Phonogenre Identification: A Perceptual Experiment with 8 Delexicalised Speaking Styles

Jean-Philippe Goldman[1], Tea Pršir[1,2], George Christodoulides[2], Anne Catherine Simon[2], Antoine Auchlin[1]

[1] Department of Linguistics, University of Geneva
[2] Centre Valibel, IL&C, Université catholique de Louvain
<jean-philippe.goldman,tea.prsir,antoine.auchlin@unige.ch
anne-catherine.simon@uclouvain.be, george@mycontent.gr>

Résumé

Cet article examine la contribution relative des marqueurs prosodiques dans l'identification de genres de parole. Il présente les résultats d'une série d'expériences de perception d'échantillons brefs en français, dans 8 genres de parole différents, en versions filtrée (pour masquer le contenu linguistique) et non filtrée, par une population francophone, et une non-francophone. Les résultats globaux montrent un effet de filtrage moins important que prévu pour les francophones, prouvant ainsi le rôle de la prosodie dans la perception et l'identification du phonogenre, ainsi qu'un effet culturel plus important qu'escompté entre les deux populations, montrant que les caractéristiques de certains genres dépendent de la langue envisagée. Enfin, les résultats détaillés mettent en avant la stabilité des caractéristiques de certains genres par rapport aux autres.

Keywords: *phonogenre, phonostyle, speaking style, perception*

1. Introduction

1.1. Prosodic approaches to speaking styles

Forty years ago, Fónagy & Fónagy (1976, 195) tested the ability of French listeners to identify speaking styles (sermons, scientific lectures, political speeches, TV news bulletins, traditional tales, theatrical plays and spontaneous conversation) by listening to filtered speech samples. The best identification was achieved for broadcast news (correctly identified by 38 subjects out of 40). This small-scale experiment allowed them to conclude that there are professional speaking styles, expressed by prosodic cues only, and to illustrate the so-called "fonction identificatrice de la prosodie" (1976, 194, see also Léon 1993, 21-22, *i.a.*).

> « Il est, en effet, parfaitement possible de reconnaître, d'une autre pièce, par les murs qui absorbent les mots et ne laissent passer que les structures ryhmiques et mélodiques, la causerie scientifique, les nouvelles du jour, le reportage sportif, sans parler du discours politique ou du sermon. On peut

> se convaincre du bien-fondé de cette impression en présentant à des auditeurs quelques passages de textes sonores –appartenant à différents genres oraux– après avoir éloigné les "paroles", c'est-à-dire, après avoir éliminé d'une façon ou d'une autre tout ce qui pourrait permettre de reconnaître les phonèmes et par conséquent, les mots du texte. » (Fónagy & Fónagy 1976, 193-194)
>
> "Someone listening to sound coming from another room, through walls that obscure words and dampen everything except the sound's rhythmic and melodic patterns, will be perfectly capable of recognising a scientific lecture, the daily news, a sports commentary, or (even more so) a political speech or a sermon. The validity of this claim can be demonstrated by presenting sound samples –of different speaking styles– to subjects, after somehow removing any content or clue that could allow them to recognise the phonemes and therefore the words spoken". [Our translation]

A long tradition of approaches to those "non-linguistic" functions of prosody developed, ranging from the prosodic cues to emotion (e.g. Banzinger & Scherer 2005; Grichkovtsova, Morel & Lacheret 2012) to speaking styles (Johns-Lewis 1986; Hirschberg 2000; Goldman, Pršir, Christodoulides & Auchlin 2014), and from regional accents (Simon 2012, Bardiaux 2014) to genre identification (Arnold 2012; Pépiot 2013). Regarding speaking style identification, most of the studies compared 2 or 3 styles (usually reading vs. conversation) and derived the prosodic cues specific to each speech situation (see Llisterri 1992 for a review). More recently, large annotated corpora of natural speech became available and allowed for new types of research, such as the automatic classification of speech samples from prosodic cues alone (e.g. Obin, Lacheret, Veaux, Rodet & Simon 2008) or the (semi-)automatic description of phonostyles, i.e. research on the prosodic correlates of speech situations (Goldman, Auchlin & Pršir 2014; Pršir, Goldman & Auchlin 2014).

Building upon studies based on the extraction of acoustic features, a perceptual approach to prosody highlights the effects of these prosodic cues to the perception and identification of regional accents, emotions, or speaking styles. For example, Castro, Frietas, Moraes & Serridge (2010) conducted an experiment on the perceptual identification of professional speaking styles in Brazilian Portuguese. Speech samples representative of 4 styles (TV news broadcasts, sermons, political debates in the senate, and interviews) were filtered and presented randomly to native speakers of Brazilian Portuguese. The identification task consisted of choosing between two speaking styles, "one corresponding to the speaking style being presented, and one chosen randomly among the other three speaking styles" (2010, 2). Religious sermons were best identified (accuracy 96.7 %), followed by TV news (90.0 %), interviews and political speeches (86.7 %).

However one should be cautious when comparing these results with the results in the present paper, due to numerous differences in methodology.

1.2. Hypotheses

This paper presents the results of a series of perceptual experiments to study the role of prosody in the identification of speaking styles (phonogenres). Previous research has shown broad differences in prosodic features between different communicative situations, so that it seems plausible to speak of "speaking styles". Are prosodic features sufficient for a listener to distinguish between speaking styles, in the absence of any other information? For example, is it possible to identify a given speech sample as belonging to a conversation, a live sports commentary, a broadcast news bulletin, or a radio show on popular science, without access to its lexical content? Are some speaking styles easier to identify because of their specific prosodic characteristics, and which speaking styles are particularly marked? To what extent is it possible to distinguish between speaking styles sharing some prosodic features? Which prosodic features are the most important in identifying speaking styles?

In order to address these questions, we conducted a perception experiment in which two groups of participants (native speakers of French, and participants without any knowledge of French) were asked to attribute a speaking style to French speech samples presented to them.

Our objective was to test the following hypotheses:

- prosodic cues convey sufficient information for the identification of speaking style by listeners;
- non-filtered (French) speech is better identified than filtered (French) speech, both by French-speaking and by non-French-speaking listeners;
- since the samples were extracted from French speech, phono-styles are better identified by French-speaking participants than by non-French-speaking participants; this should be the case both for filtered and for non-filtered samples;
- some speaking styles are better distinguished than others, because they belong to an established tradition and/or because listeners are more familiar with them (e.g. radio broadcast news, sermons, etc.).

2. Experiment design and material

Participants listened to samples from eight speaking styles: CONVERSATION (informal chat between two persons), DIDACTIC (a

radio show on popular science), MAPS (a person giving directions), NEWS (radio broadcast news), READ (reading a text), LITURGICAL (religious sermon), POLITICAL (parliamentary speeches), SPORT (live sports commentary). The experiments were conducted online (on labguistic.com) and included 40 audio extracts with an average duration of 12 seconds. For each speaking style, five different speakers were included. The stimuli were either filtered (with a band-stop filter 400-5000 Hz), in order to make it impossible to understand the words spoken (de-lexicalisation), or not filtered. Sounds were levelled in order to minimise differences arising from recording conditions. The orthographic transcription of two stimuli is given as an example:

(1) POLITICAL "...qui sont accessibles parfois huit ou dix mois plus tard et souvent trop tard / les hôpitaux qui ont été restructurés brutalement / les urgences trop éloignées à la campagne ou saturées dans les villes…"

(2) DIDACTIC "... parlons d'un organe / ou plutôt de deux organes / et ces deux organes ce sont les reins // les reins ce sont des organes très mystérieux // quand quelqu'un n'a pas l'air très franc du collier / quand on s'demande ce qu'il pense / on dit qu'il faudrait lui sonder le coeur et les reins..."

The stimuli were presented randomly and subjects could listen to them as many times as they wanted. Subjects participated either in the "filtered speech" (Filt) experiment, or in the "original samples" (nFilt) experiment. For each audio sample, they were asked to select one of the eight speaking styles or the option "I don't know".

In total, 290 subjects participated in the experiment. Table 1 summarises the number of subjects according to their knowledge of French (native French speakers, hence Fr, and speakers without any knowledge of French, hence nFr) and the two conditions (filtered and non-filtered speech samples). The mother tongue of the 80 participants who reported no knowledge of French was distributed as follows: 50 South Slavic (Serbian 29, Croatian 21), 14 English, 13 Brazilian Portuguese and, finally, one Norwegian, one Cantonese and one German L1 speaker.

	Non-filtered (nFilt)	Filtered (Filt)	Total
Native French Speakers (Fr)	31	179	210
French non-Speakers (nFr)	45	35	80
Total	76	214	290

Table 1: *Number of participants*

The percentage of "I don't know" answers for each group is shown in Table 2. These are the cases in which subjects couldn't associate any of eight proposed speaking styles to the stimuli.

% unknown	nFilt	Filt
Fr	3.87	3.98
nFr	4.11	6.47

Table 2: *Percentage of answers as "I don't know" in the four conditions*

At the end of the experiment, the subjects were asked whether they found the task easy or difficult; whether they did recognise some words (both in the filter and the non-filtered conditions). The participants also provided feedback in free text form, which was analysed to understand their reactions. Global quantitative results are presented in section 3.1, more details in section 3.2, and a summary of the participants' qualitative comments in section 3.3.

3. Results

3.1. Global results

Table 1 shows the percentage of correct phonogenre identification for the four conditions (Fr vs. nFr and Filt vs. nFilt). Note that the answer "I don't know" was excluded from these results.

% correct	nFilt	Filt
Fr	91.19	60.63
nFr	58.17	36.84

Table 3: *Percentage of correct answers in the four conditions*

On the first row of Table 3, we observe that French speakers listening to non-filtered samples (Fr-nFilt) managed to correctly recognise the speaking style in 91.19 % of the cases. This shows that even with access to the linguistic content the accuracy does not reach 100 %. A possible explanation for this would be that speaking style labels were confounded. Furthermore, the percentage of correct recognition for French speakers in the filtered condition (Fr-Filt) is 60.63 %, which is much higher than the level of chance (12.5 %). This confirms the hypothesis that prosodic cues help listeners identify speaking styles despite the absence of linguistic content. Even though the difference between the two conditions (Fr-Filt vs. Fr-nFilt) is not as high as expected, the filtering process (and the consequent absence of content) had a negative impact on the recognition accuracy: it dropped significantly, by 31 % ($\chi^2(1,8067)=417.97$, $p<0.001$). Under both experimental conditions (Fr-Filt and Fr-nFilt) 4 % of the subjects chose the "I don't know" option.

On the second row of Table 3, we notice that the non-French-speaking subjects correctly recognised the speaking style in 58.17 % of

the cases under the non-filtered condition (nFr-nFilt), and in 36.84 % of the cases under the filtered condition. A slightly greater proportion of "I don't know" answers were recorded in the filtered condition (nFilt: 4.11 % vs. Filt: 6.47 %). In these two conditions, the participants did not have access to the linguistic content because of the language barrier, but they could still identify phonogenres at a level better than chance. Filtering reduced the recognition accuracy by 22 %, still significant ($\chi^2(1,2999)=133.33$, $p<0.001$).

When comparing the two filtered conditions (Fr-Filt, nFr-Filt) we note that the recognition accuracy drops by 24 % for non-French-speaking subjects. While one might expect that hiding the linguistic content by means of filtering would put the two groups on an equal footing, this difference is statistically significant ($\chi^2(1,8148)=247.76$, $p<0.001$). This may be attributed to a cultural effect: French-speaking participants are more familiar with the prosodic features of French speaking styles, which may be language-specific. We should also note that some ambient acoustic features remained even in the filtered stimuli: e.g. an echo suggests that speech was produced in a large hall like a church (leading the participant to choose the "sermon" option), while a crowd roar is a hint that the recording took place in a stadium (leading the participant to select the "sports commentary" option). However, a similar, statistically significant difference is observed between the two non-filtered conditions (Fr-nFilt vs. nFr-nFilt: drop by 33 % with $\chi^2(1,2918)=378.57$, $p<0.001$), a finding that supports the cultural effect hypothesis.

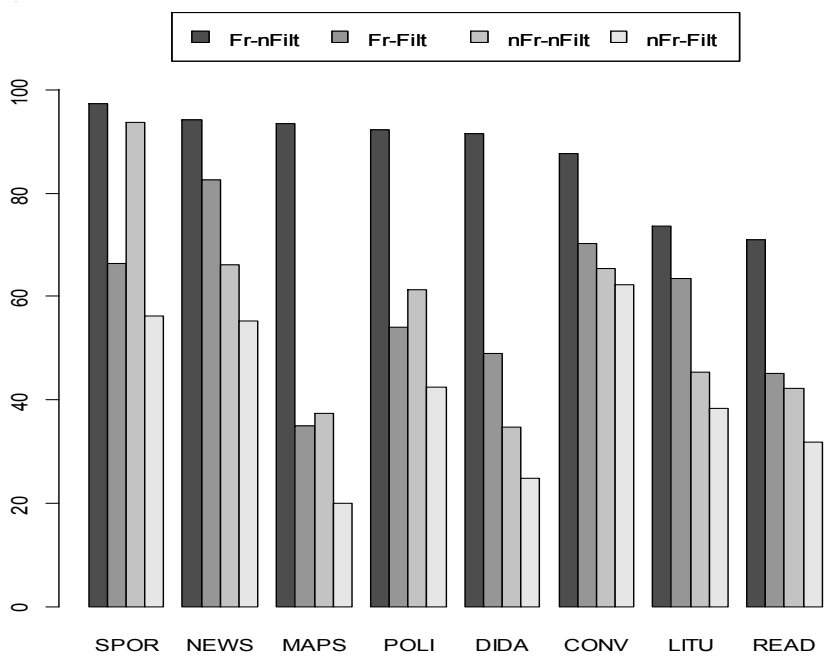
In summary, hiding the linguistic content does not prevent subjects from recognising the speaking style (phonogenre): prosodic and acoustic features convey enough information guide the listener in the identification task. The experiment's global results can be effectively explained as an interplay of two different effects: a "filter effect" that degrades the identification accuracy by 31 % (resp. 22 %) for Fr subjects (resp. nFr subjects), and a "cultural effect" that reduces the identification accuracy by 33 % (resp. 24 %) for the Filt condition (resp. the nFilt condition).

3.2. Detailed results

A closer look at the results could help finding answers to questions such as: which genres are more easily recognized? Which ones are more sensitive to the experimental condition (subject native tongue/ filtering)? Do transcultural phonogenres exist (i.e. would have a good recognition score for nFr subjects)?

Figure 1 shows the percentage of correct recognition for the 8 phonogenres and across the 4 experimental conditions (and Table 4

includes the precise figures) sorted by decreasing percentage of good recognition for the first condition (Fr-nFilt). Figure 2 shows the confusion matrix for the 4 conditions.



		CONV	LITU	MAPS	NEWS	DIDA	POLI	READ	SPOR	UNK
Fr	nFilt	87.74	73.55	93.55	94.19	91.61	92.26	70.97	97.42	3.87
Fr	Filt	70.17	63.46	34.86	82.68	49.05	53.97	45.03	66.48	6.15
nFr	nFilt	65.33	45.33	37.33	66.22	34.67	61.33	42.22	93.78	4.11
nFr	Filt	62.35	38.24	20.00	55.29	24.71	42.35	31.76	56.14	5.85

Figure 1 & Table 4: *Percentage of correct recognition (accuracy) for the 8 phonogenres, across the 4 experimental conditions*

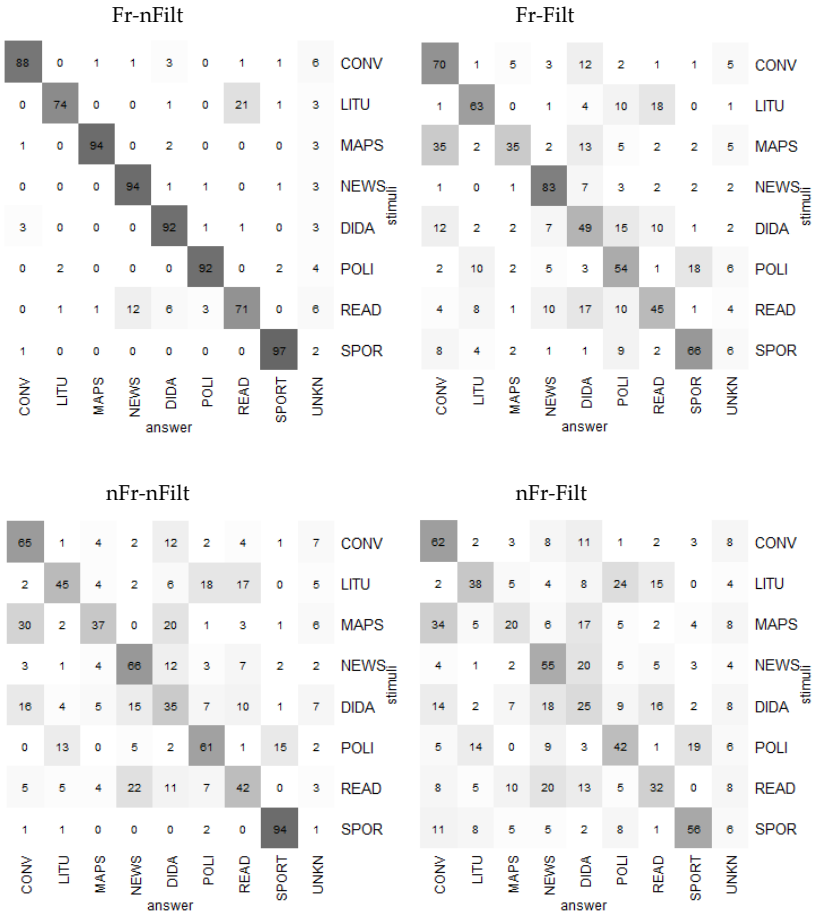


Figure 2: *Confusion table for the four conditions*

Under optimal conditions (no filtering, no language barrier i.e. Fr-nFilt), 5 phonogenres are very well recognized, with an accuracy exceeding 90 % (SPOR 97 %, NEWS 94 %, MAPS 94 %, DIDA 92 % and POLI 92 %), 1 phonogenre reaches 88 % (CONV) and 2 phonogenres are more difficult to distinguish (LITU 74 % and READ 71 %). By observing the confusion matrix, we notice that LITU is often (21 %) confounded with reading (READ), and that reading (READ) is often confounded with NEWS (12 %), or remains unidentified (6 %). Two hypotheses could explain this: first, the READ stimuli could be confusing as they are neutral reading of newspapers and could therefore be perceived as NEWS. Second, the READ genre refers to a situational feature (read vs.

non-read) which is present in several genres (e.g. LITU and NEWS are probably read speech, and therefore share properties of that production condition).

Under the filtered speech condition, French-speaking subjects (Fr-Filt) achieved the best identification accuracy for the NEWS genre (83 %). This can be explained by the fact that this phonogenre is homogeneous, i.e. the prosodic features of the samples belonging to this speaking style are very similar (Pršir, Goldman & Auchlin 2014), and by the fact that listeners are very familiar with broadcast news, while genres like parliamentary debates (POLI), or a radio show on popular science (DIDA) reach a narrower audience and are less prototypical. The next three genres, in order of recognition accuracy, are CONV (70 %), SPOR (66 %) and LITU (63 %): they are also stereotyped and listeners are exposed to such speaking situations almost daily.

It is interesting to notice that MAPS is recognised very well under the unfiltered condition, but not under the filtered condition (35 %), where many subjects confused it with CONV. This can be explained by the fact that MAPS and CONV share situational features, as they are both forms of spontaneous dialogue, and sound similar when filtered. LITU (and DIDA to a lesser extent) is confounded with READ, which is, as mentioned above, a more general genre label. We could actually consider LITU to be a sub-genre of read speech which is delivered in the context of religious service. This highlights a problem with the experimental design itself: the choice of labels for speaking styles is not unequivocal. READ is correctly identified in only 45 % of the cases, and confounded with other read styles (LITU 8 %, POLI 10 %) or prepared speech (DIDA 17 %). Finally, POLI is not identified very well (54 %), as it is often misperceived for SPOR (18 %) and LITU (10 %), two styles that are also very expressive.

Comparing the perception of participants who speak French with the perception of those who don't gives us some indication about the extent of the cultural effect. SPOR remains the genre best identified under the unfiltered speech condition (94 %), followed by NEWS (66 %), CONV (65 %) and POLI (61 %). As expected, the MAPS genre is clearly identified mainly based on its lexical content and not based on its prosodic features. It is very often confused with CONV (30 %) and even DIDA (20 %), two genres sharing situational characteristics with it (spontaneous or semi-prepared dialogue).

The three genres best identified by non-Francophone participants under the filtered condition are the same with the ones best identified under the unfiltered speech condition, but in a different order: CONV is better recognized (62 %) than SPOR (56 %), and NEWS is correctly

identified in 55 % of the cases. The trends are the same in filtered speech judged by non-Francophone subjects, but with lower scores: POLI is recognized by 42 % and confounded with SPOR (19 %) or LITU (14 %); LITU (correctly recognised in 38 % of the cases) coincides with POLI (24 %) and READ (15 %) for the reasons explained above. MAPS (20 %) and DIDA (25 %) are very poorly identified, either because the lexical content is required (MAPS is confounded with CONV in 34 % of cases and in 17 % with DIDA). Possibly these are genres that are not well-defined, or genres where differences between languages and cultures are more pronounced.

3.3. Summary of the participants' qualitative feedback

In the feedback provided after the experiment, both French-speaking and non-French-speaking participants acknowledged that they found the filtered speech condition quite difficult.

Under the filtered speech condition, only 18 out of the 179 French-speaking participants indicated whether they had recognised some words (despite the filtering) that helped them reach the decision: 9 participants claim they did and 9 say they did not. Regardless of their mother tongue, some non-Francophone participants reported that they recognised some French words.

On the other hand, under the unfiltered speech condition, 23 out of the 45 non-Francophone subjects said they did recognise some words that helped them in their decisions, and 8 said they did not.

Finally, the participants reported that it was quite easy to recognise the sport phonogenre (because they are familiar with stadium noise: loud crowd roar) and liturgical recordings (because of the typical echo at the church). Beyond prosodic features, therefore, such contextual information is crucial in recognising speaking styles. Religious sermon samples that were recorded in a studio and not in a church were often confounded with political speech. This is compatible with the findings of one of our previous studies (Pršir, Goldman & Auchlin 2014) where we showed that these two phonogenres share many prosodic features.

4. Discussion

Our findings confirm the claim made by Fónagy & Fónagy (1976), by showing that it is possible to recognise a speaking style even in the absence of linguistic content, and this with higher discrimination on a larger set of phonogenres' samples. Speech filtering prevented the participants from understanding the words spoken, yet the recognition accuracy of speaking styles was well above the level of chance.

This series of experiments assumes that "genres" are well defined by means of situational cues, and that samples are sufficiently "homogeneous" (global prosodic cues within samples vs. local episodic cues). The detailed results of the experiment show that the genres that were often confounded are the ones that share situational features. Furthermore, some speaking style labels are too vague and this hampered the identification accuracy. Some subjects reported that their decision was influenced by word they did manage to recognise despite the filtering process. Future experiments will take into consideration the subjects' reaction time, in order to evaluate which genre are more easily recognised, and manipulate prosodic cues/parameters within samples, in order to better understand which parameter trigger the identification of which genre.

By extending the experiment to participants who did not understand or speak French, we have found the impact of filtering and the impact of the language barrier to be cumulative. It is therefore possible that there is also a cultural effect in play, in the sense that non-Francophone participants are less familiar with French speaking styles. However, assuming that situational equivalency (e.g. two parliaments in two different countries/cultures) should elicit equivalent speaking style prosodic properties was an implicit assumption in the present study. This issue clearly must be addressed specifically (Christodoulides, in prep.) for large scale, cross-cultural speech style research. It might indeed be the case that there is no perfect equivalence between phonogenres across languages, and that the expectations as to genre identification greatly vary from one language to another.

Acknowledgments

This research is funded by Swiss National Science Foundation – FNS Grant n° 100012_134818.

References

Bardiaux, A. (2014). *La prosodie de quelques variétés de français en Belgique : analyse perceptive et acoustique*. Thèse de doctorat. Université catholique de Louvain, Louvain-la-Neuve.

Boula de Mareuïl, P., A. Rilliard & A. Allauzen (2012). A Diachronic Study of Initial Stress and other Prosodic Features in the French News Announcer Style: Corpus-based Measurements and Perceptual Experiments. *Language and Speech, 55(2)*, 263-293.

Castro, L., M. Freitas, J. Moraes & B. Serridge (2010). Listeners' Ability to Identify Professional Speaking Styles Based on Prosodic Cues. *Proceedings of the 7th conference on Speech Prosody*, Chicago.

Christodoulides, G. (in prep.). Phonogenres without borders? A cross-linguistic study of speaking style variation.

Fónagy, I. & J. Fónagy (1976). Prosodie professionnelle et changements prosodiques. *Le Français Moderne*, *44*, 193‑228.

Goldman, J.-P., T. Pršir, G. Christodoulides & A. Auchlin (2014). Speaking style prosodic variation: an 8-hour 9-style corpus study. *Proceedings of the 7th conference on Speech prosody*, Dublin.

Goldman, J.-P., A. Auchlin & T. Pršir (2014). C-PhonoGenre: a speech corpus in French 8-hours 8-speaking styles: relations between situational features and prosodic properties. *Proceedings of LREC*, Reykjavik.

Hirschberg, J. (2000). A corpus-based approach to the study of speaking style. In Horne, M. (Ed.), *Prosody: Theory and Experiments* (pp. 335-350). Kluwer Academic Publisher, The Netherlands.

Johns-Lewis, C. (1986). Prosodic differentiation of discourse modes. In Johns-Lewis, C. (Ed.), *Intonation in discourse* (pp. 199-219). London: Croom Helm.

Joos, M. (1968) The isolation of styles. In Fishman, J.A. (Ed.), *Readings in the Sociology of Language* (pp. 185-191). The Hague: Mouton.

Kain, A. & J. van Santen (2010). Frequency-domain delexicalization using surrogate vowels. *Proceedings of Interspeech*, Makuhari, Japan.

Léon, P. R. (1993). *Précis de phonostylistique. Parole et expressivité*. Paris: Nathan.

Llisteri, J. (1992). Speaking styles in speech research. ELSNET/ESCA/SALT. *Workshop on Integrating Speech and Natural Language* (pp. 15-17). Dublin, Ireland.

Martin, J.R. (1997). Analysing genre: Functional parameters». In Christie, F. & J.R. Martin (Eds.), *Genres and institutions* (pp. 3-39). London: Cassell.

Martti, V., A. Suni, T. Raitio, J. Nurminen, J. Järvikivi & P. Alku (2009). New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis. *Proceedings of Interspeech* (pp. 1703-1706). Brighton, United Kingdom.

Miller, C.R. (1984) Genre as social action. *Quarterly Journal of Speech, 70*, 151-167.

Obin, N., V. Dellwo, A. Lacheret &. X. Rodet (2010). Expectations for Discourse Genre Identification: a Prosodic Study. *Proceedings of Interspeech*, Makuhari, Japan.

Pagel, V. (1999). *De l'utilisation d'informations acoustiques suprasegmentales en reconnaissance de la paroles continue*. Thèse de doctorat. Université Henri Poincaré – Nancy 1.

Pršir, T., J.-Ph. Goldman & A. Auchlin (2014, in press). Prosodic features of situational variation across nine speaking styles in French. In P. Mertens & A.C. Simon (Eds.), *Prosody Discourse interface. Journal of Speech Science*.

Simon, A.C., P. Hambye, A. Bardiaux & Ph. Boula de Mareüil (2012). Caractéristiques des accents régionaux en français : que nous apprennent les approches perceptives ? In Simon, A.C. (Ed.), *La variation prosodique régionale en français* (pp. 27-40). Louvain-la-Neuve: De Boeck Duculot.