chambers (2009) by considering that occupation is a good indicator of social class.

Prominent syllables were identified independently by two experts on the basis of their perceptual judgment only, following a methodology described in Avanzi et al. (2010); Avanzi (2014). Finally, prosodic phrase boundaries were automatically identified on a separate tier. Practically, we crossed the information from the corrected DisMo annotation and the prominent syllable labeling. First, the phonological status of each word regarding lexical stress was automatically coded on the bases of the distinction between grammatical and lexical words proposed in Delais-Roussarie (1996); Mertens et al. (2001b) and Goldman et al. (2010), leading to a segmentation of the speech flow in Phonological Words. We inserted at the end of any Phonological Word whose last metrical syllable was prominent a prosodic phrase boundary, called accentual phrase (AP) in the remainder of this chapter (Avanzi, 2014).

This very rich annotation should allow for a detailed analysis of the correlations between phonetic modifications and other aspects of speech, be they linguistic or extra-linguistic. Prosodic analyses of the corpus have been carried out by Avanzi (2014) but CPROM-PFC was never studied as such in terms of phonetic variation.

11.2.4 Overview of the impact of speaking style on phonotactic variation

Phenomena extraction

We first carried out an analysis of the role played by several extralinguistic factors on the phonetic realization, namely speaking style, speaker's age, socioeconomic status and origin. The results of that study are detailed in Brognaux and Avanzi (2015b). We will here summarize the analysis of the role played by speaking style (*i.e.* spontaneous vs. read) which constitutes the core topic of this chapter. It should

be noted that the very nature of the corpus, which is balanced in terms of speaker's age, origin and socioeconomic status allows for a more precise analysis of the results, as the influence exerted by these other extra-linguistic factors may be neglected. This comes in contrast to our analysis in Chapter 9, in which the studied database was not balanced in that respect.

The methodological protocol applied in the current study is rather similar to that presented in Chapter 9. It is briefly summarized here:

- The extraction of all sites of potential schwa deletion was performed through the alignment of the phonetic realization of the corpus and the eLite automatic phonetization of its orthographic transcription, with a method described in Chapter 9. All monosyllabic function words (*i.e.* *ce, de, je, le, me, ne, que, se* and *te*) and polysyllabic words containing an initial schwa in their standard pronunciation were extracted. Based on the automatic alignment, elision phenomena were automatically identified.
- For liaisons, an automatic script considered the orthographic and phonetic transcriptions of the corpus. Potential liaisons sites were defined as words ending in a latent consonant /t,n,z,R,p/, *i.e.* ending in a t,d,n,s,x,r,p consonant, and followed by a vowel, a h or a semi-vowel. While their identification in Chapter 9 relied on these criteria only, the current study is based on a more detailed analysis of these sites. Indeed, the latent status of the final consonant was determined on the basis of a phonetic dictionary look-up relying on an enriched version of Brulex containing all derived and inflexed forms (Pagel et al., 1998). For rare cases for which two pronunciations were proposed (*i.e.* with or without the final consonant) such as the word 'plus', a manual check was performed. For all other cases, the realization of the liaison was automatically determined by analyzing the last pronounced phoneme of the word in the corpus. In addition, all potential sites were then manually checked, and liaisons were categorized by Dr. Avanzi as obligatory, optional or forbidden, based on previous works in Durand and Lyche (2008) and Delattre (1947), relying on the POS annotation⁴.
- For deletion of /l/ and /R/ in word-final obstruent-liquid clusters, all tokens whose last consonant were clusters formed by an obstruent (*i.e.* *t, d, p, b, g, k, c, f, v*) followed by a liquid (*i.e.* *l* or *r*) and a *e* (and possibly a *s* or *nt* which are the orthographic marks of plural in French and are never pronounced in isolation) were automatically extracted from the orthographic transcription. Their phonetic transcription was analyzed to assess the pronunciation of their final liquid.
- All occurrences of the singular personal clitic subject pronoun 'il' were also extracted, along with their phonetic realization. The plural form was here not considered due to its very low frequency in the reading task.

⁴It should be noted that optional liaisons can be seen as belonging to a continuous scale, ranging from highly probable to highly unlikely liaisons. Such a detailed annotation of the corpus would be however rather costly to make. It should be based on statistical analyzes of another large corpus and should consider the fact that the probabilities may be highly speaker-dependent.

Note that all items in context of disfluency were not taken into account in the analysis.

Statistics

The statistical analyses performed in the framework of the current study go further than those proposed in Chapter 9. Due to the nature of the dependent variable (presence or absence of a given phenomenon), data were analyzed by means of Generalized Estimated Equations (GEE) with repeated measures (Ghisletta and Spini, 2004). Participants were entered as random variables to discard potential influence of the pronunciation of a specific speaker. Independent variables of the model included speaking style (read vs conversational speech), speakers' age, speakers' regional origin (Belgium, France, Switzerland), speakers' socioeconomic level (WC vs HC) and all the double interactions. Bonferroni corrections were applied when examining pairwise comparisons between levels of a given predictor. The full results are discussed in Brogniaux and Avanzi (2015b). We will here mainly focus on the role played by the speaking style.

Results and discussion

The results of our analysis of the role played by speaking style on phonetic variations confirmed most observations made in Chapter 9.

In total, 2427 sites of **initial schwas in polysyllabic words** were extracted from the corpus. The analysis revealed that the speaking style has a significant effect on the realization of schwa in this position (Wald $\chi^2(2) = 54.009$, $p < .001$). In conversational speech, initial schwas are maintained in 45.3 % of the cases of the cases, against 98.52 % of the cases in read speech. 11537 sites of **schwas in monosyllabic grammatical words** were extracted from the corpus. Here again, the analysis pointed out that the speaking style plays a significant role in the presence of schwa in this position (Wald $\chi^2(2) = 418.173$, $p < .001$). In conversational speech, the vowel is maintained in 48.23 % of the cases of the cases, while it is maintained in 94.76 % of the cases in read speech.

In total, 5705 sites of **potential liaison** (including obligatory, facultative and forbidden cases) were extracted from the database. To complement the observations made in Chapter 9 their distribution was analyzed by including the consonant as an independent variable in the model⁵. This analysis first indicates an influence of the nature of the consonant (Wald $\chi^2(4) = 131.946$, $p < .001$). This is shown in Figure 11.1 which indicates the percentage of realized liaisons as a function of the consonant liaison and the category of liaison. We observe that, among facultative liaisons, large

⁵It should be noted that the amount of occurrences of each liaison type highly depends on the nature of the latent consonant. A table providing the distribution of liaisons contexts in our corpus is shown in Appendix C. It indicates, for instance, that no obligatory liaison context in *p* or *r* is found in our corpus, which explains their absence in the figure.

discrepancies are found between frequently pronounced latent consonants like *n*, *t* or *z* and rather unfrequent cases like *p* and *r*. This confirms what has been shown in many previous studies (see Durand and Lyche (2008); Durand et al. (2011) and references therein). It should however be pointed out that our corpus does not contain enough contexts in *p* to provide any statistical significance. The rate of realized facultative liaison is also close to the one reported in the literature (Durand and Lyche, 2008; Durand et al., 2011). Such information should be clearly considered when designing phonetic rewrite rules.

Interestingly, no significant influence of the speaking style was found regarding liaison realization. This lack of distinction between read and spontaneous speaking styles is rather surprising and contradicts the results of earlier studies (Fougeron and Frauenfelder, 2001; Fougeron et al., 2001; Lucci, 1983) and those obtained in Chapter 9. Further analyses have been performed in that respect and have been partly presented in the framework of Avanzi and Brognaux (2015). The use of Generalized Linear Mixed Models (GLMM) including several additional features (*e.g.* articulation rate, word frequency and the number of syllables in the cluster (*i.e.* Word 1 + Word 2)) indicated that liaisons are significantly more realized in reading than in conversation. This implies that more precise analyses are required in that respect. Such analyses are also currently in progress regarding schwa elision phenomena and should be published soon.

In total, 2622 words ending with a **post-obstruent /l/ and /R/ in word-final obstruent-liquid clusters** were extracted from the corpus. Table 11.1 reveals their distribution. The nature of the segment on the right side of the liquid (consonant [C], vowel [V] or pause [#]) was included as an independent variable. Results first revealed an effect of context (Wald $\chi^2(2) = 86.072$, $p < .001$). The liquid is more often maintained before V and # than before C. An effect of style was also found (Wald $\chi^2(1) = 143.409$, $p < .001$). In conversational speech, the liquid is maintained in only 53.22 % of the cases, while it is maintained in 87.84 % of the cases in read speech. This is further investigated in the next section through a more detailed analysis of the corpus in that respect.

Table 11.1 - *Distribution of words ending with a post-obstruent /l/ and /R/ in the corpus, according to the segment to the right of the word (followed by a consonant, C; vowel, V; or pause, #).*

	+C	+V	+#
L/R deleted	460	64	84
L/R maintained	1025	419	570
% maintained	69.02	86.74	87.15

In total, 982 *il* were extracted from the database. It appeared that the liquid was always realized before a vowel. Conversely, it was deleted in front of a consonant in 240 cases, against 254 cases of /l/ realization in that context. In the analysis, only the cases for which *il* is followed by a consonant were considered (N=494). Similarly

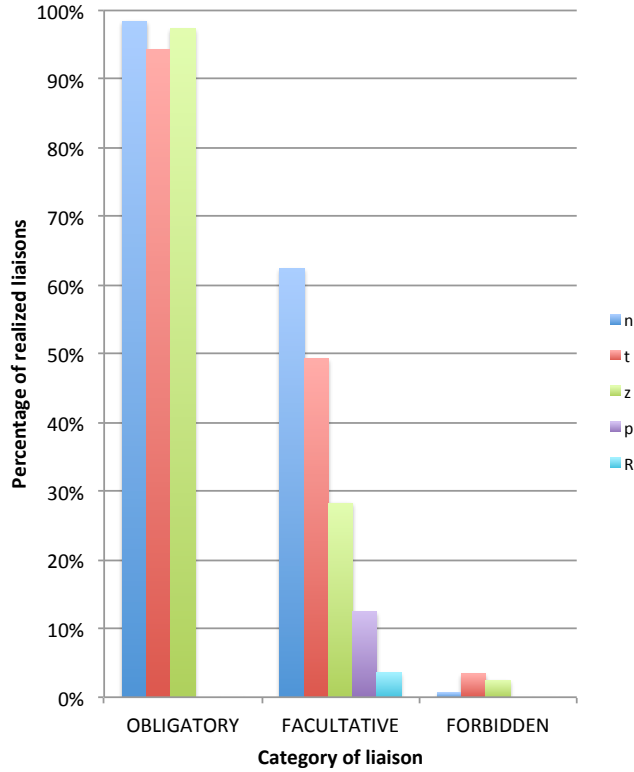


Figure 11.1 - *Percentage of realized liaisons in the corpus, according to the category of liaison (obligatory, facultative and forbidden) and the nature of the final latent consonant.*

to post-obstruent liquid deletion, the statistical analysis revealed a significant effect of style (Wald $\chi^2(1) = 174.584, p < .001$). In conversational speech, the liquid is maintained in only 5.76 % of the cases, against 96.15 % of the cases in read speech.

11.2.5 An in-depth multifactorial analysis of post-obstruent liquid elision

This second section presents a multifactorial analysis of word-final post-obstruent liquid (WFPOL) behavior, in the light of linguistic and extralinguistic contextual features.

Phenomena extraction

The extraction of this phenomenon is similar to that presented in the previous section. It however goes a step further as:

- Proper names (*i.e.* toponyms: "Naples", Napels; and first names: "Alexandre", Alexander) and regional words were discarded (N=22).
- Potentially ambiguous items, *i.e.* words finishing by <r> or <l> and followed by a word beginning with the same consonant (as in "notre roman", our novel) were discarded from the analysis (N=13), as the following word-initial liquid could be mistaken for the second member of a WFPOL.
- All compound items ("quatre-quart", pound cake; "entre-temps", in the meantime) were discarded from the analysis (N=20).
- Out of the 2737 remaining items, it appeared that 708 were followed by a schwa. Since the presence of a schwa in these contexts relies on other factors⁶, they were set aside for future analysis.

The number of clusters analyzed corresponds to 2028 items. Out of these 2028 items, 638 were pronounced without their WFPOL (31.45 %).

Statistics and variables

As discussed in Section 11.2.1, the motivations behind the reduction or the pronunciation of WFPOL in French are numerous, be they linguistic or extralinguistic. Due to the reduced size of our dataset (2028 items), we could not consider all the potential variables in the analysis. We were nevertheless able to consider a large number of factors, which are listed below:

- Phonetic
 - Nature of the segment following the WFPOL (consonant, vowel or pause)
- Lexical
 - Part of speech of the word (broadly categorized as, "noun", "verb", "modifier" (including adjective and adverbs) or "grammatical item")
 - Lexical frequency (computed on the conversational part of the corpus)
 - Number of syllables in the word
- Prosodic
 - Position of the word hosting the WFPOL in the AP
 - Local articulation rate (obtained by measuring the duration of the interpausal chunk within which the item hosting the WFPOL was produced, and by dividing this value by the number of syllables the interpausal chunk contains, giving a measure in s/syll)

⁶A strategy to resolve the issue of rising sonority in the coda, as detailed in Section 5.2, lies in the insertion of an final schwa after the cluster, as in the pronunciation of *vendre* [va~dRə] or *sembler* pronounced [sa~blə], where the liquid appears to be a part of the onset of the syllable (Dell, 1985). Since this strategy may be considered as more related to the question of schwa pronunciation, it won't be addressed in this section.

MODELING SPONTANEOUS SPEECH: THE INTEGRATION OF PHONETIC VARIANTS IN SPEECH SYNTHESIS

- Speaker's characteristics
 - Age of the speaker
 - Sex of the speaker
 - Socio-economic status of the speaker
 - Country of origin of the speaker
- Speaking style
 - Read or spontaneous speech

It should be noted that the number of syllables in the word and the local articulation rate have, to our knowledge, never been studied in the framework of WFPOL realizations, but were shown, to play a crucial role in the behavior of other phonotactic phenomena (see Bürki et al. (2011)).

Unlike Section 11.2.4, data were analyzed by means of Generalized Linear Mixed Model (GLMM) with a logit link function (Baayen et al., 2008). The advantage of GLMM compared to GEE as used in the previous section is that they allow for the consideration of multiple random variables. In addition to the speaker, we also entered items as random variables in the framework of this study, which avoids biases due to highly frequent word with a preferred WFPOL behavior. The maintain/deletion of the WFPOL was entered as the outcome, and the 14 variables as predictors in the model. For a more detailed description of the statistical processing of the results, see Avanzi and Brognaux (2015).

Results and discussion

Table 11.2 provides the list of predictors which were shown to be significant and were retained in the final model.

Table 11.2 - *Summary of the predictors having a significant effect on WFPOL behavior in the corpus.*

Predictors	<i>p</i>
Linguistic variables	
Nature of the following segment	<0.001
Articulation rate	<0.001
Nb. syll. item	<0.05
Extralinguistic variables	
Speaking style	<0.001
Dialect	<0.001
Socioeconomic level	<0.05

In terms of **linguistic variables**, we observe a major impact of the *nature of the following segment*. Figure 11.2 points out that, when the item hosting the WFPOL is followed by a consonant, the deletion rate is higher (58.9 %) than when the item is

followed by a pause or a vowel (13.1 % and 11.7 %, respectively). Post-hoc analyses show that the difference between pre-consonantal and the two other contexts is significant (both at $p < 0.0001$), while pre-pausal and pre-vocalic contexts do not differ from each other (*n.s.*). It should be noted that these results confirm the findings of the vast majority of authors who considered this variable (Chevrot et al., 2000; Hornsby, 2006; Villeneuve, 2010; Boughton, 2014).

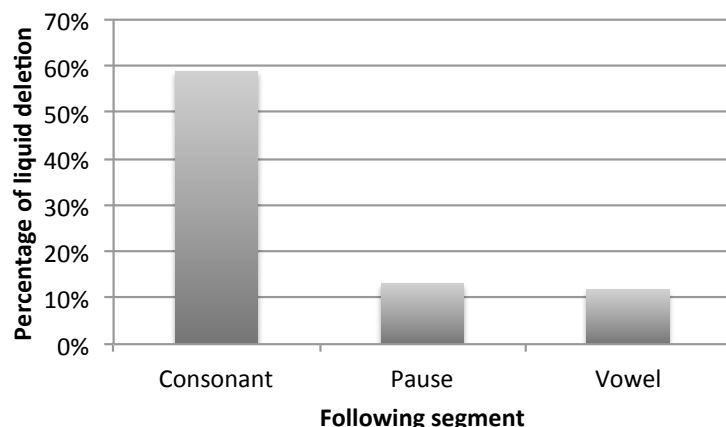


Figure 11.2 - WPFOL deletion rate as a function of the following segment (consonant, pause or vowel).

We also observed significant effects of the *articulation rate* and the *number of syllables in the item hosting the WFPOL*. The slower the articulation rate, the greater the chance for the liquid to be maintained. The liquid also tends to be more often maintained in longer words.

This first stage of our study highlights the fact that the nature of the following segment, along with the number of syllables of the word should be taken into account when predicting liquid elision in speech synthesis. It should also consider the fact that the elision rate depends on the speaking rate.

Regarding **extralinguistic factors**, the role played by the speaking style (*i.e.* read vs. spontaneous speech) is significant, which confirms our previous results. As shown in Figure 11.3, the rate of deletion is much higher in conversational speech than in reading (50.7 % vs. 18.3 %). This further underlines the need for phonetic modifications of WFPOL in the framework of conversational/spontaneous speech synthesis. An effect of the speaker's origin and socioeconomic levels is also found but falls beyond the scope of this chapter. For more details on that, see Avanzi and Brognaux (2015).

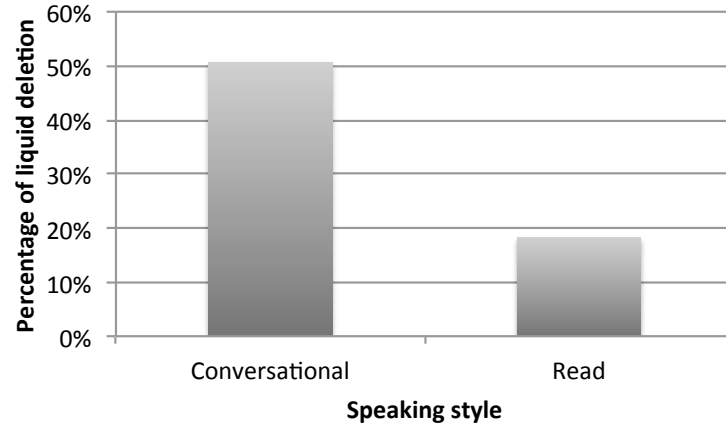


Figure 11.3 - *WFPOL deletion rate as a function of the speaking style (read or spontaneous).*

11.2.6 Conclusion

Following the results obtained in Chapter 9 which indicated that some phonetic variations should be considered when synthesizing spontaneous speech, we proposed in this section a multifactorial analysis of these phenomena in a very large corpus in French containing more than 13 hours and 120 speakers. The whole corpus was manually annotated and enriched in terms of phonotactic annotation by automatic annotation tools. This section provided further insight into the way linguistic and extra linguistic factors interact in pronunciation variation. This should facilitate the choice for relevant contextual features when designing phonetic rewrite rules in speech synthesis.

The first stage of our study focused on **extralinguistic factors** playing a role in schwa/liquid elision and liaison. Special emphasis was put on the impact exerted by the speaking style (*i.e.* spontaneous vs. read speech) with spontaneous speech synthesis in mind. In order to neutralize effects due to other extralinguistic factors (*e.g.* speaker's age, sex, origin or socioeducational level), we relied on a large corpus balanced in that respect. Results have shown an important effect of speaking style on schwa distribution at the initial of polysyllabic words and in grammatical items, as well as on liquid deletion in word-final obstruent-liquid and in 3rd personal clitic subjects pronouns. Spontaneous context gives rise to more elisions in that respect, which confirms the observations made in Chapter 9. No effect of speaking style was found on liaison distribution but further analyses considering additional features still indicated that liaisons tend to be more realized in read than in spontaneous speech.

The second stage of our study concentrated on Word-Final Post-Obstruent Liquid (WFPOL) deletion. As Chapter 9 had indicated the crucial role played by that phenomenon in the naturalness of the synthesized spontaneous speech, we investigated

in this study which factors should be considered when predicting liquid deletion. In our corpus, more than 2500 tokens presenting a WFPOL were extracted. Unlike most existing studies, our analysis considered a large set of linguistic and extra-linguistic features simultaneously. It also included factors which had never been studied in that respect, namely word length and speech rate.

The analyses relied on Generalized Mixed Linear Models, which allowed us to enter both speakers and items as random effects, in order to avoid biases due to the behavior of certain speakers or frequent items. Out of the 14 analyzed predictors, our study indicated that less than half of them was significant:

1. The nature of the following segment (pre-consonantal contexts being more favorable to WFPOL deletion than prevocalic and pre-pausal contexts, which do not differ from each other)
2. Articulation rate (more deletion occurring as the rate increases)
3. Number of syllables in the word (more deletion occurring in shorter words)
4. Speaking style, spontaneous vs. read speech (more deletion occurring in spontaneous speech)
5. Speaker's origin
6. Speaker's socioeconomic level

This first highlights the crucial role played by the speaking style, which confirms findings of the previous Section and of Chapter 9. It complements these results by indicating that some linguistic factors, be they phonotactic, lexical or prosodic, should also be considered when predicting liquid deletion in speech synthesis. Interestingly, it emphasizes the impact exerted by word length and articulation rate, which had never been shown in previous studies. The inclusion of our findings in phonetic rewrite rules for speech synthesis should lead to a more accurate pronunciation in conversational-like speech style.

A similar in-depth analysis is also currently carried out on schwa elision, which was also found to highly impact the naturalness of the speech synthesizer in Chapter 9. The use of both GLMM and additional linguistic features should provide broader insight into the modalities of appearance of this phonotactic phenomenon. This should result in an additional publication soon.

11.3 Spontaneous speech synthesis in English: Investigating the impact of the phonetization of the training corpus

11.3.1 Introduction

In Chapter 9, we have shown that the integration of phonetic variants at synthesis significantly improved the naturalness of HMM-based synthesized sports commen-

taries. In that study, the training of the synthesizer relied on a manually-annotated corpus in terms of phonetic realization. In that sense, we only tested the impact exerted by phonetic changes of the sentences to synthesize but did not assess the role played by the phonetic transcription of the training corpus. This is the topic of the current study.

In an ideal framework, synthesis training should always rely on a manually-annotated corpus in terms of phonetic realization, especially for spontaneous speech which has been shown to display a high amount of phonetic variation. This however poses two major issues:

- As previously mentioned regarding alignment and prosody, manually annotating large amounts of data (as required for HMM-based speech synthesis training) is costly and therefore prohibitive. Most current studies attempt to rely on minimal manual intervention and most stages of the front-end component have therefore been automated.
- Most available front-ends (*e.g.* Festival) do not accept a manually-checked phonetic transcription as input. As previously mentioned, eLite-HTS is an exception to that rule but applies to French corpora only. The main cause behind this restriction is that front-ends usually provide a multi-level transcription of the text (including for instance parts-of-speech, syllables and phonemes). If a phonetic transcription is imposed, the synchronization with the other transcription tiers needs to be computed. This poses issues in particular when the phonetic variations result in modifications of the syllabic structure (*i.e.* the insertion or deletion of syllables), which causes a mismatch in the alignment with the other transcription tiers. Stress information, for instance, is usually computed from a phonetic dictionary. If the phonetic transcription is changed, it may be complex to assign stress information correctly.

The training of HMM-based speech synthesis therefore usually relies on a standard forced alignment of the automatic phonetization of the training corpus, as presented in the first part of this thesis. The automatic (canonical) phonetization⁷ of the text is produced by the TTS front-end component. This phonetization may be incorrect and the phonemes wrong, *e.g.* when reductions or deletions occur. The subsequent alignment is then forced to find boundaries of phonemes that may not have been realized (and conversely). This has been considered as a minor issue in unit selection as the join cost will usually discourage any bad unit from being selected. The problem, however, arises in HMM-based speech synthesis, though it is usually assumed to have a minor effect provided the phonetization is close enough to being correct. It is then assumed that any bad unit will be averaged out as "noise" or that this will not impair the resulting synthesis as the same phonetization errors will be made at synthesis. This is what we call the "assumption of consistency".

This assumption is widely accepted in the speech community and states that being consistent across training and synthesis should allow for transcription errors

⁷Technically, it usually rather corresponds to a phonemization, but we will stick to the term "phonetization" in the remainder of this chapter, for clarity sake.

not to affect the resulting synthesis⁸. In other words, it is assumed that making the same mistakes consistently may "accidentally" have positive effects, such as appropriate phoneme reductions. In fact, Campbell (1997) argues that the annotation of fine phonetic details in the training corpus is not necessary since these phonetic modifications are determined by the linguistic context (*e.g.* high-level prosodic features). When predicting the canonical phonetic transcription in the same context, the resulting synthesis should be adequate. To the best of our knowledge, however, this assumption of consistency has never been proved, at least not in HMM-based speech synthesis. Besides, we believe that this assumption may not apply if the discrepancy between canonical and real pronunciation is high, which may be the case when considering spontaneous speech which has been shown to display a high degree of phonetic variation. A notable study in that framework is that of Kim et al. (2004), who studied the integration of phonetic modifications both at training and synthesis stages. Their results indicated that consistency should be maintained, as the best results were given by the full improved model but that improving the training stage only degraded the baseline results. This was however tested on unit-selection synthesis and considered phoneme substitutions only (*i.e.* not phoneme elisions and insertions).

The current study sets out to address that issue in the framework of HMM-based spontaneous speech synthesis. More precisely, it proposes to investigate whether better phoneme accuracy of the training corpus can improve (spontaneous) speech synthesis, despite being inconsistent across training and synthesis.

The section is organized as follows. The parallel corpus used in the current study is presented in Subsection 11.3.2. Based on this, Subsection 11.3.3 provides evidence that the canonical phonetization and forced-alignment procedure produces many, and serious mistakes in English, especially in the case of spontaneous speech. In Subsection 11.3.4, we describe a method we implemented to account for pronunciation variation of the training corpus. The improvement of the resulting alignment is objectively evaluated through comparisons with the manual annotation. Based on this, perception tests are performed in Section 11.3.5 to assess the quality of the resulting speech synthesis. They compare the baseline (*i.e.* training and synthesizing with canonical phonetization) and the improved version, based on the improved phonetization of the training corpus but using the canonical phonetization at synthesis, thereby breaking the consistency assumption between training and synthesis. Finally, Subsection 11.3.6 concludes the section.

11.3.2 Material: A parallel corpus

The current study relies on a parallel corpus presented in Dall et al. (2014). It consists of 50 sentences that were pronounced spontaneously, in the framework of a normal conversation, by a British English female speaker. A few days after the

⁸Miller et al. (1997) and Miller (1998b,a) also insist on the importance of consistency between the database and the predicted phonetization. Her study however supposes that the database has been correctly phonetized. Similar assumptions are made in Fackrell et al. (2003)

recording, the same speaker was asked to re-read the sentences she had previously pronounced. The value of this corpus is that both sub-corpora contain the very same sentences pronounced by the same speaker. This allows for assessing the impact exerted by the speaking style (*i.e.* spontaneous vs. read speech) while neutralizing all other potential influences. A similar corpus was used in Laan (1997) but was only studied in terms of intonation, duration and spectral features. Interestingly, no attention was paid to phonetic variation, even though this question was briefly mentioned. The major advantage of our corpus lies in its manual annotation in terms of phonetic realization (see below).

Based on the orthographic transcription of the corpus, the Festival front-end (Black et al., 2001) was used to produce an automatic phonetization of the text, with the Received Pronunciation (RP) British English version of the Combilex phonemic dictionary (Richmond et al., 2009, 2010). This automatic transcription will be referred to as "Standard" in the remainder of this section. Combilex is a phonemic dictionary which makes use of the phoneme set detailed in Appendix D. It should be noted that it provides a rather narrow transcription however, as some allophones are explicitly transcribed⁹. The transcriptions were automatically aligned based on average models adapted to 1176 read or 1146 spontaneous sentences from the voice talent. The transcription also included pausing as it helped alleviate the cascading issue (see below). Two annotators¹⁰ then independently corrected the automatic phonetization. It should be noted that this correction went a step further than that performed in Chapter 9 as substitutions were also given close attention. The voicing state of the consonants, for example, was considered in the transcription.

All disagreement cases were discussed during joint sessions, so that the two annotators agreed upon a final phonetization. If the disagreement was whether to keep the original automatic phonetization or change it, the automatic phonetization was preferred. In that sense, the resulting transcription is rather conservative and is therefore biased towards the automatic canonical phonetization. This is inevitable since human labellers tend to be influenced by the original starting phonetization (Van Bael, 2007). It should be noted that the focus was put on phoneme identity and not on phoneme boundaries. The boundaries were however corrected if grossly wrong (*e.g.* following the elision or the insertion of a phoneme). This manually-corrected transcription will be referred to as the "Gold Standard" in the remainder of this section.

11.3.3 Phonetic analysis of the corpus

In order to support the claim that spontaneous speech conforms less to standard phonetization than read speech, we investigated the degree of discrepancy between

⁹The distinction between the flapped "t" and standard "t" or the distinction between dark and clear "l" is made explicit in the phoneme set. However, it does not distinguish between aspirated and non-aspirated stops.

¹⁰Rasmus Dall and myself.

11.3 SPONTANEOUS SPEECH SYNTHESIS IN ENGLISH: INVESTIGATING THE IMPACT OF THE PHONETIZATION OF THE TRAINING CORPUS

the Gold Standard annotation and the automatic Standard annotation provided by Festival. To that end, we computed the mean percentage deviation in Levenshtein distance, based on the alignment technique presented in Chapter 9. This will be referred to as the Phoneme Error Rate (PER) in the remainder of this section. This measure offers a quick overview of major errors made by the front-end. It should however not be interpreted as such as the manual annotation was highly biased towards the automatic transcription.

Table 11.3 shows the comparison between manual and automatic phonetizations. The latter is surprisingly bad, getting over 19 % of the phonemes wrong for the "simple" case of clear read speech. This even questions the use of such automatic phonetization in standard neutral read speech synthesis. As expected, the phonetization is even worse for spontaneous speech, which yields about 25 % of erroneous phonemes. Unsurprisingly, the real pronunciation displays a high rate of elision, compared to the canonical phonetization. Few insertions have been made by the manual annotators, which is in line with the observations made in Chapter 9 for French.

Table 11.3 - *The overall differences between the agreed manual annotation (i.e. the Gold Standard) and the automatic (i.e. Standard) phonetization by Festival based on Combilex. Discrepancies are defined in terms of insertions (Ins), deletions (Del), or substitutions (Sub). The Phoneme Error Rate (PER) is also given.*

	Ins	Del	Sub	Total	PER
Read	10	149	151	294	19.1 %
Spontaneous	17	202	180	399	25.2 %

A detailed analysis of the discrepancies points at the following phenomena (percentages relate to comparisons between the Standard and Gold standard annotations of the spontaneous part of the corpus and correspond to the percentage of errors the phenomenon accounts for) :

- Deletion of "t" (15 %)
- Substitution of "t" by a glottal stop (9.8 %)
- Substitution of "z" by "s" (6.5 %)
- Vowel reduction to schwa ([ə]) (6 %)
- Deletion of "d" (5 %)

The same phenomena are highlighted when computing the comparison on read speech. A qualitative analysis of the results indicates that "t/d" elisions and "z" devoicing mainly appear in end-of-word position. This is in line with previous studies, such as Miller et al. (1997) who mentioned t/d deletion and consonant glottalization in the list of post-lexical modifications learned by their neural network.

The important role played by "t" glottalization and vowel reduction to schwa was also pointed out by Kim et al. (2004). It should however be noted that the glottalization could be considered as rather speaker specific and non-standard RP English (though common in many dialects).

Noteworthy is also the fact that producing a large number of phonetic errors naturally induces an important degradation of the alignment. In fact, due to the high amount of deletions in spontaneous speech, some utterances experience a cascade of alignment errors, as illustrated in Figure 11.4. The excerpt shows the alignment of the two words "basically because" in read and spontaneous speech. While the automatic transcription is appropriate in the case of read speech, a high reduction of both words is observed in spontaneous speech, *i.e.* bas'ly 'cause. This does not only cause non-existent phonemes to be "found", but also the more serious issue of phonemes being "pushed" later into the utterance, putting every single phoneme further down the line out of alignment. The resulting sample phonemes for model training are therefore grossly wrong.

While the issue regarding because/'cause is arguably a lexicalized difference which could be accounted for at transcription time, the issue of basically/bas'ly is not and similar situations will degrade the trained models. It should be noted that lexicalized differences like because/'cause were considered as orthographic transcription errors and were therefore not considered when computing deletion rates in the framework of our study. It was also found that providing silence transcription to the aligner improved the resulting alignment by confining the cascading issue to interpausal units only. This was therefore done for the automatic transcription which was manually corrected to ease the transcription.

11.3.4 Integrating pronunciation variation in forced alignment

Following the aforementioned observations, attention was paid to improve the automatic phonetization process. To that aim, a lattice-based forced alignment was implemented. This allows for the integration of phonetic variants in the transcription of the training corpus, the most probable phoneme sequence being determined at alignment stage¹¹. Lattice-based pronunciation modeling has long been applied in automatic speech recognition (ASR) and forced alignment (*e.g.* Wightman and Talkin (1997); Adda-Decker and Lamel (1999); Kessens et al. (1999); Binnenpoorte et al. (2004); Bürki et al. (2008)). While such enrichment of the phonetic transcription usually relies on phonetic dictionaries only, our system is based on two sources of information:

- The use of the pre-encoded variants integrated in the Combilex pronunciation dictionary

¹¹It should be noted that other ways to include phonetic variation have been proposed, as in Paulo and Oliveira (2005) who enrich a grapheme-to-phoneme and syllable-to-phoneme transcription on which they train weighted finite state transducers, with no use of phonetic dictionaries. This system was however not tested in speech synthesis.

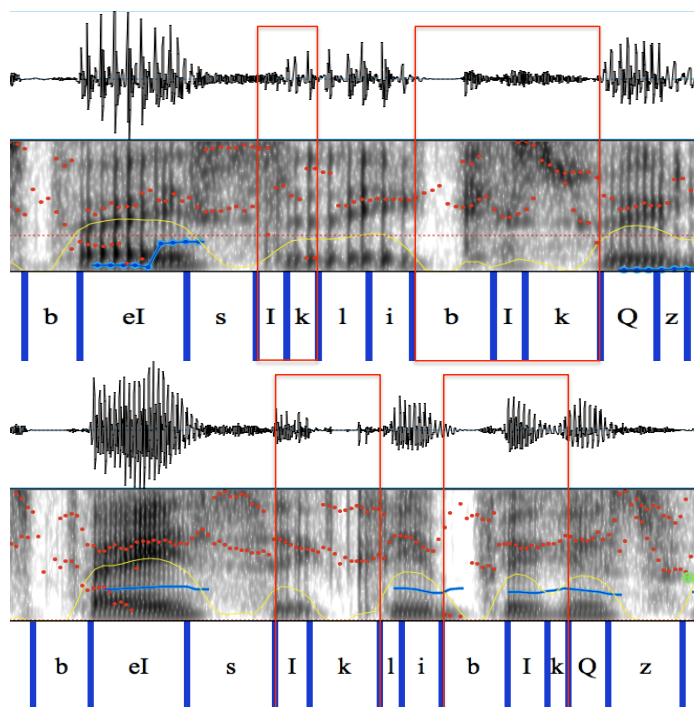


Figure 11.4 - Segment of the same sentence produced spontaneously (top) or as a read prompt (bottom), showing alignment issues due to deletions. The red squares highlight the same segments in both.

- Hand-written variant options driven from the analysis carried out in the previous section

Combilex proposes, for each word, one (or several) canonical/full pronunciations and specifies, if any, its phonetic variants/reduced versions¹². In a standard phonetization procedure, as produced by Festival in the previous section, only the first canonical form is retained for each word, disregarding potential pronunciation variations. Our improved system is based on the underlying context-dependent rewrite rules of Combilex as described in Richmond et al. (2007). The system automatically extracts all variants from the dictionary and populates the lattices (realized as finite state transducers (FSTs)) with them. It should be noted that this may also degrade the resulting phonetization as it is well-known that multiplying the number of variants complicates the correct recognition (Kim et al., 2004).

Additionally, a set of expert-written variant options was designed, based on the discrepancies that were found between the automatic and manual annotation of our corpus. The list of the 14 implemented rules is shown in Appendix E. These options

¹²Combilex contains 1.82 variants per word on average.

MODELING SPONTANEOUS SPEECH: THE INTEGRATION OF PHONETIC VARIANTS IN SPEECH SYNTHESIS

are context-dependent rewrite rules written in terms of regular expressions. They take the form of a phoneme, with its left and right context and the resulting variant option. This allows for modeling substitutions and deletions. Insertions were not considered as they were found to be rare in our analysis. They are although supported by the system. In relation to Section 11.2, it should be noted that the rules consider various levels of contextual features including surrounding phonemes and their characteristics, word, phrase and sentence boundaries. The lattices are automatically enriched by applying these rules on the predicted phonetic transcriptions. No weight is assigned to the lattices and each variant is therefore considered as equiprobable.

Material

To train our system, read and spontaneous corpora of 1176 and 1146 sentences respectively were used. They were recorded in hemi-anechoic studio at 96kHz, 32 bit, then downsampled to 48 kHz, 16 bit. The spontaneous sentences consisted in free conversation between an experimenter and the voice talent. These were manually orthographically transcribed by one transcriber and checked by a second. The read corpus was from the same speaker, under the same recording conditions except the speaker was reading normal phonetically-balanced sets of prompts from a screen. The whole corpus was automatically phonetized and aligned with Festival, based on the RP British English version of the Combilex dictionary. For spontaneous and read speech corpora, the respective 50 sentences of the parallel corpus used in the previous section were also included. This means that 50 sentences of the training set also served as test data. This is not problematic as the system does not consider the alignment of the corpus and relies on a flat-start initialization. As in Part I of the thesis, the aligner is (partly) trained on the corpus that will be aligned in a second stage.

The alignment procedure is very similar to that described in Chapter 2. A flat-start alignment is performed to initialize the models. After a first series of embedded training, short pause models are introduced. A second series of embedded trainings is then performed.

Because lattices cannot be applied directly to the flat start models, they are integrated before the second series of embedded training, *i.e.* simultaneously to the introduction of short pause models which are directly integrated in the lattices. As the lattices are initially created as FSTs these were converted to the HTK SLF format for alignment. The standard procedure was followed by a non-standard pause removal step, where short pauses under 40 ms were removed from the labels and models re-aligned and estimated. This was found to improve performance.

Results

We propose here a preliminary evaluation of the system based on the comparison with the agreed manual phonetization, *i.e.* the Gold Standard. The evaluation is based on a Levenshtein distance, as previously described. While not the suggested method of Van Bael et al. (2006) it is standard (Section 1.32 of Van Bael (2007)).

11.3 SPONTANEOUS SPEECH SYNTHESIS IN ENGLISH: INVESTIGATING THE IMPACT OF THE PHONETIZATION OF THE TRAINING CORPUS

As previously mentioned, the manual annotation is biased towards the standard automatic annotation on which it was based. To alleviate that issue, the whole corpus was manually checked in terms of phonemic realization by a third transcriber¹³, based on the output of the full model, which includes both expert-driven rules and pre-encoded variants. Obviously, that manual annotation is biased towards the full model but still helps in identifying the impact exerted by the two sources of pronunciation variation.

The results are shown in Table 11.4 and indicate an improvement when integrating rules and pre-encoded variants, although the integration of rules achieves higher improvement¹⁴. Improvement is found on both read and spontaneous speech. The integration of both rules and pre-encoded variants achieves the best results. Using lattices particularly reduces the number of phoneme additions, meaning they are more likely to delete phonemes compared to the traditional method. This is important as the cascading errors shown in the previous section occur because of additional phonemes in the automatic transcription.

It should still be remembered that the hand-driven rules were derived from our analysis of the test corpus and that the results may be biased in that sense. Besides, the annotation of the third transcriber was based on the full system and is therefore biased toward that improved automatic annotation. No conclusion is therefore drawn here. Our evaluation will rely on a perceptual evaluation of the resulting speech synthesis as proposed in the next section, which can be seen as a task-based evaluation (Van Bael, 2007) and which constitutes the ultimate goal of our study.

Table 11.4 - Overall PER differences between the manual annotation by the third annotator and the different automatic annotators: the standard phonetizer, the application of the hand-written rules, the application of the pre-encoded dictionary variants or both.

	Ins	Del	Sub	Total	PER
Read					
Standard	15	169	179	363	22.6 %
Lattice w/ Combilex	6	154	111	271	16.9 %
Lattice w/ rules	10	111	102	223	13.9 %
Lattice w/ both	8	102	36	146	9.1 %
Spontaneous					
Standard	15	296	165	476	32 %
Lattice w/ Combilex	3	268	117	388	26.1 %
Lattice w/ rules	9	201	100	310	20.8 %
Lattice w/ both	5	193	48	246	16.5 %

¹³Cassia Valentini Botinhao from the CSTR.

¹⁴A comparison of the improvement methods with our Gold standard annotation is also shown in Appendix F. It also indicates improvement with expert-driven rules but none when adding pre-encoded variants. As previously mentioned, this annotation is however biased towards the standard annotation.

11.3.5 Synthesis evaluation

This section tackles the main objective of our study, *i.e.* assessing whether breaking the consistency by improving the annotation of the database without modifying the transcription at synthesis stage, still improves the resulting synthesis.

To that aim, eight speech synthesizers were trained, relying on the implementation of the HTS toolkit (version 2.3 beta) (Zen et al., 2007a). For each speaking style (*i.e.* read and spontaneous), four models were trained based on :

- The standard automatic annotation (*Std*)
- The standard automatic annotation enriched by lattices based on hand-driven rules (*Man*)
- The standard automatic annotation enriched by lattices based on pre-encoded Combilex variants (*Pron*)
- The standard automatic annotation enriched by lattices based on both sources of phonetic variation (*All*)

The corpora described in the previous section were used for training. The 50 sentences from our parallel corpus were however excluded from the synthesis training and were used for testing.

Method

A side-by-side Mean Opinion Score (MOS) test using a sliding scale from 1-100 was run¹⁵. The natural version of both read and spontaneous speech was included in the comparisons, which means that 10 syntheses were compared simultaneously. It should be remembered that read and spontaneous sentences were identical and phonetized with the same transcription, based on Combilex. The listener was asked to rate each stimulus according to how natural it sounded. At least one synthesis had to be rated at 100 and the others in relation to it. A random subset of 30 sentences from the 50 test sentences was selected. This was divided into two sets of 15. One practice sentence was also selected.

Each participant was presented with one practice and 15 critical slides. 32 paid participants were recruited and performed the experiment in a sound-insulated booth wearing Beyerdynamic DT770 PRO headphones. In total this gives 16 evaluations of each of the 30 critical sentences for a total of 480 datapoints for each system.

Results

Figure 11.5 shows the results of the test. Table 11.5 shows all system pairs compared using a Wilcoxon signed-rank test, after Holm-Bonferroni correction to avoid false positives. Natural speech is, unsurprisingly, rated much higher than synthetic speech.

¹⁵All test materials, including synthetic samples, analysis scripts and raw results are available here: <http://www.dall.dk/rasmus/Materials/ICASSP2016.zip>

11.3 SPONTANEOUS SPEECH SYNTHESIS IN ENGLISH: INVESTIGATING THE IMPACT OF THE PHONETIZATION OF THE TRAINING CORPUS

Interestingly there is no significant difference between the spontaneous and read natural speech, in contrast to Dall et al. (2014). However, this is probably due to the large differences compared to the synthetic samples. The read speech based voices were rated significantly higher than the spontaneous speech-based. For the read speech the standard alignment produces significantly higher rated speech than the other types, and the Combilex variants significantly lower, although the effect size is quite small. Conversely, for spontaneous speech, the lattice system with only Combilex rules was rated significantly higher than the others, with no significance between the other synthesizers.

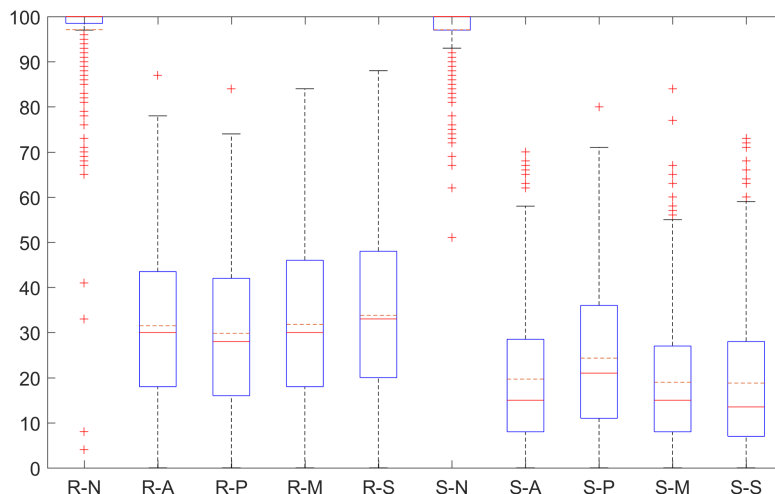


Figure 11.5 - Boxplot of the side-by-side MOS test. *R* = Read speech; *S* = Spontaneous speech. Then, for each speaking style: *N* = Natural (i.e. original version); *M* = Lattice with manual rules; *P* = Lattice with Combilex variants; *A* = Lattice with both manual rules and Combilex variants; *S* = Standard automatic phonetization. Solid lines are medians, stapled means and boxes 25 and 75 % quantiles.

This indicates that the consistency assumption between training and synthesis only partly holds.

For read speech synthesis, our results tend to confirm this assumption as the standard method achieves higher ratings. Subjective listenings of the output synthesis of our "improved" systems suggest that they produce hyper-articulated speech, which could make them less natural-sounding. Arguably, however, we are getting what we ask for. At synthesis time we *ask* for the hyper-articulated version of the sentence, though the reason why we do not normally get it is because of the accidental reductions obtained through the consistency assumption. If we truly wish to control our output synthesis, this however indicates that the acoustic models are more representative of their respective phonemes, which should be appreciated. Based on this, methods for controllable, perhaps even gradable, reduction of a

Table 11.5 - *Adjusted p-values after Holm-Bonferroni correction for the Wilcoxon signed-rank test. $\ast=p < 0.001$. Labels as in Figure 11.5.*

System	R-N	R-A	R-P	R-M	R-S	S-N	S-A	S-P	S-M	S-S
S-S	*	*	*	*	*	*	1	*	1	-
S-M	*	*	*	*	*	*	1	*	-	
S-P	*	*	*	*	*	*	*	-		
S-A	*	*	*	*	*	*	-			
S-N	1	*	*	*	*	-				
R-S	*	<0.005	*	<0.05	-					
R-M	*	1	<0.05	-						
R-P	*	<0.05	-							
R-A	*	-								
R-N	-									

phonetic transcription of the sentence to synthesize should be developed, *e.g.* by taking advantage of the reduced variants already encoded in phonetic dictionaries.

Spontaneous speech synthesis, in contrast, seems to benefit from the use of pronunciation variants at training. It is encouraging that simply applying pre-encoded pronunciation variants helps learn a better model, particularly on difficult data such as spontaneous speech where a fully pronounced alignment is very inappropriate. While the manual rules achieved the best accuracy results on their own and combined when compared to the manual phonetization, they did not increase perceived naturalness in synthesis. This may be due to them being overfitted to the test sentences, and thus not entirely generalizable to the training data.

It should also be noted that read speech-based synthesis models are considered more natural than models trained on spontaneous speech. This is not surprising as spontaneous data is much more varied and harder to model, as exemplified by the much less accurate alignments. However using pronunciation variant forced alignment helped spontaneous speech toward closing this gap.

No perceptual evaluation was carried out here with the improved phonetization at synthesis stage, instead of the standard phonetization. However, preliminary subjective evaluation indicates that the results in that respect are promising. It points out that integrating such phonetic variations at synthesis could improve synthetic speech, which would confirm the results obtained in Chapter 9.

11.3.6 Conclusion

We have reported on an investigation of the consistency assumption between training and synthesis in TTS. This assumption states that errors in the annotation of the training corpus should not impair the resulting synthesis as long as the same errors are made at synthesis stage. To address that question, a pronunciation variant based forced-alignment method was implemented and its application to speech synthesis

11.3 SPONTANEOUS SPEECH SYNTHESIS IN ENGLISH: INVESTIGATING THE IMPACT OF THE PHONETIZATION OF THE TRAINING CORPUS

evaluated. Relying on the observations made in Chapter 9, which indicated that spontaneous speech displayed a high degree of phonetic variation, the consistency assumption was tested in the framework of spontaneous speech synthesis.

Relying on a parallel corpus, containing both read and spontaneous speech, we first highlighted the high degree of discrepancy between automatic canonical phonetization as produced by TTS front-ends and the pronunciation of the speaker. This led us to implement methods to enhance the forced alignment stage by considering potential phonetic variants, thereby improving the phonetization of the training corpus. As that improved phonetic transcription cannot be predicted for new sentences to synthesize, this however breaks the assumption of consistency.

The naturalness of the resulting synthesis was assessed through perception tests, of spontaneous-based and read-based synthesis. It was found that read speech standard synthesis did not benefit from these variants, although the underlying acoustic models were arguably more correct. For spontaneous speech, however, improvement was found when including pronunciation variation extracted from a phonetic dictionary to phonetize the training corpus. This points out that the consistency assumption between training and synthesis only holds when minor errors are made.

An obvious improvement lies in the development of automatic modifications of the phonetic transcription of the sentences to synthesize. Such improvement could benefit from the pre-encoded phonetic dictionary but their probability of appearance should be trained on a large corpus considering all relevant contextual features, as detailed in the previous section for French. Such features may also integrate, for instance, n-gram models as the probability of the word was shown to influence reduction phenomena in English (Jurafsky et al., 2001b,a). That factor was considered in Werner and Hoffmann (2006), along with the sequence of variants. It was shown to improve the resulting unit-selection synthesis. This will be considered in future work and is discussed in Section 12.2.

Chapter 12

General conclusion

Contents

12.1 Conclusion and contributions	287
12.2 Limitations and perspectives	294

12.1 Conclusion and contributions

The present PhD research focused on the improvement of HMM-based expressive speech synthesis in French. It comes as an answer to the strong demand for more natural speech synthesizers, able to adapt their speaking style to express emotions and affect, as humans do in everyday conversation. While the field of expressive synthesis has been rapidly growing in the last decade, we have indicated that current studies present a certain number of drawbacks.

In the framework of our research, we aimed at improving three stages of the front-end component used for the training of HMM-based speech synthesis:

- The phonetic alignment of the corpus with the corresponding speech signal
- The prosodic annotation of the corpus
- The phonetic annotation of the corpus

How the two last types of modifications could be integrated in the front-end at synthesis is also explored.

We postulated that these stages should be improved by integrating (i) in-depth linguistic analyses of the language (here French) and (ii) characteristics of expressive speech. This drove us, in the introduction, to formulate three major hypotheses, related to the three annotation levels. Our PhD research tested and validated these hypotheses and provided tentative answers to the related research questions.

HYPOTHESIS 1. The automatic alignment of expressive (spontaneous) corpora by generic HMM models of the language poses alignment problems and could be automatically improved

This hypothesis was tested and validated in Part I of the thesis. To that end, our study relied on a very large database containing 15 corpora, varying in size, language and speaking style. Major findings in this respect are listed below:

- In Chapter 2, we have indicated that most available alignment tools provide the user with pre-trained speaker-independent HMM-based models of the language. This poses a major constraint, namely the fact that corpora in rare languages or specific speaking styles cannot be aligned because no corresponding model is provided. The alignment of an expressive corpus in English by several existing models of English was shown to achieve low quality results.
- To address this constraint, we set out to analyze the alignment rates obtained by a widely-used method in which the models are directly trained on the corpus to align. The evaluation on our database indicated that it achieved results similar to the use of generic models of the language. A major asset is that the method is speaking-style and language independent while performing well on corpora as small as 2 minutes. This method was made available as an online tool presented in Chapter 4.
- As this method was still shown to achieve poor alignment rates on corpora with high acoustic variability (*e.g.* expressive or spontaneous speech), we provided a detailed analysis of the typical alignment errors made by our method. It pointed out at the poor alignment of silences, semi-vowels and plosives and indicated imprecise boundaries in vowel-to-vowel transitions.
- These typical errors drove us to implement three fully-automatic improvement methods that were described in Chapter 3:
 - To improve silence alignment, we proposed a better initialization of the silence models, based on a Voice Activity Detection algorithm (*VAD*)
 - As systematic errors were found (some phonemes being nearly always too late or too early), we proposed to consider the forced-alignment of the reversed sound (*Reverse*)
 - Because some transitions may not be correctly represented by the cepstral coefficients, we explored the use of supplementary acoustic features (*AddFeat*)

AddFeat was shown to achieve the best results in isolation. The performance of *VAD*, however, was pointed out as striking on very small corpora, correcting more than 60 % of the errors greater than 40 ms. The combination of all three improvement methods yielded the best alignment rates, with very low variability across the corpora, regardless of their size, style and language. This combined

technique was shown to outperform available speaker-independent models of the language, improving the alignment rate by 8 to 10 % absolute, at a 20 ms threshold. The alignment of expressive corpora was shown to benefit from these improvements, which reduced the amount of errors superior to 40 ms by 10 and 37 % for a spontaneous expressive corpus in French and a read expressive corpus in English, respectively.

- Finally, Chapter 11 pointed out that the quality of the alignment significantly impacted the naturalness of the resulting HMM-based synthesis. This was shown in the framework of spontaneous speech synthesis in English. However, that alignment quality was highly impacted by the phonetic annotation of the training corpus, phonetization errors resulting in gross misalignment of the corpus. This is therefore discussed under HYPOTHESIS 3.

HYPOTHESIS 2. The integration of a prosody annotation considering phonological aspects of French and characteristics of expressive speech should enhance HMM-based expressive speech synthesis in French

This hypothesis was tested and validated in the framework of HMM-based speech synthesis of basketball live commentaries in Part II. Further experiments on audiobook synthesis in Part III confirmed our observations. Here is a brief summary of the main findings of our research in that respect:

- In Chapter 6, the prosodic analysis of sports commentaries drove us to distinguish between two prosodic levels, leading to a **two-tier annotation protocol**. The set of labels is both related to a specific function and characterized by a distinct acoustic realization which should facilitate their prediction from text and/or from the speech signal. Ways to automatically predict the labels from text were presented in Chapter 8.
 1. **A local prosodic annotation** refers to accentual phenomena and assigns a label to each syllable. Seven labels indicate initial (emphatic) and final stresses and the degree of boundary the latter determine. While emphatic stress has been shown to display various acoustic realizations, perceptual experiments in Chapter 7 indicated that the use of a *single emphatic label* did not impair the resulting HMM-based synthesis. This obviously facilitates their prediction from text at synthesis.
 2. **A global prosodic annotation** classifies sequences of words into five distinct sub-genres, or speaking styles, defined in terms of valence and arousal. This was dictated by the nature of the corpus which contains various levels of excitation (*e.g.* after a basket) but also varies from positive to negative emotions (*e.g.* the frustration following a missed basket).

Reasonable inter-annotator rates were achieved for both annotation levels.

- **Regarding the integration of the prosody annotation in HMM-based speech synthesis**, local labels were considered as contextual features for the tree clustering process, as this was shown to achieve the best results in earlier studies. For global prosodic changes, three configurations were investigated: (i) the training of style-independent models on each sub-corpus, (ii) the adaptation of a neutral model to each sub-genre and (iii) the adaptation of an average model to each sub-genre. The third option was shown to outperform the two other methods in terms of quality of the resulting synthesis.
- **Results from subjective perceptual tests** indicated, in Chapter 6, that the integration of local labels improves the quality, the intelligibility and the expressivity of the resulting voice. Surprisingly, the integration of global prosody alone did not lead to enhanced synthesized voice. However, its combination with local labels was pointed out as achieving the best results, the expressivity of the improved system being preferred in 51 % of the case, against 33 % of preferences for the baseline. This also resulted in an enhanced intelligibility of the synthesized message.
- While the lack of improvement when integrating global prosody only was unexpected, a detailed analysis of the annotation suggested that the human annotator may have been influenced by the propositional content of the sentences, thereby neglecting their acoustic realization. We therefore investigated in Chapter 8 the possibility to take advantage of the manual annotation in excitation levels to generalize global tendencies and **automatically correct annotation outliers**. To that aim, two classification models were trained, based on Gaussian mixtures and on logistic regression. New synthesizers were built on the resulting corrected annotation. Subjective experiments highlighted that the synthesizers exploiting the corrected annotation, and especially that obtained with Gaussian mixture models, achieved a much better discrimination between the various levels of excitation.
- In Chapter 10, the integration of a local prosody annotation was further investigated in the framework of audiobook speech synthesis in French. A new annotation protocol was developed, based on an autosegmental approach of prosody. It aimed at **automatically providing a detailed description of French accentual patterns**, relying on a set of 17 syllable-based labels. We pointed out that the combination of prominence, tone and phonetic and prosodic information could serve to predict our annotation. Based on this, an automatic annotator was implemented relying on existing tools: (i) a prominence detector, (ii) a tonal annotator and (iii) a synthesis front-end. Its application to a large multi-style corpus achieved high prosody description quality. The integration of this annotation in audiobook synthesis resulted in enhanced naturalness of the synthesized voice. A reduction of the set of labels was also proposed, based on a functional analysis of the labels. That set was similar to that proposed in the framework of sports commentaries. However, a subjective evaluation of the resulting synthesis indicated no significant improvement over the baseline.

HYPOTHESIS 3. The consideration of phonetic modification phenomena in the annotation of the training corpus and for new sentences to synthesize should improve the naturalness of HMM-based spontaneous expressive speech synthesis

This hypothesis was tested and validated in Part II of the thesis, in the framework of HMM-based speech synthesis of sports commentaries. Further experiments in Part III provided avenues for future work in that direction. Our main findings are that:

- Spontaneous speech was shown to display a high level of phonetic variability, compared to the standard read canonical phonetization provided by existing front-ends. This was shown in Chapter 9, based on a large multi-style corpus in French. Media speech was also shown to be characterized by typical phonetic features, notably the high rate of epenthetic schwas. Expressive speech was not found to be characterized by specific phonetic features but this may have been due to the nature of our classification which merged various emotions with different valences.
- In Chapter 9, we further highlighted that the integration of those phonetic variations in the HMM-based speech synthesis of sports commentaries significantly improved the naturalness of the synthesized voice. Schwa and word-final liquid elisions, in particular, were pointed out as playing a significant role in that respect.
- The prediction of these modifications for new sentences to synthesize should rely on a detailed analysis of the linguistic and extralinguistic factors influencing phonetic variation in French. Such statistical analyses were provided in Chapter 11 and focused on spontaneous speech which was shown to display a high rate of phonetic variation. Our analysis was based on a 13-hour corpus containing read and spontaneous speech, balanced in terms of speaker's origin, age, sex and socioeconomic level to neutralize potential effect of these extra-linguistic factors. Preliminary results confirmed the observations made in Chapter 9 by highlighting a significantly higher rate of schwa and liquid elision and less realized liaisons in spontaneous speech. An in-depth analysis of word-final post-obstruent liquid deletion was also carried out as this phenomenon was shown in Chapter 9 to greatly impact the naturalness of the synthesized speech. Our results indicate that this reduction phenomenon is also dictated by linguistic factors like the nature of the following segment, the number of syllables in the word and the speaking rate. These features should be considered when predicting phonetic modifications with French spontaneous speech synthesis in mind.
- Finally, we investigated in Chapter 11 the role played by the phonetization of the training corpus with a view to HMM-based spontaneous speech synthesis in English. In that framework, we explored the widely-acknowledged consistency assumption, which states that errors in the phonetization of the training corpus should not affect the synthesis, as long as the same errors are made at synthesis.

GENERAL CONCLUSION

Analyses of a parallel corpus containing identical sentences in read and spontaneous speech indicated the higher rate of phonetic modifications in spontaneous speech in English. An improvement of the automatic phonetization of the training corpus was proposed by including phonetic variants in the forced alignment. Perceptual evaluations of the resulting synthesized voice indicates that, while the assumption of consistency holds in the case of read speech, the improvement of the phonetization of the training corpus is shown to significantly improve the synthesis of spontaneous speech, despite breaking the consistency assumption.

Compared to the state of the art, the main contributions of the thesis are the following:

- ✓ We provided a systematic step-by-step study of a widely used HMM-based alignment technique which consists in training the models directly on the corpus to align. We studied the role played by various training settings (*e.g.* models configurations, number of training iterations, use of bootstrap) and developed and evaluated our system on 15 corpora varying in size, language (some being under-resourced) and speaking style.
- ✓ Based on typical alignment errors, we proposed 3 fully automatic improvement methods of the basic HMM-based alignment described above. Unlike most existing techniques, they do not require manually-annotated data. The combination of the three improvement methods yields high reduction of the alignment errors, also for expressive corpora, reducing errors greater than 40 ms by 28 %, on average.
- ✓ Experiments on expressive speech synthesis relied on the recording of a corpus of basketball commentaries in French. This corpus presents several assets, lacking in many studies on expressive speech:
 - It is spontaneous
 - It contains "natural" expressivity (no prompt was indicating to the speaker which type of expressivity he should display)
 - It is recorded in high quality conditions to allow for the training of a speech synthesizer
- ✓ We pointed out the need for two prosody annotation levels, describing accentual and macro-prosodic changes. This resulted in enhanced speech synthesis of basketball commentaries, the preference for the improved synthesizer in terms of expressivity reaching 51 % against 33 % for the baseline.
- ✓ We indicated that it may not be necessary to explicitly distinguish between different types of emphatic stresses in French, when annotating accentual patterns with expressive HMM-based speech synthesis in mind.

- ✓ Based on the combination of existing prosody annotation techniques, we proposed an automatic prosody annotator describing accentual patterns in French. To our knowledge, no such annotator had been previously developed. The integration of the resulting annotation in an HMM-based speech synthesizer of audiobooks in French achieved significant improvement in terms of naturalness, being preferred in 59 % of the sentences against 31 % for the baseline.
- ✓ We indicated that phonetic modifications significantly impacted the naturalness of the resulting synthesized voice, in the framework of speech synthesis of sports commentaries, while this had been neglected in most current studies on expressive speech synthesis. Preferences for the real phonetizations achieved 45 % against 31 % for the baseline. Liquid and [ə] elisions were shown to play a crucial role in that respect.
- ✓ We provided an in-depth linguistic study of phonotactic phenomena in French, investigating *simultaneously* a wide range of linguistic and extralinguistic factors. This should serve as a preliminary basis for the definition of phonetic rewrite rules for spontaneous speech synthesis.
- ✓ We questioned the assumption of consistency, stating that phonetization errors in the annotation of the training corpus should not affect the synthesized speech as long as the same errors are made at synthesis. We have shown that this does not hold if a large amount of phonetic errors is made, notably in the case of spontaneous speech synthesis. This highlights the importance of improving the phonetization of such corpora.
- ✓ We developed available online tools to allow for HMM-based expressive speech synthesis in French:
 - Train&Align (http://cental.fltr.ucl.ac.be/train_and_align/) is an online platform to align any corpus with its phonetic transcription. It does not require manually-annotated data and works on any language and speaking style.
 - eLite-HTS (<http://cental.uclouvain.be/elitehts/v1/>) is a web service which provides a complete front-end for French. It was realized in collaboration with Sophie Roekhaut and derived from eLite as developed by the Multitel Group. It automatically transforms the text into HTS labels (with all required linguistic annotation levels) for HMM-based speech synthesis and can be combined with Train&Align. It also allows for the integration of prosodic information at the syllable-level and accepts a manually-checked phonetic transcription as input, which facilitates its use for HMM-based expressive and/or spontaneous speech synthesis.

Several publications resulted from this PhD thesis and are listed in Appendix G.

12.2 Limitations and perspectives

The current study presents a certain number of limitations which should lead to complementary studies to validate and/or generalize the presented results:

- In Chapter 6, the integration of global labels alone was shown to provide no improvement in the naturalness and expressivity of the resulting synthesized voice. As pointed out in Chapter 8 this may partly be due to the highly subjective manual annotation of our corpus with macro-prosodic labels. A relabelling of the database by several annotators, with precise guidelines indicating the importance to consider acoustic features only, may improve the resulting synthesis. Besides, the *ExRise* sub-genre may need to be modeled differently. It could benefit from a gradual interpolation from a neutral to a very excited model, to rightly reproduce the gradual rise in excitation.
- In Chapter 9, the study of the role played by phonetic variation in the synthesis of sports commentaries was only assessed in terms of naturalness and expressivity of the synthesized voice. We believe that further experiments should also evaluate the resulting intelligibility, which is known to be sensitive to phonetic modifications (Larreur and Sorin, 1990).
- The corpus of audiobooks used to validate our automatic prosody annotator in Chapter 10 was not ideally suited for that task. Post-hoc analyzes have indeed indicated that:
 - It contains many whispered and creaky segments which prevent from the extraction of tonal information
 - It displays very complex prosodic patterns which may seem unnatural, especially when generated by a synthetic voice
 - It is not as expressive as expected, being uttered with a rather solemn voice

The annotation process presented in that chapter should then be validated on another expressive corpus, as the one used in Part II.

- The statistical analysis of phonetic variations in French proposed in Chapter 11 only provided preliminary results. A more detailed analysis of word-final post-obstruent liquid deletion indicated the benefit of using Generalized Linear Mixed Models in that framework to integrate both speakers and items as random effects. It also included a wide range of linguistic and extralinguistic factors. Such analyses should be carried out on the other phonotactic phenomena which were shown to positively impact synthesized speech, namely [ə] elision and liquid elision in the ‘il’ pronoun.
- In Chapter 11 we have explored the consistency assumption by comparing synthesis based on an automatic phonetization at synthesis and training, with synthesis based on an improved phonetization of the training corpus only. To provide further insight into that question, tests should also include a version in which improved phonetization is integrated at both training and synthesis stages.

As research is an ongoing process, our study also gave rise to new research questions, thereby providing directions for future work. In line with our main research topic, perspectives are first proposed in terms of annotation of the training corpus. We then provide avenues for future work with respect to the annotation of the new sentences to synthesize. As mentioned in the introduction, our research did not focus on that second stage of the synthesis, even though the improved annotation proposed at training was developed with the objective to be easily predictable from text. For each topic, we propose two main research questions which we briefly discuss.

It should also be noted that more generic perspectives include the evaluation of our improvements on other languages and speaking styles. We have for instance indicated that the various acoustic realizations of the emphatic stress in French may not need to be explicitly annotated in the training corpus, with speech synthesis in mind. Further studies should be carried out to investigate whether this is applicable to other languages.

Another interesting research direction is the application of our findings, and mainly our automatic accentual annotator, to speech recognition. Recognizers may indeed benefit from prosodic information such as contrast or emphasis in the speaker's message¹.

1. Annotation of the training corpus

RESEARCH QUESTION 1. How could we improve the automatic phonetization of the training corpus in the framework of spontaneous (expressive) speech synthesis ? How could this be integrated in our forced alignment tool ?

In Chapter 11, we have indicated that improving the phonetization of spontaneous training corpora resulted in enhanced naturalness of the synthesized voice. We have, in that framework, proposed a very basic improvement method which integrated pre-encoded phonetic variants and expert-driven rules to produce phonetic lattices which were considered during the forced alignment. These lattices were however shown to solve only a small part of the phonetization errors. Further work should investigate how this phonetization could be improved by (i) selecting the right variants to integrate in the lattices and (ii) associating probabilities to the phonetic variants. A selective choice of the variants to integrate, according to the speaking style or other extralinguistic factors, should improve the alignment, as it has been shown that including too many variants degrades the alignment (Kim et al., 2004). Associated probabilities could benefit from an in-depth statistical analysis of linguistic and extralinguistic factors as proposed in Chapter 11 for French. The a

¹An interesting work in that respect is that of Geoffrois (1995) who showed that the consideration accentual information (by means of Mertens's (1987) tones) in French speech recognition helped in solving syntactic ambiguities. In English, Shriberg and Stolcke (2004) also pointed out that integrating prosodic information in ASR helps in sentence and topic segmentation but also dialog act labeling, besides improving word recognition rates.

GENERAL CONCLUSION

priori probability of a variant would be determined on the basis of its linguistic and extralinguistic contextual features.

The resulting improved phonetization modeling could be, in a second stage, directly integrated in the eLite-HTS front-end. Modifications of our alignment tool Train&Align should then be carried out to allow for the introduction of phonetic variants (possibly with weights) in the forced alignment procedure.

RESEARCH QUESTION 2. How could global macro-prosodic changes be automatically annotated in any new training corpus ?

In Chapter 10 we have proposed an automatic annotator of local (accentual) prosody annotation in French. We have however not dealt with the question of automatically annotating macro-prosodic labels from speech. As pointed out in Chapter 8, this objective first requires defining minimal units. We suggested to rely on rhythmic groups but more adequate units may be defined, based on the detection of major prosodic changes in speech. In that respect, De Looze (2010b,a) developed ADoReVA and ADoTeVa, two tools that automatically identify major changes in terms of register and tempo, respectively. Acoustically-similar speech segments could then be automatically clustered based on Zhao et al. (2006); Székely et al. (2011); Eyben et al. (2012). The main questions that should then be addressed are :

- How can we determine the optimal number of clusters (and resulting sub-genres) in a given corpus ?
- How can we assign specific functions to the sub-genres to allow for their prediction at synthesis ?
- Does the integration of such sub-genres in HMM-based speech synthesis results in significant enhancement of the synthesis ?

In a first stage, studies could focus on the automatic annotation in a fixed number of excitation degrees only, based on acoustic features of excitation emphasized in Chapter 6 and 8. This would provide an easy answer to the two first questions to concentrate on the integration of such labels in speech synthesis.

2. Annotation of the sentences to synthesize

RESEARCH QUESTION 3. How could we automatically annotate any sentence to synthesize in terms of local accentual prosody ?

The integration of automatic accentual prosody annotation proposed in Chapter 10 was shown to improve the resulting synthesis. The experiments however relied on the fact that the test sentences had been previously recorded and could then be

automatically annotated with that protocol, based on their acoustic realization. In an attempt to reduce the set of labels, we have proposed a functional definition of each label, to facilitate their prediction from text. The integration of this reduced set of labels in synthesis achieved however no improvement over the baseline.

Further research should be carried out in that respect. It could rely on functional analyses of French accentual patterns, as provided in Mertens (2008). Expert-driven rules could then be defined to automatically predict the labels from the linguistic annotation of the text provided by current front-ends in French, such as eLite-HTS. They should determine whether these rules should rely on more in-depth syntactic analyses of the text, as proposed in Mertens et al. (2001b).

RESEARCH QUESTION 4. How could we efficiently predict phonetic variations for new sentences to synthesize according to the speaking style ?

This constitutes, in our view, the core perspective of our work. This thesis has highlighted that integrating phonetic variation in the annotation of both the training corpus and the sentences to synthesize significantly improved the naturalness of the synthesized speech, in the case of conversational speech in English and sports commentaries in French, respectively. Here again, the phonetization of the test sentences benefitted from the fact that they had been previously recorded. The definition of post-phonetic modification rules still remains a major challenge and should consider various linguistic and extralinguistic factors, as pointed out in Chapter 11. Most linguistic information could be automatically obtained by the front-end component.

Studies in that perspective could rely on Miller's work (Miller et al., 1997; Miller, 1998b,a) who investigated the prediction of speaker-specific phonetic variants. He proposed to automatically learn phonetic post-lexical rules based on the training of a neural network on aligned lexical and post-lexical² English pronunciations. The model is based on a contextual window of nine phones and a set of features for each phone including the distance to the word, phrase, clause and sentence boundaries. An interesting aspect of that work is that it makes use of a "recurrent buffer" to take into account the previously emitted post-lexical phones. In Miller (1998b), stress and accent information is also considered by the model. Other studies implementing phonetic modification in speech synthesis include Jande (2008) for Swedish (based on the training of a decision tree on information from various linguistic level) and Van Bael et al. (2004) for Dutch who automatically learns context-sensitive optional rewrite rules. However, the latter rely on the phonetic context only while we believe that other linguistic levels, such as stress information and rhythmic group boundaries should also be considered, potentially with the use of multilevel

²Lexical pronunciation as coded in the dictionary v.s. post lexical variation which occurs in connected speech.

GENERAL CONCLUSION

rewrite rules (Boite et al., 2000). Regrettably, none of these studies has evaluated the perceived naturalness of the resulting speech synthesis. Only the integration of phonetic reduction in Swedish was assessed in Jande (2003a,b) for speech syntheses with various speaking rates. Perceptive tests highlighted that the reduced version is perceived as significantly more natural in fast synthesized speech.

We believe that studies on this topic should first focus on spontaneous speech synthesis, as this speaking style was shown to display a high amount of phonetic variation. It should be noted that the training of post-phonetic rules in French could benefit from the rich manual annotation of the CPROM-PFC corpus as described in Chapter 11.

It should finally be underlined that both the prediction of local prosody information and of phonetic changes from text should obviously benefit from semantic analyses of the text to synthesize. While concept-to-speech systems could rely on information provided by the NLG component, text provided to simple TTS systems may require to go through a first set of semantic analysis modules, *e.g.* to detect topic changes and better predict where emphatic stresses (and potentially resulting phonetic changes) should be predicted.

Bibliography

- Abe, M. (1997). Speaking styles: Statistical analysis and synthesis by a text-to-speech system. In Van Santen, J., Sproat, R., Olive, J., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 495–510. Springer-Verlag. (Quoted on page 110.)
- Adda-Decker, M. (2007). Problèmes posés par le schwa en reconnaissance et en alignement automatiques de la parole. In *Proc. of the 5èmes Journées Linguistiques de Nantes*, pages 211–216. (Quoted on page 87.)
- Adda-Decker, M., Boula de Mareüil, P., and Lamel, L. (1999). Pronunciation variants in French: schwa and liaisons. In *Proc. of the International Congress of Phonetic Sciences (ICPhS)*, pages 2239–2242. (Quoted on pages 202, 209 and 262.)
- Adda-Decker, M. and Lamel, L. (1999). Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29(2):83–98. (Quoted on pages 201 and 278.)
- Adda-Decker, M. and Snoeren, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, 39(3):261–270. (Quoted on page 202.)
- Adell, J., Bonafonte, A., and Escudero-Mancebo, D. M. (2010). Modelling filled pauses prosody to synthesise disfluent speech. In *Proc. of Speech Prosody*, pages 1–4. (Quoted on pages 257 and 258.)
- Adell, J., Bonafonte, A., Gómez, J. A., and Castro, M. J. (2005). Comparative study of automatic phone segmentation methods for TTS. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 309–312. (Quoted on pages 25, 27, 31, 32, 52 and 56.)
- Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience, New York, 2nd edition. (Quoted on page 192.)
- Allen, J., Hunnicutt, M. S., Klatt, D. H., Armstrong, R. C., and Pisoni, D. B. (1987). *From Text to Speech, The MITALK System*. Cambridge University Press. (Quoted on pages 83 and 104.)
- Amir, N., Kerret, O., and Karlinski, D. (2001). Classifying emotions in speech: A comparison of methods. In *Proc. of Interspeech*, pages 127–130. (Quoted on page 187.)

BIBLIOGRAPHY

- Ananthakrishnan, S. and Narayanan, S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio Speech and Language Processing*, 16(1):216–228. (Quoted on pages 146, 181, 229, 235 and 239.)
- Anderson, M., Pierrehumbert, J., and Liberman, M. (1984). Synthesis by rule of English intonation patterns. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 77–80. (Quoted on page 108.)
- Andersson, S., Georgila, K., Traum, D., Aylett, M., and Clark, R. (2010a). Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Proc. of Speech Prosody*, pages 1–4, Chicago (USA). (Quoted on page 257.)
- Andersson, S., Yamagishi, J., and Clark, R. (2010b). Utilising spontaneous conversational speech in HMM-based speech synthesis. In *Proc. of the 7th ISCA Workshop on Speech Synthesis (SSW7)*, pages 173–178. (Quoted on page 257.)
- Andersson, S., Yamagishi, J., and Clark, R. (2012). Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication*, 54(2):175–188. (Quoted on page 257.)
- André, E., Herzog, G., and Rist, T. (1988). On the simultaneous interpretation of real world image sequences and their natural language description: the system SOCCER. In *Proc. of the 8th European Conference on Artificial Intelligence (ECAI)*, pages 449–454. (Quoted on page 131.)
- Argod-Dutard, F. (1996). *Eléments de phonétique appliquée*. Armand Colin, Paris. (Quoted on page 212.)
- Armstrong, N. (1996). Variable deletion of French /l/: Linguistic, social and stylistic factors. *Journal of French Language Studies*, 6(1):1–22. (Quoted on pages 86 and 262.)
- Armstrong, N. (2001). *Social and stylistic variation in spoken French: A comparative approach*. John Benjamins, Amsterdam/New York. (Quoted on page 262.)
- Astesano, C. (1999). *Rythme et discours : Invariance et sources de variabilité des phénomènes accentuels en français*. PhD thesis, Université d’Aix-Marseille I. (Quoted on page 90.)
- Audrit, S., Pršir, T., Auchlin, A., and Goldman, J.-P. (2012). Sport in the media: A contrasted study of three sport live media reports with semi-automatic tools. In *Proc. of Speech Prosody*, pages 127–130. (Quoted on pages 143, 153 and 204.)
- Avanzi, M. (2012). *L’interface prosodie/syntaxe en français*. Peter Lang, Brussels. (Quoted on pages 89 and 103.)
- Avanzi, M. (2014). A corpus-based approach to French regional prosodic variation. *Nouveaux Cahiers de Linguistique Française (Proceedings of the 3rd Swiss Workshop in Prosody)*, 31:309–323. (Quoted on pages 226, 240, 263 and 264.)

- Avanzi, M. and Brognaux, S. (2015). A multifactorial analysis of French liaisons. In *Proc. of Phonetics and Phonology in Europe (PAPE 2015)*, Cambridge. (Quoted on pages 261, 262, 263, 267, 270 and 271.)
- Avanzi, M., Lacheret-Dujour, A., and Victorri, B. (2008). Analor. A tool for semi-automatic annotation of French prosodic structures. In *Proc. of Speech Prosody*, pages 119–122. (Quoted on pages 230 and 239.)
- Avanzi, M., Obin, N., Lacheret, A., and Victorri, B. (2011). Toward a continuous modeling of French prosodic structure: Using acoustic features to predict prominence location and prominence degree. In *Proc. of Interspeech*, pages 2033–2036. (Quoted on pages 230 and 239.)
- Avanzi, M., Simon, A. C., Goldman, J. P., and Auchlin, A. (2010). C-PROM. An annotated corpus for French prominence study. In *Proc. of Speech Prosody*. (Quoted on pages 204, 234 and 264.)
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412. (Quoted on page 270.)
- Badino, L., Andersson, J., Yamagishi, J., and Clark, R. (2009). Identification of contrast and its emphatic realization in HMM-based speech synthesis. In *Proc. of Interspeech*, pages 520–523. (Quoted on pages 124, 137 and 169.)
- Bailly, G. (1989). Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, 8(2):137–146. (Quoted on page 94.)
- Bänziger, T. and Scherer, K. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46(3):252–267. (Quoted on page 110.)
- Bardiaux, A. (2014). *La prosodie de quelques variétés de français en Belgique. Analyse perceptive et acoustique*. PhD thesis, Université catholique de Louvain (UCL). (Quoted on pages 92, 94 and 99.)
- Bartkova, K., Delais-Roussarie, E., and Santiago-Vargas, F. (2012). PROSOTRAN: A tool to annotate prosodically non-standard data. In *Proc. of Speech Prosody*, pages 55–58. (Quoted on pages 181 and 231.)
- Batliner, A., Steidl, S., Seppi, D., and Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach. *Advances in Human-Computer Interaction*. (Quoted on pages 4, 186 and 187.)
- Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41 (1):164–171. (Quoted on pages 29 and 33.)

BIBLIOGRAPHY

- Beaufort, R. (2008). *Application des Machines à Etats Finis en Synthèse de la Parole. Sélection d'unités non-uniformes et Correction orthographique*. PhD thesis, Facultés Universitaires Notre-Dame de la Paix (FUNDP), Namur. (Quoted on pages 127, 183 and 187.)
- Beaufort, R., Dutoit, T., and Pagel, V. (2002). Analyse syntaxique du français. Pondération par trigrammes lissés et classes d'ambiguïtés lexicales. In *Proc. of the Journées d'étude de la Parole (JEP)*, pages 133–136. (Quoted on page 127.)
- Beaufort, R. and Ruelle, A. (2006). eLite : Système de synthèse de la parole à orientation linguistique. In *Proc. of the Journées d'étude sur la parole (JEP)*, pages 509–512. (Quoted on page 126.)
- Béchet, F. (2001). LIA PHON: un système complet de phonétisation de textes. *Traitement automatique des langues*, 42(1):47–67. (Quoted on page 125.)
- Beckman, M. and Hirschberg, J. (1994). *The ToBI annotation conventions*. Ohio State University. (Quoted on page 97.)
- Beckman, M. E. (1992). Evidence for speech rhythms across languages. In Tohkura, Y., Vatikiotis-Bateson, E., and Sagisaka, Y., editors, *Speech Perception, Production and Linguistic Structure*, pages 457–463. IOS Press, Amsterdam. (Quoted on page 92.)
- Beckman, M. E. and Pierrehumbert, J. B. (1986). Japanese prosodic phrasing and intonation synthesis. In *Proc. of the Twenty-Fourth Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 173–180. (Quoted on pages 123 and 142.)
- Beller, G. (2007). Influence de l'expressivité sur le degré d'articulation. In *Proc. of the Rencontre Jeunes Chercheurs Parole (RJCP)*, pages 24–27. (Quoted on page 40.)
- Beller, G. (2010). Espresso: Transformations of expressivity in speech. In *Proc. of Speech Prosody*. (Quoted on pages 3 and 4.)
- Bennett, C. and Black, A. (2005). Prediction of pronunciation variations for speech synthesis: A data-driven approach. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 297–300. (Quoted on pages 88, 202, 257, 258 and 260.)
- Bertenstam, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., de Serpa-Leitão, A., and Ström, N. (1995). The WAXHOLM application database. In *Proceedings of Eurospeech*, pages 833–836. (Quoted on page 29.)
- Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., and Rauzy, S. (2008). Le CID - corpus of interactional data. annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, 49(3):1–30. (Quoted on page 41.)

- Biber, D. and Conrad, S. (2009). *Register, Genre and Style*. Cambridge University Press. (Quoted on page 204.)
- Bigi, B. (2012). The SPPAS participation to Evalita 2011. In *Proc. of the Workshop on Evaluation of NLP and speech tools for Italian (EVALITA)*, pages 1–6. (Quoted on page 41.)
- Bigi, B. (2014). The SPPAS participation to Evalita 2014. In *Proc. of the Workshop on Evaluation of NLP and speech tools for Italian (EVALITA)*. (Quoted on page 41.)
- Bigi, B. and Hirst, D. (2012). Speech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In *Proc. of Speech Prosody*, pages 1–4. (Quoted on pages 25, 27, 29, 36, 41, 43, 51, 61 and 65.)
- Binnenpoorte, D., Cucchiarini, C., Strik, H., and Boves, L. (2004). Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modelling. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 681–684. (Quoted on page 278.)
- Black, A. (2003). Unit selection and emotional speech. In *Proc. of Eurospeech*, pages 1649–1652. (Quoted on pages 8 and 119.)
- Black, A., Bunnell, H., Dou, Y., Muthukumar, P., Metze, F., Perry, D., Polzehl, T., Prahallad, K., Steidl, S., and Vaughn, C. (2012). Articulatory features for expressive speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4005–4008. (Quoted on page 121.)
- Black, A. and Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. In *Proc. of Eurospeech 1995*, pages 581–584. (Quoted on page 123.)
- Black, A. and Hunt, A. (1996). Generating F0 contours from ToBI labels using linear regression. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 1385–1388. (Quoted on page 108.)
- Black, A., Taylor, P., Caley, R., Clark, R., Richmond, K., King, S., Strom, V., and Zen, H. (2001). The Festival speech synthesis system, version 1.4.2. Unpublished document available at <http://www.festvox.org/festival>. (Quoted on pages 118, 125 and 276.)
- Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1229–1232. (Quoted on pages 3 and 7.)
- Blanche-Benveniste, C., Delofoeu, J., Stefanini, J., and Van de Eynde, K. (1984). *Pronom et syntaxe. L’approche pronominale et son application au français*. SELAF, Paris. (Quoted on page 107.)
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott international*, 5(9/10):341–345. (Quoted on pages 24, 40, 128, 148 and 264.)

BIBLIOGRAPHY

- Boite, R., Boulard, H., Dutoit, T., Hancq, H., and Leich, H. (2000). *Traitement de la parole*. Presses Polytechniques et Universitaires Romandes. (Quoted on pages 80, 88, 103, 105 and 298.)
- Boughton, Z. (2013). Locating variation in French: Geolinguistic patterns, levelling and the ‘french exception’. In Jones, M. C. and Hornsby, D. C., editors, *Language and Social Structure in Urban France*, pages 119–132. Legenda, Oxford. (Quoted on page 262.)
- Boughton, Z. (2014). Social class, cluster simplification and following context: Sociolinguistic variation in word-final post-obstruent liquid deletion in French. *Journal of French Language Studies*, 25(1):1–21. (Quoted on pages 86, 262, 264 and 271.)
- Boula de Mareüil, P., Adda-Decker, M., and Gendner, V. (2003). Liaisons in French: a corpus-based study using morpho-syntactic information. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*. (Quoted on pages 201 and 212.)
- Boula de Mareüil, P. and d’Alessandro, C. (1998). Text chunking for prosodic phrasing in French. In *Proc. of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pages 127–132. (Quoted on page 104.)
- Boula de Mareüil, P. and Soulage, B. (2001). Input/output normalisation and linguistic analysis for a multilingual text-to-speech synthesis system. In *Proc. of the 4th ISCA Workshop on Speech Synthesis (SSW4)*. (Quoted on page 105.)
- Bozkurt, B., Doval, B., D’Alessandro, C., and Dutoit, T. (2004). Improved differential phase spectrum processing for formant tracking. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*. (Quoted on page 55.)
- Brandt, A. (1983). Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio tests. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1017–1020. (Quoted on pages 27 and 52.)
- Braunschweiler, N. (2006). The prosodizer - automatic prosodic annotations of speech synthesis databases. In *Proc. of Speech Prosody*, Dresden (Germany). (Quoted on pages 181, 230 and 235.)
- Braunschweiler, N., Gales, M. J., and Buchholz, S. (2010). Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Proc. of Interspeech*, pages 2222–2225. (Quoted on pages 26, 142 and 231.)
- Briony, W. and Isard, S. D. (1997). A key vowel approach to the synthesis of regional accents of English. In *Proc. of Eurospeech*. (Quoted on pages 88 and 203.)
- Brognaux, S. and Avanzi, M. (2015a). Dialectal variation of French schwa in the light of speaker-related factors. In *Proc. of Phonetics and Phonology in Europe 2015 (PAPE 2015)*.

- Brognaux, S. and Avanzi, M. (2015b). Sociophonetics of phonotactic phenomena in French. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*. (Quoted on pages 262, 264 and 266.)
- Brognaux, S. and Drugman, T. (2014a). Phonetic variations: Impact of the communicative situation. In *Proc. of Speech Prosody*, pages 428–432.
- Brognaux, S. and Drugman, T. (2014b). Variations phonétiques: Impact de la situation de communication. *Nouveaux Cahiers de la Linguistique Française*, 31:63–76.
- Brognaux, S. and Drugman, T. (2016). HMM-based speech segmentation: Improvements of fully automatic approaches. *IEEE Transactions on Acoustics Speech and Language Processing*, 24(1):5–15.
- Brognaux, S., Drugman, T., and Beaufort, R. (2012a). Automatic detection and correction of syntax-based prosody annotation errors. In *Proc. of the IEEE Workshop on Spoken Language Technologies (SLT)*, pages 410–415. (Quoted on page 182.)
- Brognaux, S., Drugman, T., and Beaufort, R. (2012b). Automatic phone alignment. A comparison between speaker-independent models and models trained on the corpus to align. *Lecture Notes in Computer Science*, 7614:300–311.
- Brognaux, S., Drugman, T., and Saerens, M. (2014a). Synthesizing sports commentaries: One or several emphatic stresses? In *Proc. of Speech Prosody*, pages 270–274.
- Brognaux, S., Picart, B., and Drugman, T. (2013). A new prosody annotation protocol for live sports commentaries. In *Proc. of Interspeech*, pages 1554–1558. (Quoted on page 39.)
- Brognaux, S., Picart, B., and Drugman, T. (2014b). La synthèse de commentaires sportifs: intégration d’une annotation prosodique à deux niveaux dans un synthétiseur HMM. *Nouveaux Cahiers de la Linguistique Française*, 31:223–236.
- Brognaux, S., Picart, B., and Drugman, T. (2014c). Speech synthesis in various communicative situations: Impact of pronunciation variations. In *Proc. of Interspeech*, pages 1524–1528.
- Brognaux, S., Roekhaut, S., Drugman, T., and Beaufort, R. (2012c). Train&Align: A new online tool for automatic phonetic alignments. In *IEEE Workshop on Spoken Language Technologies (SLT)*, pages 416–421.
- Brognaux, S., Roekhaut, S., Drugman, T., and Beaufort, R. (2014d). Train&Align: Un outil d’alignement phonétique automatique disponible en ligne. In *Proc. of the Journées d’étude sur la parole (JEP)*, pages 412–420.
- Brugnara, F., Falavigna, D., and Omologo, M. (1993). Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, 12(4):357–370. (Quoted on pages 27, 43, 44, 46, 51 and 52.)

BIBLIOGRAPHY

- Burkhardt, F. (2005). Emofilt: The simulation of emotional speech by prosody-transformation. In *Proc. of Interspeech*, pages 509–512. (Quoted on pages 8 and 119.)
- Burkhardt, F. and Stegmann, J. (2009). Emotional speech synthesis: Applications, history and possible future. In *Proc. of the Electronic Speech Signal Processing Conference (ESSV)*. (Quoted on pages 3 and 118.)
- Bürki, A., Ernestus, M., Gendrot, C., Fougeron, C., and Frauenfelder, U. H. (2011). What affects the presence versus absence of schwa and its duration: A corpus analysis of French connected speech. *The journal of the Acoustical Society of America*, 130 (6):3980–3991. (Quoted on pages 86, 201, 202, 209, 261, 263 and 270.)
- Bürki, A., Gendrot, C., Gravier, G., Linarès, G., and Fougeron, C. (2008). Aligement automatique et analyse phonétique: Comparaison de différents systèmes pour l’analyse du schwa. *Traitement Automatique des Langues*, 49(3):165–197. (Quoted on pages 36, 56 and 278.)
- Cabral, J. P., Renals, S., Richmond, K., and Yamagishi, J. (2007). Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In *Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, pages 113–118. (Quoted on page 115.)
- Cabral, J. P., Renals, S., Richmond, K., and Yamagishi, J. (2008). Glottal spectral separation for parametric speech synthesis. In *Proc. of Interspeech*, pages 1829–1832. (Quoted on page 115.)
- Caelen-Haumont, H. and Bel, B. (2003). Subjectivité et émotion dans la prosodie de parole et du chant: Espace, coordonnées et paramètres. In Colletta, J.-M. and Tcherkasoof, A., editors, *Les émotions. Cognition, langage et développement*, pages 187–196. Mardaga. (Quoted on pages 91 and 93.)
- Cahn, J. E. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19. (Quoted on page 119.)
- Campbell, N. (1992). Syllable-based segmental duration. In Bailly, G., Benoit, C., and Sawallis, T., editors, *Talking Machines: Theories, Models and Designs*, pages 211–224. Elsevier Science Publishers. (Quoted on page 107.)
- Campbell, N. (2000). Databases of emotional speech. In *Proc. of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 34–38. (Quoted on page 5.)
- Campbell, N. (2005). Getting to the heart of the matter: Speech as the expression of affect; rather than just text or language. *Language Resources and Evaluation*, 39(1):109–118. (Quoted on page 257.)
- Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Acoustics Speech and Language Processing*, 14(4):1171–1178. (Quoted on pages 2, 3, 5, 9, 122 and 142.)

- Campbell, N. (2007). Towards conversational speech synthesis; lessons learned from the expressive speech processing project. In *Proc. of the 7th ISCA Workshop on Speech Synthesis (SSW7)*, pages 22–27. (Quoted on page 257.)
- Campbell, N. and Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 2417–2420. (Quoted on page 6.)
- Campbell, W. N. (1997). Synthesizing spontaneous speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pages 165–186. Springer, New York. (Quoted on page 275.)
- Campione, E., Hirst, D., and Véronis, J. (2000). Automatic stylisation and symbolic coding of F0: implementations of the INTSINT models. In Botinis, A., editor, *Intonation: Analysis, Modelling and Technology*, pages 185–208. Kluwer Academic Publishers, Dordrech. (Quoted on pages 98, 181 and 231.)
- Campione, E. and Véronis, J. (2001). Etiquetage prosodique semi-automatique des corpus oraux. In *Proc. of the Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 123–132. (Quoted on pages 181, 229 and 231.)
- Candea, M. (2002). Le e d’appui parisien : Statut actuel et progression. In *Proc. of the Journées d’étude sur la parole (JEP)*, pages 185–188. (Quoted on pages 201, 209 and 210.)
- Cangemi, F., Cutugno, F., Ludusan, B., Seppi, D., and Van Compernelle, D. (2011). Automatic speech segmentation for Italian (ASSI) : Tools, models, evaluation and applications. In *Proc. of the 7th Conference of the Italian Association for Speech Sciences (AISV)*, pages 337–344, Lecce (Italy). (Quoted on pages 29, 46, 56 and 65.)
- Carlson, R., Gustafson, K., and Strangert, E. (2006). Modelling hesitation for synthesis of spontaneous speech. In *Proc. of Speech Prosody*. (Quoted on page 257.)
- Carr, P. (2012). *English Phonetics and Phonology: An Introduction*. John Wiley & Sons, Malden/Oxford. (Quoted on page 82.)
- Carton, F. (1974). *Introduction à la phonétique du français*. Bordas, Paris. (Quoted on page 92.)
- Carton, F. (1999). L’épithèse vocalique et son développement en français parlé. *Faits de langue*, 13:35–45. (Quoted on pages 209 and 210.)
- Cernak, M. and Honnet, P.-E. (2015). An empirical model of emphatic word detection. Technical report, IDIAP (EPFL-Report - 209098). (Quoted on page 128.)
- Chambers, J. K. (2009). *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Wiley-Blackwell, Oxford. (Quoted on page 264.)
- Chambers, J. K. and Trudgill, P. (1998). *Dialectology*. Cambridge University Press, Cambridge. (Quoted on page 264.)

BIBLIOGRAPHY

- Charfuelan, M. and Steiner, I. (2013). Expressive speech synthesis in MARY TTS using audiobook data and emotionML. In *Proc. of Interspeech*, pages 1564–1568. (Quoted on page 231.)
- Charonnat, L., Vidal, G., and Boeffard, O. (2008). Automatic phone segmentation of expressive speech. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 2376–2379. (Quoted on page 29.)
- Chen, K., Hasegawa-Johnson, M., and Cohen, A. (2004a). An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic models. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 509–512. (Quoted on pages 181 and 229.)
- Chen, L., Liu, Y., Harper, M., Maia, E., and McRoy, S. (2004b). Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 759–762. (Quoted on page 29.)
- Chevrot, J.-P., Beaud, L., and Varga, R. (2000). L’apprentissage des unités phonologiques variables : l’exemple du /R/ post-consonantique final en français. *Linux*, 42:89–100. (Quoted on page 271.)
- Christodoulides, G. and Avanzi, M. (2014). An evaluation of machine learning methods for prominence detection in French. In *Proc. of Interspeech*, pages 116–119. (Quoted on pages 183, 230, 239 and 240.)
- Christodoulides, G., Avanzi, M., and Goldman, J.-P. (2014). DisMo: A morphosyntactic, disfluency and multi-word unit annotator. An evaluation on a corpus of French spontaneous and read speech. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 3902–3907. (Quoted on page 264.)
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. (Quoted on page 149.)
- Colotte, V. and Beaufort, R. (2005). Linguistic features weighting for a text-to-speech system without prosody model. In *Proc. of Interspeech*, pages 2549–2552. (Quoted on pages 42, 62, 105 and 126.)
- Content, A., Mousty, P., and Radeau, M. (1990). BRULEX. Une base de données lexicales informatisée pour le français écrit et parlé. *L’année Psychologique*, 90(4):551–566. (Quoted on pages 83 and 206.)
- Cooke, M., King, S., Garnier, M., and Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer, Speech and Language*, 28(2):543–571. (Quoted on page 201.)
- Cosi, P., Falavigna, D., and Omologo, M. (1991). A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In *Proc. of Eurospeech*, pages 693–696. (Quoted on page 32.)

- Côté, M.-H. (2004). Consonant cluster simplification in Québec French. *Probus*, 16(2):151–201. (Quoted on page 86.)
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80. (Quoted on page 148.)
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press. (Quoted on page 175.)
- Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge University Press. (Quoted on page 96.)
- Daelemans, W. and Van den Bosch, A. (1993). Tabtalk: Reusability in data-oriented grapheme-to-phoneme conversion. In *Proc. of Eurospeech*, pages 1459–1466. (Quoted on page 84.)
- d’Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer, Speech and Language*, 9(3):257–288. (Quoted on pages 95, 149 and 234.)
- Dall, R., Yamagishi, J., and King, S. (2014). Rating naturalness in speech synthesis: The effect of style and expectation. In *Proc. of Speech Prosody*. (Quoted on pages 5, 226, 257, 275 and 283.)
- De Looze, C. (2010a). ADoReVA and ADoTeVA, two PRAAT plugins for the automatic detection of register and tempo variations. In *Sao Paulo School of Advanced Studies in Speech Dynamics*. (Quoted on pages 189 and 296.)
- De Looze, C. (2010b). *Analyse et interprétation de l’empan temporel des variations prosodiques en français et en anglais*. PhD thesis, Laboratoire Parole et Langage, UMR 6057 CNRS, Université de Provence, Aix-en-Provence, France. (Quoted on pages 109, 154, 189 and 296.)
- de Moraes, J. A. and Rilliard, A. (In Press). Prosody and emotion. In Armstrong, M., Henriks, N., and del Mar, M., editors, *Interdisciplinary approaches to intonational grammar in Ibero-Romance*. John Benjamins Publisher, Amsterdam. (Quoted on page 93.)
- de Reuse, J. W. (1987). La phonologie du français de la région de Charleroi (Belgique) et ses rapports avec le wallon. *La linguistique*, 23:99–115. (Quoted on pages 86 and 212.)
- Delais-Roussarie, E. (1995). *Pour une approche probabiliste de la structure prosodique, étude de l’organisation prosodique et rythmique de la phrase française*. PhD thesis, Université Toulouse-le-Mirail. (Quoted on page 89.)
- Delais-Roussarie, E. (1996). Phonological phrasing and accentuation in French. In Nespor, M. and Smith, N., editors, *Dan Phonology*, pages 1–38. Holland Academic Graphics, La Haye. (Quoted on page 264.)

BIBLIOGRAPHY

- Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldgausen, I., Jun, S.-A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R., and Yoo, H.-Y. (2015). Intonational phonology of French: Developing a ToBI system for French. In Frota, S. and Prieto, P., editors, *Intonational variation in Romance*, pages 63–100. Oxford University Press. (Quoted on pages 98, 241 and 248.)
- Delais-Roussarie, E., Yoo, H., and Post, B. (2011). Quand frontières prosodiques et frontières syntaxiques se rencontrent. *Langue française*, 170(2):29–44. (Quoted on pages 89 and 103.)
- Delattre, P. (1947). La liaison en français. Tendances et classifications. *The French Review*, 12(2):148–157. (Quoted on pages 261 and 265.)
- Delattre, P. (1966). Les dix intonations de base du français. *The French Review*, 40:1–14. (Quoted on pages 96 and 97.)
- Dell, F. (1985). *Les règles et les sons*. Hermann, Paris, 2nd edition. (Quoted on page 269.)
- Dell, F. (1995). Consonant clusters and phonological syllables in French. *Lingua*, 95(1):5–26. (Quoted on page 86.)
- Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 1970–1973. (Quoted on pages 39, 42, 61 and 62.)
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithms. *Journal of the Royal Statistical Society*, 39(1):1–38. (Quoted on page 111.)
- Demuyne, K. and Laureys, T. (2002). A comparison of different approaches to automatic speech segmentation. In *Text, Speech and Dialogue*, pages 277–284. Springer Berlin Heidelberg. (Quoted on pages 46, 52 and 54.)
- Demuyne, K., Roelens, J., Van Compernelle, D., and Wambacq, P. (2008). SPRAAK: An open source "Speech Recognition and Automatic Annotation Kit". In *Proc. of Interspeech*, page 495. (Quoted on pages 24 and 65.)
- Deroo, O., Dutoit, T., Malfrère, F., and Ris, S. (2003). Phonetic alignment: Speech synthesis-based vs. Viterbi-based. *Speech Communication*, 40:503–515. (Quoted on page 27.)
- Devillers, L. and Vasilescu, I. (2003). Prosodic cues for emotion characterization in real-life spoken dialogs. In *Proc. of Interspeech*. (Quoted on page 110.)
- Di Cristo, A. (1975). Recherches sur la structuration prosodique de la phrase française. In *Proc. of the Journées d'étude sur la parole (JEP)*, pages 95–116. (Quoted on page 97.)
- Di Cristo, A. (1985). *De la microprosodie à l'intonosyntaxe*. PhD thesis, Université de Provence, Aix-en-Provence. (Quoted on page 171.)

- Di Cristo, A. (1999). Vers une modélisation de l’accentuation du français : Première partie. *Journal of French Language Studies*, 9(2):143–179. (Quoted on pages 90, 92, 96 and 146.)
- Di Cristo, A. (2000). Vers une modélisation de l’accentuation du français : Deuxième partie. *Journal of French Language Studies*, 10(1):27–44. (Quoted on pages 90, 91, 146 and 147.)
- Di Cristo, A. and Hirst, D. (1996). L’accentuation non emphatique en français: Stratégies et paramètres. In Perrot, J., editor, *Polyphonie pour Ivan Fonagy*, pages 71–102. L’Harmattan, Paris. (Quoted on page 90.)
- Digalakis, V., Rtischev, D., and Neumeyer, L. (1995). Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3(5):357–366. (Quoted on page 159.)
- D’Imperio, M., Bertrand, R., Di Cristo, A., and Portes, C. (2007). Investigating phrasing levels in French : Is there a difference between nuclear and prenuclear accents ? In Camacho, J., Deprez, V., Flores, N., and Sanchez, L., editors, *Selected Papers from the 36th Linguistic Symposium on Romance Languages (LSRL)*, pages 97–110. John Benjamins Publishing Company. (Quoted on pages 89, 90 and 92.)
- Dister, A. and Simon, A. C. (2008). La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé. *Arena Romanistica*, 1(1):54–79. (Quoted on page 24.)
- Doukhan, D. (2013). *Synthèse de parole expressive au delà du niveau de la phrase : Le cas du conte pour enfant*. PhD thesis, Université Paris Sud. (Quoted on page 231.)
- Drugman, T. and Alwan, A. (2011). Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proc. of Interspeech*, pages 1973–1976. (Quoted on pages 55, 150 and 170.)
- Drugman, T., Dubuisson, T., and Dutoit, T. (2011). Phase-based information for voice pathology detection. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4612–4615. (Quoted on page 55.)
- Drugman, T. and Dutoit, T. (2012). The deterministic plus stochastic model of the residual signal and its applications. *IEEE Transactions on Audio, Speech and Language Processing*, 20(3):968–981. (Quoted on pages 7, 115 and 155.)
- Drugman, T., Kane, J., and Gobl, C. (2012). Resonator-based creaky voice detection. In *Proc. of Interspeech*, pages 1592–1595. (Quoted on page 147.)
- Durand, J., Laks, B., Calderone, B., and Tchobanov, A. (2011). Que savons-nous de la liaison aujourd’hui ? *Langue Française*, 169:103–135. (Quoted on page 267.)
- Durand, J., Laks, B., and Lyche, C. (2009). *Phonologie, variation et accents du français*. Hermes, Paris. (Quoted on pages 240 and 263.)

BIBLIOGRAPHY

- Durand, J. and Lyche, C. (2008). French liaison in the light of corpus data. *Journal of French Language Studies*, 18:33–66. (Quoted on pages 262, 265 and 267.)
- Dusterhoff, K., Black, A., and Taylor, P. (1999). Using decision trees within the Tilt intonation model to predict F0 contours. In *Proc. of Eurospeech*. (Quoted on page 108.)
- Dutoit, T. (1997a). High-quality text-to-speech synthesis: An overview. *Journal of Electrical and Electronics Engineering Australia*, 17:25–36. (Quoted on page 80.)
- Dutoit, T. (1997b). *An Introduction to Text-to-Speech synthesis*. Kluwer Academic Publishers. (Quoted on pages 80, 83, 84, 85, 88, 94, 97, 98, 101, 103 and 104.)
- Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and Van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 1393–1396. (Quoted on pages 6 and 126.)
- Dutoit, T. and Stylianou, Y. (2003). Text-to-speech synthesis. In Mitkov, R., editor, *Handbook of Computational Linguistics*, pages 323–338. (Quoted on page 81.)
- Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., and Pitrelli, J. (2004). A corpus-based approach to expressive speech synthesis. In *Proc. of the 5th ISCA Workshop on Speech Synthesis (SSW5)*, pages 79–84. (Quoted on page 119.)
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200. (Quoted on pages 4 and 122.)
- El Ayadi, M., Kamel, S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587. (Quoted on page 186.)
- Encrevé, P. (1988). *La liaison avec et sans enchaînement: Phonologie tridimensionnelle et usages du français*. Seuil, Paris. (Quoted on pages 86 and 262.)
- Eyben, F., Bucholz, S., Braunschweiler, N., Latorre, J., Wan, V., Gales, M., and Knill, K. (2012). Unsupervised clustering of emotion and voice styles for expressive TTS. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4009–4012. (Quoted on pages 8, 9, 120, 121, 122, 123, 142, 144, 158, 159, 186, 187, 189, 231, 244 and 296.)
- Fackrell, J., Skut, W., and Hammervold, K. (2003). Improving the accuracy of pronunciation prediction for unit selection TTS. In *Proc. of Interspeech*, pages 2473–2476. (Quoted on page 275.)
- Ferguson, J. (1980). Variable duration models for speech. In *Proc. of the Symposium on Application of Hidden Markov Models to Text and Speech*, pages 143–179. (Quoted on page 159.)

- Fernandez, R. and Ramabhadran, B. (2007). Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis. In *Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, pages 34–39. (Quoted on pages 124 and 169.)
- Fitt, S. (1997). The generation of regional pronunciations of English for speech synthesis. In *Proc. of Eurospeech 97*. (Quoted on pages 88 and 203.)
- Fónagy, I. (1949). *Déplacement de l'accent d'intensité dans les langues romanes (en hongrois)*. PhD thesis, Université L. Eötvös, Budapest. (Quoted on page 93.)
- Fónagy, I. (1979). L'accent français : Accent probabilitaire. In Fónagy, I., Léon, P., and Carton, F., editors, *L'accent en français contemporain*, volume 15 of *Studia Phonetica*, pages 123–232. Marcel Didier Ltée, Ottawa. (Quoted on pages 6, 90, 93, 110 and 142.)
- Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278. (Quoted on page 28.)
- Fougeron, C., Goldman, J.-P., Dart, A., Guélat, L., and Jeager, C. (2001). Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français. In *Proc. of the Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 173–182. (Quoted on pages 86, 201, 203, 212 and 267.)
- Fougeron, C. Goldman, J.-P. and Frauenfelder, U. H. (2001). Liaison and schwa deletion in French: an effect of lexical frequency and competition ? In *Proc. of Interspeech*, pages 639–642. (Quoted on pages 6, 86, 87, 201, 202, 209, 212, 262 and 267.)
- Freund, Z. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139. (Quoted on page 230.)
- Fujisaki, H. (1992). The role of quantitative modeling in the study of intonation. In *Proc. of the International Symposium on Japanese Prosody*, pages 163–174. (Quoted on pages 94 and 121.)
- Fujisaki, H. (2004). Information, prosody, and modeling with emphasis on tonal features of speech. In *Proc. of Speech Prosody*. (Quoted on page 95.)
- Fujisaki, H. and Kawai, H. (1982). Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. In *Proc. of the Working group on Intonation, 13th International Congress of Linguists*. (Quoted on pages 94 and 121.)
- Gales, M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer, Speech and Language*, 12(2):75–98. (Quoted on page 159.)
- Gandour, J. T. (1978). The perception of tone. In Fromkin, V. A., editor, *Tone: A Linguistic Survey*. Academic Press, New York. (Quoted on page 92.)

BIBLIOGRAPHY

- Garde, P. (1968). *L'accent*. Presses Universitaires de France, Paris. (Quoted on page 89.)
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (1990). The DARPA-TIMIT acoustic-phonetic continuous speech corpus. Technical report, NASA STI/Recon. CD-ROM. (Quoted on page 82.)
- Gaudinat, A. and Wehrli, E. (1997). Analyse syntaxique et synthèse de la parole: Le système FipsVox. *Traitement Automatique des Langues*, 38(1):121–130. (Quoted on pages 105 and 107.)
- Gauvain, J. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298. (Quoted on page 116.)
- Geoffrois, E. (1995). *Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole*. PhD thesis, Université Paris XI Orsay. (Quoted on pages 229, 231, 236, 241 and 295.)
- Ghisletta, P. and Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics*, 29(4):421–437. (Quoted on page 266.)
- Gobl, C. and Ni Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1):189–212. (Quoted on page 6.)
- Goldman, J.-P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In *Proc. of Interspeech*, pages 3233–3236. (Quoted on pages 25, 26, 29, 32, 39, 41, 43, 46, 51, 61, 65, 206 and 264.)
- Goldman, J.-P. (2012). ProsoDyn: a graphical representation of macroprosody for phonostylistic ambiance change detection. In *Proc. of Speech Prosody*. (Quoted on pages 110, 143, 153, 154 and 189.)
- Goldman, J.-P., Auchlin, A., Roekhaut, S., Simon, A. C., and Avanzi, M. (2010). Prominence perception and accent detection in French. A corpus-based account. In *Proc. of Speech Prosody*. (Quoted on pages 148, 181, 234 and 264.)
- Goldman, J.-P., Auchlin, A., and Simon, A. C. (2009). Discrimination de styles de parole par analyse prosodique semi-automatique. In *Proc. of the Prosody-Discourse Interface Conference (IDP)*, pages 207–221. (Quoted on pages 110, 143, 201, 202, 204 and 216.)
- Goldman, J.-P., Avanzi, M., Auchlin, A., and Simon, A. C. (2012). A continuous prominence score based on acoustic features. In *Proc. of Interspeech*. (Quoted on pages 151, 152, 169, 171, 183, 230 and 239.)

- Goldman, J.-P., Avanzi, M., Lacheret-Dujour, A., Simon, A. C., and Auchlin, A. (2007). A methodology for the automatic detection of perceived prominent syllables in spoken French. In *Proc. of Interspeech*, pages 2414–2417. (Quoted on pages 151, 171, 183, 189, 215, 230 and 239.)
- Goldman, J.-P. and Schwab, S. (2011). EasyAlign Spanish: An (semi-) automatic segmentation tool under Praat. In *Proc. of 5th Congreso de Fonología Experimental (CFE)*. (Quoted on pages 25, 31, 46 and 66.)
- Gonzalvo, X., Socoró, J. C., Iriondo, I., Monzo, C., and Martínez, E. (2007). Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish. In *Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, pages 362–367. (Quoted on page 115.)
- Goslin, J., Content, A., Goldman, J.-P., and Frauenfelder, U. (1999). Syllable boundary placement in French. Man and machine. A comparison. In *Proc. of the Journées d’Etudes Linguistiques (JEL)*. (Quoted on page 264.)
- Grabe, E. and Post, B. (2002). Intonational variation in the British Isles. In *Proc. of Speech Prosody*, pages 343–346, Aix-en-Provence. (Quoted on page 110.)
- Grammont, M. (1971). *Traité de phonétique*. Delagrave, Paris. (Quoted on page 92.)
- Grancharov, V. and Kleijn, W. (2008). Speech quality assessment. In Benesty, J., Sondhi, M. M., and Huang, Y., editors, *Springer Handbook of Speech Processing*, pages 83–100. Springer. (Quoted on page 156.)
- Greenberg, S., Carvey, H., and Hitchcock, L. (2002). The relation between stress accent and pronunciation variation in spontaneous American English discourse. In *Proc. of Speech Prosody*. (Quoted on page 146.)
- Grice, M., Reyelt, M., Benz Müller, R., Mayer, J., and Batliner, A. (1996). Consistency in transcription and labeling of German intonation with GToBI. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 1716–1719. (Quoted on page 146.)
- Halliday, M. A. K. (1967). *Intonation and Grammar in British English*. Mouton, The Hague. (Quoted on page 96.)
- Hambye, F. (2005). *La prononciation du français contemporain en Belgique. Variation, normes et identités*. PhD thesis, Université catholique de Louvain, Belgique. (Quoted on pages 86, 201, 209 and 262.)
- Hambye, P. and Simon, A. C. (2012). The variation of pronunciation in Belgian French: from segmental phonology to prosody. In Gess, R., Lyche, C., and Meisenburg, T., editors, *Phonological Variation in French: Illustrations from Three Continents*, pages 129–149. John Benjamins, Amsterdam. (Quoted on page 262.)
- Hamza, W., Eide, E., Bakis, R., Picheny, M., and Pitrelli, J. F. (2004). The IBM expressive speech synthesis systems. In *Proc. of Interspeech*, pages 2577–2580. (Quoted on page 119.)

BIBLIOGRAPHY

- Hansen, A. B. (1991). The covariation of [schwa] with style in Parisian French: an empirical study of ‘e caduc’ and pre-pausal [schwa]. In *Proc. of the ESCA Workshop on Phonetics and Phonology of Speaking Styles*. (Quoted on pages 6, 86, 87, 201, 202, 209, 211, 214 and 219.)
- Hansen, A. B. (1994). Etude du e caduc - stabilisation en cours et variations lexicales. *Journal of French Language Studies*, 4:25–54. (Quoted on page 209.)
- Hansen, A. B. (2000). Le e caduc interconsonantique en tant que variable sociolinguistique. Une enquête en région parisienne. *Linx*, 42(1):45–58. (Quoted on page 262.)
- Hansen, A. B. and Hansen, M.-J. (2003). Le schwa prépausal et l’interaction. In Hansen, A. and Hansen, M.-J., editors, *Structures linguistiques et interactionnelles dans le français parlé.*, pages 89–109. Museum Tusculanum Press. (Quoted on pages 209, 210 and 219.)
- Hayes, A. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89. (Quoted on page 192.)
- Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proc. of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79. (Quoted on page 193.)
- Hinterleitner, F., Neitzel, G., Möller, S., and Norrenbrock, C. (2011). An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Proc. of the Blizzard challenge workshop*. (Quoted on page 231.)
- Hirose, K., Eto, M., and Minematsu, N. (2002). Improved corpus-based synthesis of fundamental frequency using generation process model. In *Proc. of Interspeech*, pages 2085–2088. (Quoted on page 108.)
- Hirose, K., Minematsu, N., and Kawanami, H. (2000). Analytical and a preceptual study on the role of acoustic features in realizing emotional speech. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 369–372. (Quoted on pages 6 and 110.)
- Hirose, K., Sato, K., Asano, Y., and Minematsu, N. (2005). Synthesizing of F0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis. *Speech Communication*, 46(3):385–404. (Quoted on pages 9, 121 and 123.)
- Hirose, K., Sato, K., and Minematsu, N. (2004). Emotional speech synthesis with corpus-based generation of F0 contours using generation process model. In *Proc. of Speech Prosody*, pages 417–420. (Quoted on pages 9, 121, 122, 123 and 142.)
- Hirschberg, J. (1990). Accent and discourse in context: Assigning pitch accent in synthetic speech. In *Proc. of the American Association of Artificial Intelligence*, pages 952–957. (Quoted on page 169.)

- Hirschberg, J. (1993). Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63 (1):305–340. (Quoted on page 105.)
- Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1):31–43. (Quoted on page 109.)
- Hirschberg, J. and Rambow, O. (2001). Learning prosodic features using a tree representation. In *Proc. of Eurospeech*, pages 1175–1178. (Quoted on page 105.)
- Hirst, D. (2005). Form and function in the representation of speech prosody. *Speech Communication*, 46:334–347. (Quoted on page 170.)
- Hirst, D. (2007). A Praat plugin for Momel and Intsint with improved algorithms for modelling and coding intonation. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 1233–1236. (Quoted on page 230.)
- Hirst, D. and Di Cristo, A. (1998). *Intonation Systems: A Survey of Twenty Languages*. Cambridge University Press, Cambridge. (Quoted on pages 98 and 230.)
- Hirst, D. and Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 15:75–85. (Quoted on pages 98 and 109.)
- Hitzeman, J., Black, A., Mellish, C., Oberlander, J., Poesio, M., and Taylor, P. (1999). An annotation scheme for concept-to-speech synthesis. In *Proc. of the European Workshop on Natural Language Generation*, pages 59–66. (Quoted on page 134.)
- Hitzeman, J., Black, A. W., Taylor, P., Mellish, C., and Oberlander, J. (1998). On the use of automatically generated discourse-level information in a concept-to-speech synthesis systems. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 2763–2766. (Quoted on page 134.)
- Hornsby, D. (2006). *Redefining Regional French: Koinéization and Dialect Levelling in Northern France*. Legenda, Oxford. (Quoted on page 271.)
- Hosom, J.-P. (2000). *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. PhD thesis, University of Massachusetts at Amherst. (Quoted on pages 27, 52 and 53.)
- Hosom, J.-P. (2002). Automatic phoneme alignment based on acoustic-phonetic modeling. In *Proc. of Interspeech*, pages 357–360. (Quoted on pages 27, 31 and 52.)
- Hosom, J.-P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51(4):352–368. (Quoted on pages 32 and 52.)
- Hovy, D., Anumanchipalli, G. K., Parlikar, A., Vaughn, C., Lammert, A., Hovy, E., and Black, A. (2013). Analysis and modeling of "focus" in context. In *Proc. of Interspeech*, pages 402–406. (Quoted on pages 105 and 169.)

BIBLIOGRAPHY

- Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 373–376. (Quoted on pages 6 and 24.)
- Iida, A., Campbell, N., Iga, S., Higuchi, F., and Yasumura, M. (2000). A speech synthesis system with emotion for assisting communication. In *Proc. of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 167–172. (Quoted on pages 5, 8 and 119.)
- Imai, S., Sumita, K., and Furuichi, C. (1983). Mel log spectrum approximation (MLSA) filter for speech synthesis. *Electronics and Communications in Japan*, 66(2):10–18. (Quoted on page 115.)
- Inanoglu, Z. and Young, S. (2007). A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality. In *Proc. of Interspeech*, pages 490–493. (Quoted on page 119.)
- Izenman, A. (2008). *Modern multivariate statistical techniques*. Springer, New York. (Quoted on page 170.)
- Jande, P.-A. (2003a). Evaluating rules for phonological reduction in Swedish. In *Proc. of Fonetik*, pages 149–152. (Quoted on pages 202, 214, 219, 258, 259, 260 and 298.)
- Jande, P.-A. (2003b). Phonological reduction in Swedish. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 2557–2560. (Quoted on pages 202, 258, 259, 260 and 298.)
- Jande, P.-A. (2008). Spoken language annotation and data-driven modelling of phone-level pronunciation in discourse context. *Speech Communication*, 50(2):126–141. (Quoted on pages 125, 202, 214, 260 and 297.)
- Jarifi, S., Pastor, D., and Rosec, O. (2008). A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. *Speech Communication*, 50(1):67–80. (Quoted on pages 44, 45 and 52.)
- Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press, USA. (Quoted on page 24.)
- Jeon, J. H. and Liu, Y. (2012). Automatic prosodic event detection using a novel labeling and selection method in co-training. *Speech Communication*, 54(3):445–458. (Quoted on pages 146, 181, 229 and 235.)
- Jilka, M., Möhler, G., and Dogil, G. (1999). Rules for the generation of ToBI-based American English intonation. *Speech Communication*, 28(2):83–108. (Quoted on page 108.)
- Johnson, R. and Wichern, D. (2002). *Applied multivariate statistical analysis*. Prentice Hall, 5th edition. (Quoted on page 170.)

- Jun, S.-A. and Fougeron, C. (1995). The accentual phrase and the prosodic structure of French. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 722–725. (Quoted on page 98.)
- Jun, S.-A. and Fougeron, C. (2000). A phonological model of French intonation. In Botinis, A., editor, *Intonation: Analysis, modelling and technology*, pages 209–242. Kluwer, Boston. (Quoted on page 89.)
- Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001a). The effect of language model probability on pronunciation reduction. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 801–804. (Quoted on pages 85 and 285.)
- Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001b). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, 45:229–254. (Quoted on pages 86, 201 and 285.)
- Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing*. Pearson Education International, 2nd edition. (Quoted on pages 87, 88, 89, 92 and 258.)
- Kane, J., Drugman, T., and Gobl, C. (2013). Improved automatic detection of creak. *Computer, Speech and Language*, 27(4):1028–1047. (Quoted on page 184.)
- Katsamanis, A., Black, M., Georgiou, P. G., Goldstein, L., and Narayanan, S. (2011). Sailalign : Robust long speech-text alignment. In *Proc. of the Workshop on New Tools and Methods for Very Large Scale Research in Phonetic Science*, pages 28–31. (Quoted on pages 25, 26, 52 and 65.)
- Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3):187–207. (Quoted on pages 7 and 115.)
- Kawahara, H. and Morise, M. (2011). Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *SADHANA - Academy Proceedings in Engineering Sciences*, 36(5):713–727. (Quoted on pages 7 and 115.)
- Kawai, H. and Toda, T. (2004). An evaluation of automatic phone segmentation for concatenative speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 677–680. (Quoted on pages 24, 29, 36, 44, 45, 46 and 52.)
- Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H., and Shikano, K. (1999). GMM-based voice conversion applied to emotional speech synthesis. In *Proc. of Eurospeech*, pages 2401–2404. (Quoted on page 119.)
- Kern, F. (2010). Speaking dramatically. The prosody of live radio commentary of football matches. In Barth-Weingarten, D., Reber, E., and Selting, M., editors, *Prosody in Interaction*, pages 217–238. John Benjamins. (Quoted on pages 110, 142, 143, 144, 148 and 153.)

BIBLIOGRAPHY

- Kessens, J. M., Wester, M., and Strik, H. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication*, 29(2):193–207. (Quoted on page 278.)
- Kim, Y. J., Syrdal, A., and Jilka, M. (2004). Improving TTS by higher agreement between predicted versus observed pronunciations. In *Proc. of the 5th ISCA Workshop on Speech Synthesis (SSW5)*, pages 127–132. (Quoted on pages 3, 86, 88, 202, 275, 278, 279 and 295.)
- Klabbers, E., Odijk, J., De Pijper, J., and Theune, M. (1996). Goalgetter: Football results, from teletext to speech. *IPO Annual Progress Report*, 31:66–75. (Quoted on pages 131 and 142.)
- Klatt, D. (1979). Synthesis by rule of segmental durations in English sentences. In Lindblom, B. and Ohman, S., editors, *Frontiers of Speech Communication Research*, volume 1, pages 287–299. Academic. (Quoted on page 107.)
- Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3):971–995. (Quoted on page 6.)
- Koch, P. and Oesterreicher, W. (2001). Langage oral et langage écrit. In Holthus, G., editor, *Lexicon der Romanistischen Linguistik*, volume tome 1-2, pages 584–627. Max Niemeyer, Tübingen. (Quoted on page 204.)
- Koehn, P., Abney, S., Hirschberg, J., and Collins, M. (2000). Improving intonational phrasing with syntactic information. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1289–1290. (Quoted on page 105.)
- Koriyama, T., Nose, T., and Kobayashi, T. (2011). On the use of extended context for HMM-based spontaneous conversational speech synthesis. In *Proc. of Interspeech*, pages 2657–2660. (Quoted on pages 257 and 258.)
- Krstulović, S., Hunecke, A., and Schröder, M. (2007). An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. In *Proc. of Interspeech*, pages 1897–1900. (Quoted on pages 123, 142, 143 and 202.)
- Laan, G. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22(1):43–65. (Quoted on pages 5, 257 and 276.)
- Labov, W. (1972). The internal evolution of linguistic rules. In Stockwell, R. P. and Macaulay, R. K. S., editors, *Linguistic Change and Generative Theory*, pages 101–171. Indiana University Press. (Quoted on page 86.)
- Labov, W. (2006). *The Social Stratification of English in New York City*. Cambridge University Press. (Quoted on page 86.)

- Lacheret, A., Kahane, S., Belaïo, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., and Tchobanov, A. (2014a). Rhapsodie: A prosodic-syntactic treebank for spoken French. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 295–301. (Quoted on page 241.)
- Lacheret, A., Kahane, S., and Pietrandrea, P. (2014b). *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French*. Studies in corpus linguistics. Benjamins, Amsterdam. (Quoted on page 241.)
- Lacheret-Dujour, A. (1991). Le débit de la parole: un filtre utilisé pour la génération des variantes de prononciation en français parisien. In *Proc. of the XIIème Congrès International des Sciences Phonétiques*, pages 194–197. (Quoted on pages 86, 201, 203, 214 and 215.)
- Lacheret-Dujour, A. and Beaugendre, F. (1999). *La prosodie du français*. CNRS Editions, Paris. (Quoted on pages 90, 94, 98, 109, 151 and 183.)
- Ladd, D. (1996). *Intonational Phonology*. Cambridge University Press, Cambridge. (Quoted on page 97.)
- Ladefoged, P. and Johnson, K. (2014). *A Course in Phonetics*. Cengage Learning. (Quoted on page 82.)
- Laks, B. (1977). Contribution empirique à l’analyse socio-différentielle de la chute des /r/ dans les groupes consonantiques finals. *Langue Française*, 34:109–125. (Quoted on page 262.)
- Lamel, L., Rosset, S., Bennacef, S., Bonneau-Maynard, H., Devillers, L., and Gauvain, J.-L. (1995). Development of spoken language corpora for travel information. In *Proc. of Eurospeech*, pages 1961–1964. (Quoted on page 241.)
- Lanchantin, P., Degottex, G., and Rodet, X. (2010). A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4630–4633. (Quoted on page 106.)
- Lanchantin, P., Morris, A., Rodet, X., and Veaux, C. (2008). Automatic phoneme segmentation with relaxed textual constraints. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*. (Quoted on pages 24, 36, 65 and 242.)
- Larreur, D. and Sorin, C. (1990). Quality evaluation of French text-to-speech synthesis within a task. The importance of the mute "e". In *Proc. of the ESCA workshop on speech synthesis*, pages 91–96. (Quoted on pages 125, 219 and 294.)
- Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press, Cambridge. (Quoted on page 6.)
- Lebart, L., Morineau, A., and Piron, M. (1995). *Statistique exploratoire multidimensionnelle*. Dunod, 4th edition. (Quoted on page 172.)

BIBLIOGRAPHY

- Lee, A., Kawahara, T., and Shikano, K. (2001). Julius — An open source real-time large vocabulary recognition engine. In *Proc. of Eurospeech*, pages 1691–1694. (Quoted on pages 29 and 65.)
- Lee, C.-H., Wu, C.-H., and Guo, J.-C. (2010). Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4826–4829. (Quoted on pages 125, 201, 202, 257 and 258.)
- Leggetter, C. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer, Speech and Language*, 9(2):171–185. (Quoted on pages 25 and 116.)
- Léon, P. (1970). Prosodic feature analysis: Analyse des faits prosodiques. In Leon, P., Faure, G., and Rigault, A., editors, *Systématique des fonctions expressives de l'intonation.*, Studia Phonetica 3, pages 57–74. Marcel Didier Ltée, Ottawa. (Quoted on page 110.)
- Léon, P. (1992). *Phonétisme et prononciations du français*. Nathan, Paris. (Quoted on pages 6, 82, 86 and 262.)
- Léon, P. (2005). *Précis de phonostylistique : parole et expressivité*. Colin, Paris. (Quoted on page 91.)
- Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. (Quoted on page 207.)
- Liberman, M. and Church, K. (1992). Text analysis and word pronunciation in text-to-speech synthesis. In Furui, S. and Sondhi, M., editors, *Advances in Speech Signal Processing*, pages 791–831. Dekker, New York. (Quoted on pages 104, 127, 183, 187, 196 and 243.)
- Liscombe, J., Venditti, J., and Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. In *Proc. of Eurospeech*, pages 725–728. (Quoted on pages 148 and 186.)
- Ljolje, A., Hirschberg, J., and Van Santen, J. (1997). Automatic speech segmentation for concatenative inventory selection. In Van Santen, J., Sproat, R., Olive, J., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 305–311. Springer. (Quoted on pages 24, 30 and 46.)
- Ljolje, A. and Riley, M. D. (1993). Automatic segmentation of speech for TTS. In *Proc. of Eurospeech*, pages 1445–1448. (Quoted on page 36.)
- Lo, H. and Wang, H. (2007). Phonetic boundary refinement using support vector machine. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 933–936. (Quoted on pages 52 and 53.)
- Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., and Montero, J. M. (2013). Towards speaking style transplantation in speech synthesis. In *Proc. of the 8th ISCA Workshop on Speech Synthesis (SSW8)*. (Quoted on pages 123 and 142.)

- LT-Innovate (2012). The global language technology market - innovation for the next generation of ICT. Technical report, LT-Innovate.eu. (Quoted on page 2.)
- Lucci, V. (1983). *Etude phonétique du français contemporain à travers la variation situationnelle (débit, rythme, accent, intonation, e muet, liaisons, phonèmes)*. Publications de l'Université des Langues et Lettres de Grenoble. (Quoted on pages 6, 87, 202, 212, 262 and 267.)
- Maeno, Y., Nose, T., Kobayashi, T., Ijima, Y., Nakajima, H., Mizuno, H., and Yoshioka, O. (2011). HMM-based emphatic speech synthesis using unsupervised context labeling. In *Proc. of Interspeech*, pages 1849–1852. (Quoted on pages 123 and 124.)
- Malfrère, F. and Dutoit, T. (1997). High-quality speech synthesis for phonetic speech segmentation. In *Proc. of Eurospeech*, pages 2631–2634. (Quoted on pages 27 and 105.)
- Malfrère, F., Dutoit, T., and Mertens, P. (1998). Fully automatic prosody generator for text-to-speech. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 1395–1398. (Quoted on pages 106 and 108.)
- Martin, P. (1979). Une théorie syntaxique de l'accentuation en français. In Fónagy, I., Léon, P., and Carton, F., editors, *L'accent en français contemporain*, volume 15 of *Studia Phonetica*, pages 1–12. Marcel Didier Ltée, Montreal. (Quoted on pages 92 and 96.)
- Martin, P. (1997). L'intonation. Analyse instrumentale et modèles. *Pre-print*, 4. (Quoted on page 96.)
- Martin, P. (1998). Association prosodie-syntaxe: validation par synthèse. In *Proc. of the Journées d'étude sur la parole (JEP)*, pages 119–122. (Quoted on page 96.)
- Martin, P. (2001). ToBi: L'illusion scientifique ? In *Proc. of Journées de la prosodie*. (Quoted on page 146.)
- Martin, P. (2009). *Intonation du français*. Armand Colin, Paris. (Quoted on page 94.)
- Martin, P. (2012). Automatic detection of voice creak. In *Proc. of Speech Prosody*, Shanghai. (Quoted on page 184.)
- Martinet, A. (1965). *Eléments de linguistique générale*. Armand Colin. (Quoted on page 91.)
- Martinet, A. (1971). *La prononciation du français contemporain*. Librairie Droz, Geneva. (Quoted on pages 86 and 201.)
- Masuko, T. (2002). *HMM-Based Speech Synthesis and Its Applications*. PhD thesis, Nagoya Institute of Technology. (Quoted on page 111.)
- Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. (1997). Voice characteristics conversion for HMM-based speech synthesis systems. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1611–1614. (Quoted on page 116.)

BIBLIOGRAPHY

- Mehrabian, A. and Russel, J. A. (1974). *An Approach to Environmental Psychology*. MIT Press, Cambridge, MA. (Quoted on pages 4, 148 and 159.)
- Mertens, P. (1987). *L'intonation du français. De la description linguistique à la reconnaissance automatique*. PhD thesis, University of Leuven (Belgium). (Quoted on pages 92, 94, 95, 96, 98, 99, 100, 101, 146, 149, 231, 232, 236, 238, 241, 248 and 295.)
- Mertens, P. (1990). L'intonation. In Blanche-Benveniste, C., Bilger, M., Rouget, C., Van den Eynde, K., Mertens, P., and Willems, D., editors, *Le français parlé*, pages 159–176. Editions du CNRS, Paris. (Quoted on pages 98, 100, 101, 236 and 248.)
- Mertens, P. (1991). Local prominence of acoustic and psychoacoustic functions and perceived stress in French. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 218–221. (Quoted on pages 89, 90 and 238.)
- Mertens, P. (1993). Intonational grouping, boundaries, and syntactic structure in french. In *Proc. of an ESCA Workshop on Prosody*, pages 156–159. (Quoted on page 89.)
- Mertens, P. (2004a). The Prosogram: Semi-automatic transcription of prosody based on a tonal perception models. In *Proc. of Speech Prosody*. (Quoted on pages 150, 152, 184 and 230.)
- Mertens, P. (2004b). Quelques allers-retours entre la prosodie et son traitement automatique. *Le français moderne*, 72 (1):39–57. (Quoted on pages 100, 104, 232, 244 and 248.)
- Mertens, P. (2005). *Phonétique française*. Acco, Leuven. (Quoted on page 6.)
- Mertens, P. (2006). A predictive approach to the analysis of intonation in discourse in French. In Kawaguchi, Y., Fonagy, I., and Moriguchi, T., editors, *Prosody and Syntax*, volume 3 of *Usage-Based Linguistic Informatics*, pages 64–101. John Benjamins. (Quoted on pages 89, 98, 100 and 248.)
- Mertens, P. (2008). Syntaxe, prosodie et structure informationnelle: Une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de Linguistique*, 56(1):87–124. (Quoted on pages 98, 100, 101, 107, 146, 235, 238, 247, 248, 249 and 297.)
- Mertens, P. (2012). La prosodie des clivées. In Caddéo, S., Roubaud, M.-N., Rouquier, M., and Sabio, F., editors, *Penser les langues avec Claire Blanche-Benveniste*, number 20 in *Collection Langues et langage*, pages 127–139. Presses Universitaires de Provence, Aix-en-Provence. (Quoted on pages 101 and 248.)
- Mertens, P. (2013a). Automatic labelling of pitch levels and pitch movements in speech corpora. In *Proc. of Tools and Resources for the Analysis of Speech Prosody (TRASP2013)*, pages 42–46. (Quoted on pages 181, 226, 231 and 234.)

- Mertens, P. (2013b). A classification of french adverbs based on distributional, syntactic and prosodic criteria. *Linguisticae Investigationes*, 36(2):201–228. (Quoted on page 235.)
- Mertens, P. (2013c). From pitch stylization to automatic tonal annotation of speech corpora. In Lacheret, A., Kahane, S., and Pietrandrea, P., editors, *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French. Studies in Corpus Linguistics*. Benjamins, Amsterdam. (Accepted for publication). (Quoted on pages 226 and 234.)
- Mertens, P. (2014). Polytonia: a system for the automatic transcription of tonal aspects in speech corpora. *Journal of Speech Sciences*, 4(2):17–57. (Quoted on pages 144, 181, 226, 234 and 235.)
- Mertens, P., Auchlin, A., and Goldman, J.-P. (2001a). Intonation du discours et synthèse de la parole: Premiers résultats d’une approche par balises. *Cahiers de Linguistique Française*, 23:189–209. (Quoted on page 107.)
- Mertens, P., Beaugendre, F., and d’Alessandro, C. (1997). Comparing approaches to pitch contour stylization for speech synthesis. In van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 347–363. Springer Verlag, New York. (Quoted on page 234.)
- Mertens, P., Goldman, J.-P., Wehrli, E., and Gaudinat, A. (2001b). La synthèse de l’intonation à partir de structures syntaxiques riches. *Traitement Automatique des Langues*, 42 (1):142–195. (Quoted on pages 89, 98, 101, 105, 107, 108, 236, 248, 264 and 297.)
- Mertens, P. and Simon, A. C. (2013). Towards automatic detection of prosodic boundaries in spoken French. In *Proc. of the Prosody-Discourse Interface Conference (IDP)*, pages 81–87. (Quoted on pages 171 and 231.)
- Mertens, P. and Vercammen, F. (1998). FONILEX manual. Technical report, K.U.Leuven. (Quoted on page 83.)
- Michaux, M.-C., Brognaux, S., and Christodoulides, G. (2014). The production and perception of L1 and L2 Dutch stress. In *Proc. of Speech Prosody*, pages 462–466.
- Miller, C. (1998a). Individuation of postlexical phonology for speech synthesis. In *Proc. of the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, pages 133–136. (Quoted on pages 87, 202, 259, 260, 275 and 297.)
- Miller, C., Karaali, O., and Massey, N. (1997). Variation and synthetic speech. In *Proc. of New Ways of Analyzing Variation (NWAV 26)*. (Quoted on pages 87, 202, 259, 260, 275, 277 and 297.)
- Miller, C. A. (1998b). *Pronunciation Modeling in Speech Synthesis*. PhD thesis, Factulties of the University of Pennsylvania. (Quoted on pages 85, 87, 202, 259, 260, 275 and 297.)

BIBLIOGRAPHY

- Miyanaga, K., Masuko, T., and Kobayashi, T. (2004). A style control technique for HMM-based speech synthesis. In *Proc. of Interspeech*, pages 1437–1440. (Quoted on pages 120, 122 and 201.)
- Monaghan, A. I. C. (1989). Phonological domains for intonation in speech synthesis. In *Proc. of Eurospeech*, pages 1502–1505. (Quoted on page 92.)
- Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enriquez, E., Aguilera, S., and Pardo, J. M. (1998). Emotional speech synthesis: From speech database to TTS. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 923–926. (Quoted on pages 3, 8 and 119.)
- Moreno, P. J. and Alberti, C. (2009). A factor automaton approach for the forced alignment of long speech recordings. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4869–4872. (Quoted on page 26.)
- Moreno, P. J., Joerg, C., Van Thong, J.-M., and Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 2711–2714. (Quoted on page 26.)
- Morin, Y. and Kaye, J. (1982). The syntactic bases for French liaison. *Journal of Linguistics*, 18(2):291–330. (Quoted on page 262.)
- Morizane, J., Nakamura, K., Toda, T., Saruwatari, H., and Shikano, K. (2009). Emphasized speech synthesis based on hidden Markov models. In *Proc. of Oriental COCOSA International Conference on Speech Database and Assessment*, pages 76–81. (Quoted on page 124.)
- Morlec, Y., Bailly, G., and Aubergé, V. (2001). Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication*, 33(4):357–371. (Quoted on page 109.)
- Moulines, E. and Charpentier, F. (1989). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5):453–467. (Quoted on page 6.)
- Mounin, G. (1974). *Dictionnaire de la linguistique*. Presses universitaires de France, Paris. (Quoted on page 6.)
- Mporas, I., Ganchev, T., and Fakotakis, N. (2008). Phonetic segmentation using multiple speech features. *International Journal of Speech Technologies*, 11(2):73–85. (Quoted on pages 52 and 53.)
- Murray, I. and Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93(2):1097–1108. (Quoted on page 187.)

- Murray, I. and Arnott, J. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4):369–390. (Quoted on pages 8, 110 and 119.)
- Murray, I. and Arnott, J. (2008). Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. *Computer, Speech and Language*, 22(2):107–129. (Quoted on page 110.)
- Nguyen, N. and Espesser, R. (2004). Méthodes et outils pour l’analyse acoustique des systèmes vocaliques. *Bulletin PFC (Phonologie du Français Contemporain)*, 3:77–85. (Quoted on pages 45 and 46.)
- Nguyen, N., Wauquier-Gravelines, S., and Durand, J. (2005). *Phonologie et phonétique. Forme et substance*. Hermès, Paris. (Quoted on page 82.)
- Nose, T. and Kobayashi, T. (2011). Recent development of HMM-based expressive speech synthesis and its applications. In *Proc. of the Annual Summit and Conference (ASC) of the Asia-Pacific Signal and Information Processing Association (APSIPA)*. (Quoted on page 120.)
- Obin, N. (2011). *MeLos: Analysis and Modelling of Speech Prosody and Speaking Style*. PhD thesis, Ircam-Upmc, Université Pierre et Marie Curie - Paris VI. (Quoted on pages 232 and 242.)
- Obin, N., Dellwo, V., Lacheret, A., and Rodet, X. (2010a). Expectations for discourse genre identification: A prosodic study. In *Proc. of Interspeech*, pages 3070–3073. (Quoted on page 143.)
- Obin, N., Lanchantin, P., Avanzi, M., Lacheret, A., and Rodet, X. (2010b). Towards improved HMM-based speech synthesis using high-level syntactical features. In *Proc. of Speech Prosody*, pages 2000–2004. (Quoted on pages 105 and 106.)
- Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011). Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation. In *Proc. of Interspeech*, pages 2785–2788. (Quoted on pages 123, 142, 143, 147 and 201.)
- Obin, N., Rodet, X., and Lacheret, A. (2009). A multi-level context-dependent prosodic model applied to durational modeling. In *Proc. of Interspeech*, pages 512–515. (Quoted on pages 107 and 109.)
- O’Connor, J. D. and Arnold, G. F. (1973). *Intonation of Colloquial English: A practical Handbook*. Longman. (Quoted on page 96.)
- Odgen, R. and Walker, G. (2001). We speak prosodies and we listen to them. In *Proc. of the Symposium on Prosody and Interaction*. (Quoted on pages 110 and 143.)
- Ostendorf, M., Byrne, B., Bacchiani, M., Finke, M., Gunawardana, A., Ross, K., Roweis, S., Shriberg, E., Talkin, D., Waibel, A., Wheatley, B., and Zeppenfeld, T. (1996). Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *Proc. of the LVCSR Summer Research Workshop*. (Quoted on pages 87, 202 and 258.)

BIBLIOGRAPHY

- Paeschke, A., Kienast, M., and Sendlmeier, W. F. (1999). F0-contours in emotional speech. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 929–932. (Quoted on pages 6 and 93.)
- Paeschke, A. and Sendlmeier, W. F. (2000). Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In *Proc. of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 75–80. (Quoted on pages 6 and 93.)
- Pagel, V., Lenzo, K., and Black, A. (1998). Letter to sound rules for accented lexicon compression. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 252–255. (Quoted on pages 127, 206 and 265.)
- Pan, S. (2002). *Prosody Modeling in Concept-to-Speech Generation*. PhD thesis, Columbia University. (Quoted on page 134.)
- Park, S. S. and Kim, N. S. (2007). On using multiple models for automatic speech segmentation. *IEEE Transactions on Audio Speech and Language Processing*, 15(8):2202–2212. (Quoted on page 52.)
- Parlikar, A. and Black, A. (2011). A grammar-based approach to style specific phrase prediction. In *Proc. of Interspeech*, pages 2149–2152. (Quoted on page 105.)
- Paulo, S. and Oliveira, L. C. (2004). Automatic phonetic alignment and its confidence measures. In *Proc. of the 4th International Conference EsTAL*, pages 36–44. (Quoted on page 27.)
- Paulo, S. and Oliveira, L. C. (2005). Generation of word alternative pronunciations using weighted finite state transducers. In *Proc. of Interspeech*, pages 1157–1160. (Quoted on page 278.)
- Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, Institut de Recherche et Coordination Acoustique/Musique (IRCAM). (Quoted on pages 55 and 150.)
- Picart, B. (2013). *Statistical Parametric Speech Synthesis Based on the Degree of Articulation*. PhD thesis, University of Mons. (Quoted on pages 106, 117, 118 and 135.)
- Picart, B., Brognaux, S., and Drugman, T. (2013). HMM-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation. In *Proc. of the 8th ISCA Workshop on Speech Synthesis (SSW8)*, pages 19–24.
- Picart, B., Drugman, T., and Dutoit, T. (2010). Analysis and synthesis of hypo and hyperarticulated speech. In *Proc. of the 7th ISCA Workshop on Speech Synthesis (SSW7)*, pages 270–275. (Quoted on pages 40, 86 and 201.)
- Picart, B., Drugman, T., and Dutoit, T. (2014). Analysis and HMM-based synthesis of hypo and hyperarticulated speech. *Computer, Speech and Language*, 28(2):687–707. (Quoted on page 125.)

- Pickering, B., Williams, B., and Knowles, G. (1996). Analysis of transcriber differences in the SEC. In Knowles, G., Wichmann, A., and Alderson, P., editors, *Working with Speech: perspectives and research into the Lancaster/IBM Spoken English Corpus*, pages 59–105. Longman, London/New-York. (Quoted on page 181.)
- Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology (MIT), Cambridge. (Quoted on page 97.)
- Pitrelli, J., Beckman, M., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 123–126. (Quoted on pages 97, 98, 146 and 149.)
- Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95. (Quoted on page 32.)
- Pooley, T. (1996). *Chtimi: The Urban Vernaculars of Northern France*. Multilingual Matters, Clevedon. (Quoted on page 262.)
- Portes, C. (2004). *Prosodie et économie du discours: Spécificité phonétique, écologie discursive et portée pragmatique de l'intonation d'implication*. PhD thesis, Université de Provence, Aix-Marseille 1. (Quoted on page 41.)
- Post, B. (2000). *Tonale and Phrasal Structure in French Intonation*. Holland Academic Graphics, The Hague. (Quoted on pages 89 and 98.)
- Post, B., Delais-Roussarie, E., and Simon, A. C. (2006). IVTS, un système de transcription pour la variation prosodiques. *Bulletin de la Phonologie du Français Contemporain (PFC)*, 6:51–68. (Quoted on page 110.)
- Povey, D., Goshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silvosky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 1–4. (Quoted on page 25.)
- Prahallad, K., Black, A., and Mosur, R. (2006). Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 853–856. (Quoted on pages 125, 201, 202, 257 and 258.)
- Prahallad, K., Toth, A. R., and Black, A. W. (2007). Automatic building of synthetic voices from large multi-paragraph speech databases. In *Proc. of Interspeech*, pages 2901–2904. (Quoted on pages 26 and 231.)
- Price, P. (1991). *An Introduction to French Pronunciation*. Blackwell Publishing. (Quoted on page 97.)

BIBLIOGRAPHY

- Prince, E. (1978). A comparison of wh-clefts and it-clefts in discourse. *Language*, 54:883–906. (Quoted on page 248.)
- Proust, M. (1913). *Du côté de chez Swann*. Grasset. (Quoted on page 242.)
- Pršir, T., Goldman, J.-P., and Auchlin, A. (2013). Variation prosodique situationnelle: Étude sur corpus de huit phonogenres en français. In *Proc. of the Prosody-Discourse Interface Conference (IDP)*, pages 107–111. (Quoted on pages 86, 201 and 202.)
- Putska, E. (2011). Le conditionnement lexical de l’élision des liquides en contexte post-consonantique finals. *Langue Française*, 169:19–38. (Quoted on pages 86 and 262.)
- Qin, L., Ling, Z.-H., Wu, Y.-J., Zhang, B.-F., and Wang, R.-H. (2006). HMM-based emotional speech synthesis using average emotion models. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 233–240. (Quoted on pages 8, 120, 122, 142, 144 and 159.)
- Quene, H. and Kager, R. (1989). Automatic accentuation and prosodic phrasing for Dutch text-to-speech conversion. In *Proc. of Eurospeech*, pages 214–217. (Quoted on page 106.)
- Racine, I. and Andreassen, H. (2012). A phonological study of a Swiss French variety: Data from the canton of Neuchâtel. In Gess, R., Lyche, C., and Meisenburg, T., editors, *Phonological Variation in French: Illustrations from Three Continents*, pages 173–207. John Benjamins, Amsterdam. (Quoted on page 262.)
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American statistical Association*, 66(336):846–850. (Quoted on page 173.)
- Raux, A. and Black, A. (2003). A unit selection approach to F0 modeling and its application to emphasis. In *Proc. of the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, pages 700–705. (Quoted on page 169.)
- Rea, L. M. and Parker, R. A. (1992). *Designing and Conducting Survey Research*. Jossey-Bass, USA. (Quoted on page 175.)
- Richardson, L. and Ruby, S. (2008). *RESTful web services*. O’Reilly Media, Inc., USA. (Quoted on page 128.)
- Richmond, K., Clark, R., and Fitt, S. (2009). Robust LTS rules with the Combilex speech technology lexicon. In *Proc. of Interspeech*, pages 1295–1298. (Quoted on page 276.)
- Richmond, K., Clark, R., and Fitt, S. (2010). On generating Combilex pronunciations via morphological analysis. In *Proc. of Interspeech*, pages 1974–1977. (Quoted on page 276.)
- Richmond, K., Strom, V., Clark, R. J., Yamagishi, J., and Fitt, S. (2007). Festival multisyn voices for the 2007 blizzard challenge. In *Proc. of the Blizzard Challenge Workshop*, Bonn (Germany). (Quoted on page 279.)

- Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., and Zavaliagkos, G. (1999). Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29(2):209–224. (Quoted on pages 87, 202 and 258.)
- Riley, M. D. (1990). Tree-based modelling for speech synthesis. In *Proc. of the ESCA Workshop on Speech Synthesis*, pages 229–232. (Quoted on page 107.)
- Rillard, A. (2014). Prosodie et interaction homme-machine: Étude de la variation démarcative, diatopique, diachronique et expressive. Thèse d’habilitation à diriger des recherches, Université Paris Sud. (Quoted on pages 4, 11, 93 and 232.)
- Roach, P. (2009). *English Phonetics and Phonology. A Practical Course*. Cambridge University Press, Cambridge, 4th edition. (Quoted on page 82.)
- Roekhaut, S., Brognaux, S., Beaufort, R., and Dutoit, T. (2014a). eLite-HTS. a NLP tool for French HMM-based speech synthesis. In *Proc. of Interspeech Show and Tell*, pages 2136–2137.
- Roekhaut, S., Brognaux, S., Beaufort, R., and Dutoit, T. (2014b). eLite-HTS: un outil TAL pour la génération de synthèse en français. In *Démonstration aux Journées d’étude sur la parole (JEP)*.
- Roekhaut, S., Goldman, J.-P., and Simon, A. C. (2010). A model for varying speaking style in TTS systems. In *Proc. of Speech Prosody*. (Quoted on pages 86, 123, 125, 202, 203 and 210.)
- Rosenberg, A. (2010). AuToBI -a tool for automatic ToBI annotation. In *Proc. of Interspeech*, pages 146–149. (Quoted on pages 181, 184, 230 and 235.)
- Rosenberg, A., Cooper, E., Levitan, R., and Hirschberg, J. (2012). Cross-language prominence detection. In *Speech Prosody*. (Quoted on page 230.)
- Rositzke, H. A. (1939). Vowel-length in general American speech. *Language*, 15(2):99–109. (Quoted on page 171.)
- Ross, K. and Ostendorf, M. (1996). Prediction of abstract prosodic labels for speech synthesis. *Computer, Speech and Language*, 10(3):155–185. (Quoted on page 106.)
- Rossi, M. (1980). Le français, langue sans accent ? In Fónagy, I., Léon, P., and Carton, F., editors, *L’accent en français contemporain*, volume 15 of *Studia Phonetica*, pages 13–51. Marcel Didier Ltée. (Quoted on pages 90 and 92.)
- Rossi, M. (1985). L’intonation et l’organisation de l’énoncé. *Phonetica*, 42:135–153. (Quoted on pages 90 and 91.)
- Rossi, M. (1999). *L’intonation, le système du français: description et modélisation*. Ophrys, Paris. (Quoted on pages 91 and 94.)
- Rossi, M., Di Cristo, A., Hirst, D., Martin, P., and Nishinuma, Y. (1981). *L’intonation. De l’acoustique à la sémantique*. Klincksieck, Paris. (Quoted on pages 89, 92 and 96.)

BIBLIOGRAPHY

- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65. (Quoted on page 173.)
- Russell, J. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178. (Quoted on pages 4 and 148.)
- Sagisaka, Y., Kaiki, N., Iwahashi, N., and Mimura, K. (1992). ATR V-talk speech synthesis systems. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 483–486. (Quoted on page 6.)
- Saito, T. (1998). On the use of F0 features in automatic segmentation for speech synthesis. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 145–148. (Quoted on page 56.)
- Saraclar, M., Nock, H., and Khudanpur, S. (2000). Pronunciation modeling by sharing gaussian densities across phonetic models. *Computer, Speech and Language*, 14(2):137–160. (Quoted on page 87.)
- Scharenborg, O., Wan, V., and Ernestus, M. (2010). Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. *The Journal of Acoustical Society of America*, 127 (2):1084–1095. (Quoted on pages 26, 27 and 46.)
- Schiel, F., Kipp, A., and Tillmann, H. G. (1998). Statistical modelling of pronunciation: It’s not the model, it’s the data. In *Proc. of the ESCA Tutorial and Research Workshop (ETRW) on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 131–136. (Quoted on pages 24 and 52.)
- Schröder, M. (2001). Emotional speech synthesis: A review. In *Proc. of Interspeech*, pages 561–564. (Quoted on pages 118 and 164.)
- Schröder, M. (2004). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In André, E., Dybkjaer, L., Minker, W., and Heisterkamp, P., editors, *Affective Dialogue Systems*. Springer, Berlin. (Quoted on pages 8, 119 and 164.)
- Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proc. of Interspeech*, pages 87–90. (Quoted on pages 148, 186 and 187.)
- Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087. (Quoted on pages 4, 11 and 187.)
- Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 401–404. (Quoted on page 187.)
- Séguinot, A. (1976). L’accent d’insistance en français standard. In Carton, F., editor, *L’accent d’insistance*, volume 12 of *Studia Phonetica*, pages 1–58. Didier. (Quoted on pages 91 and 151.)

- Sejnowski, R. and Rosenberg, C. (1987). Parallel networks that learn to pronounce english text. *Complex Systems*, 1(1):145–168. (Quoted on page 84.)
- Selkirk, E. (1986). *Phonology and Syntax: The relation between Sound and Structure*. MIT Press, Cambridge. (Quoted on pages 86 and 89.)
- Sethy, A. and Narayanan, S. (2002). Refined speech segmentation for concatenative speech synthesis. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 149–152. (Quoted on page 52.)
- Shriberg, E. and Stolcke, A. (2004). Prosody modeling for automatic speech recognition and understanding. In Johnson, M., Khudanpur, S., Ostendorf, M., and Rosenfeld, R., editors, *Mathematical Foundations of Speech and Language Processing*, pages 105–114. Springer, New York. (Quoted on page 295.)
- Silipo, R. and Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous English discourse. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 2351–2354. (Quoted on page 239.)
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 867–870. (Quoted on pages 92, 97, 149, 175 and 232.)
- Simon, A. (2004). *La structuration prosodique du discours en français : une approche multidimensionnelle et expérientielle*. Peter Lang, Berne. (Quoted on page 109.)
- Simon, A. C., Auchlin, A., Avanzi, M., and Goldman, J.-P. (2009). Les phonostyles: une description prosodique des styles de parole en français. In Abecassis, M. . G. L., editor, *Les voix des Français. En parlant, en écrivant*, pages 71–88. Peter Lang, Bern. (Quoted on pages 86, 103, 202 and 215.)
- Sjölander, K. (2001). Automatic alignment of phonetic segments. Working Papers 49, Lund University (Department of Linguistics). (Quoted on page 29.)
- Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik*, pages 93–96. (Quoted on pages 24, 27 and 65.)
- Sjölander, K. and Beskow, J. (2000). Wavesurfer - an open-source speech tool. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 464–467. (Quoted on page 24.)
- Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6 (1):1–3. (Quoted on page 54.)
- Soto, V., Cooper, E., Rosenberg, A., and Hirschberg, J. (2013). Cross-language phrase boundary detection. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8460–8464. (Quoted on page 230.)

BIBLIOGRAPHY

- Sproat, R., Hirschberg, J., and Yarowsky, D. (1992). A corpus-based synthesizer. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 563–566. (Quoted on page 106.)
- SPTK Working Group (2009). Speech signal processing toolkit (SPTK), available at : <http://sp-tk.sourceforge.net>. (Quoted on page 118.)
- Sridhar, V., Bangalore, S., and Narayanan, S. (2008). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy frameworks. *IEEE Transactions on Audio, Speech and Language Processing*, 16(4):797–811. (Quoted on pages 181 and 229.)
- Stan, A., Mamiya, Y., Yamagishi, J., Bell, P., Watts, O., Clark, R. A. J., and King, S. (2016). Alisa: An automatic lightly supervised speech segmentation and alignment tool. *Computer, Speech and Language*, 35:116–133. (Quoted on pages 26 and 128.)
- Strik, H. and Cucchiaroni, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29(2):225–246. (Quoted on pages 87, 202 and 258.)
- Strom, V., Clark, R., and King, S. (2006). Expressive prosody for unit-selection speech synthesis. In *Proc. of Interspeech*. (Quoted on pages 105, 106, 107, 124, 169 and 170.)
- Strom, V., Nenkova, A., Clark, R., Vazquez-Alvarez, Y., Brenier, J., King, S., and Jurafsky, D. (2007). Modelling prominence and emphasis improves unit-selection synthesis. In *Proc. of Interspeech*, pages 1282–1285. (Quoted on page 169.)
- Sundaram, S. and Narayanan, S. (2003). An empirical text transformation method for spontaneous speech synthesizers. In *Eurospeech 2003*, pages 1221–1224. (Quoted on pages 123 and 257.)
- Syrdal, A., Hirschberg, J., McGory, J., and Beckman, M. (2001). Automatic ToBI prediction and alignment to speech manual labeling of prosody. *Speech Communication*, 33(1):135–151. (Quoted on pages 106, 181 and 230.)
- Syrdal, A. and McGorgy, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 235–238. (Quoted on page 146.)
- Székely, E., Cabral, J. P., Cahill, P., and Carson-Berndsen, J. (2011). Clustering expressive speech styles in audiobooks using glottal source parameters. In *Proc. of Interspeech*, pages 2409–2412. (Quoted on pages 110, 186, 187, 189, 231 and 296.)
- ’t Hart, J. R., Collier, R., and Cohen, A. (1990). *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*. Cambridge University Press, Cambridge. (Quoted on page 95.)
- Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transactions on Information and Systems*, E98-D no. 11:2484–2491. (Quoted on pages 121, 143 and 165.)

- Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2006). A style adaptation technique for speech synthesis using HSMM and suprasegmental features. *IE-ICE Transactions on Information and Systems*, E89-D, no. 3:1092–1099. (Quoted on page 120.)
- Takahashi, J. and Sagayama, S. (1995). Vector-field-smoothed bayesian learning for incremental speaker adaptation. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–699. (Quoted on page 116.)
- Tamburini, F. (2003). Automatic prosodic prominence detection in speech using acoustic features : An unsupervised systems. In *Proc. of Eurospeech*, pages 129–132. (Quoted on page 239.)
- Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (1998). Speaker adaptation for HMM-based speech synthesis system using MLLR. In *Proc. of the Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, pages 273–276. (Quoted on pages 116 and 117.)
- Tato, R., Santos, R., Kompe, R., and Pardo, J. M. (2002). Emotional space improves emotion recognition. In *Proc. of Interspeech*, pages 2029–2032. (Quoted on page 186.)
- Taylor, P. (1994). *A phonetic model of English Intonation*. PhD thesis, University of Edinburgh. (Quoted on page 94.)
- Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107(3):1697–1714. (Quoted on pages 94 and 134.)
- Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge. (Quoted on pages 6, 82, 83, 88, 98, 101, 103, 104, 105, 107, 109, 134 and 169.)
- Taylor, P. and Black, A. (1998). Assigning phrase breaks from part of speech sequences. *Computer, Speech and Language*, 12(2):99–117. (Quoted on pages 104 and 105.)
- Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, 89(4):1768–1776. (Quoted on page 89.)
- Tesnière, L. (1959). *Eléments de syntaxe structurale*. Klincksieck, Paris. (Quoted on page 107.)
- Theune, M. and Klabbers, E. (1998). System demonstration. goalgetter: Generation of spoken soccer reports. In *Proc. of the 9th International Workshop on Natural Language Generation*, pages 292–295. (Quoted on pages 131 and 142.)
- Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 229–232. (Quoted on page 111.)

BIBLIOGRAPHY

- Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (2002a). Multi-space probability distribution HMMs. *IEICE Transactions on Information and Systems*, 85(3):455–464. (Quoted on page 111.)
- Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K. (2013). Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, 101(5):1234–1252. (Quoted on page 111.)
- Tokuda, K., Zen, H., and Black, A. (2002b). An HMM-based speech synthesis system applied to English. In *Proc. of the IEEE Speech Synthesis Workshop*, pages 227–230. (Quoted on pages 111 and 155.)
- Toledano, D. and Gómez, L. (2002). HMMs for automatic phonetic segmentation. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*. (Quoted on pages 27, 36 and 52.)
- Traber, C. (1993). Syntactic processing and prosody control in the SVOX TTS system for German. In *Proc. of Eurospeech*, pages 2099–2102. (Quoted on page 104.)
- Trouvain, J. (2011). Between excitement and triumph - Live football commentaries in radio vs. TV. In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 2022–2025. (Quoted on pages 135, 142, 143, 144, 148 and 153.)
- Trouvain, J. and Barry, W. (2000). The prosody of excitement in horse race commentaries. In *Proc. of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 86–91. (Quoted on pages 143 and 153.)
- Tsurumi, Y. and Nakagawa, S. (1994). An unsupervised speaker adaptation method for continuous parameter HMM by maximum a posteriori probability estimation. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 431–434. (Quoted on page 116.)
- Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M., and Narayanan, S. (2004). Constructing emotional speech synthesizers with limited speech database. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 1185–1188. (Quoted on pages 9, 121, 122, 158, 201 and 202.)
- Türk, O. and Schröder, M. (2008). A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis. In *Proc. of Interspeech*, pages 2282–2285. (Quoted on page 119.)
- Türk, O., Schröder, M., Bozkurt, B., and Arslan, L. M. (2005). Voice quality interpolation for emotional text-to-speech synthesis. In *Proc. of Interspeech*, pages 797–800. (Quoted on page 119.)
- Vaissière, J. (1983). Language-independent prosodic features. In Cutler, A. and Ladd, D. R., editors, *Prosody: Models and measurements*, pages 53–66. Springer-Verlag, Heidelberg. (Quoted on page 171.)
- Vaissière, J. (1991). Perceiving rhythm in French ? In *Proc. of the International Conference on Phonetic Sciences (ICPhS)*, pages 258–261. (Quoted on page 92.)

- Van Bael, C. (2007). *Validation, Automatic Generation and Use of Broad Phonetic Transcriptions*. PhD thesis, Radboud University Nijmegen, Nijmegen. (Quoted on pages 276, 280 and 281.)
- Van Bael, C., Boves, L., van den Heuvel, H., and Strik, H. (2006). Automatic phonetic transcription of large speech corpora. In *Proc. of Interspeech*, pages 1085–1088. (Quoted on page 280.)
- Van Bael, C., van den Heuvel, H., and Strik, H. (2004). Investigating speech style specific pronunciation variation in large spoken language corpora. In *Proc. of Interspeech*, pages 2793–2796. (Quoted on pages 87, 202, 206, 208, 209, 260 and 297.)
- van Niekirk, D. and Barnard, E. (2009). Phonetic alignment for speech synthesis in under-resourced languages. In *Proc. of Interspeech*, pages 880–883. (Quoted on pages 25, 27, 30, 39 and 43.)
- Van Santen, J. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer, Speech and Language*, 8(2):95–128. (Quoted on page 107.)
- Van Santen, J. and Sproat, R. (1999). High-accuracy automatic segmentation. In *Proc. of Eurospeech*, pages 2809–2812. (Quoted on page 26.)
- Veaux, C., Beller, G., Schwarz, D., and Rodet, X. (2008). IrcamCorpusTools: An extensible platform for speech corpora exploitation. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*. (Quoted on pages 126, 230 and 239.)
- Véronis, J., Di Cristo, A., Courtois, F., and Chaumette, C. (1998). A stochastic model of intonation for text-to-speech synthesis. *Speech Communication*, 26(4):233–244. (Quoted on pages 106 and 108.)
- Villeneuve, A.-J. (2010). Word-final cluster simplification in Vimeu French: A preliminary analysis. *University of Pennsylvania Working Papers in Linguistics*, 15(2):133–144. (Quoted on pages 262 and 271.)
- Wagner, A. (2009). Analysis and recognition of accentual patterns. In *Proc. of Interspeech*, pages 2427–2430. (Quoted on pages 181 and 230.)
- Wagner, M. (1981). Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1156–1159. (Quoted on page 27.)
- Wang, M. Q. and Hirschberg, J. (1992). Automatic classification of intonational phrase boundaries. *Computer, Speech and Language*, 6(2):175–196. (Quoted on pages 105, 181, 230 and 235.)
- Ward, J. (1963). Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association*, 58(301):236–244. (Quoted on page 172.)

BIBLIOGRAPHY

- Warnant, L. (1996). *Orthographe et prononciation en français. Les 12000 mots qui ne se prononcent pas comme ils s'écrivent*. De Boeck-Duculot. (Quoted on pages 209 and 212.)
- Watts, O., Stan, A., Mamiya, Y., Suni, A., Burgos, J. M., and Montero, J. M. (2013). The simple4all entry to the blizzard challenge 2013. In *Proc. of the Blizzard Challenge Workshop*. (Quoted on page 9.)
- Weide, R. (1998). The CMU pronunciation dictionary, release 0.6. Carnegie Mellon University. (Quoted on page 83.)
- Weintraub, M., Taussig, K., Hunicke-Smith, K., and Snodgrass, A. (1996). Effect of speaking style on LVCSR performance. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 16–19. (Quoted on page 87.)
- Werner, S. and Hoffmann, R. (2006). Pronunciation variant selection for spontaneous speech synthesis - A summary of experimental results. In *Proc. of Speech Prosody*. (Quoted on pages 125, 201, 202, 257, 258 and 285.)
- Wesenick, M.-B. and Kipp, A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 129–132. (Quoted on pages 32 and 46.)
- Wester, M. (2003). Pronunciation modeling for ASR - Knowledge-based and data-derived methods. *Computer, Speech and Language*, 17(1):69–85. (Quoted on pages 87, 202 and 258.)
- Wightman, C. (2002). ToBI or not ToBI ? In *Proc. Speech Prosody*. (Quoted on pages 12, 146 and 169.)
- Wightman, C. and Ostendorf, M. (1992). Automatic recognition of intonational features. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 221–224. (Quoted on pages 129 and 239.)
- Wightman, C. and Ostendorf, M. (1994). Automatic labelling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481. (Quoted on pages 146, 169, 181, 229 and 235.)
- Wightman, C. and Rose, R. (1999). Evaluation of an efficient prosody labeling system for spontaneous speech utterances. In *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 333–336. (Quoted on pages 146 and 169.)
- Wightman, C., Syrdal, A., Stemmer, G., Conkie, A., and Beutnagel, M. (2000). Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 71–74. (Quoted on page 107.)

- Wightman, C. W. and Talkin, D. T. (1997). The aligner: Text-to-speech alignment using markov models. In Van Santen, J., Sproat, R., Olive, J., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 313–323. Springer-Verlag, New York. (Quoted on pages 24, 27, 29, 45, 65 and 278.)
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 1556–1559. (Quoted on page 24.)
- Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., and Cowie, R. (2008). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. of Interspeech*, pages 597–600. (Quoted on page 186.)
- Yamagishi, J. (2006). *Average-Voice-Based Speech Synthesis*. PhD thesis, Tokyo Institute of Technology. (Quoted on page 158.)
- Yamagishi, J. and Kobayashi, T. (2007). Average-voice-based speech synthesis using HMM-based speaker adaptation and adaptive training. *IEICE Transactions Information and Systems*, 90(2):533–543. (Quoted on pages 111 and 159.)
- Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009a). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):66–83. (Quoted on page 118.)
- Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., and Renals, S. (2009b). Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1208–1230. (Quoted on pages 117, 158 and 159.)
- Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IECE Transactions on Information and Systems*, E88-D(3):502–509. (Quoted on pages 8, 9, 120, 121, 122, 142, 143, 158, 201 and 202.)
- Yamagishi, J., Tachibana, M., Masuko, T., and Kobayashi, T. (2004). Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5–8. (Quoted on pages 8, 9, 120, 121, 122 and 158.)
- Yoshimura, T. (2002). *Simultaneous modeling of phonetic and prosodic parameters, and characteritic conversion for HMM-based Text-To-Speech Systems*. PhD thesis, Nagoya Institute of Technology. (Quoted on page 114.)
- Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1998). Duration modeling in HMM-based speech synthesis. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 29–32. (Quoted on page 114.)

BIBLIOGRAPHY

- Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. of Eurospeech*, pages 2347–2350. (Quoted on pages 3, 7, 8, 109, 111, 112 and 115.)
- Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2001). Mixed excitation for HMM-based speech synthesis. In *Proc. of Eurospeech*, pages 2263–2266. (Quoted on page 115.)
- Young, S. and Fallside, F. (1979). Speech synthesis from concept: A method for speech output from information systems. *Journal of the Acoustical Society of America*, 66(3):685–695. (Quoted on page 134.)
- Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2000). *The HTK Book (for HTK Version 3)*. Cambridge University. (Quoted on pages 25, 28, 29, 30, 32, 33, 34, 37, 52, 65, 111 and 117.)
- Yu, K., Mairesse, F., and Young, S. (2010). Word-level emphasis modelling in HMM-based speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4238–4241. (Quoted on pages 124, 137, 169 and 170.)
- Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proc. of Acoustics '08*, pages 5687–5690. (Quoted on pages 24, 29, 41, 46, 61 and 65.)
- Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (2007a). The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, pages 294–299. (Quoted on pages 8, 107, 113, 118, 125, 155, 176, 193, 217, 243 and 282.)
- Zen, H., Tokuda, K., and Black, A. (2007b). Statistical parametric speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (Quoted on pages 111 and 112.)
- Zen, H., Tokuda, K., and Black, A. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064. (Quoted on pages 3, 7, 24, 111, 116, 117, 144 and 176.)
- Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2007c). A hidden semi-Markov model-based speech synthesis system. *IEICE Transactions on Information and Systems*, 90(5):825–834. (Quoted on page 113.)
- Zhao, Y., Peng, D., Wang, L., Chu, M., Chen, Y., Yu, P., and Guo, J. (2006). Constructing stylistic synthesis databases from audiobooks. In *Proc. of Interspeech*, pages 1750–1753. (Quoted on pages 231 and 296.)

Appendices

Appendix A

HTS description file

**An example of context-dependent label format
for HMM-based speech synthesis in English**

Keiichiro Oura

July 7, 2011

$p_1 \sim p_2 \sim p_3 + p_4 \sim p_5 @ p_6 \sim p_7$
 $/A: a_1 \cdot a_2 \cdot a_3 / B: b_1 \sim b_2 \sim b_3 @ b_4 \sim b_5 \& b_6 \sim b_7 \# b_8 \sim b_9 \$ b_{10} \sim b_{11} ! b_{12} \sim b_{13} ; b_{14} \sim b_{15} | b_{16} / C: c_1 + c_2 + c_3$
 $/D: d_1 \cdot d_2 / E: e_1 + e_2 @ e_3 + e_4 \& e_5 + e_6 \# e_7 + e_8 / F: f_1 \cdot f_2$
 $/G: g_1 \cdot g_2 / H: h_1 \sim h_2 \sim h_3 = h_4 | h_5 / I: i_1 \cdot i_2$
 $/J: j_1 + j_2 \sim j_3$

p_1	the phoneme identity before the previous phoneme
p_2	the previous phoneme identity
p_3	the current phoneme identity
p_4	the next phoneme identity
p_5	the phoneme after the next phoneme identity
p_6	position of the current phoneme identity in the current syllable (forward)
p_7	position of the current phoneme identity in the current syllable (backward)
a_1	whether the previous syllable stressed or not (0: not stressed, 1: stressed)
a_2	whether the previous syllable accented or not (0: not accented, 1: accented)
a_3	the number of phonemes in the previous syllable
b_1	whether the current syllable stressed or not (0: not stressed, 1: stressed)
b_2	whether the current syllable accented or not (0: not accented, 1: accented)
b_3	the number of phonemes in the current syllable
b_4	position of the current syllable in the current word (forward)
b_5	position of the current syllable in the current word (backward)
b_6	position of the current syllable in the current phrase (forward)
b_7	position of the current syllable in the current phrase (backward)
b_8	the number of stressed syllables before the current syllable in the current phrase
b_9	the number of stressed syllables after the current syllable in the current phrase
b_{10}	the number of accented syllables before the current syllable in the current phrase
b_{11}	the number of accented syllables after the current syllable in the current phrase
b_{12}	the number of syllables from the previous stressed syllable to the current syllable
b_{13}	the number of syllables from the current syllable to the next stressed syllable
b_{14}	the number of syllables from the previous accented syllable to the current syllable
b_{15}	the number of syllables from the current syllable to the next accented syllable
b_{16}	name of the vowel of the current syllable
c_1	whether the next syllable stressed or not (0: not stressed, 1: stressed)
c_2	whether the next syllable accented or not (0: not accented, 1: accented)
c_3	the number of phonemes in the next syllable
d_1	gpos (guess part-of-speech) of the previous word
d_2	the number of syllables in the previous word
e_1	gpos (guess part-of-speech) of the current word
e_2	the number of syllables in the current word
e_3	position of the current word in the current phrase (forward)
e_4	position of the current word in the current phrase (backward)
e_5	the number of content words before the current word in the current phrase
e_6	the number of content words after the current word in the current phrase
e_7	the number of words from the previous content word to the current word
e_8	the number of words from the current word to the next content word
f_1	gpos (guess part-of-speech) of the next word
f_2	the number of syllables in the next word
g_1	the number of syllables in the previous phrase
g_2	the number of words in the previous phrase
h_1	the number of syllables in the current phrase
h_2	the number of words in the current phrase
h_3	position of the current phrase in this utterance (forward)
h_4	position of the current phrase in this utterance (backward)
h_5	TOBI endtone of the current phrase
i_1	the number of syllables in the next phrase
i_2	the number of words in the next phrase
j_1	the number of syllables in this utterance
j_2	the number of words in this utterance
j_3	the number of phrases in this utterance

Appendix B

List of eLite phonetic symbols

LIST OF ELITE PHONETIC SYMBOLS

Table B.1 - *List of phonemes used in eLite*

Phoneme	Location	Articulation	Lips	Voicing	Aperture
Vowels					
i	Palatal	Oral	Stretched	Voiced	First
e	Palatal	Oral	Stretched	Voiced	Second
E	Palatal	Oral	Stretched	Voiced	Third
e~	Palatal	Nose	Stretched	Voiced	Third
a	Palatal	Oral	Stretched	Voiced	Fourth
A	Palatal	Oral	Stretched	Voiced	Fourth
y	Post-palatal	Oral	Rounded	Voiced	First
eu	Post-palatal	Oral	Rounded	Voiced	Second
oe	Post-palatal	Oral	Rounded	Voiced	Third
oe~	Post-palatal	Nose	Rounded	Voiced	Third
@	Post-palatal	Oral	Rounded	Voiced	Fourth
u	Velar	Oral	Rounded	Voiced	First
o	Velar	Oral	Rounded	Voiced	Second
O	Velar	Oral	Rounded	Voiced	Third
o~	Velar	Nose	Rounded	Voiced	Third
a~	Velar	Nose	Rounded	Voiced	Fourth
Semivowels					
w	Velar	Oral	Rounded	Voiced	First
H	Post-palatal	Oral	Rounded	Voiced	First
j	Palatal	Oral	Stretched	Voiced	First
Consonants					
p	Bilabial	Plosive	PLP-Undefined	Unvoiced	AP-undefined
b	Bilabial	Plosive	PLP-Undefined	Voiced	AP-undefined
m	Bilabial	Nose	PLP-Undefined	Voiced	AP-undefined
f	Labiodental	Constrictive	PLP-Undefined	Unvoiced	AP-undefined
v	Labiodental	Constrictive	PLP-Undefined	Voiced	AP-undefined
t	Alveodental	Plosive	PLP-Undefined	Unvoiced	AP-undefined
d	Alveodental	Plosive	PLP-Undefined	Voiced	AP-undefined
n	Alveodental	Nose	PLP-Undefined	Voiced	AP-undefined
s	Alveolar	Constrictive	PLP-Undefined	Unvoiced	AP-undefined
z	Alveolar	Constrictive	PLP-Undefined	Voiced	AP-undefined
S	Post-alveolar	Constrictive	PLP-Undefined	Unvoiced	AP-undefined
Z	Post-alveolar	Constrictive	PLP-Undefined	Voiced	AP-undefined
k	Velar	Plosive	PLP-Undefined	Unvoiced	AP-undefined
g	Velar	Plosive	PLP-Undefined	Voiced	AP-undefined
N	Velar	Nose	PLP-Undefined	Voiced	AP-undefined
G	Palatal	Nose	PLP-Undefined	Voiced	AP-undefined
R	Post-velar	Liquidly	PLP-Undefined	Voiced	AP-undefined
l	Lateral	Liquidly	PLP-Undefined	Voiced	AP-undefined

Appendix C

Distribution of liaisons in the CPROM-PFC corpus

Nature of the final latent consonant	Number of occurrences of that context in the corpus			
	Facultative	Obligatory	Forbidden	Total
n	24	413	714	1151
p	8	2	0	10
r	194	24	0	218
t	1170	465	34	1669
z	905	522	1230	2657

Figure C.1 - *Distribution of liaisons in the CPROM-PFC corpus, according to the type of liaison and the nature of the latent consonant*

Appendix D

List of Combilex phonetic symbols

Table D.1 - *List of phonemes used in Combilex*

Symbol	Type	Keyword
p	cons	PAT
t	cons	TAT
?	cons	
t	cons	
k	cons	CAT
b	cons	BAT
d	cons	DAB
g	cons	GAP
m	cons	MAT
n	cons	NAP
N	cons	PANG
T	cons	MATH
D	cons	THAT
f	cons	FAT
v	cons	VAT
s	cons	SAT
z	cons	ZAP
S	cons	SHAM
Z	cons	BEIGE
tS	cons	CHAT
dZ	cons	JAB
h	cons	HAT
l	cons	LAD
lw	cons	

LIST OF COMBILLEX PHONETIC SYMBOLS

Symbol	Type	Keyword
r	cons	RAT
x	cons	LOCH
j	cons	YAP
w	cons	WAG
W	cons	WHACK
K	cons	LLANELLI
m!	cons	SPASM
n!	cons	GARDEN
l!	cons	CATTLE
E	vowel	DRESS
a	vowel	TRA
A	vowel	PALM
Ar	vowel	START
Er	vowel	PERT
@@r	vowel	NURSE
@@	vowel	DIEU
@U	vowel	GOAT
Q	vowel	LOT
QO	vowel	CLOTH
O	vowel	THOUGHT
Or	vowel	NORTH
@Ur	vowel	FORCE
i	vowel	FLEECE
iy	vowel	HAPPY
I	vowel	KIT
@	vowel	COMMA
@r	vowel	LETTER
V	vowel	STRUT
U	vowel	FOOT
u	vowel	GOOSE
eI	vowel	WAIST

Symbol	Type	Keyword
aI	vowel	PRICE
ae	vowel	TIED
aer	vowel	FIRE
OI	vowel	CHOICE
OIr	vowel	COIR
aU	vowel	MOUTH
aUr	vowel	HOUR
i@	vowel	IDEA
I@	vowel	NEARING
E@	vowel	SQUARING
U@	vowel	JURY
jU@	vowel	NEURAL
e	vowel	VIN
o	vowel	BON
9	vowel	BRUN

LIST OF COMBILED PHONETIC SYMBOLS

Appendix E

The set of hand-written rules
exploited to populate the
lattices

THE SET OF HAND-WRITTEN RULES EXPLOITED TO POPULATE THE LATTICES

Rule	Description
ANY < t WORD BOUNDARY > ANY -> ?	Any end of word t can be glottalised.
ANY < z > ANY -> s	Any z can be devoiced.
ANY < STOP BOUNDARY > STOP -> DELETION	Any two stops in a row across a boundary, the first can be deleted.
ANY < STOP WORD BOUNDARY > ANY -> DELETION	Any word final stop can be deleted.
PHRASE BOUNDARY < VOWEL > ANY -> REDUCTION	Any phrase-initial vowel can be reduced.
ANY < d WORD BOUNDARY > ANY -> ?	Any word final d can be glottal.
ANY < g WORD BOUNDARY > ANY -> ?	Any word final g can be glottal.
n < k WORD BOUNDARY > ANY -> g	A k after an n before a word boundary can be voiced.
j < u > ANY -> DELETION	A u preceded by a j can be deleted.
ANY < d > D -> DELETION	A d followed by a D can be deleted.
ANY < t > s -> DELETION	A t followed by a s can be deleted.
CONSONANT NO BOUNDARY < I > CONSONANT NO BOUNDARY -> DELETION	A I between two consonants in a word can be deleted.
f < @ > STOP -> DELETION	A @ after an f and before a stop can be deleted.
ANY < VOICED CONSONANT > SEN- TENCE BOUNDARY -> VOICELESS CONSONANT	Any voiced consonant at the end of a sen- tences can be voiceless.

Appendix F

PER differences between the improved systems and the Gold Standard annotation

Table F.1 - *PER differences between the improved systems and the Gold Standard annotation.*

	Ins	Del	Sub	Total	PER
Read					
Standard	10	149	151	310	19.1 %
Lattice w/ Combilex	6	139	184	329	20.2 %
Lattice w/ rules	20	106	120	246	15.2 %
Lattice w/ both	22	101	142	265	16.3 %
Spontaneous					
Standard	17	202	180	399	25.2 %
Lattice w/ Combilex	9	178	199	386	24.4 %
Lattice w/ rules	37	133	134	304	19.2 %
Lattice w/ both	38	130	145	313	19.7 %

PER DIFFERENCES BETWEEN THE IMPROVED SYSTEMS AND THE GOLD
STANDARD ANNOTATION

-

Appendix G

Our publications and related talks

G.1 Regular papers in journals

1. **S. Brognaux**, T. Drugman. (2016). HMM-Based Speech segmentation: Improvements of Fully Automatic Approaches, *IEEE Transactions on Audio, Speech and Language Processing*, 24(1):5-15.
2. M. Avanzi, **S. Brognaux**. (2016). A Multifactorial Analysis of Word-Final Post-Obstruent Liquid Behavior in French, *Journal of Phonetics* (*Under review*).
3. **S. Brognaux**, T. Drugman. (2014). Variations phonétiques: Impact de la situation de communication, *Cahiers de la linguistique française*, 31, pp. 63-76.
4. **S. Brognaux**, B. Picart, T. Drugman. (2014). La synthèse de commentaires sportifs: intégration d'une annotation prosodique à deux niveaux dans un synthétiseur HMM, *Cahiers de la linguistique française*, 31, pp. 223-236.
5. **S. Brognaux**, S. Roekhaut, T. Drugman, R. Beaufort. (2012). Automatic Phone Alignment. A Comparison between Speaker-Independent Models and Models Trained on the Corpus to Align, *Lecture Notes in Computer Science*, 7614, pp. 300-311.

G.2 Papers in conference proceedings

1. R. Dall, **S. Brognaux**, K. Richmond, C. Valentini-Botinhao, G. E. Henter, J. Hirschberg, J. Yamagishi, S. King. (2016). Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis, *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shangai (China) (*Under review*).

2. **S. Brognaux**, M. Avanzi. (2015). Sociophonetics of phonotactic phenomena in French, *Proc. of the International Conference on phonetic sciences (ICPhS)*, Glasgow (United Kingdom).
3. B. Picart, **S. Brognaux**, S. Dupont. (2014). Analysis and automatic recognition of human beatbox sounds: A comparative study, *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane (Australia).
4. **S. Brognaux**, B. Picart, T. Drugman. (2014). Speech synthesis in various communicative situations: Impact of pronunciation variations, *Proc. of Interspeech*, Singapore, Singapore [*ISCA Best Student Paper award*].
5. S. Roekhaut, **S. Brognaux**, R. Beaufort, T. Dutoit. (2014). eLite-HTS. a NLP tool for French HMM-based speech synthesis, *Proc. of Interspeech Show & Tell*, Singapore (Singapore).
6. **S. Brognaux**, S. Roekhaut, T. Drugman, R. Beaufort. (2014). Train&Align: Un outil d'alignement phonétique automatique disponible en ligne, *Proc. of the Journées d'étude sur la parole (JEP)*, Le Mans (France).
7. S. Roekhaut, **S. Brognaux**, R. Beaufort, T. Dutoit. (2014). eLite-HTS : un outil TAL pour la génération de synthèse HMM en français, *Proc. of the Démonstrations aux Journées d'étude de la parole (JEP)*, Le Mans (France).
8. M.-C. Michaux, **S. Brognaux**, G. Christodoulides. (2014). The production and perception of L1 and L2 Dutch stress, *Proc. of Speech Prosody*, Dublin (Ireland).
9. **S. Brognaux**, T. Drugman. (2014). Phonetic variations: Impact of the communicative situation, *Proc. of Speech Prosody*, Dublin (Ireland).
10. **S. Brognaux**, T. Drugman, M. Saerens. (2014). Synthesizing sports commentaries: One or several emphatic stresses ?, *Proc. of Speech Prosody*, Dublin (Ireland).
11. B. Picart, **S. Brognaux**, T. Drugman. (2013). HMM-based Speech Synthesis of Live Sports Commentaries: Integration of a Two-Layer Prosody Annotation, *Proc. of the 8th ISCA Speech Synthesis Workshop (SSW8)*, Barcelone (Spain).
12. **S. Brognaux**, T. Drugman. (2013). A New Prosody Annotation Protocol for Live Sports Commentaries, *Proc. of Interspeech*, Lyon (France).
13. **S. Brognaux**, S. Roekhaut, T. Drugman, R. Beaufort. (2012). Train&Align: A New Online Tool for Automatic Phonetic Alignment, *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, Miami (USA).
14. **S. Brognaux**, T. Drugman, R. Beaufort. (2012). Automatic Detection of Syntax-based Prosody Annotation Errors, *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, Miami (USA).

G.3 Talks and posters (with abstract)

1. **S. Brognaux**, M. Avanzi. (2015). Dialectal Variation of French Schwa in the light of Speaker-related Factors, Phonetics and Phonology in Europe 2015 (PAPE 2015), Cambridge (United Kingdom).
2. M. Avanzi, **S. Brognaux**. (2015). A Multifactorial Analysis of French Liaison, Phonetics and Phonology in Europe 2015 (PAPE 2015), Cambridge (United Kingdom).
3. **S. Brognaux**, S. Roekhaut, T. Drugman, R. Beaufort. (2014). Train&Align : A New Online Tool for Automatic Phonetic Alignment, *Le langage dans tous ses états: 6ème Journée Scientifique du Pôle Hainuyer (JSPH6)*.
4. **S. Brognaux**, S. Roekhaut, T. Drugman, R. Beaufort. (2013). Train&Align : A New Online Tool for Automatic Phonetic Alignment. *ICTEAM Young Researcher's day (UCL)*.
5. **S. Brognaux**, S. Roekhaut. (2012). Train&Align: Outil d'alignement phonétique de corpus oraux. *Journée Linguistique du Cercle Belge de Linguistique (CBL)*, Université de Liège (ULg).

G.4 Invited talks (without abstract)

1. **S. Brognaux**. (September 3rd 2014). *Speech synthesis in various communicative situations: Impact of pronunciation variations*, Speech Lab Talks (University of Columbia in the city of New York).
2. **S. Brognaux**. (May 22nd 2013). *Train&Align : A New Online Tool for Automatic Phonetic Alignment*, Phonetics Labs Talks (Universiteit Leiden).
3. **S. Brognaux**. (April 16th 2013). *Train&Align : un nouvel outil d'alignement phonétique automatique disponible en ligne*, Cental Lab Seminars and ILC Workshops, Université catholique de Louvain (UCL).

G.5 Technical reports

1. S. Roekhaut, **S. Brognaux**, R. Beaufort. (2012). *Syllabation graphémique automatique à l'aide d'un dictionnaire phonétique aligné*, CENTAL, Groupe Norme, Louvain-la-Neuve (Belgium).

