



Université catholique de Louvain
Secteur des sciences de la santé
Institut de Neuroscience
Ecole doctorale en sciences biomédicales

Auditory Processing and Categorization: Development, Organization and Reorganization

Siddharth Talwar

2025

Thèse de doctorat présentée en vue de l'obtention du grade de Docteur en
Sciences biomédicales et pharmaceutiques

Promoter

Prof. Dr. Olivier Collignon (UCLouvain)

Supervising Committee

Prof. Dr. Valérie Goffaux (UCLouvain)

Prof. Dr. André Mouraux (UCLouvain)

Prof. Dr. Sylvie Nozaradan (UCLouvain)

President

Prof. Dr. Valérie Goffaux (UCLouvain)

External Members of the Jury

Prof. Dr. Hans Op De Beeck (KU Leuven)

Prof. Dr. Nathan Weisz (Paris Lodron Universität Salzburg)

On the cover: image by Nandana Talwar

Table of Contents

CHAPTER 1- Introduction	15
1.1 General Introduction	15
1.2 Auditory Processing and Categorization	16
1.2.1 Relationship between Acoustic Properties and Categories	18
1.2.2 Are voices special?	21
1.2.3 Auditory Emotions	24
1.3 Hierarchical vs Parallel Processing	27
1.4 Development of Auditory Processing and Categorization	28
1.5 Auditory Processing and Categorization in Blind People	31
1.6 Methodology Techniques	33
1.6.1 Fast Periodic Auditory Stimulation (FPAS)	33
1.6.2 Multivariate Pattern Analysis (MVPA)	34
1.7 Research Goals	36
1.8 Bibliography	37
CHAPTER 2 - Automatic Brain Categorization of Discrete Auditory Emotion Expressions	51
2.1 Introduction	52
2.2 Materials and Methods	54
2.2.1 EEG Experiment	54
2.2.2 Cochlear Model and Envelope	61
2.2.3 Behavioral Experiment	61
2.3 Results	63
2.3.1 Experiment 1- EEG	63
2.3.2 Cochlear Model and Envelope	66
2.3.3 Experiment 2 - Behavior	67
2.4 Discussion	68
2.5 Supplementary Information	72
2.6 Bibliography	77

CHAPTER 3 - Temporal Dynamics of Natural Sound Representation in the Infant Brain.....85

3.1 Introduction.....	85
3.2 Methods.....	87
3.2.1 Stimul.....	87
3.2.2 Participants.....	88
3.2.3 Data Acquisition.....	88
3.2.4 Analysis	89
3.3 Results.....	95
3.3.1 Sound Decoding.....	95
3.3.2 Representation Similarity Analysis (RSA)	96
3.3.3 Deep Neural Networks (DNN)	99
3.3.4 Shared Neural Representations Between Infants and Adults	103
3.4 Discussion	104
3.5 Supplementary Figures	108
3.6 Bibliography.....	109

CHAPTER 4 - Temporal Dynamics of Sound Representation in Sighted and Blind.....115

4.1 Introduction.....	115
4.2 Methods.....	117
4.2.1 Stimuli.....	117
4.2.2 Participants.....	118
4.2.3 Data Acquisition.....	118
4.2.4 Analysis	120
4.3 Results.....	125
4.3.1 Enhanced Sound Decoding in Blind Individuals	125
4.3.2 Disentangling Acoustic, Behavior and Categorical Representations...	127
4.3.3 Deep Neural Network - YAMNet.....	131
4.3.5 Shared Representations Within and Across Populations	136

4.4 Discussion	138
4.5 Supplementary Figures	144
4.6 Bibliography	148
CHAPTER 5 - Discussion	155
5.1 Auditory Emotion Categories in Neurotypical Population	155
5.2 Auditory Representations - Development and (Re)Organization	160
5.2.1 Typical Development - Infants and Neurotypical Adults	160
5.2.2 Reorganization - Blind and Sighted	166
5.2.3 Atypical Development - Infants and Blinds	174
5.4 Linking Results with Theories of Categorization.....	176
5.5 Deep Neural Network (DNN): YAMNet.....	177
5.6 Frequency Tagging vs Multivariate Pattern Analysis.....	180
5.7 Bibliography	181

List of Figures

Figure 1. 1 Acoustic Representations in MEG.....	20
Figure 1. 2 Temporal Voice Area.....	22
Figure 1. 3 Voice selectivity in human adults	24
Figure 1. 4 Mapping of emotion categories	26
Figure 1. 5 Voice selective response in infants.	31
Figure 1. 6 Categorical representation in blind people	33
Figure 1. 7 Pipeline of multivariate pattern analysis	36
Figure 2. 1 Stimuli and Sequence.	56
Figure 2. 2 EEG Results.....	64
Figure 2. 3 EEG time domain results.....	66
Figure 2. 4 Behavior Results	67
Figure S2. 1 Individual baseline subtracted amplitude.....	76
Figure S2. 2 Contrasting ERP between intact and scrambled conditions	77
Figure 3. 1 Stimuli.....	88
Figure 3. 2 Decoding in infants and adults	95
Figure 3. 3 Representation Similarity Analysis	97
Figure 3. 4 Deep Neural Network	100
Figure 3. 5 Time generalization	104
Figure S3. 1 RSA of 4- and 2-category model with EEG.....	108
Figure S3. 2 Correlations between EEG and individual models.....	109
Figure 4. 1 Stimuli.....	118
Figure 4. 2 Decoding pairwise sounds in blinds and sighted.....	126
Figure 4. 3 Representation Similarity Analysis	130
Figure 4. 4 Deep Neural Network	134
Figure 4. 5 Within and across-group comparisons	138
Figure S4. 1 RSA with acoustic models	144
Figure S4. 2 RSA between EEG and all DNN layers	145
Figure S4. 3 Source space RSA with all DNN layers	146
Figure S4. 4 Event Related Potentials.....	147
Figure S4. 5 Cross Category Decoding.....	148
Figure 5. 1 MTF acoustic representation of affective vocalizations.....	157

Figure 5. 2 Emotion categorical and dimensional representations in MEG.....	159
Figure 5. 3 Decoding and RSA within the theta band in Blind and Sighted..	169
Figure 5. 4 Representation of human model with EEG in Blind and Sighted	171
Figure 5. 5 Evidence for cortico-cortical connections	173
Figure 5. 6 Comparison of Infants and Blind Adults.	175
Figure 5. 7 Representation of DNN Layer 4 in EEG of all three populations.....	176
Figure 5. 8 Comparison of DNN results with previous study.	178
Figure 5. 9 Representation of auditory models in space and time in sighted	179

List of Tables

Table S2. 1: Z-scores for all significant channels	72
Table S2. 2 Isolation task : Statistics of scrambled sounds.....	75
Table S2. 3 Isolationtask : Statistics of intact sound	75
Table 3. 1. Peak latencies of all external representational models in EEG	103
Table 4. 1. Peak latencies and 95% confidence intervals for Blind and Sighted..	136

Abstract

Humans possess a remarkable ability to effortlessly categorize natural sounds into ecologically relevant domains. But how does the brain transform sound into meaning? Does this process rely primarily on low-level acoustic features specific to sound categories, or are higher-order auditory mechanisms at play that support categorization and guide behaviour? In this thesis, I use electroencephalography (EEG) to investigate auditory processing and categorization across different populations. First, I show that neurotypical adults can automatically categorize affective vocalizations into discrete emotion categories, the representations of which are at least partially independent of their acoustic properties. Next, I explore how auditory processes and sound categorization of natural sounds emerge in early development by comparing EEG responses of infants aged 4 to 7 months with those of adults. Finally, I present how congenital blindness impacts auditory function showing that while early acoustic processing remains intact, blindness enhances the representations of intermediate stages of auditory processing and, critically, sharpen their brain's categorical and behavioural representation. Collectively, these studies provide new insights into auditory processing and categorization across developing (infants), typical (neurotypical adults), and reorganized (congenitally blind) brains.

Résumé

L'être humain possède une capacité remarquable à catégoriser des sons naturels. Mais comment le cerveau attribue-t-il leurs signification aux sons? Ce processus repose-t-il principalement sur des caractéristiques acoustiques de bas niveau spécifiques aux catégories de sons, ou des mécanismes auditifs d'ordre supérieur sont-ils en jeu pour soutenir la catégorisation et guider le comportement ? Dans cette thèse, j'utilise l'électroencéphalographie (EEG) pour étudier le traitement auditif et la catégorisation dans différentes populations. Tout d'abord, je montre que les adultes neurotypiques peuvent automatiquement classer les vocalisations affectives dans des catégories d'émotions distinctes, dont les représentations sont au moins partiellement indépendantes de leurs propriétés acoustiques. Ensuite, j'explore comment les processus auditifs et la catégorisation des sons naturels émergent au cours du développement précoce en comparant les réponses EEG de nourrissons âgés de 4 à 7 mois avec celles d'adultes. Enfin, je présente l'impact de la cécité congénitale sur la fonction auditive, en montrant que si le traitement acoustique précoce reste intact, la cécité améliore les représentations cérébrales des étapes intermédiaires du traitement auditif et, surtout, améliore les représentations cérébrales catégorielles et comportementales. Collectivement, ces études donnent un nouvel aperçu du traitement auditif et de la catégorisation dans les cerveaux en développement (nourrissons), typiques (adultes neurotypiques) et réorganisés (aveugles congénitaux).

Acknowledgements

No long journey is ever completed alone, and I am deeply grateful for the guidance, support, and encouragement I have received from many along the way.

I would like to express my heartfelt thanks to my promoter, Professor Olivier Collignon, whose unwavering guidance made all of this work possible. I am truly grateful for the opportunity to be part of such a dynamic lab, for the freedom to think, experiment, and express, for the insightful discussions that helped me develop a sense of scientific acuity, and for instilling in me the belief that I could grow as a researcher. Thank you for your confidence, patience, and trust.

I would also like to thank the members of my scientific committee—Professor André Mouraux, Professor Sylvie Nozaradan, Professor Valerie Goffaux, Professor Hans Op de Beeck, and Professor Nathan Weisz. Your feedback, support, and guidance have been instrumental in helping me improve and refine my work.

Throughout this journey, many paths diverged and converged, each leaving a meaningful mark on both me and this work. I would like to thank, in no particular order: Francesca, for her contributions to this work and for the inspiration and knowledge I always drew from her; Marco, for tolerating me in the same office and for our wide-ranging discussions; Iqra, for the cathartic conversations and for guiding me through bureaucratic mazes; Mohamed, for his profound wisdom; Stefania, for her critical contributions to this work and her patience and availability in answering my many questions; Roberta, for her invaluable support and essential contributions; Eléonore, for taking on the huge responsibility of recruiting and testing blind participants; Ceren, for her affability and perceptiveness; Alice, for her smiling support through our parallel PhD journeys; Filippo, for his humour and witty outlook; Rémi, for introducing me to data standardization and for easing tense moments with a laugh; Ying, for her kindness; Jacek, for his critical thinking and constant willingness to help; Federica, for the engaging discussions; Jyothi, for her warmth and welcoming presence; Ane and Virginie, for their geniality; Marie-Ange, Giorgia, Giada, Jeanne, Michele, Irene, Yu Feng, Giulia, Tanisha, and Teodora, for the pleasant and friendly conversations; and Sofia and Rumeysa, who helped me discover a mentor within myself.

I would also like to thank my collaborators—Professor Adelaide de Heering and Laura Bourgaux—for their work in collecting EEG data from infants. A major component of this thesis would not have been possible without their help.

My gratitude also extends to members of other labs: Mrittika, Marius, Jolien, Umut, Kirsten, Hélène, Chloe, Cathy, Veena, Sara, Antoine, Marie, Magda and others for making this experience memorable. Thank you as well to the IPSY community and the administrative staff for their support throughout this journey.

This thesis was made possible not only by scientific but also by non-scientific contributions. Thank you, Papa, Ma, and Nandana—you are the reason I have reached this milestone. Thank you, Francesca, for being my support system and for standing by me through the good times and the bad. And to all my friends in India and Belgium, thank you for making this journey enjoyable.

Finally, I am sincerely grateful to all the participants who took part in the experiments. This thesis would not have been possible without them.

ॐ

Chapter 1

INTRODUCTION

1.1 General Introduction

The pursuit in the field of auditory neuroscience has been to trace the flow of the auditory signal from the moment it reaches the ear to the emergence of behavior. While the functions of the ear and subcortical regions are better known, how the resulting auditory signal is transformed into different representations leading to behavior at the cortical level remains ambiguous. These questions become trickier to answer because humans are able to extract multiple dimensions of information from a single sound and integrate them to a unified percept, such as linguistic content, identity, emotion and location. Moreover, the auditory system communicates with other brain regions that typically perform other functions such as vision and motor tasks, as well as non-sensory cognitive regions such as parts of the frontal cortex and ventral occipital temporal cortex etc, forming large-scale networks for different functions. Adding to this complexity, attention is known to modulate certain auditory processes adapting behavior. Lastly, the interplay of inter-individual variability and generalizations across populations vary across studies hindering the development of a comprehensive framework for understanding the functioning of the auditory cortex. These issues encapsulate the complexity of sound processing in humans and the compelling challenge of unraveling its underlying mechanisms.

One approach to disentangle these issues is by investigating how the auditory system and other brain regions organize their structure and function in developing, typical and atypical populations. While the former reveals the innate auditory architecture and how it develops, the others show how varied sensory experiences interact with this architecture and biological constraints to shape auditory functions. In this context, neuroplasticity becomes the guiding principle influencing the emergence of typical auditory processing, and enhanced or deviant patterns often observed in atypical populations (Heimler et al., 2014). For instance, the most consistent reports of enhanced behavior performance and neural activations to sounds come from blind people (Sabourin et al., 2022). Due to lack of visual input, the visual cortex begins to represent auditory information such as motion, linguistic and categorical representations (Bedny et al., 2011; Collignon et al., 2011; Mattioni et al., 2020). By investigating and comparing these populations, we can better grasp the continuum of auditory processing, from development to organization and/or reorganization, and reveal the mechanisms that guide them at these different stages.

Most of the previous research has predominantly focused on isolating and addressing one or two dimensions of auditory processing at a time while assuming the separation between the dimension in focus and other auditory related functions. For instance, while most human participants perform well in auditory emotion recognition tasks, studies consider only a subset of sound properties that are characteristic to an emotion category (Talwar et al., 2023). Other studies rely on mathematical subtraction of two conditions to isolate an effect, while the subtracted components may have information regarding the concerned effect (Frühholz, S & Belin, P., 2024). Generally, many studies exhibit limited stimuli and participants due to testing constraints and can lack generalization in their results, causing more variability in research (Bauer, 2023). These issues are analogous to the parable of the 'Elephant in the dark', where everyone describes the same animal differently as they are exposed to distinct and isolated parts of the elephant. Nevertheless in recent years, researchers have been able to disentangle multiple auditory dimensions within the same experiment through more robust experimental designs and advanced analysis techniques, such as frequency tagging and multivariate pattern analysis in conjunction with representational similarity analysis, respectively.

This thesis contributes to the field in two ways - firstly, I will investigate development, organization and reorganization of auditory functions in infants, sighted adults and blind adults, respectively; secondly, I will introduce two novel electrophysiological methods to investigate auditory processing and categorization. In Chapter 2, a technique called Fast Periodic Auditory Stimulation (FPAS, Barbero et al., 2021) is introduced to acquire categorical responses to non-verbal emotional vocalizations in typical hearing adult population. Chapter 3 will discuss the temporal dynamics of auditory object processing in infants and adults highlighting to what extent the auditory functions are already developed in 4-7 month old infants and if they already exhibit categorical selectivity that may be independent of the acoustic features of the sounds. Chapter 4 will show how varied sensory experience impacts the temporal dynamics of auditory object processing in blind and sighted adults.

1.2 Auditory Processing and Categorization

Humans can remarkably extract meaning from the rich and dynamic sounds around them. The brain can seamlessly isolate distinct auditory streams, group them into coherent auditory objects, and categorize these objects by associating them with relevant concepts. Such capabilities not only allow humans to navigate complex auditory environments but also highlight the remarkable adaptability and efficiency of the auditory system in shaping perception.

From an evolutionary perspective, identification and categorisation of living and non-living objects have been central for survival. For instance, the sound of predators evoked fight or flight responses, a cry informed that someone may be in pain, environmental sounds were used to detect danger etc. Evolution has played a crucial role in the development of the auditory system observed in humans today. Early humans developed enhanced frequency selectivity patterns that were likely shaped by the demands of speech and language development. They became better at recognizing and categorizing different natural sounds, partly due to the expansion of the cochlea and its frequency range (Manley, 2017). Additionally, the development of higher-order cortical regions involved in cognition led to more sophisticated categorization of sounds.

A common question refers to the difference in perception and categorization. Perception involves the initial detection and analysis of sensory stimuli, such as discerning pitch, loudness, and timbre in sounds. Categorization, on the other hand, entails assigning these perceived sounds to specific groups or categories based on shared characteristics or learned associations. If perception can be conceptualised as a multi-dimensional continuum, categorization are the boundaries in that space bounding various objects differently. Therefore, one can argue that perception may be a prerequisite for categorization and perception alone is possible without categorization. For instance, subtle acoustic differences can be detected without necessarily assigning them to known phonemic categories (Holt & Lotto, 2010). Nevertheless, the boundaries of categorization begin to form very early in infancy, especially to speech sounds, and dynamically evolve over time (Kuhl, 1991) and become integral in the way we navigate life. These boundaries are also found to be modulated in different contexts and recalls (Higgins & Lurie, 1983).

The essence of categorization lies in its ability to maintain invariance across diverse sensory instances i.e., the brain can group or associate sounds despite differentiating them based on their stimuli features. How does the brain do that? Few theories have been suggested over the years. The 'prototype theory' proposes that a "perceptual magnet" effectively warps sensory-feature space to exert influence over neighboring sounds from a mental reference point (Kuhl, 1991, 2004). In contrast, the 'decision-bound theory' suggests that existing mental representations of the listeners invoke specific decision boundaries to differentiate between categories (Smits et al., 2006). Although this theory originated in vision research, perception-based experiments have demonstrated its applicability to audition as well (Smits et al., 2006). Alternatively, another theory discusses the statistical relationship between different auditory objects (Smits et al., 2006). It posits that the listeners make statistical associations between sounds based on occurrence, acoustic properties etc. Furthermore, for speech perception, motor patterns related

to the speaker's articulatory gestures potentially mediate categorical perception of speech phonemes (Liberman et al., 1967). Although these theories contribute to a more nuanced understanding of auditory categorization, neurophysiological research on these differing schools of thought is necessary to deepen our understanding of them.

Past electrophysiological research in experimental auditory neuroscience could predominantly be divided into two classes. The first ones are the classical auditory neurophysiologists who investigated responses to pure tones to probe the auditory cortex. Their viewpoint lies on the fact that the auditory system decomposes all sounds to its constituent frequencies (similar to a function of cochlea), having found neural ensembles tuned to different frequencies bands and temporal scales. The second camp investigated the link between behavior and natural sounds, and discovered parts of the auditory cortex that respond selectively to classes of natural sounds but not necessarily to its constituent frequencies. The former group argued that the stimulus response function describing neural tuning in these neurons is primarily driven by the frequency tuning curve, while the latter attributed it to non-linear or contextual effects (Theunissen & Elie, 2014). Despite ample evidence supporting the two camps, the 'reductionist' group debated that the 'ethologists' lacked explanations for the processing of complex acoustic features, whereas the latter gave support for evolutionary arguments. Recently, with the introduction of more advanced and sensitive techniques, researchers are suggesting the merging of the two camps (Theunissen & Elie, 2014). Although this thesis will not resolve this debate, it will attempt to provide more clarity on the relationship between the two camps using complex models that can explain the low-level features of the sounds to an extent while characterizing higher-order cognitive processes of perception and categorization.

1.2.1 Relationship between Acoustic Properties and Categories

Maybe one of the most prominent and intuitive evidence for the relationship between acoustic features and perception is the 'bouba-kiki' effect (Ramachandran and Hubbard, 2005). While people associate round shapes with the sound of 'bouba', 'kiki' is often associated with spiky shapes, suggesting a deep link between stimuli features and perception. Although human vocalizations carry a wealth of information (Belin et al., 2004; Fröhholz, S & Belin, P., 2024), sensory impressions are not only limited to them, but also material sounds showing consistent associations between perception and sound spectrograms and geometric properties. For instance, high and low spectral center of gravity of artificial sounds can elicit spiky and round shapes respectively (Gurman et al., 2021). Synthetic "rising pitch sweeps" might evoke an approaching object and "descending glides" can be perceived as receding

objects (Tajadura-Jiménez et al., 2010). Thus, in addition to human sounds, many natural sounds can evoke multi-dimensional perception.

A wealth of literature suggests that the responses from the auditory cortex to sounds can be well differentiated based on many acoustic properties of the sounds such as *fundamental frequency* (Allen et al., 2017; Giordano et al., 2013; Patterson et al., 2002), *spectro-temporal modulation* (S. Norman-Haignere et al., 2015; Ogg et al., 2020; Schönwiesner & Zatorre, 2009), *spectral flatness* (Alluri et al., 2012; Lewis et al., 2012), *spectral centroid* (Allen et al., 2017; Alluri et al., 2012; Giordano et al., 2013; Ogg et al., 2020), *spectral envelope* (Ogg et al., 2019; Warren et al., 2005), *loudness* (Giordano et al., 2013; Langers et al., 2007), *noisiness* (Giordano et al., 2013; Lewis et al., 2009), *attack time*: (Menon et al., 2002) and *decay* (Aramaki et al., 2011). With such an overwhelming number of properties influencing auditory response, it may appear challenging to characterize acoustic representations and how they lead to higher-order categorical representations.

One recent MEG study used multivariate methods and representation similarity analysis (RSA, Kriegeskorte et al., 2008) where the emergence of most of these properties and categorical selectivity to auditory objects were investigated (Ogg et al., 2020). Specifically, Ogg and colleagues aimed to investigate how the brain recognizes auditory objects before the sound fully unfolds e.g., an approaching car, emergency siren etc. They included environmental, instruments and speech sounds in their stimuli set. They found that most of the acoustic features fluctuated over time with spectral envelope and modulation spectra exerting their greatest influence soon after sound onset, while spectral centroid and temporal centroid had a greater influence later in the epoch (Figure 1.1). Spectral variability and aperiodicity could be significantly decoded throughout the length of the epoch. In fact, the correlation of the MEG with spectral variability, centroid and aperiodicity peaked at ~170 ms which coincides with auditory P200 ERP in response to sounds (Crowley & Colrain, 2004; Roth et al., 1976). In response to voices, the ERP in this time range is also called frontotemporal positivity to voices (FTPV, also see section 1.2.1.1), known to encode paralinguistic features (Charest et al., 2009). Other studies also showed that evoked responses in this time range are affected by acoustic differences among stimuli (Caclin et al., 2006; Giard et al., 1995; Näätänen & Picton, 1987; Roberts et al., 2000; Rosburg, 2003) and appear to support representations of speech (Bidelman & Walker, 2017; Näätänen & Picton, 1987; Roberts et al., 2000).

Furthermore, Ogg and colleagues found categorical representation to peak around the same time (~170 ms) by pooling trials of different examples of environment (E), instruments (I) and speech (S) sounds and consequently decoding them.

Interestingly, the peak shares the same latency of the N170 ERP component that is also known to index face processing in vision (Rossion & Jacques, 2012). However,

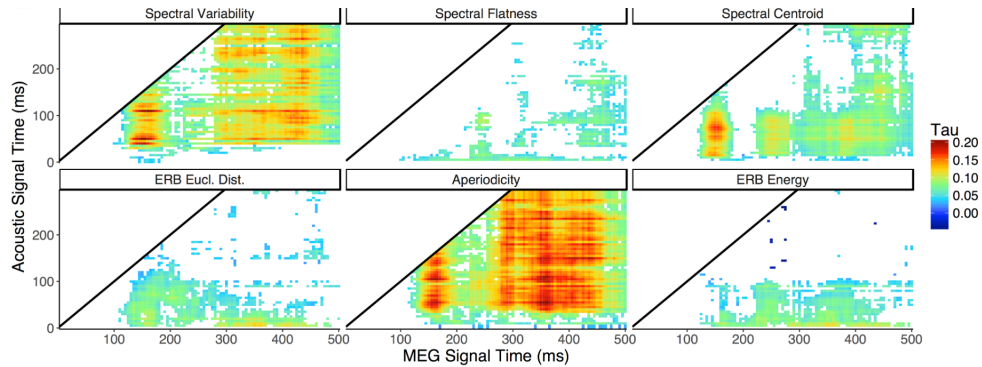


Figure 1. 1 Acoustic Representations in MEG, adapted from Ogg *et al.*, 2020: The correlation between neural models and acoustic models throughout the epoch plotted as averages of subject level correlations. Spectral variability, centroid and aperiodicity peak at ~170 ms.

it is unclear whether the observed categorical representation is entirely independent of the processing of acoustic features, since the sounds across all three categories exhibit distinct acoustic properties (e.g., aperiodicity high in E, less in I and S, temporal envelope different across all three, phonemes in S are characterized with a distinct spectral modulations compared to E and I etc). This study demonstrates that disentangling acoustic processing from categorical representations is challenging, as responses to acoustics often overlap with categorical responses.

Another approach to isolate these processes is to examine how the auditory cortex processes fragments of sounds across multiple time scales. A study using intracranial EEG recordings by Norman-Haignere and colleagues pointed out that parts of the auditory cortex process certain acoustic features based on different timescales, therefore exhibiting distinct functional properties (Norman-Haignere *et al.*, 2022). Auditory regions with shorter-integration windows (<200 ms) revealed functional tuning to spectrotemporal modulations of the sounds, while areas with longer-integration windows (>200 ms) showed categorical selectivity. Based on these findings, the 200 ms time window following sound onset appears to be a critical period for auditory processing, where information cascading from different auditory regions and multiple time-scales converge. In this thesis, I will present similar results in blind and sighted individuals, where the sound decoding also peaked around 200 ms, and discuss the underlying reasons for this observed time range.

While some studies provide evidence of acoustic processing scaffolding higher order categorical representations (Lewis *et al.*, 2009; Moerel *et al.*, 2012; Santoro *et al.*,

2014) others found that some conspecifics such as human voices elicit enhanced responses in the auditory cortex in comparison to artificial sounds, that are proven to be at least partially independent of acoustic properties (Barbero et al., 2021; Belin et al., 2000). Some studies also argue the emergence of post primary complex representations of sounds, that can be abstracted from the acoustic properties enabling perception and behavior (Giordano et al., 2013, 2023). As a result, these questions arise - To what extent are categories based on specific inherent characteristics of sounds (Moerel et al., 2012; Ogg et al., 2020; Staeren et al., 2009)? And to what extent are they based on higher-order cognitive and abstract associations? In the following sections we address these questions separately for human voices and information derived from voices - emotion.

1.2.2 Are voices special?

We are exposed to voices frequently in our everyday lives making them the most relevant set of natural sounds. Voices are the carrier of speech as well as non-verbal information enabling social communication. A wealth of information can be extracted from voices such as their age, gender, emotional state, identity, physical size and health status (Belin et al., 2004, 2011; Formisano et al., 2008; Frühholz, S & Belin, P., 2024; Klasen et al., 2012; Pernet & Belin, 2012; Whiteside, 1998; Zäske et al., 2013). The acoustic properties of animal and human vocalizations are unique due to the physiological apparatus to produce sounds. In brief, voices are produced by passing air through the vocal cords that can be stretched or contracted modulating the fundamental frequency (F_0) of the sounds. The vocal tract and following formation of the passage (jaws, tongue, teeth, lips) cause resonance and shapes the sound characterizing it with formants (F_n), and providing the control to produce different sounds. The unique acoustic quality of voices is the periodicity or harmonicity which gives rise to the perception of pitch. With the exception of musical instruments, voices have higher harmonicity than artificial sounds.

A plethora of studies on voice perception have revealed enhanced responses to voices compared to artificial sounds in the auditory cortex (Barbero et al., 2021; Belin et al., 2000; Charest et al., 2009; Frühholz, S & Belin, P., 2024). For instance, Charest and colleagues identified an electrophysiological distinction between vocal and non-vocal sounds emerging as early as 164 ms after stimulus onset and peaking around 200 ms, referred to as the “fronto-temporal positivity to voices” (FTPV, Charest et al., 2009). The brain sources eliciting FTPV were identified to be located in the mid-anterior superior temporal sulcus (STS) bilaterally and in the right planum temporale (Capilla et al., 2013). fMRI studies have consistently identified these regions as exhibiting unique selectivity to voices, called the Temporal Voice Areas (TVA, Figure 1.2, Agus et al., 2017; Belin et al., 2000, 2002; Kriegstein & Giraud, 2004; Pernet et al., 2015; Roswadowitz et al., 2020). This preference for voices

has been demonstrated in the primate brain in a manner analogous to face processing, with cortical patches exhibiting selective responses to conspecific vocalizations (Bodin & Belin, 2020; Petkov et al., 2008). Furthermore, speaker identity and speech content can be decoded from these voice-sensitive areas of the auditory cortex (AC) in humans (Formisano et al., 2008). However, it is debated whether the processing of harmonic structure of sounds (Lewis et al., 2009) or the regions of the auditory cortex that show a response bias to low-frequency sounds (Moerel et al., 2012), both key characteristics of voices, play a role in shaping the responses obtained from TVA. Additionally, neuronal regions in the secondary auditory cortex exhibit specific sensitivity to the spectro-temporal modulations typical of speech aiding voice selective responses (Santoro et al., 2014). Many attempts have been made to disentangle the higher-order categorical response to voices from the acoustic processes with methods including subtraction of neural responses to voices and non-voices, regressing out the acoustic properties from the neural signal. However, separating the voice-selective neural signal from acoustic features is challenging, as some of the deleted components may contain information related to voice selectivity (Frühholz, S & Belin, P., 2024).

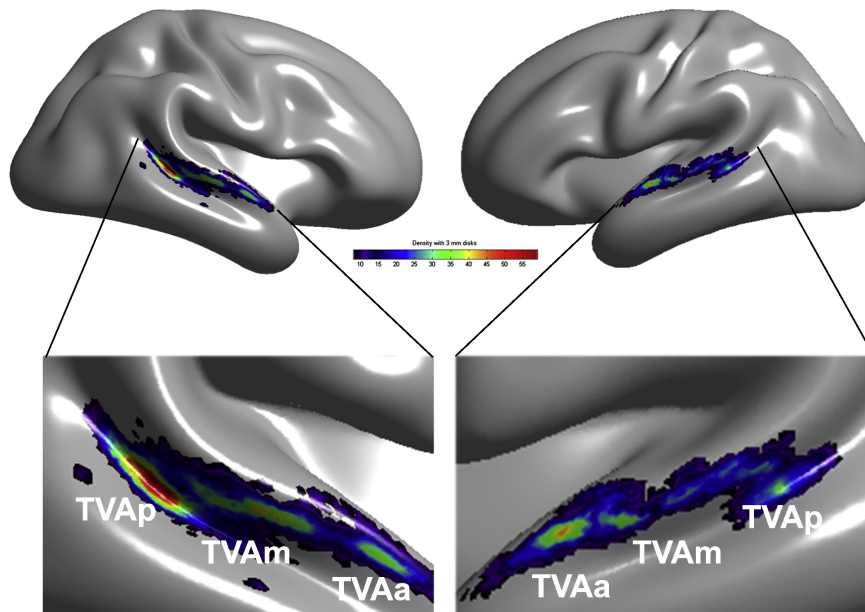


Figure 1. 2 Temporal Voice Area (TVA), adapted from *Pernet et al., 2015*. Three voice selective areas were found in the superior temporal region: posterior (p), medial (m), anterior (a).

In my lab, I was part of one such study which attempted to disentangle the voice-selective response from the acoustic features using a novel technique called 'frequency tagging' (Figure 1.3, Barbero et al., 2021, see the details of the method

in section 6.1). Specifically, we presented heterogenous sounds of voices and objects of length of 250 ms in a periodic stream at 4 Hz. Unbeknown to the participant, only voices were placed at every third position, thus being presented at 1.33 Hz (target rate). Testing EEG on human adults, we found a neural signal at 4 Hz depicting the segregation of different successive sounds. Interestingly, we also found a significant response at 1.33 Hz corresponding to the voice presentation rate exhibiting the *discrimination* of voices from object sounds, but also the *generalization* over multiple instances of voices. However the response at the superior temporal channels and regions at the target rate also encompassed distinct voice associated acoustic processes. Therefore, we presented a scrambled condition, where the spectral information of the sounds were scrambled in successive windows of 200 Hz. Scrambling the sounds resulted in similar spectral content and temporal information compared to the original sounds, however it disrupted the harmonicity and intelligibility of the sounds. We found a significantly higher response at 1.33 Hz in the standard condition compared to the scrambled condition, suggesting that the former response does not depend on the spectral-temporal information of the sounds. To ensure that the responses in the standard condition did not arise due to differences in harmonicity between voices and object sounds, another control condition was introduced which consisted of presentation of voices amongst musical instruments matched for harmonicity and pitch. In this control condition, we found a significant response at the voice presentation rate indicating that the voice-selective response is at least partially independent from low-level acoustic features (spectral and temporal information, center of gravity, harmonicity-to-noise ratio and pitch).

The frequency tagging approach enabled us to record voice-selective responses independent from low-level acoustic cues in 4 min of recording only, making this approach highly promising for studying auditory perception in children and clinical populations. Similarly, frequency tagging can be used to investigate brain selectivity to more nuanced information conveyed by voices such as identity, age and emotion. In this thesis, I adapted this technique to investigate emotion categorical selectivity to auditory non-verbal expressions in humans (Chapter 2).

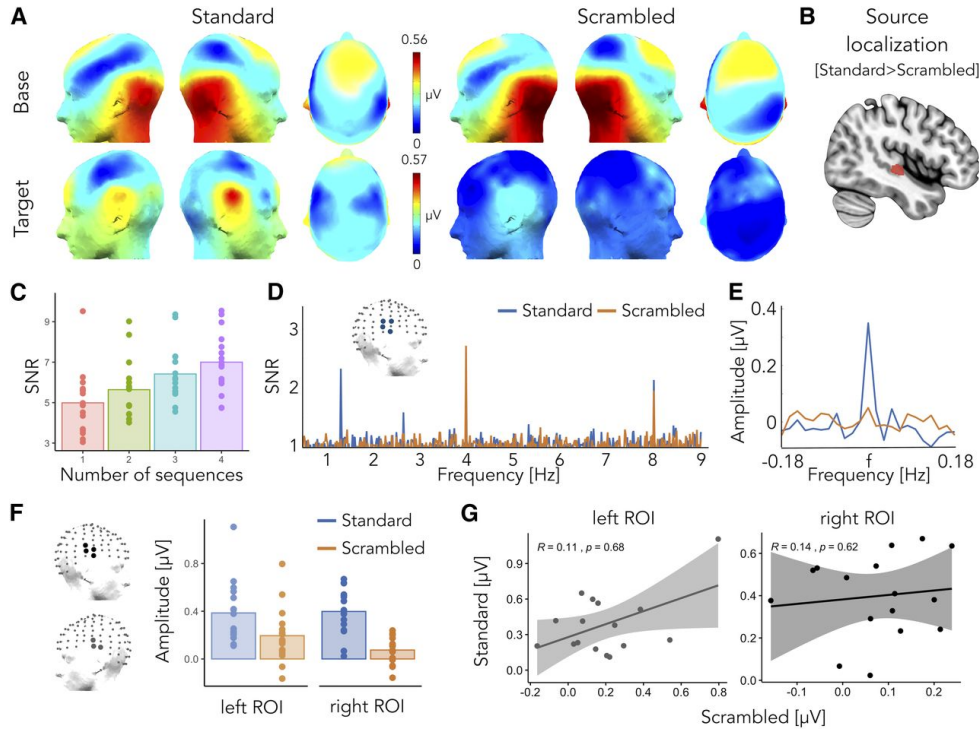


Figure 1. 3 Voice selectivity in human adults, *Barbero et al., 2021*: Voice versus object sounds. **A**, Responses at the base (top) and at the target frequency (bottom) for the standard (left) and scrambled (right) conditions as topographic head plots. Topographic head plots were obtained by summing the baseline subtracted amplitude at the significant harmonics. **B**, Source localization analysis revealed stronger coherence values at the target frequency for the standard condition (standard > scrambled) over right superior and middle temporal areas (voxels with intensity above the 95th percentile are highlighted in red). **C**, Signal-to-noise ratio (SNR) of voice-selective responses obtained by averaging SNR values of ROIvoice electrodes as a function of number of stimulation sequences/minutes of recording for the standard condition. **D**, SNR averaged spectra of ROIvoice electrodes. **E**, Sum of the baseline corrected amplitude (considering 10 bins on each side, excluding the immediately adjacent ones) of voice selective responses at the four significant harmonics for the standard (blue) and scrambled condition (orange). **F**, Responses at the target frequency for the standard (blue) and scrambled (orange) conditions as the sum of baseline subtracted amplitude at four harmonics in the left and right ROIs. **G**, There is no correlation between responses at the left and right ROIs across the standard and scrambled conditions.

1.2.3 Auditory Emotions

Although visual emotions have received more attention in the past, mounting evidence points that discrete emotion expressions can be delivered and decoded through vocal expressions with a high accuracy in humans (Cornew et al., 2010; Falagiarda & Collignon, 2019; Pell & Kotz, 2011). Emotions can be inferred from voice in three different formats - through speech, the prosody that accompanies the speech and non-verbal expressions such as laughs, grunts etc. While the former two

co-occur together, non-verbal vocal expressions primarily convey prosodic content and emotional cues through modulations of acoustics. In this thesis, I will focus only on categorizing emotion from non-verbal expressions.

Humans can successfully infer primary emotions (Ekman, 2009) such as Fear, Anger, Disgust, Happiness, Sadness and Surprise from non-verbal expressions and speech prosody, where the former is argued to be uniform across cultures (Elfenbein & Ambady, 2002; Juslin & Laukka, 2003; Sauter et al., 2010; Sauter & Eimer, 2010; Scherer et al., 2001). However, affective vocalizations can also be mapped on valence-arousal continua with descriptions that are consistent expression of facial emotions in the two dimensions (Jin & Wang, 2005). Till date it remains ambiguous whether the brain conceptualizes these vocalizations as distinct categories and/or as dimensional features.

More recently, Cowen and Keltner used a more data driven approach to question how many categories of emotions humans can typically perceive (Cowen & Keltner, 2017). They applied large-scale data collection to categorize 2032 emotional vocal bursts by 1105 English speakers, finding 24 distinct categories that can be mapped as a continua (Figure 1.4). They also found emotion categories (sympathy, awe) to organize emotion recognition more than affective dimensions (including valence and arousal). An interactive map can be found on <https://s3-us-west-1.amazonaws.com/vocs/map.html#>

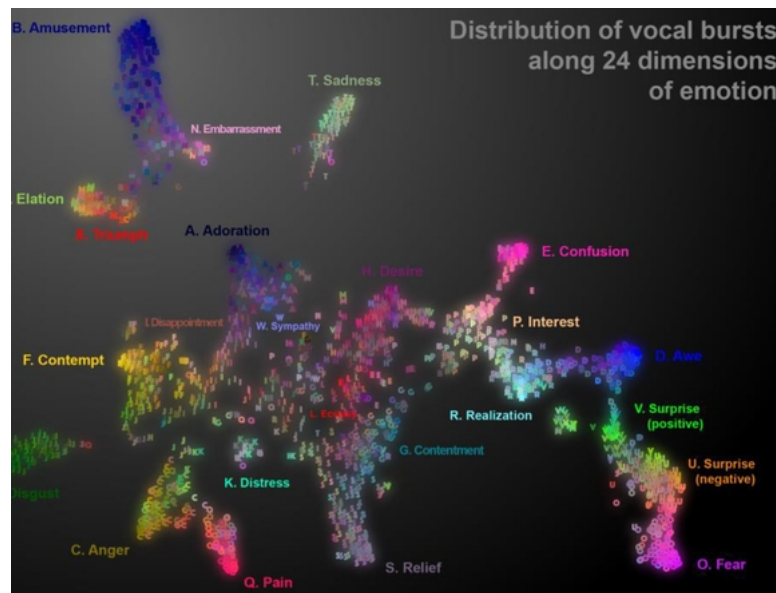


Figure 1. 4 Mapping of emotion categories using behavior ratings, adapted from *Cowen et al., 2019*.

Verbal expressions of different emotion categories can be characterized by both shared and distinct acoustic properties (Banse & Scherer, 1996). For instance, the Fo and pitch of angry, fearful and joyful sounds is usually higher and similar but different compared to sadness; the temporal envelope of laughter is distinct from other categories etc. Thus, whether researchers capture neural signals from the processing of distinct acoustic features characteristic to an emotion category, or the categorical representation itself is a matter of debate.

Nevertheless, the network subserving emotion processing and recognition have been extensively researched. Although sensory processes in the primary auditory cortex can occur as early as 100 ms and are usually driven by dominant acoustic features (Salvia et al., 2014), evidence of emotion categorical representations in the human brain have been shown in both fMRI (Ethofer et al., 2012; Frühholz & Grandjean, 2013) and EEG studies (Frühholz et al., 2011; Jessen & Kotz, 2011, 2011; Paulmann & Kotz, 2008), that may be partially independent of acoustic amplitude and frequency cues (Giordano et al., 2021; Grandjean et al., 2005). In fact, the existence of Emotional Voice Areas (EVA) have been suggested, which overlap with the Temporal Voice Areas (TVA) that functionally respond more to affective vocalizations compared to neutral utterances (Ethofer et al., 2012). These specialized areas are further connected to inferior frontal cortices (IFC), forming an emotion processing network (Ethofer et al., 2012). The rostral-ventral IFC is involved in categorizing emotional vocalizations, while the caudo-dorsal IFC might be specifically sensitive to their temporal features (Frühholz & Grandjean, 2013). In fact, it is suggested that a fronto-temporo-striatal network including anterior temporal areas, such as anterior superior temporal cortex (STC), anterior insula and IFC is activated for emotion evaluation, categorization and appreciation (Frühholz & Grandjean, 2013; Kotz et al., 2013; Petrides, 2005; Schirmer et al., 2004). A potential ERP indexing these processes may be the P200 (~200 ms) that was found to be modulated by emotional utterances more than neutral utterances (Liu et al., 2012; Paulmann & Kotz, 2008).

The STC has been consistently identified as a key region in all auditory emotion studies. Different regions of the STC have been found to exhibit distinct functional roles in processing of both voices and emotional vocalizations. For instance, middle STS is linked with acoustic properties of the sounds (Grandjean et al., 2005; Kriegstein & Giraud, 2004). Right aSTG is connected with the paralinguistic aspects of voice processing (Belin et al., 2004) and voice recognition (Kriegstein & Giraud, 2004). Also, the right pSTS/STG responds to voice identity (Kriegstein and Giraud, 2004). These functionally diverse regions potentially encode suprasegmental

information of the sounds facilitating voice and emotion perception which are mediated by higher-order regions in the auditory network (Kotz et al., 2013).

A drawback of most studies aiming to isolate the emotion selective responses from the brain is to compare neural responses to emotional vocalizations with neutral utterances. This is problematic, as both can vary in acoustic features, valence and arousal, raising doubts about whether they truly isolate emotion-specific processes. More recently, without using neutral sound, multivariate methods have shown that information encoded within both local neural ensembles and whole-brain activation patterns can be effectively used to predict affective dimensions and discrete emotions with a high degree of specificity (Kragel & LaBar, 2016).

Another enduring debate is whether distinct emotion categories activate different brain substrates. A meta-analysis method called the activation likelihood estimation (ALE) has been used to find overlapping yet discriminable brain patterns for primary emotion categories across independent studies (see Hamann, 2012; Vytal & Hamann, 2010). However, divergent results across studies are now leading the researchers to give up on one-to-one mapping of emotion and brain area (Hamann, 2012; Kragel & LaBar, 2016). For instance, amygdala was popularly linked with fearful sounds, yet others found it to be activated when manipulating the arousal and valence of the sounds (see Kragel & LaBar, 2016 for meta-analysis on 200+ studies). This has resulted in a shift towards 'network based hypotheses' which require more sensitive yet robust methods in terms of design and analysis to extract brain networks underlying auditory emotion perception. In this thesis, we introduce a new method - frequency tagging using EEG to extract brain responses to auditory emotion categories, that are independent of the acoustic properties, and reveal whether discrete emotion categories evoke different neural representations (Chapter 2).

1.3 Hierarchical vs Parallel Processing

Effectively processing diverse acoustic signals and then categorizing sounds collectively may indicate a more hierarchical approach to auditory processing. Neurophysiological studies have affirmed with this idea finding that the primary auditory cortex receives and processes filtered spectral and temporal information from the cochlea and subcortical regions earlier in time, which is further relayed to higher-order auditory regions such as middle and superior regions expressing abstract representations of sounds and categories at later stages (Cohen et al., 2007; Cowen & Keltner, 2017; Giordano et al., 2013; Leaver & Rauschecker, 2010; Lowe et al., 2021; Russ et al., 2007). This auditory pathway in conjunction with the ventrolateral prefrontal cortex which is also involved in sound perception, is often referred to as the 'ventral auditory pathway', found in both primates (Tian et al., 2001)

and humans (Arnott et al., 2004; Bizley & Cohen, 2013). More recent computational evidence for hierarchical processing comes from the studies using sound based Deep Neural Networks (DNN) that perform categorization of various natural sounds, musical sounds and speech (Giordano et al., 2023; A. J. E. Kell et al., 2018; A. J. Kell & McDermott, 2019; Rupp et al., 2024). The layers of deep neural networks (DNNs) progressively encode increasingly abstract sound representations, analogously predicting neural responses from progressively hierarchical regions of the auditory network. For instance, these studies found the early layers of the DNN that correlate higher with low-level acoustic properties of the sounds predict neural responses from the primary belt of the auditory cortex, while the later DNN layers which encode higher-order associations and categorical information can well predict secondary auditory regions (A. J. E. Kell et al., 2018; Rupp et al., 2024).

Opposing the hierarchical theory, a distributed view of the auditory networks has been suggested where different brain regions within the processing network support auditory perception. In this case, the feedforward and feedback connections between these regions would facilitate the auditory categorization (Bizley & Cohen, 2013). This perspective stems from the observation that, beyond the activation of auditory regions in response to sounds, other brain regions are also activated e.g., during pitch and timbre perception tasks (see Bizley & Cohen, 2013 for a review). In Chapter 4 and the general discussion, in addition to finding results that support the claim of hierarchical processing, I also present a perspective to discuss parallel processing of natural sounds in the human brain.

1.4 Development of Auditory Processing and Categorization

Auditory development is known to begin during the gestational period. Experiences start to shape this architecture already in the prenatal stage, as vibrations travel through the amniotic fluid and reach the infant's skull from the mother. Further evidence of auditory development is provided by the fetal movements triggered by low-frequency sounds (e.g., 100–500 Hz) that can be observed through ultrasound as early as the 19th gestational week (GW), while responses to higher frequencies (1000 Hz and 3000 Hz) begin to appear after the 27th GW (Shahidullah & Hepper, 1994). Extremely preterm infants demonstrate cortical responses to sound (Graziani et al., 1968), suggesting they possess the fundamental anatomical structures for hearing.

Most research in infants has focussed on speech perception and processing, since language is a crucial ability for humans and its unique complexity sets humans apart from all other species. However the origin of these skills have sparked debates on the topic of nature vs nurture. While behaviorists have supported that stimulus-

response mechanisms and reinforcement may facilitate language acquisition, nativists argue that apart from the influence of experience, humans at birth may possess innate architecture for language abilities (Chomsky, 1959). A plethora of behavioral and electrophysiological studies now support innate predispositions that are shaped by experience. It is suggested that the brain's predisposition to process linguistic information appears long before language acquisition (Dehaene-Lambertz, 2000; Dehaene-Lambertz et al., 2002), as neonates can encode consonants in noisy conditions (Cabrera & Gervain, 2020) and vowels (Aldridge et al., 2001), show sensitivity towards phonemes (Telkemeyer et al., 2009) and male vs female identity (Mahmoudzadeh et al., 2013), setting the foundation for the acquisition of language skills. It has also been shown that neonates can recognize native language from foreign language (Mehler et al., 1988). Moreover, neonates show higher selectivity to their mother's voice (Beauchemin et al., 2011; Purhonen et al., 2005) as well as their father's voice (Lee & Kisilevsky, 2014) in comparison to strangers' voices, suggesting that they can already develop familiarity to features of voices and differentiate identity. Although these remarkable skills are developed early in life, the mechanisms governing them are still under research.

During early development, an infant's brain structure is fine-tuned based on their experiences to support more effective cognitive functions. This tuning, also termed as synaptic pruning, involves excess and unused synapses to be eliminated over time and creation of a large number of synapses allowing learning and adaptation to the environment. For instance in the context of auditory categorization, several studies have reported that, from birth to 3–4 months, there is no selective neural processing for human speech over primate vocalizations, even though older infants and adults show clear preference of the former over the latter (Carr et al., 2021; Ferry et al., 2013; Perszyk & Waxman, 2016; Vouloumanos et al., 2010). Carr and colleagues found that this neural adaptation develops over time. They demonstrated 4-9 Hz EEG responses in 3-month-old and 6-month-old infants showing a decrease in neural selectivity to non-human primate vocalizations and an increase in neural response to infant-directed speech when comparing the two age-groups (Carr et al., 2021). It was suggested - *“A broad biological template initially encompasses vocalizations of human and nonhuman primates and within 6 months, this link to cognition is tuned specifically to human vocalizations. At 3 and 4 months, nonhuman primate vocalizations promote object categorization, mirroring precisely the advantages conferred by human vocalizations, but by 6 months, nonhuman primate vocalizations no longer exert this advantageous effect.”* (Ferry et al., 2013, also see Polver et al., 2024 for review).

These reports demonstrate that the auditory functions can develop similar to adults within the first few months of life, however the underlying mechanisms governing

this development remains an active area of research. Furthermore, whether infants develop higher-order selectivity, as observed in adults, depends on the processing of spectral and temporal cues remains to be determined. In our lab, we explored if the voice-selective response in adults that is at least partially independent of the acoustic properties of the sounds (Barbero et al., 2021) can already be found in 3-4 month infants (Calce et al., 2024). Using frequency tagging and a similar experimental paradigm as explained in section 2.2, a voice topography was observed in infants resembling the same of adults found in Barbero et al., 2021, which was at least partially independent from the acoustic features of the sounds (Figure 1.5). This indicates that this neural selectivity in early infancy supports efficient categorization of voices, allowing the development of social communication, speech perception, and early language acquisition. In this thesis, I extend the exploration to find representations and selectivity to the subordinate categories of human vocalizations - human verbalizations and non-verbalizations, as well as to natural sounds. In comparison to adults, we reveal the temporal dynamics of acoustic processing and representations of various categories of natural sounds in infants (Chapter 3).

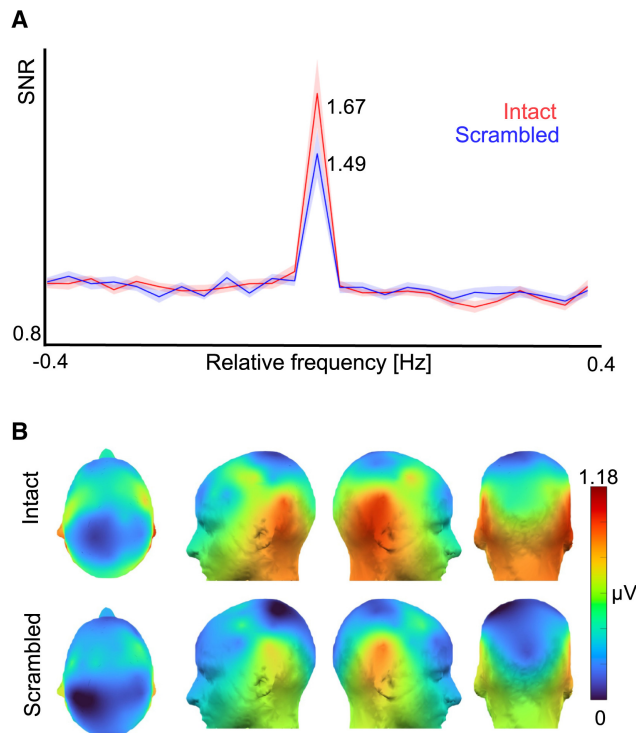


Figure 1. 5 Voice selective response in infants, adapted from *Calce et al., 2024*. 4-7 month old infants show an enhanced voice-selective response to the intact or standard condition in contrast to the control frequency scrambled condition originating from bilateral temporal channels.

1.5 Auditory Processing and Categorization in Blind People

Investigating auditory functions in sensory deprived populations can highlight how sensory experience interacts with intrinsic biological constraints to shape auditory processes. In this regard, comparing neural responses to sounds in blind people and normal sighted individuals represents a unique model to investigate the (re)organization of auditory processes and categorization in the absence and presence of visual input, respectively. In blind people, while some studies have found enhanced responses in the temporal cortex (Elbert et al., 2002; Gougoux et al., 2009; Manjunath et al., 1998; Naveen et al., 1998; Pascual-Leone & Torres, 1993; Rauschecker, 2002; Röder et al., 2002), others found reduced responses in the same regions (Bedny et al., 2015; Burton et al., 2002; Pietrini et al., 2004; Ricciardi et al., 2009; Stevens & Weaver, 2009; Striem-Amit et al., 2012; Wallmeier et al., 2015; Mattioni et al., 2022). However in contrast to mixed results in the temporal cortex, increased involvement of the occipital cortex in auditory functions have been consistently reported, indicating a structural and functional reorganization in the brain's circuitry through cross-modal plasticity (CMP, (Bavelier & Neville, 2002; Heimler et al., 2014; Merabet & Pascual-Leone, 2010).

CMP is not a stochastic process and is known to be adaptive for behavior (Heimler et al., 2014). For instance, correlations have been reported between CMP and enhanced non-visual behavior both at functional (Gougoux et al., 2009) and structural levels (Voss et al., 2014). However, debates on the guiding mechanisms of this plasticity still exist. One perspective suggests that the occipital cortex is pluripotent, and in the absence of visual input, its functional organization cannot develop based on its proto-organization, leading it to adapt to more distant computations (Bedny, 2017). In contrast, many studies have shown that specific parts of the occipital cortex preserve their native computation prototype while switching to process non-visual input instead. For instance, the lateral occipital cortex (LOC) shows functional selectivity to process information related to shapes (audition: Amedi et al., 2007 touch: Amedi et al., 2010; Pietrini et al., 2004), hMT + V5 for motion (hMT+V5; audition: Poirier et al., 2006; Wolbers et al., 2011; touch: Ricciardi et al., 2007, right dorsal extrastriate visual cortex for localization of stimuli in space (Collignon et al., 2011; Renier et al., 2010), visual word form area (VWFA) for letters (audition: Büchel et al., 1998; Striem-Amit et al., 2012) and extrastriate body area for body shapes (EBA; audition: Striem-Amit & Amedi, 2014), ventral/medial fusiform gyrus for categorizing animate and inanimate objects (audition: He et al., 2013; Mattioni et al., 2022); touch: (Pietrini et al., 2004). All these

studies suggest a reorganization in the functional roles of specific occipital regions while maintaining their computation specificity. An alternative perspective challenges the use of the term "reorganization," proposing that neural ensembles are biologically predisposed to perform a specific set of functions. Following injury or the loss of visual input, occipital regions may not fundamentally reorganize but instead upregulate their existing capacity to process non-visual information while simultaneously downregulating their visual processing capabilities (Makin & Krakauer, 2023).

Whether the recruitment of visual areas in the aforementioned studies was due to an overlap in categorical information shared between visual and non-visual sensory modalities, or simply due to similarities in the processing of low-level sensory features was hugely debated. Addressing this question, studies found enhanced categorical selectivity to sounds in the visual cortex of blind people compared to sighted controls, which are independent of the acoustic features of the sounds (Dormal et al., 2018; Mattioni et al., 2020, 2022; Peelen et al., 2014; Schepers et al., 2012; van den Hurk et al., 2017; Vetter et al., 2020; Wang et al., 2015). In fact it is suggested that the occipital cortex of the blind encodes higher-order categorical structure, similar to the manner implemented for vision in the sighted (Dormal et al., 2018; Mattioni et al., 2020, 2022; Peelen et al., 2014; Striem-Amit & Amedi, 2014; van den Hurk et al., 2017; Vetter et al., 2020; Wang et al., 2015). For instance, Mattioni and colleagues conducted an fMRI experiment where they presented sounds to blind and sighted people belonging to different categories. Additionally, sighted individuals also performed a visual task using the images of the same concepts. They found that the ventral occipital temporal cortex (VOTC) of the blind encodes sound categories similar to the manner sighted people represent categories in vision in the same brain region (Figure 1.6). This auditory categorical representation in the VOTC of the blind was independent of acoustic properties such as pitch. While the VOTC in both groups demonstrated significant decoding accuracies across sound categories, the early visual cortex (EVC) of blind individuals also displayed significant decoding accuracy, unlike that of sighted individuals. A follow-up study by Mattioni et al., showed that concomitant to the enhanced responses to natural sounds in the occipital cortex of the blind, reduced responses were observed in their temporal regions. However, no differences were observed in the encoding of sound features in the temporal regions (Mattioni et al., 2022). These findings collectively suggest that blindness exerts varying effects across brain regions, influencing networks involved in higher-order categorization of natural sounds by recruiting more posterior neural substrates that maintain their computational roles, while leaving early sound processing stages unaffected. Chapter 4 of this thesis is an extension of this work where I investigated the temporal dynamics of processes subverting auditory categorization in blind and sighted

individuals. Similar evidence will be presented showing that early auditory processing remains unaffected in congenitally blind individuals. However, more sophisticated auditory models reveal that intermediate stages of auditory processing, as well as categorical representations, are enhanced by blindness.

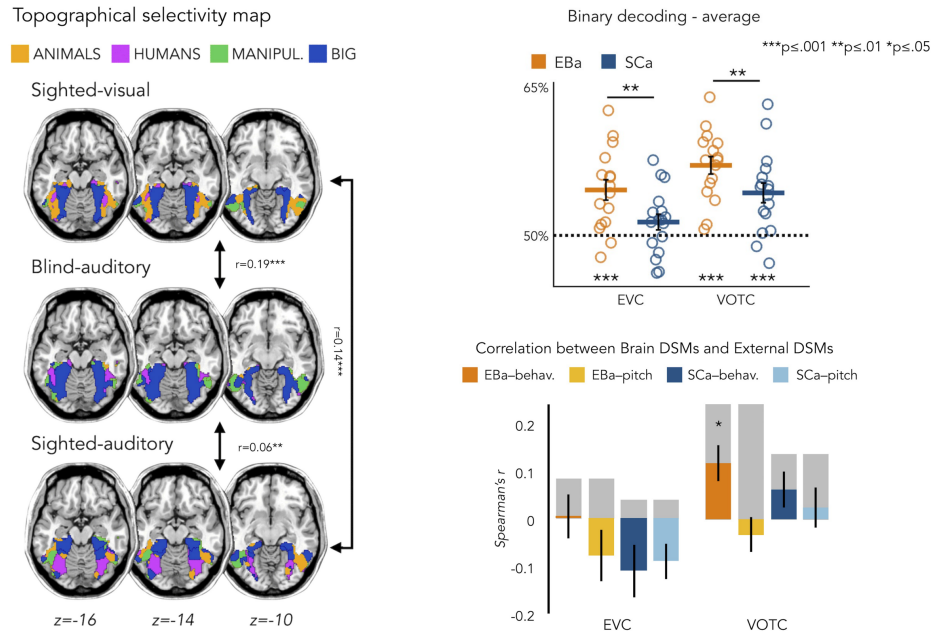


Figure 1. 6 Categorical representation in blind people, adapted from *Mattioni et al., 2020*. Auditory representation of categories in the ventral occipital cortex (VOTC) in the blind (EB) are more similar to the visual categorical representation in sighted (SC), rather than the auditory representation in the sighted in the VOTC. Auditory categorical decoding analysis shows above significant accuracies in the VOTC of both blind and sighted, however the early visual cortex (EVC) of only blind individuals was found to exhibit categorical information as well. Behavior models correlate well with the neural model of the VOTC only in the blind but do not align with the pitch model, suggesting that the categorical information in this region is at least partially independent of acoustic properties.

1.6 Methodology Techniques

Two techniques compatible with electrophysiology are introduced in this thesis - Fast Periodic Auditory Stimulation (FPAS, Barbero et al., 2021) and a multivariate approach in conjunction with Representation Similarity Analysis (RSA, Kriegeskorte et al., 2008).

1.6.1 Fast Periodic Auditory Stimulation (FPAS)

FPAS was adapted from frequency tagging studies in vision, termed better as fast periodic visual stimulation (FPVS) that investigated responses to faces (Retter &

Rossion, 2016; Rossion et al., 2015) letter/word categorization (Lochy et al., 2015), and numerosity (Marlair et al., 2024), as well as auditory studies investigating rhythm (Nozaradan et al., 2017). The principle underlying frequency tagging is constituted by the steady state evoked responses: under periodic external stimulation to one specific sensory input, the brain regions coding for this input elicit a response that oscillates at the exact same frequency. Thus, knowing the stimulation frequency, which in a research setting is manipulated by the experimenter, it is possible to track the brain response with EEG. This method usually consists of presenting a rhythmic stimulus at a frequency X, which is the low-level presentation frequency. However, within this periodic presentation another pattern is hidden at a frequency Y based on a conceptual similarity. Analyzing the brain response elicited by the stimuli, a pattern X will reflect basic sensory processing, but pattern Y emerges only if the participant's brain successfully *generalizes* across the stimuli of the same category. In comparison to other EEG methods, such as the Event Related Potential (ERP) where subtraction or regression is commonly used to contrast the concerned activity with the baseline, frequency tagging provides a direct measure of conceptual processes. Secondly, the paradigm is behavior-free, thus attentional and decisional processes do not contaminate the brain's perceptual categorization, making FPAS an objective approach. Thirdly, this technique also relies on the possibility to record data with high signal-to-noise ratio (SNR) in a very short period of time without the necessity for the participant to be involved in a task, making this technique particularly suitable for the study of babies and special populations. As described previously in section 2.2, FPAS was developed in the lab and used to find voice selective responses in adults (Barbero et al., 2021) and 4-month old infants (Calce et al., 2024). More recently, this technique was used to identify intermodulation frequencies for audio-visual integration of fearful emotion (Barbero et al., 2025). I extended the use of FPAS to investigate categorical selectivity for auditory emotions in adults (Chapter 2).

1.6.2 Multivariate Pattern Analysis (MVPA)

MVPA is an increasingly popular choice for analyzing neuroimaging data. In contrast to univariate methods in which each variable is considered independent, MVPA relies on the relationship between multiple variables (e.g., channels of EEG, voxels in fMRI). The term 'decoding' is popularly used in reference to creation of a model using empirical data. Inversely, 'encoding' approaches refer to prediction of data from a model. Generally for decoding, a machine learning classifier is used e.g., support vector machines (SVM), linear discriminant analysis (LDA), logistic regression etc to identify patterns in the data that correspond to the concerned experimental parameters. As an example, if we want to find patterns that differentiate 2 conditions of an EEG experiment, we first divide the data into training sets and testing sets. This division, also called cross-validation strategy, can depend upon

the classifier used, the amount of data and noise present (Grootswagers et al., 2017). In this thesis, a 'leave-one-trial-out' approach is used, where one exemplar trial from each condition is placed in the testing set while the remaining trials are used for training, ensuring a strong emphasis on the training process. The classifier creates a model from the labelled data in the training set considering all channels as features and predicts the labels of the remaining trials providing a decoding accuracy. An accuracy of 50% refers to chance-level prediction, while a significantly higher accuracy indicates that the classifier can successfully differentiate both conditions. This method can be applied to channel magnitudes at each time-point, within a time range or even the whole trial in which all time-points multiplied by channels could be considered as features.

Once MVPA identifies class-specific information within the data, further investigation can be conducted using Representation Similarity Analysis (RSA, Kriegeskorte et al., 2008). RSA allows us to abstract responses from the native space to similarity space using a principle that stimuli with closer neural representations are similar, while those with distinct representations are easier to decode. The essence of RSA is the dissimilarity matrix (DSM) which consists of distances between pair-wise conditions which can be denoted by a metric of distance such as, 1-correlation, decoding accuracy or even simple subtraction between conditions. The empirical DSM then can be used to correlate with or be predicted by other DSMs derived from other empirical data, stimuli features, or theoretical models. This allows us to test whether representations in a region, or at a time point, follow a predictable structure.

As an example, we conducted an experiment where participants were presented 24 sounds (Figure 1.7, also see methods in Chapter 3 and 4). We performed pairwise sound decoding (sound 1 vs sound 2, sound 1 vs sound 3 etc) and created neural DSMs at each EEG time-point using the decoding metric to quantify the pairwise distances between sounds. Thus for 24 sounds, a neural model will consist of a 24x24 matrix where each cell depicts the distance between pairwise sounds individually at each timepoint. Separately and similarly, external representational models can be created using pairwise distances of acoustic properties of the sounds, apriori categorical models and behavior responses. Thus, under the framework of RSA, we can find the temporal emergence of different dimensions of sound processing in the EEG.

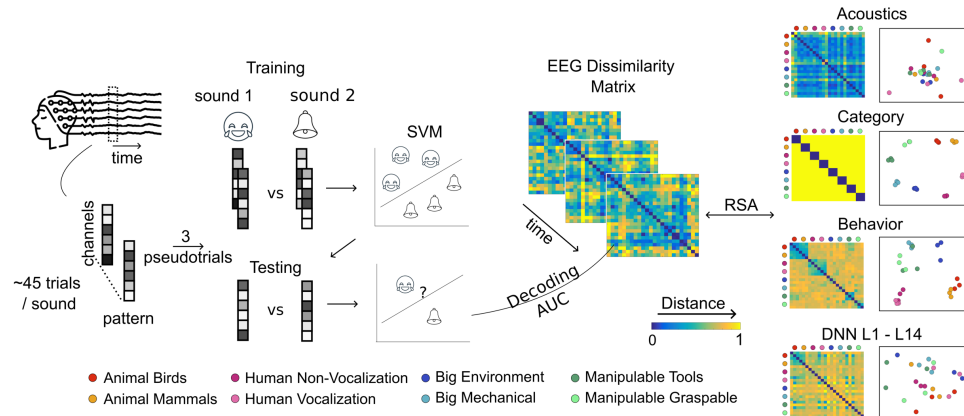


Figure 1. 7 Pipeline of multivariate pattern analysis: Illustration of a pipeline for multivariate decoding analysis and representation similarity analysis (RSA) from chapters 3 and 4.

1.7 Research Goals

During development, auditory functions are shaped by experience within the constraints of biology. But this experience can vary across people, such as in the neurotypical population and in people with congenital blindness. Audition and categorization till now has been mostly studied in these populations separately and this thesis aims to combine investigations across them. This unique perspective will help us to understand the continuum of auditory functions, providing clarity on how they develop in infants, organize in neurotypical people and reorganize in blind individuals.

The central theme of this thesis is the exploration of auditory categorization, focusing on isolating higher-order semantic and conceptual processing from acoustic related representations of the sounds. This includes the categorization of natural sounds, such as artificial objects, voices (speech and non-speech), and the information that can be derived from voices such as emotions.

Being part of previous projects involving frequency tagging (Barbero et al., 2021; Calce et al., 2024), I extended the use of this technique in Chapter 2 to investigate whether auditory emotion categories can be categorized in the neurotypical population. Using rigorous methods, we aim to resolve a) whether the brain generalizes a group of vocalizations as a category, b) if this categorical representation is dependent on the acoustic properties of the vocalization and, c) whether different neural substrates or processes support distinct emotion categories. Although this technique has not yet been implemented in other populations, this chapter lays the groundwork for its expansion beyond auditory

emotions to identifying and categorizing information from voices such as identity and gender, as well as, other natural sounds.

In chapter 3 and 4, using multivariate approach with RSA, I aim to characterize and disentangle various stages in the acoustic to category transformation of the auditory signal in three populations- infants, sighted adults and blind adults using EEG activity in response to natural sounds and compare these neural representations across groups and time. Collectively, these explorations can reveal the extent to which auditory processes are inherent in infants, and how distinctly or similarly these processes mature through different sensory experiences. The focus will be on to estimate the extent to which auditory categories abstract from acoustic properties, as well as investigate how the dynamics of different representations of sound processing are shaped by typical and atypical development (blind individuals).

1.8 Bibliography

Agus, T. R., Paquette, S., Suied, C., Pressnitzer, D., & Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Scientific Reports*, 7(1), Article 1. <https://doi.org/10.1038/s41598-017-11684-1>

Aldridge, M. A., Stillman, R. D., & Bower, T. G. R. (2001). Newborn categorization of vowel-like sounds. *Developmental Science*, 4(2), 220–232. <https://doi.org/10.1111/1467-7687.00167>

Allen, E. J., Burton, P. C., Olman, C. A., & Oxenham, A. J. (2017). Representations of Pitch and Timbre Variation in Human Auditory Cortex. *Journal of Neuroscience*, 37(5), 1284–1293. <https://doi.org/10.1523/JNEUROSCI.2336-16.2016>

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, 59(4), 3677–3689. <https://doi.org/10.1016/j.neuroimage.2011.11.019>

Amedi, A., Raz, N., Azulay, H., Malach, R., & Zohary, E. (2010). Cortical activity during tactile exploration of objects in blind and sighted humans. *Restorative Neurology and Neuroscience*, 28(2), 143–156. <https://doi.org/10.3233/RNN-2010-0503>

Amedi, A., Stern, W. M., Camprodon, J. A., Bermpohl, F., Merabet, L., Rotman, S., Hemond, C., Meijer, P., & Pascual-Leone, A. (2007). Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. *Nature Neuroscience*, 10(6), 687–689. <https://doi.org/10.1038/nn1912>

Aramaki, M., Besson, M., Kronland-Martinet, R., & Ystad, S. (2011). Controlling the Perceived Material in an Impact Sound Synthesizer. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2), 301–314. *IEEE Transactions on Audio, Speech, and Language Processing*.

<https://doi.org/10.1109/TASL.2010.2047755>

Arnott, S. R., Binns, M. A., Grady, C. L., & Alain, C. (2004). Assessing the auditory dual-pathway model in humans. *NeuroImage*, 22(1), 401–408. <https://doi.org/10.1016/j.neuroimage.2004.01.014>

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3), 614.

Barbero, F. M., Calce, R. P., Talwar, S., Rossion, B., & Collignon, O. (2021). Fast Periodic Auditory Stimulation Reveals a Robust Categorical Response to Voices in the Human Brain. *eNeuro*, 8(3). <https://doi.org/10.1523/ENEURO.0471-20.2021>

Barbero, F. M., Talwar, S., Calce, R. P., Rossion, B., & Collignon, O. (2025). Intermodulation frequencies reveal common neural assemblies integrating facial and vocal fearful expressions. *Cortex*, 184, 19–31. <https://doi.org/10.1016/j.cortex.2024.12.008>

Bauer, P. J. (2023). Recognizing limits on the generalizability of findings of psychological science research. *Journal of Applied Research in Memory and Cognition*, 12(2), 199–202. <https://doi.org/10.1037/mac0000125>

Bavelier, D., & Neville, H. J. (2002). Cross-modal plasticity: Where and how? *Nature Reviews Neuroscience*, 3(6), 443–452. <https://doi.org/10.1038/nrn848>

Beauchemin, M., González-Frankenberger, B., Tremblay, J., Vannasing, P., Martínez-Montes, E., Belin, P., Béland, R., Francoeur, D., Carceller, A.-M., Wallois, F., & Lassonde, M. (2011). Mother and stranger: An electrophysiological study of voice processing in newborns. *Cerebral Cortex (New York, N.Y.: 1991)*, 21(8), 1705–1711. <https://doi.org/10.1093/cercor/bhq242>

Bedny, M. (2017). Evidence from Blindness for a Cognitively Pluripotent Cortex. *Trends in Cognitive Sciences*, 21(9), 637–648. <https://doi.org/10.1016/j.tics.2017.06.003>

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, 108(11), 4429–4434. <https://doi.org/10.1073/pnas.1014818108>

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding Voice Perception. *British Journal of Psychology*, 102(4), 711–725. <https://doi.org/10.1111/j.2044-8295.2011.02041.x>

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3), 129–135. <https://doi.org/10.1016/j.tics.2004.01.008>

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Brain Research. Cognitive Brain Research*, 13(1), 17–26. [https://doi.org/10.1016/s0926-6410\(01\)00084-2](https://doi.org/10.1016/s0926-6410(01)00084-2)

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-

selective areas in human auditory cortex. *Nature*, 403(6767), 309–312. <https://doi.org/10.1038/35002078>

Bidelman, G. M., & Walker, B. S. (2017). Attentional modulation and domain-specificity underlying the neural organization of auditory categorical perception. *The European Journal of Neuroscience*, 45(5), 690–699. <https://doi.org/10.1111/ejn.13526>

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10), Article 10. <https://doi.org/10.1038/nrn3565>

Bodin, C., & Belin, P. (2020). Exploring the cerebral substrate of voice perception in primate brains. *Philosophical Transactions of the Royal Society B*. <https://doi.org/10.1098/rstb.2018.0386>

Büchel, C., Price, C., Frackowiak, R. S., & Friston, K. (1998). Different activation patterns in the visual cortex of late and congenitally blind subjects. *Brain: A Journal of Neurology*, 121 (Pt 3), 409–419. <https://doi.org/10.1093/brain/121.3.409>

Cabrera, L., & Gervain, J. (2020). Speech perception at birth: The brain encodes fast and slow temporal information. *Science Advances*, 6(30), eaba7830. <https://doi.org/10.1126/sciadv.aba7830>

Caclin, A., Brattico, E., Tervaniemi, M., Näätänen, R., Morlet, D., Giard, M.-H., & McAdams, S. (2006). Separate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience*, 18(12), 1959–1972. <https://doi.org/10.1162/jocn.2006.18.12.1959>

Calce, R. P., Rekow, D., Barbero, F. M., Kiseleva, A., Talwar, S., Leleu, A., & Collignon, O. (2024). Voice categorization in the four-month-old human brain. *Current Biology: CB*, 34(1), 46-55.e4. <https://doi.org/10.1016/j.cub.2023.11.042>

Capilla, A., Belin, P., & Gross, J. (2013). The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cerebral Cortex (New York, N.Y.: 1991)*, 23(6), 1388–1395. <https://doi.org/10.1093/cercor/bhs119>

Carr, K. W., Perszyk, D. R., Norton, E. S., Voss, J. L., Poeppel, D., & Waxman, S. R. (2021). Developmental changes in auditory-evoked neural activity underlie infants' links between language and cognition. *Developmental Science*, 24(6), e13121. <https://doi.org/10.1111/desc.13121>

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.-P., & Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, 10. <https://doi.org/10.1186/1471-2202-10-127>

Chomsky, Noam. (1959). *A Review of BF Skinner's Verbal Behavior—University of Northwestern*. <https://librarysearch.unwsp.edu>

Cohen, Y. E., Theunissen, F., Russ, B. E., & Gill, P. (2007). Acoustic

Features of Rhesus Vocalizations and Their Representation in the Ventrolateral Prefrontal Cortex. *Journal of Neurophysiology*, 97(2), 1470–1484. <https://doi.org/10.1152/jn.00769.2006>

Collignon, O., Vandewalle, G., Voss, P., Albouy, G., Charbonneau, G., Lassonde, M., & Lepore, F. (2011). Functional specialization for auditory–spatial processing in the occipital cortex of congenitally blind humans. *Proceedings of the National Academy of Sciences*, 108(11), 4435–4440. <https://doi.org/10.1073/pnas.1013928108>

Cornew, L., Carver, L., & Love, T. (2010). There's more to emotion than meets the eye: A processing bias for neutral content in the domain of emotional prosody. *Cognition and Emotion*, 24(7), 1133–1152. <https://doi.org/10.1080/02699930903247492>

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38), E7900–E7909. <https://doi.org/10.1073/pnas.1702247114>

Crowley, K. E., & Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: Age, sleep and modality. *Clinical Neurophysiology*, 115(4), 732–744. <https://doi.org/10.1016/j.clinph.2003.11.021>

Dehaene-Lambertz, G. (2000). Cerebral specialization for speech and non-speech stimuli in infants. *Journal of Cognitive Neuroscience*, 12(3), 449–460. <https://doi.org/10.1162/089892900562264>

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional Neuroimaging of Speech Perception in Infants. *Science*, 298(5600), 2013–2015. <https://doi.org/10.1126/science.1077066>

Dormal, G., Pelland, M., Rezk, M., Yakobov, E., Lepore, F., & Collignon, O. (2018). Functional Preference for Object Sounds and Voices in the Brain of Early Blind and Sighted Individuals. *Journal of Cognitive Neuroscience*, 30(1), 86–106. https://doi.org/10.1162/jocn_a_01186

Ekman, P. (2009). Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), 3449–3451. <https://doi.org/10.1098/rstb.2009.0189>

Elbert, T., Sterr, A., Rockstroh, B., Pantev, C., Müller, M. M., & Taub, E. (2002). Expansion of the Tonotopic Area in the Auditory Cortex of the Blind. *The Journal of Neuroscience*, 22(22), 9941–9944. <https://doi.org/10.1523/JNEUROSCI.22-22-09941.2002>

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128(2), 203–235. <https://doi.org/10.1037/0033-2909.128.2.203>

Ethofer, T., Breitscher, J., Gschwind, M., Kreifelts, B., Wildgruber, D., & Vuilleumier, P. (2012). Emotional Voice Areas: Anatomic Location, Functional

Properties, and Structural Connections Revealed by Combined fMRI/DTI. *Cerebral Cortex*, 22(1), 191–200. <https://doi.org/10.1093/cercor/bhr113>

Falagiarida, F., & Collignon, O. (2019). Time-resolved discrimination of audio-visual emotion expressions. *Cortex*, 119, 184–194. <https://doi.org/10.1016/j.cortex.2019.04.017>

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(38), 15231–15235. <https://doi.org/10.1073/pnas.1221166110>

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). “Who” Is Saying “What”? Brain-Based Decoding of Human Voice and Speech. *Science*, 322(5903), 970–973. <https://doi.org/10.1126/science.1164318>

Frühholz, S., & Belin, P. (2024). Oxford Handbook of Voice Perception | Request PDF. *ResearchGate*. <https://doi.org/10.1177/0301006620938229>

Frühholz, S., & Grandjean, D. (2013). Processing of emotional vocalizations in bilateral inferior frontal cortex. *Neuroscience & Biobehavioral Reviews*, 37(10, Part 2), 2847–2855. <https://doi.org/10.1016/j.neubiorev.2013.10.007>

Frühholz, S., Jellinghaus, A., & Herrmann, M. (2011). Time course of implicit processing and explicit processing of emotional faces and emotional words. *Biological Psychology*, 87(2), 265–274. <https://doi.org/10.1016/j.biopsycho.2011.03.008>

Giard, M. H., Lavikainen, J., Reinikainen, K., Perrin, F., Bertrand, O., Pernier, J., & Näätänen, R. (1995). Separate Representation of Stimulus Frequency, Intensity, and Duration in Auditory Sensory Memory: An Event-Related Potential and Dipole-Model Analysis. *Journal of Cognitive Neuroscience*, 7(2), 133–143. <https://doi.org/10.1162/jocn.1995.7.2.133>

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26(4), 664–672. <https://doi.org/10.1038/s41593-023-01285-9>

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract Encoding of Auditory Objects in Cortical Activity Patterns. *Cerebral Cortex*, 23(9), 2025–2037. <https://doi.org/10.1093/cercor/bhs162>

Giordano, B. L., Whiting, C., Kriegeskorte, N., Kotz, S. A., Gross, J., & Belin, P. (2021). The Representational Dynamics of Perceived Voice Emotions Evolve from Categories to Dimensions. *Nature Human Behaviour*, 5(9), 1203–1213. <https://doi.org/10.1038/s41562-021-01073-0>

Gougoux, F., Belin, P., Voss, P., Lepore, F., Lassonde, M., & Zatorre, R. J. (2009). Voice perception in blind persons: A functional magnetic resonance imaging study. *Neuropsychologia*, 47(13), 2967–2974. <https://doi.org/10.1016/j.neuropsychologia.2009.06.027>

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). The voices of wrath: Brain responses to angry prosody in meaningless speech. *Nature Neuroscience*, 8(2), 145–146. <https://doi.org/10.1038/nn1392>

Graziani, L. J., Weitzman, E. D., & Velasco, M. S. (1968). Neurologic maturation and auditory evoked responses in low birth weight infants. *Pediatrics*, 41(2), 483–494.

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, 29(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Gurman, D., McCormick, C. R., & Klein, R. M. (2021). *Crossmodal Correspondence Between Auditory Timbre and Visual Shape*. <https://doi.org/10.1163/22134808-bja10067>

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: Controversies and consensus. *Trends in Cognitive Sciences*, 16(9), 458–466. <https://doi.org/10.1016/j.tics.2012.07.006>

He, C., Peelen, M. V., Han, Z., Lin, N., Caramazza, A., & Bi, Y. (2013). Selectivity for large nonmanipulable objects in scene-selective visual cortex does not require visual experience. *NeuroImage*, 79, 1–9. <https://doi.org/10.1016/j.neuroimage.2013.04.051>

Heimler, B., Weisz, N., & Collignon, O. (2014). Revisiting the adaptive and maladaptive effects of crossmodal plasticity. *Neuroscience*, 283, 44–63. <https://doi.org/10.1016/j.neuroscience.2014.08.003>

Jessen, S., & Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *NeuroImage*, 58(2), 665–674. <https://doi.org/10.1016/j.neuroimage.2011.06.035>

Jin, X., & Wang, Z. (2005). An Emotion Space Model for Recognition of Emotions in Spoken Chinese. In J. Tao, T. Tan, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction* (pp. 397–402). Springer. https://doi.org/10.1007/11573548_51

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5), 770–814. <https://doi.org/10.1037/0033-2909.129.5.770>

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), 630–644.e16. <https://doi.org/10.1016/j.neuron.2018.03.044>

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in*

- Neurobiology*, 55, 121–132. <https://doi.org/10.1016/j.conb.2019.02.003>
- Klasen, M., Chen, Y.-H., & Mathiak, K. (2012). Multisensory emotions: Perception, combination and underlying neural processes. *Reviews in the Neurosciences*, 23(4), 381–392. <https://doi.org/10.1515/revneuro-2012-0040>
- Kotz, S. A., Kalberlah, C., Bahlmann, J., Friederici, A. D., & Haynes, J.-D. (2013). Predicting vocal emotion expressions from the human brain. *Human Brain Mapping*, 34(8), 1971–1981. <https://doi.org/10.1002/hbm.22041>
- Kragel, P. A., & LaBar, K. S. (2016). Decoding the Nature of Emotion in the Brain. *Trends in Cognitive Sciences*, 20(6), 444–455. <https://doi.org/10.1016/j.tics.2016.03.011>
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. <https://doi.org/10.3389/neuro.06.004.2008>
- Kriegstein, K. V., & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, 22(2), 948–955. <https://doi.org/10.1016/j.neuroimage.2004.02.020>
- Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2), 93–107. <https://doi.org/10.3758/bf03212211>
- Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5(11), 831–843. <https://doi.org/10.1038/nrn1533>
- Langers, D. R. M., van Dijk, P., Schoenmaker, E. S., & Backes, W. H. (2007). fMRI activation in relation to sound intensity and loudness. *NeuroImage*, 35(2), 709–718. <https://doi.org/10.1016/j.neuroimage.2006.12.013>
- Leaver, A. M., & Rauschecker, J. P. (2010). Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. *Journal of Neuroscience*, 30(22), 7604–7612. <https://doi.org/10.1523/JNEUROSCI.0296-10.2010>
- Lee, G. Y., & Kisilevsky, B. S. (2014). Fetuses respond to father’s voice but prefer mother’s voice after birth. *Developmental Psychobiology*, 56(1), 1–11. <https://doi.org/10.1002/dev.21084>
- Lewis, J. W., Talkington, W. J., Tallaksen, K. C., & Frum, C. A. (2012). Auditory object salience: Human cortical processing of non-biological action sounds and their acoustic signal attributes. *Frontiers in Systems Neuroscience*, 6. <https://doi.org/10.3389/fnsys.2012.00027>
- Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., & Brefczynski-Lewis, J. A. (2009). Human Cortical Organization for Processing Vocalizations Indicates Representation of Harmonic Structure as a Signal Attribute. *The Journal of Neuroscience*, 29(7), 2283–2296. <https://doi.org/10.1523/JNEUROSCI.4145-08.2009>
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M.

(1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461. <https://doi.org/10.1037/h0020279>

Liberman, A. M., & Mattingly, I. G. (n.d.). *A Specialization for Speech Perception*. 243.

Liu, Y., Huang, H., McGinnis-Deweese, M., Keil, A., & Ding, M. (2012). Neural Substrate of the Late Positive Potential in Emotional Processing. *Journal of Neuroscience*, 32(42), 14563–14572. <https://doi.org/10.1523/JNEUROSCI.3109-12.2012>

Lochy, A., Van Belle, G., & Rossion, B. (2015). A robust index of lexical representation in the left occipito-temporal cortex as evidenced by EEG responses to fast periodic visual stimulation. *Neuropsychologia*, 66, 18–31. <https://doi.org/10.1016/j.neuropsychologia.2014.11.007>

Lowe, M. X., Mohsenzadeh, Y., Lahner, B., Charest, I., Oliva, A., & Teng, S. (2021). Cochlea to categories: The spatiotemporal dynamics of semantic auditory representations. *Cognitive Neuropsychology*, 38(7–8), 468–489. <https://doi.org/10.1080/02643294.2022.2085085>

Mahmoudzadeh, M., Dehaene-Lambertz, G., Fournier, M., Kongolo, G., Goudjil, S., Dubois, J., Grebe, R., & Wallois, F. (2013). Syllabic discrimination in premature human infants prior to complete formation of cortical layers. *Proceedings of the National Academy of Sciences*, 110(12), 4846–4851. <https://doi.org/10.1073/pnas.1212220110>

Makin, T. R., & Krakauer, J. W. (2023). Against cortical reorganisation. *eLife*, 12, e84716. <https://doi.org/10.7554/eLife.84716>

Manjunath, N. K., Srinivas, R., Nirmal, K. S., Nagendra, H. R., Kumar, A., & Telles, S. (1998). Shorter Latencies of Components of Middle Latency Auditory Evoked Potentials in Congenitally Blind Compared to Normal Sighted Subjects. *International Journal of Neuroscience*, 95(3–4), 173–181. <https://doi.org/10.3109/00207459809003339>

Manley, G. A. (2017). Comparative Auditory Neuroscience: Understanding the Evolution and Function of Ears. *JARO: Journal of the Association for Research in Otolaryngology*, 18(1), 1–24. <https://doi.org/10.1007/s10162-016-0579-3>

Marlair, C., Lochy, A., & Crollen, V. (2024). Frequency-tagging EEG reveals the effect of attentional focus on abstract magnitude processing. *Psychonomic Bulletin & Review*, 31(5), 2266–2274. <https://doi.org/10.3758/s13423-024-02480-w>

Mattioni, S., Rezk, M., Battal, C., Bottini, R., Cuculiza Mendoza, K. E., Oosterhof, N. N., & Collignon, O. (2020). Categorical representation from sound and sight in the ventral occipito-temporal cortex of sighted and blind. *eLife*, 9, e50732. <https://doi.org/10.7554/eLife.50732>

Mattioni, S., Rezk, M., Battal, C., Vadlamudi, J., & Collignon, O. (2022). Impact of blindness onset on the representation of sound categories in occipital and temporal cortices. *eLife*, 11, e79370. <https://doi.org/10.7554/eLife.79370>

- Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoni, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29(2), 143–178. [https://doi.org/10.1016/0010-0277\(88\)90035-2](https://doi.org/10.1016/0010-0277(88)90035-2)
- Menon, V., Levitin, D. J., Smith, B. K., Lember, A., Krasnow, B. D., Glazer, D., Glover, G. H., & McAdams, S. (2002). Neural Correlates of Timbre Change in Harmonic Sounds. *NeuroImage*, 17(4), 1742–1754. <https://doi.org/10.1006/nimg.2002.1295>
- Merabet, L. B., & Pascual-Leone, A. (2010). Neural Reorganization Following Sensory Loss: The Opportunity Of Change. *Nature Reviews. Neuroscience*, 11(1), 44–52. <https://doi.org/10.1038/nrn2758>
- Moerel, M., Martino, F. D., & Formisano, E. (2012). Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. *Journal of Neuroscience*, 32(41), 14205–14216. <https://doi.org/10.1523/JNEUROSCI.1388-12.2012>
- Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology*, 24(4), 375–425. <https://doi.org/10.1111/j.1469-8986.1987.tb00311.x>
- Naveen, K. V., Srinivas, R., Nirmala, K. S., Nagarathna, R., Nagendra, H. R., & Telles, S. (1998). Differences between Congenitally Blind and Normally Sighted Subjects in the P1 Component of Middle Latency Auditory Evoked Potentials. *Perceptual and Motor Skills*, 86(3_suppl), 1192–1194. <https://doi.org/10.2466/pms.1998.86.3c.1192>
- Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, 88(6), 1281–1296. <https://doi.org/10.1016/j.neuron.2015.11.035>
- Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, 6(3), 455–469. <https://doi.org/10.1038/s41562-021-01261-y>
- Nozaradan, S., Mouraux, A., & Cousineau, M. (2017). Frequency tagging to track the neural processing of contrast in fast, continuous sound sequences. *Journal of Neurophysiology*, 118(1), 243–253. <https://doi.org/10.1152/jn.00971.2016>
- Ogg, M., Carlson, T. A., & Slevc, L. R. (2020). The Rapid Emergence of Auditory Object Representations in Cortex Reflect Central Acoustic Attributes. *Journal of Cognitive Neuroscience*, 32(1), 111–123. https://doi.org/10.1162/jocn_a_01472
- Ogg, M., Moraczewski, D., Kuchinsky, S. E., & Slevc, L. R. (2019). Separable neural representations of sound sources: Speaker identity and musical

- timbre. *NeuroImage*, 191, 116–126.
<https://doi.org/10.1016/j.neuroimage.2019.01.075>
- Pascual-Leone, A., & Torres, F. (1993). Plasticity of the sensorimotor cortex representation of the reading finger in Braille readers. *Brain: A Journal of Neurology*, 116 (Pt 1), 39–52. <https://doi.org/10.1093/brain/116.1.39>
- Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, 36(4), 767–776. [https://doi.org/10.1016/s0896-6273\(02\)01060-7](https://doi.org/10.1016/s0896-6273(02)01060-7)
- Paulmann, S., & Kotz, S. A. (2008). Early emotional prosody perception based on different speaker voices. *Neuroreport*, 19(2), 209–213. <https://doi.org/10.1097/WNR.0b013e3282f454db>
- Peelen, M. V., He, C., Han, Z., Caramazza, A., & Bi, Y. (2014). Nonvisual and Visual Object Shape Representations in Occipitotemporal Cortex: Evidence from Congenitally Blind and Sighted Adults. *Journal of Neuroscience*, 34(1), 163–170. <https://doi.org/10.1523/JNEUROSCI.1114-13.2014>
- Pell, M. D., & Kotz, S. A. (2011). On the Time Course of Vocal Emotion Recognition. *PLOS ONE*, 6(11), e27256. <https://doi.org/10.1371/journal.pone.0027256>
- Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, 3. <https://doi.org/10.3389/fpsyg.2012.00023>
- Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G., Watson, R. H., Fleming, D., Crabbe, F., Valdes-Sosa, M., & Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, 119, 164–174. <https://doi.org/10.1016/j.neuroimage.2015.06.050>
- Perszyk, D. R., & Waxman, S. R. (2016). Listening to the calls of the wild: The role of experience in linking language and cognition in young infants. *Cognition*, 153, 175–181. <https://doi.org/10.1016/j.cognition.2016.05.004>
- Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience*, 11(3), 367–374. <https://doi.org/10.1038/nn2043>
- Petrides, M. (2005). Lateral prefrontal cortex: Architectonic and functional organization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 781–795. <https://doi.org/10.1098/rstb.2005.1631>
- Pietrini, P., Furey, M. L., Ricciardi, E., Gobbini, M. I., Wu, W.-H. C., Cohen, L., Guazzelli, M., & Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 101(15), 5658–5663. <https://doi.org/10.1073/pnas.0400707101>
- Poirier, C., Collignon, O., Scheiber, C., Renier, L., Vanlierde, A., Tranduy, D., Veraart, C., & De Volder, A. G. (2006). Auditory motion perception activates

visual motion areas in early blind subjects. *NeuroImage*, 31(1), 279–285. <https://doi.org/10.1016/j.neuroimage.2005.11.036>

Polver, S., Miller-Viacava, N., Fraticelli, M., Gervain, J., & Lorenzi, C. (2024). Developmental origins of natural sound perception. *Frontiers in Psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1474961>

Purhonen, M., Kilpeläinen-Lees, R., Valkonen-Korhonen, M., Karhu, J., & Lehtonen, J. (2005). Four-month-old infants process own mother's voice faster than unfamiliar voices—Electrical signs of sensitization in infant brain. *Brain Research. Cognitive Brain Research*, 24(3), 627–633. <https://doi.org/10.1016/j.cogbrainres.2005.03.012>

Rauschecker, J. P. (2002). Cortical map plasticity in animals and humans. In *Progress in Brain Research* (Vol. 138, pp. 73–88). Elsevier. [https://doi.org/10.1016/S0079-6123\(02\)38072-5](https://doi.org/10.1016/S0079-6123(02)38072-5)

Renier, L. A., Anurova, I., Volder, A. G. D., Carlson, S., VanMeter, J., & Rauschecker, J. P. (2010). Preserved Functional Specialization for Spatial Processing in the Middle Occipital Gyrus of the Early Blind. *Neuron*, 68(1), 138–148. <https://doi.org/10.1016/j.neuron.2010.09.021>

Retter, T. L., & Rossion, B. (2016). Uncovering the neural magnitude and spatio-temporal dynamics of natural image categorization in a fast visual stream. *Neuropsychologia*, 91, 9–28. <https://doi.org/10.1016/j.neuropsychologia.2016.07.028>

Ricciardi, E., Vanello, N., Sani, L., Gentili, C., Scilingo, E. P., Landini, L., Guazzelli, M., Bicchi, A., Haxby, J. V., & Pietrini, P. (2007). The Effect of Visual Experience on the Development of Functional Architecture in hMT+. *Cerebral Cortex*, 17(12), 2933–2939. <https://doi.org/10.1093/cercor/bhm018>

Roberts, T. P. L., Ferrari, P., Stufflebeam, S. M., & Poeppel, D. (2000). Latency of the Auditory Evoked Neuromagnetic Field Components: Stimulus Dependence and Insights Toward Perception. *Journal of Clinical Neurophysiology*, 17(2), 114.

Röder, B., Stock, O., Bien, S., Neville, H., & Rösler, F. (2002). Speech processing activates visual cortex in congenitally blind humans. *The European Journal of Neuroscience*, 16(5), 930–936. <https://doi.org/10.1046/j.1460-9568.2002.02147.x>

Rosburg, T. (2003). Left hemispheric dipole locations of the neuromagnetic mismatch negativity to frequency, intensity and duration deviants. *Cognitive Brain Research*, 16(1), 83–90. [https://doi.org/10.1016/S0926-6410\(02\)00222-7](https://doi.org/10.1016/S0926-6410(02)00222-7)

Rossion, B., & Jacques, C. (2012). *The N170: Understanding the time course of face perception in the human brain*. APA PsycNET. <https://psycnet.apa.org/record/2013-01016-005>

Rossion, B., Torfs, K., Jacques, C., & Liu-Shuang, J. (2015). Fast periodic presentation of natural images reveals a robust face-selective electrophysiological

- response in the human brain. *Journal of Vision*, 15. <https://doi.org/10.1167/15.1.18>
- Roswadowitz, C., Swanborough, H., & Fröhholz, S. (2020). Categorizing human vocal signals depends on an integrated auditory-frontal cortical network. *Human Brain Mapping*, 42(5), 1503–1517. <https://doi.org/10.1002/hbm.25309>
- Roth, W. T., Ford, J. M., Lewis, S. J., & Kopell, B. S. (1976). Effects of Stimulus Probability and Task-Relevance on Event-Related Potentials. *Psychophysiology*, 13(4), 311–317. <https://doi.org/10.1111/j.1469-8986.1976.tb03082.x>
- Rupp, K. M., Hect, J. L., Harford, E. E., Holt, L. L., Ghuman, A. S., & Abel, T. J. (2024). A hierarchy of processing complexity and timescales for natural sounds in human auditory cortex. *bioRxiv: The Preprint Server for Biology*, 2024.05.24.595822. <https://doi.org/10.1101/2024.05.24.595822>
- Russ, B. E., Lee, Y.-S., & Cohen, Y. E. (2007). Neural and behavioral correlates of auditory categorization. *Hearing Research*, 229(1), 204–212. <https://doi.org/10.1016/j.heares.2006.10.010>
- Sabourin, C. J., Merrikhi, Y., & Lomber, S. G. (2022). Do blind people hear better? *Trends in Cognitive Sciences*, 26(11), 999–1012. <https://doi.org/10.1016/j.tics.2022.08.016>
- Salvia, E., Bestelmeyer, P. E. G., Kotz, S. A., Rousselet, G. A., Pernet, C. R., Gross, J., & Belin, P. (2014). Single-subject analyses of magnetoencephalographic evoked responses to the acoustic properties of affective non-verbal vocalizations. *Frontiers in Neuroscience*, 8. <https://www.frontiersin.org/article/10.3389/fnins.2014.00422>
- Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*, 10(1), e1003412. <https://doi.org/10.1371/journal.pcbi.1003412>
- Sauter, D. A., & Eimer, M. (2010). Rapid Detection of Emotion from Human Vocalizations. *Journal of Cognitive Neuroscience*, 22(3), 474–481. <https://doi.org/10.1162/jocn.2009.21215>
- Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6), 2408–2412. <https://doi.org/10.1073/pnas.0908239106>
- Schepers, I. M., Hipp, J. F., Schneider, T. R., Röder, B., & Engel, A. K. (2012). Functionally specific oscillatory activity correlates between visual and auditory cortex in the blind. *Brain*, 135(3), 922–934. <https://doi.org/10.1093/brain/aws014>
- Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32(1), 76–92.

- Schirmer, A., Zysset, S., Kotz, S. A., & Yves von Cramon, D. (2004). Gender differences in the activation of inferior frontal cortex during emotional speech perception. *NeuroImage*, 21(3), 1114–1123. <https://doi.org/10.1016/j.neuroimage.2003.10.048>
- Schönwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34), 14611–14616. <https://doi.org/10.1073/pnas.0907682106>
- Shahidullah, S., & Hepper, P. G. (1994). Frequency discrimination by the fetus. *Early Human Development*, 36(1), 13–26. [https://doi.org/10.1016/0378-3782\(94\)90029-9](https://doi.org/10.1016/0378-3782(94)90029-9)
- Smits, R., Sereno, J., & Jongman, A. (2006). Categorization of sounds. *Journal of Experimental Psychology. Human Perception and Performance*, 32(3), 733–754. <https://doi.org/10.1037/0096-1523.32.3.733>
- Staeren, N., Renvall, H. L. M., de Martino, F., Goebel, R., & Formisano, E. (2009). Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex. *Current Biology*, 19(6), 498–502. <https://doi.org/10.1016/j.cub.2009.01.066>
- Striem-Amit, E., & Amedi, A. (2014). Visual cortex extrastriate body-selective area activation in congenitally blind people “seeing” by using sounds. *Current Biology: CB*, 24(6), 687–692. <https://doi.org/10.1016/j.cub.2014.02.010>
- Striem-Amit, E., Cohen, L., Dehaene, S., & Amedi, A. (2012). Reading with Sounds: Sensory Substitution Selectively Activates the Visual Word Form Area in the Blind. *Neuron*, 76(3), 640–652. <https://doi.org/10.1016/j.neuron.2012.08.026>
- Tajadura-Jiménez, A., Väljamäe, A., Asutay, E., & Västfjäll, D. (2010). Embodied auditory perception: The emotional impact of approaching and receding sound sources. *Emotion*, 10(2), 216–229. <https://doi.org/10.1037/a0018422>
- Talwar, S., Barbero, F. M., Calce, R. P., & Collignon, O. (2023). Automatic Brain Categorization of Discrete Auditory Emotion Expressions. *Brain Topography*, 36(6), 854–869. <https://doi.org/10.1007/s10548-023-00983-8>
- Telkemeyer, S., Rossi, S., Koch, S. P., Nierhaus, T., Steinbrink, J., Poeppel, D., Obrig, H., & Wartenburger, I. (2009). Sensitivity of Newborn Auditory Cortex to the Temporal Structure of Sounds. *The Journal of Neuroscience*, 29(47), 14726–14733. <https://doi.org/10.1523/JNEUROSCI.1246-09.2009>
- Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, 15(6), 355–366. <https://doi.org/10.1038/nrn3731>
- Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional Specialization in Rhesus Monkey Auditory Cortex. *Science*, 292(5515), 290–293. <https://doi.org/10.1126/science.1058911>
- van den Hurk, J., Van Baelen, M., & Op de Beeck, H. P. (2017). Development

of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences of the United States of America*, 114(22), E4501–E4510. <https://doi.org/10.1073/pnas.1612862114>

Vetter, P., Bola, L., Reich, L., Bennett, M., Muckli, L., & Amedi, A. (2020). Decoding Natural Sounds in Early “Visual” Cortex of Congenitally Blind Individuals. *Current Biology*, 30(15), 3039–3044.e2. <https://doi.org/10.1016/j.cub.2020.05.071>

Voss, P., Pike, B. G., & Zatorre, R. J. (2014). Evidence for both compensatory plastic and disuse atrophy-related neuroanatomical changes in the blind. *Brain*, 137(4), 1224–1240. <https://doi.org/10.1093/brain/awu030>

Vouloumanos, A., Hauser, M. D., Werker, J. F., & Martin, A. (2010). The Tuning of Human Neonates’ Preference for Speech. *Child Development*, 81(2), 517–527. <https://doi.org/10.1111/j.1467-8624.2009.01412.x>

Vytal, K., & Hamann, S. (2010). Neuroimaging Support for Discrete Neural Correlates of Basic Emotions: A Voxel-based Meta-analysis. *Journal of Cognitive Neuroscience*, 22(12), 2864–2885. <https://doi.org/10.1162/jocn.2009.21366>

Wang, X., Peelen, M. V., Han, Z., He, C., Caramazza, A., & Bi, Y. (2015). How Visual Is the Visual Cortex? Comparing Connectional and Functional Fingerprints between Congenitally Blind and Sighted Individuals. *Journal of Neuroscience*, 35(36), 12545–12559. <https://doi.org/10.1523/JNEUROSCI.3914-14.2015>

Warren, J. D., Jennings, A. R., & Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *NeuroImage*, 24(4), 1052–1057. <https://doi.org/10.1016/j.neuroimage.2004.10.031>

Whiteside, S. P. (1998). The Identification of a Speaker’s Sex from Synthesized Vowels. *Perceptual and Motor Skills*, 87(2), 595–600. <https://doi.org/10.2466/pms.1998.87.2.595>

Wolbers, T., Zahorik, P., & Giudice, N. A. (2011). Decoding the direction of auditory motion in blind humans. *NeuroImage*, 56(2), 681–687. <https://doi.org/10.1016/j.neuroimage.2010.04.266>

Zäske, R., Skuk, V. G., Kaufmann, J. M., & Schweinberger, S. R. (2013). Perceiving vocal age and gender: An adaptation approach. *Acta Psychologica*, 144(3), 583–593. <https://doi.org/10.1016/j.actpsy.2013.09.0>

CHAPTER 2

Automatic Brain Categorization of Discrete Auditory Emotion Expressions

Siddharth Talwar ¹, Francesca M. Barbero ¹, Roberta P. Calce ¹, Olivier Collignon ^{1,2}

¹ Institute for Research in Psychology (IPSY) & Neuroscience (IoNS), Louvain Bionics, University of Louvain (UCLouvain), Louvain, Belgium

² School of Health Sciences, HES-SO Valais-Wallis, The Sense Innovation and Research Center, Lausanne and Sion, Switzerland

Published article: Brain Topography (2023) 36:854–869

Abstract

Seamlessly extracting emotional information from voices is crucial for efficient interpersonal communication. However, it remains unclear how the brain categorizes vocal expressions of emotion beyond the processing of their acoustic features. In our study, we developed a new approach combining electroencephalographic recordings (EEG) in humans with a frequency-tagging paradigm to ‘tag’ automatic neural responses to specific categories of emotion expressions. Participants were presented with a periodic stream of heterogeneous non-verbal emotional vocalizations belonging to five emotion categories: anger, disgust, fear, happiness and sadness at 2.5 Hz (stimuli length of 350 ms with a 50 ms silent gap between stimuli). Importantly, unknown to the participant, a specific emotion category appeared at a target presentation rate of 0.83 Hz that would elicit an additional response in the EEG spectrum only if the brain *discriminates* the target emotion category from other emotion categories and *generalizes* across heterogeneous exemplars of the target emotion category. Stimuli were matched across emotion categories for harmonicity-to-noise ratio, spectral center of gravity and pitch. Additionally, participants were presented with a scrambled version of the stimuli with identical spectral content and periodicity but disrupted intelligibility. Both types of sequences had comparable envelopes and early auditory peripheral processing computed via the simulation of the cochlear response. We observed that in addition to the responses at the general presentation frequency (2.5 Hz) in both intact and scrambled sequences, a greater peak in the EEG spectrum at the target emotion presentation rate (0.83 Hz) and its harmonics emerged in the intact sequence in comparison to the scrambled sequence. The greater response at the target frequency in the intact sequence, together with our stimuli matching procedure, suggest that the categorical brain response elicited by a specific emotion is at least partially independent from the low-level acoustic features of the sounds. Moreover, responses at the fearful and happy vocalizations presentation rates elicited different

topographies and different temporal dynamics, suggesting that different discrete emotions are represented differently in the brain. Our paradigm revealed the brain's ability to automatically categorize non-verbal vocal emotion expressions objectively (at a predefined frequency of interest), behavior-free, rapidly (in few minutes of recording time) and robustly (with a high signal-to-noise ratio), making it a useful tool to study vocal emotion processing and auditory categorization in general and in populations where behavioral assessments are more challenging.

2.1 Introduction

In humans and other animals, efficient categorization of emotion expressions is crucial for effective social interactions and survival. In Darwin's milestone book "The Expression of the Emotions in Man and Animals" published in 1872, attention was mostly focused on the face as the carrier of emotion expressions. Since then, research on facial emotion expressions has led to the suggestion that at least six basic emotions- anger, disgust, fear, happiness, surprise and sadness (Ekman 1993) can be expressed using specific facial movements as coded by the Facial Acting Coding System (FACS, Ekman & Friesen 1978; Waller et al. 2020). These emotion categories are expressed similarly across different cultures (Elfenbein & Ambady 2002), arise very early in development (Flom & Bahrack 2007; Poncet et al. 2022), and can be found in our evolutionary ancestors (Darwin 1872; Waller & Micheletta 2013).

Although Darwin mentioned the importance of vocalizations as a carrier of affective signals (Darwin 1872), it was suggested that voice was yet to be proven an effective source for detecting discrete emotion categories (Ekman 2009). However, mounting evidence shows that discrete emotion expressions can be delivered and decoded through vocal expressions with a high accuracy in humans (Cornew et al. 2010; Falagiarda and Collignon 2019; Pell and Kotz 2011). In fact, the same discrete emotion categories as those found for faces (Ekman 1993) also express similarly through voices across cultures (Juslin and Laukka 2003; Sauter et al. 2010; Sauter and Eimer 2010; Scherer et al. 2001) and emotion detection from the voice also develops early in infancy (Izard et al. 1980; Mehler et al. 1978; Zhao et al. 2021).

It is well known that the distinct acoustic features that characterize different discrete categories of vocal emotions activate different patches in the core region of the auditory cortex that is known to be tonotopically organized (Talavage et al. 2004). However, functional magnetic resonance imaging (fMRI) studies revealed that some temporal regions show distinct activations for diverse affective vocalizations (Frühholz and Grandjean 2013b; Ethofer et al. 2012), partially independent of acoustic amplitude and frequency cues (Giordano et al. 2021; Grandjean et al. 2005). In addition to auditory cortices, other brain regions are commonly activated across a wide range of auditory emotion categories, such as the amygdala (Fecteau

et al. 2007; Wiethoff et al. 2009) and the medial prefrontal cortex (Etkin et al. 2011; Kober et al. 2008). Despite the existence of these regions involved in the processing of various emotion expressions, it is still debated whether discrete emotions may recruit separate brain areas (Calder et al. 2001; Ethofer et al. 2009; Frühholz and Grandjean 2013a; Johnstone et al. 2006; Kober et al. 2008; Kotz et al. 2013; Mauchand & Zhang 2022; Phan et al. 2002; Phillips et al. 1998; Vytal and Hamann 2010).

Affective vocalizations have been extensively studied using event related potentials (ERPs). Components as early as ~ 100 ms which usually reflect acoustic processing were found to be modulated by emotional non-speech utterances compared to neutral vocalizations (Jessen et al. 2012). This may suggest that the difference in early ERPs are driven by acoustic features (Salvia et al. 2014). In addition, there is also evidence of enhancement of later ERP components such as the early posterior negativity (EPN: 200–350 ms) and late positive potential (LPP: ~ 400 ms) when emotional utterances are contrasted with neutral vocalizations (Frühholz et al. 2011; Jessen and Kotz 2011). These later differences are thought to index the mechanism of affective categorization (Schirmer and Kotz 2006) that may be partially independent from acoustic differences. However in most electrophysiological studies, the responses to emotional vocalizations are compared to neutral vocalizations rather than to different emotions, thus leaving unanswered whether the observed differences are due to acoustic differences (Banse and Scherer 1996; but see Bostanov and Kotchoubey 2004; Pell et al. 2015) or change in valence or arousal rather than the categorisation of discrete emotion expressions.

In this study, we aimed to develop a novel approach combining electroencephalographic recordings (EEG) in humans with a frequency-tagging paradigm to provide an objective measure of automatic categorization of vocal emotion expressions beyond the processing of acoustic features. The frequency-tagging approach (Regan 1989) relies on the fact that under external stimulation of periodic stimuli, the brain regions that encode the stimuli synchronize at the exact same frequency (Norcia et al. 2015). We adapted the Fast Periodic Auditory Stimulation paradigm (FPAS, Barbero et al. 2021) to present different exemplars of discrete auditory emotion categories. The stimuli were presented periodically at 2.5 Hz (e.g. the stimulus length is 350 ms with a 50 ms silence/gap). Importantly, within the sequence of affective vocalizations, there is also embedded another periodically occurring target emotion category (e.g. different exemplars of Fear presented at every third position at 0.83 Hz). Thus, the brain will elicit a response at the general rate of sound presentation and its harmonics (2.5 Hz, 5 Hz, 7.5 Hz etc.) if it can segregate different affective sounds. Crucially, we will be able to observe a response at the target frequency and its harmonics (0.83 Hz, 1.66 Hz, 3.32 Hz etc.) only if the participants' brain can discriminate the vocalization of the target category from

sounds of other frequent non-target categories as well as generalize all target vocalizations to one common emotion category (see Barbero et al. 2021).

From a fundamental point of view, our study aimed to investigate whether the human brain categorizes discrete auditory emotion expressions partially independently from their acoustic properties. This was implemented by the careful selection of non-verbal vocalizations with similar acoustic properties (spectral center of gravity, harmonicity-to-noise ratio, pitch) and introducing a second control stimulation sequence of frequency-scrambled sounds with similar spectro-temporal profile (frequency content, sound's envelope) to the intact sounds and identical periodic constraints but disrupted intelligibility (Barbero et al. 2021; Dormal et al. 2018). In addition to the frequency domain analysis, we conducted time-locked analyses to investigate the time course of the response underlying emotion categorization. Finally, our study also aimed to provide a powerful tool to investigate the brain's ability to categorize auditory emotion expressions objectively (the response lies at predefined frequencies of interest), robustly (with a high signal-to-noise ratio), and automatically (does not need an explicit request to process affective vocalizations).

2.2 Materials and Methods

2.2.1 EEG Experiment

2.2.1.1 Participants

Twenty-four participants (12 females, mean age: 22.29, S.D.: 2.33, range: 19–28 years) participated in the study. All participants reported no history of neurological or audiological disorders and were right-handed. Auditory deficits were self-reported by the participants since there are strong associations of self-reported and audiometric hearing loss, especially within the age range of our sample set (Gomez et al. 2001; Hannula et al. 2011; Kiely et al. 2012). The experiment was approved by the local ethical committee of the Catholic University of Louvain (UCLouvain, project 2016–25). All participants provided written informed consent and received financial compensation for their participation.

2.2.1.2 Stimuli

We selected sounds of five primary emotion categories: anger, disgust, fear, happiness and sadness. The stimuli were extracted from clips in which professional actors and actresses performed emotion expressions in varied styles and intensities without any linguistic content (Belin et al. 2008). Additionally, a few stimuli were selected from a database of non-verbal vocalizations depicting several distinct varieties of emotion categories (Cowen and Keltner 2017). A total of 96 heterogeneous sounds were cropped to a length of 350 ms to allow periodic

presentation. All sounds were equalized in overall energy (root mean square, RMS) and 10 ms ramps were applied at the start and at the end of the stimuli to avoid clicking. To make sure that the empirical EEG responses to emotion vocalizations are not driven by different acoustic features across different emotion categories, we confirmed that the sounds have comparable spectral center of gravity (COG; $F = 1.1882$, $p = 0.3224$, $\eta^2 = 0.0567$), harmonicity-tonoise ratio (HNR; $F = 1.6930$, $p = 0.1599$, $\eta^2 = 0.0789$) and pitch ($F = 1.7618$, $p = 0.1448$, $\eta^2 = 0.0819$) across the five emotion categories (Figure 2.1a). The acoustic properties were computed using custom scripts in Praat (Boersma and Weenink 2001).

To select emotional stimuli eliciting a reliable recognition of the intended emotion, we conducted a behavioral experiment where 10 participants (mean age = 27.7, S.D. = 4.6200; who did not participate in the EEG experiment) categorized each sound as pertaining to one of the five selected emotion categories using a five-alternative forced choice task. 3 blocks were presented in a randomized fashion where each block was composed of 96 stimuli with 1 repetition per block (anger, disgust, sadness: 14 each; fear and happiness: 27 each). Among those sounds, 84 sounds across 5 categories were selected (anger, disgust, sadness: 12 each; fear, happiness: 24 each). All selected sounds achieved 80% or more recognition accuracy across all participants (chance level: 20%) and were subsequently used to build the sequences for the EEG session (Figure 2.1b). We required double number of stimuli for Fear and Happiness categories to present an equal number of unique stimuli of each category in the sequence (More details in Sect. “Sequences and Procedure”). Each emotion category consisted of an equal number of sounds by male and female actors to incorporate heterogeneous spectro-temporal profiles across the stimuli set.

Chapter 2. Categorization of Auditory Emotion Expressions

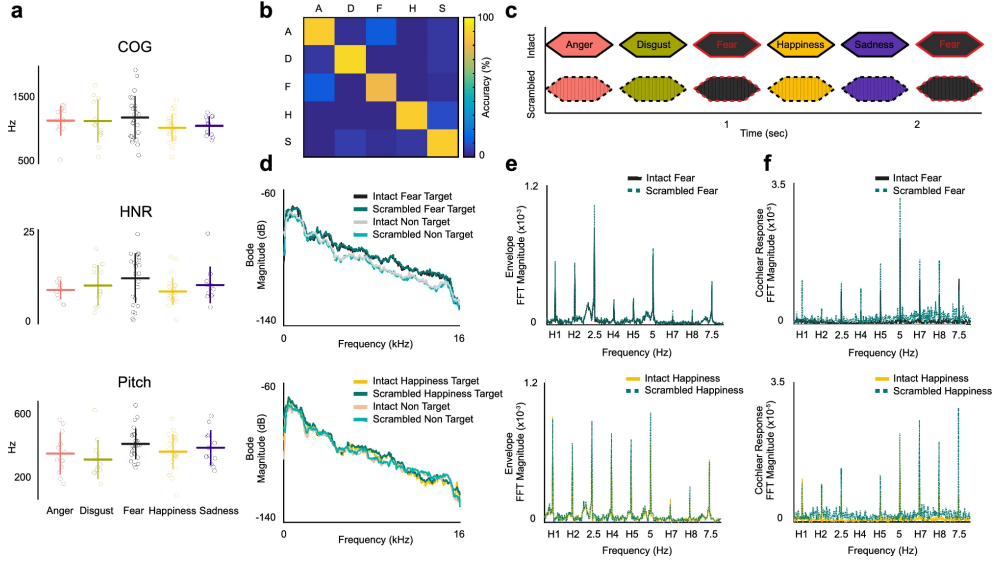


Figure 2. 1 Stimuli and Sequences: A The selected sounds of five emotion categories have similar spectral properties: center of gravity (COG), harmonicity to noise ratio (HNR) and pitch. B Behavioral experiment validated 84 short non-verbal vocalizations to depict the appropriate emotion category. A = Anger, D = Disgust, F = Fear, H = Happiness, S = Sadness. C Sounds were presented periodically at 2.5 Hz. The target emotion (e.g. fear in this illustration) repeated at every third position leading to a target presentation rate of 0.83 Hz while other emotion categories were presented randomly. D Bode plot shows similar averaged spectral power of intact and scrambled stimuli for target and non-target sounds for each emotion condition. E Similar averaged FFT distribution of the envelope between intact and scrambled sequences indicates similar temporal profiles across the two types of sequences in each emotion condition. F No significant difference found in the averaged FFT magnitudes of the simulated cochlear response to intact and scrambled sequence, verifying similar spectro-temporal profile of all sequences as well as the processing of acoustic features at the cochlear level.

To create the control sequences, the selected stimuli were frequency-scrambled (as in Dormal et al. 2018) to preserve the overall frequency content of the original sounds while disrupting their harmonicity and intelligibility. Specifically, we extracted the sound envelope of each sound using the Hilbert transform and computed a Fast Fourier Transform (FFT) of each sound to shuffle the magnitude and phase of frequency bins within consecutive windows of 200 Hz. Then, we computed the inverse FFT and we applied the original sound envelope to the resultant scrambled waveform. This procedure led to a disruption in harmonicity and intelligibility of the scrambled sounds (as confirmed in a behavior experiment: see Sect. “Experiment 2-Behavior”), while keeping their spectro-temporal structure almost identical to the original/intact sounds (Figure 2.1d, e). The overall energy (RMS) of all scrambled sounds were equalized and ramps of 10 ms were applied same as for the intact sounds.

2.2.1.3 Sequences and Procedure

The affective sounds were placed one after the other in a periodic fashion with an inter-stimulus interval (ISI) of 50 ms to facilitate sound segregation and discrimination. Thus, the general rate of presentation or base rate was 2.5 Hz (i.e., $1/(0.350 + 0.050)$ s). Importantly, each third sound in the sequence belonged to a specific emotion category which was presented periodically at a target frequency of 0.833 Hz ($2.5 \text{ Hz} / 3$), while all sounds of non-target categories were presented non-periodically. The non-target emotion categories were presented once before being repeated. Two emotion conditions were created, where Fear and Happiness were used as the target category, respectively. Fear was chosen due to its salience and evolutionary function (Stanley 1984). Happiness was selected due to its different valence amongst the other chosen emotions with negative valence. For each emotion condition, unique sequences were constructed such that the order of stimuli in each sequence was randomized to increase generalizability. The order of stimuli presentation of each frequency intact sequence was replicated in a corresponding frequency-scrambled sequence. As a result, each intact sequence had an identically ordered control scrambled sequence (Figure 2.1c). With 24 unique targets of 1 emotion category (Fear or Happiness) and 48 unique non-targets across the other 4 emotion categories in any sequence, each sound was presented twice in a sequence (total sounds: 144), leading to a sequence length of 57.6 s. Each sequence included 2 s of fade-in and fade-out where the volume gradually increased from 0 to maximum amplitude of the sounds at the start and vice-versa at the end of the sequence, in order to avoid abrupt movements of the participants at the onset of each sequence which could have introduced artifacts in the recordings.

Therefore, four different conditions were presented: two different sequence conditions (intact sequence with conditions: Fear and Happiness as target) and to control for low-level acoustic confounds (scrambled frequency sequences), namely Fear Intact, Fear Scrambled, Happiness Intact, Happiness Scrambled. Ten instances of each sequence group (total: 40) were created prior to the EEG session and presented to each participant in a pseudo-randomized fashion. The speakers were placed at 1 m distance and all sounds were presented at around 60 dB measured from the ears of the participants. The participants kept their eyes closed during the sequence presentation and were instructed to press a button when they heard a sound lower in volume as compared to the other sounds. The lower volume was implemented by reducing the sounds' root mean square (RMS) value by a factor of 10. This attention target was presented 6 times per sequence in a randomized fashion, excluding the 2 s fade-in and fade-out at the start and end of the sequence. Examples of the sequences are available at: https://github.com/sid0309/freqtag_emotion. The experiment was designed on

MATLAB R2016b (Mathworks) using the Psychtoolbox and extensions (Brainard 1997; Kleiner et al. 2007; Pelli 1997).

2.2.1.4 EEG Acquisition and Analysis

Data was analyzed using Letswave 6 ([https://github.com/ NOCIONS/Letswave6](https://github.com/NOCIONS/Letswave6)) and the Fieldtrip toolbox (Oostenveld et al. 2011) running on MATLAB R2016b (Mathworks) with custom built scripts in MATLAB and on Rstudio (Rstudio Team 2020). Subsequent sections of data preprocessing and the frequency domain analysis followed a pipeline that has been adopted in many frequency-tagging studies in vision (Bottari et al. 2020; Dzhelyova et al. 2017; Retter and Rossion 2016; Rossion et al. 2015, 2020; Volfart et al. 2021) and audition (Barbero et al. 2021).

Data was analyzed using Letswave 6 ([https://github.com/ NOCIONS/Letswave6](https://github.com/NOCIONS/Letswave6)) and the Fieldtrip toolbox (Oostenveld et al. 2011) running on MATLAB R2016b (Mathworks) with custom built scripts in MATLAB and on Rstudio (Rstudio Team 2020). Subsequent sections of data preprocessing and the frequency domain analysis followed a pipeline that has been adopted in many frequency-tagging studies in vision (Bottari et al. 2020; Dzhelyova et al. 2017; Retter and Rossion 2016; Rossion et al. 2015, 2020; Volfart et al. 2021) and audition (Barbero et al. 2021).

2.2.1.4.1 Preprocessing

Raw continuous EEG data was filtered using a fourth order Butterworth band-pass filter from 0.1 to 100 Hz and a notch filter at 50 and 100 Hz with a width of 0.5 Hz to attenuate the power line noise. The data was then down-sampled to 256 Hz to facilitate data handling and storage. Continuous data was then segmented from 2 s before the onset of the sequences to 2 s after the end of the sequence, resulting in a trial length of 61.6 s. Subsequently, the ICA matrix of each segmented trial was computed using RUNICA (Makeig et al. 1995) and the resulting components for each participant were visually inspected. Artefactual frontal components due to facial movements, when present, were deleted with at most one component deleted in 18 out of the 24 participants. Visual inspection was then conducted to detect potential noisy trials and/or electrodes. Trials containing residual artefactual activity were deleted (on average 1.6 trials out of the 40 trials per participants, maximum number of trials deleted was 5 in two participants). Noisy channels were linearly interpolated using the three closest neighboring electrodes (2 channels for 5 subjects—FT8, FC5; T8h, AF7; FPz, AFF2; AF3, P9; PPO5, AFF2 respectively—and 1 channel for 4 other subjects—I1; I1; C6; CPP5h). The trials were subsequently re-referenced to the average reference and divided into two emotion conditions (Fear and Happiness) and its two sequence types (Intact and Scrambled): Fear Intact, Fear Scrambled, Happiness Intact and Happiness Scrambled.

2.2.1.4.2 Frequency Domain Analysis

Considering the presentation rate of the target emotion category and the frequency resolution ($1/\text{duration of the sequence}$), all trials were re-segmented from 2 s after sequence onset (to remove fade-in) to a length of 52.8 s, to contain an integer number of target presentation cycles (0.833 Hz). Trials were then averaged in the time domain separately by emotion condition (Fear and Happiness) and sequence type (Intact and Scrambled) to attenuate the noise and brain responses that were not time-locked to the stimuli (Luck 2014). The resulting averaged trials for the four conditions were then grand-averaged across participants. Consequently, a Fast Fourier Transform (FFT) was applied to the averaged trials. Amplitude spectra extended from 0 to 128 Hz with a frequency resolution of 0.0189 Hz, allowing us to isolate the base response (general presentation rate) at 2.5 Hz and target response at 0.833 Hz, along with their respective harmonics. The harmonic responses in addition to fundamental frequency (in this case 2.5 Hz or 0.83 Hz) can be accounted for in relation to complex responses of the brain, corresponding to the principles of frequency domain analysis of periodic signals (Retter et al. 2021). Since the summation of harmonics in the FFT can indicate the overall responses in single values (Retter et al. 2021; Retter and Rossion 2016), we applied a criterion to select significant harmonics to include in further analysis. Significance of harmonics was determined by first pooling all 128 channels together and computing the z-score at each frequency bin, considering 12 bins at each side of the frequency bin of interest, excluding the immediate adjacent bins and the maximum and minimum of the entire window (Retter and Rossion 2016). For each target emotion condition (Fear and Happiness) and sequence type (Intact and Scrambled), consecutive harmonics that displayed a z-score > 2.32 ($p < 0.01$, 1-tail, signal $>$ noise) were considered significant. While considering significant harmonics for the target emotion conditions, the base frequency bins (i.e., 2.5, 5, 7.5 Hz etc.) were not taken into account. The chosen number of consecutive harmonics for Intact and Scrambled sequences were equalized by considering the highest number of significant consecutive harmonics in any of the two types of sequences, knowing that adding responses at non-significant harmonics is not detrimental for the calculation of the response (i.e., adding zeros, Retter et al., 2018). To quantify the responses at base and target frequencies, we computed the baseline subtracted amplitudes on the grand-averaged FFT of each condition and type, where the window for this calculation was defined similarly to the computation of z-scores (12 bins on either side except maxima, minima and adjacent bins). Baseline subtraction takes into account the different noise profiles across different frequency bands (Luck 2014), especially higher noise at low frequencies (< 1 Hz) in EEG recordings. Consequently, we summed the baseline subtracted amplitude at the selected significant harmonics to finally represent them as topographical maps. For illustration purpose, the signal-to-noise ratio (SNR) was

computed using the same criteria to estimate the noise (12 bins on either side, excluding the immediate bins on either side, the maximum and minimum).

To find the electrode(s) of interest involved in emotion categorization for Fear and Happiness that could be independent of low-level acoustic processes, we subtracted the grand averaged FFT spectra of the frequency-intact and frequency-scrambled sequences (intact—scrambled) for each target emotion condition separately. Then, we extracted chunks of the resultant spectra, where each chunk consisted of 25 frequency bins (central bin of the chosen harmonic and 12 neighboring bins on each side). The number of chunks was defined as the number of chosen harmonics in each emotion condition. The sum of the chunks and then, the z-score were computed at the harmonic (13th bin) for each channel (Volfart et al. 2021). Finally, we performed FDR (Benjamini and Hochberg 1995) correction across 128 channels for multiple comparisons, for each emotion condition. Additionally, to demonstrate the response of each individual and consider inter-individual variability, we repeated the analysis by applying the FFT to the average of all trials of each condition, type and participant. Baseline subtracted amplitudes were calculated at the target frequency and its significant harmonics (that were found as per the grand-averaged analysis for each emotion condition) and consequently summed.

2.2.1.4.3 Time Domain Analysis

We conducted time-locked analysis to characterize the time course of emotion discrimination process. The raw data was filtered through the Butterworth bandpass and the notch filter with the same parameters as in frequency domain analysis. An additional notch filter was applied to the filtered data to remove activity related to the sound presentation rate (at 2.5, 5, 7.5 Hz) which would reflect general auditory processes not linked to emotion discrimination. Stimulus-locked epochs of 800 ms ranging from the onset of one stimulus prior to the target stimulus to the end of the target stimulus were extracted such that each sequence yielded 44 epochs. The first 2 and last 2 epochs were deleted to exclude the fade in and out of the sequences. Noisy epochs with amplitude deflections exceeding $\pm 100 \mu\text{V}$ in any channel were deleted. The resultant epochs were re-referenced to averaged mastoids, then equalized in number across conditions and types separately and averaged for each subject. Baseline correction was implemented by subtracting the signal within 400 ms pre-stimulus activity, corresponding to 1 cycle of base rate, i.e. the prior epoch to the target epochs (Dzhelyova et al. 2017). Finally, the baseline subtracted epochs were grand-averaged across all subjects for each condition and type separately.

To compare the temporal activity of emotion categorization, we conducted a time point-by-time point, onetailed t-test between frequency intact and scrambled sequences across subjects for each electrode and emotion condition. FDR correction was applied across 128 channels to correct for multiple comparisons

(Benjamini and Hochberg 1995). Segments of data were considered significantly different if the two conditions (intact vs scrambled) were different for more than 25 ms, i.e. > 13 consecutive time-points (Chen et al. 2021).

2.2.2 Cochlear Model and Envelope

Cochlear Model and Envelope To ensure that potential differences in the EEG responses elicited by different conditions could not be merely explained by differences in the temporal structure of the target emotion categories of interest, we compared the temporal envelopes of intact and scrambled sequences. We prepared unique sequences for each participant before the EEG session (40 per individual) and extracted the envelope of each sequence using the Hilbert Transform. Then, we computed the FFT of all envelopes and averaged them across each sequence type (intact and scrambled) and condition (Fear and Happiness) for each participant. Further, we chose the harmonics of interest in the FFT based on the empirical EEG data (see Sect. “Frequency Domain Analysis” for the procedure and Sect. “Frequency Domain: Base Response” for EEG results) and summed the FFT magnitudes across the chosen bins for every subject. Finally, a 1-tailed t-test was calculated across all participants to contrast intact from scrambled sequences. Additionally, to assess how spectral and temporal characteristics of sound alone can influence early responses processed in the cochlea, we employed gammatone filters using Auditory Toolbox (Slaney 1994) to simulate a cochlear response to a given sequence. Cochlear simulation was computed on each sequence before the EEG session. The FFT was applied on the simulated cochlear response and each sequence type (intact and scrambled) and condition (Fear and Happiness). Then, we chose the harmonics of interest in the FFT of the cochlear response based on the empirical EEG data (see Sect. “Frequency Domain Analysis” for the procedure and Sect. “Frequency Domain: Base Response” for EEG results) and summed the FFT magnitudes across the identified bins of interest for each subject. Lastly, we compared the summed FFT magnitudes of intact and scrambled sequences across all participants using a 1-tailed t-test (intact > scrambled) for each condition separately.

2.2.3 Behavioral Experiment

We conducted a behavioral experiment to evaluate whether the participants could efficiently identify emotion from a stream of short bursts of affective voices, classify vocalizations to discrete emotion categories and to confirm the unintelligibility of the scrambled sounds.

2.2.3.1 Participants

21 out of 24 participants (11 women, age: 22.28, SD: 2.36, range: 19–28 years) who took part in the EEG experiment participated in the behavioral experiment. This

session was scheduled after the EEG experiment to avoid familiarity with the sounds during the EEG session. All participants reported no history of neurological or audiological disorders and were all right-handed. The experiment was approved by the local ethical committee of Catholic University of Louvain (UCLouvain, project 2016–25). All participants provided written informed consent and received financial compensation for their participation.

2.2.3.2 Procedure

Stimuli from the EEG session were used in the behavioral experiment (see Sect. “Stimuli”). The behavioral session was divided into two tasks: (1) a “Sequence task” where subjects were asked to identify a target emotion category amongst sounds presented in a short sequence and (2) an “Isolation task”, where each sound was presented separately. The sequence task was divided into four blocks, each block consisting of 48 short sequences and each sequence comprising of five emotional sounds. Two of the four blocks required the participants to identify if a Fearful vocalization was present amongst the sounds presented while in the other two blocks, participants were instructed to detect whether a Happy vocalization was present in the sequence. To mimic the structure of the sequences used in EEG experiment, we inserted the target emotion at the third place in each short sequence. Subjects were asked to perform a two-alternative forced-choice task (target emotion present or not) with half of the sequences in each block (24/48) consisting of the target emotion category (Fear or Happiness according to the block). All sequences and blocks were presented in a random order. Later, participants performed the “Isolation task” which was divided into four blocks, each consisting of 84 sounds. In two blocks, subjects were asked to listen to a single affective vocalization and answer three questions- (a) classify the sound to one of the five emotion categories: anger, disgust, fear, happiness or sadness in a five-alternatives forced choice task; (b) rate the valence of each sound on a scale from 1 to 5 (1 = most negative, 3 = neutral, 5= most positive); (c) rate the arousal evoked by the stimuli on a scale from 1 to 5 (1 = not aroused at all, 5 = most aroused). In the other two blocks, the scrambled version of the same sounds were presented and the subjects were asked to classify them to an emotion category in a five-alternatives forced choice task. All sounds and blocks were presented in a randomized order. The experiment was implemented on Psychtoolbox and extensions (Brainard 1997; Kleiner et al. 2007; Pelli 1997) running on MATLAB R2016b (Mathworks). For analysis, sensitivity indices were calculated for each task using D-primes (d') constituting an unbiased quantification of performance in detection tasks considering both hits and false alarms (Hautus et al. 2021; Tanner and Swets 1954).

2.3 Results

2.3.1 Experiment 1- EEG

2.3.1.1 Frequency Domain: Target Response

The target response was quantified as the sum of consecutive significant harmonics, i.e., 5 harmonics for Fear (0.833, 1.666, 3.333, 4.166 and 5.83 Hz) and 4 for Happiness (0.833, 1.666, 3.333 and 4.166 Hz). Presence of a response in each condition and type was assessed by computing the z-scores in all channels. (Fear Intact > 0: 107 channels, maximum at TP8: $z = 10.2913$, $p = 3.8550 \times 10^{-25}$; Fear Scrambled > 0: 77 channels, maximum at PO10: $z = 6.3015$, $p = 1.4731 \times 10^{-10}$; Happiness Intact > 0: 110 channels, maximum at TP8: $z = 10.0758$, $p = 3.5342 \times 10^{-24}$; Happiness Scrambled > 0: 100 channels, maximum at PPO6: $z = 9.7109$, $p = 1.3553 \times 10^{-10}$, all p-values are FDR corrected). Crucially, after pooling all channels in each condition separately, the sum of baseline subtracted amplitudes at the harmonic bins of interest revealed a higher amplitude for intact sequences type in comparison to scrambled for both Fear and Happiness conditions (Fear Intact: mean = 0.0484 μ V, S.D. = 0.0613; Fear Scrambled: mean = 0.0292 μ V, S.D. = 0.0333; Happiness Intact: mean = 0.0561 μ V, S.D. = 0.0516; Happiness Scrambled: mean = 0.0379 μ V, S.D. = 0.0357; topoplots depicted in Figure 2.2a).

To find the channels contributing to the emotion-selective EEG responses at the target frequency and harmonics, we first computed the difference between the grand averaged Fast Fourier spectra of frequency intact and scrambled conditions at each electrode and computed the z-score for the 2 conditions: Fear and Happiness (see methods Sect. "Frequency Domain Analysis"). We found 42 significant channels for Fear selective response in bilateral temporal areas (Figure 2.2c; Fear intact—Fear scrambled > 0: maximum in right temporal FT8h: $z = 5.3219$, $p = 5.1345 \times 10^{-8}$; in left temporal TP7h: $z = 4.7200$, $p = 1.1792 \times 10^{-6}$) and fronto-central electrodes (maximum at FFC3: $z = 5.4294$, $p = 2.8272 \times 10^{-8}$). 16 significant channels for Happiness condition were clustered in right temporal (Happiness intact – Happiness scrambled > 0: maximum at C6h: $z = 4.8022$, $p = 7.8466 \times 10^{-7}$) and left frontal area (maximum at AFF5: $z = 4.3399$, $p = 7.1274 \times 10^{-6}$). All p-values are FDR corrected and other significant channels with their z-scores are tabulated in Table S2.1 in supplementary material.

Chapter 2. Categorization of Auditory Emotion Expressions

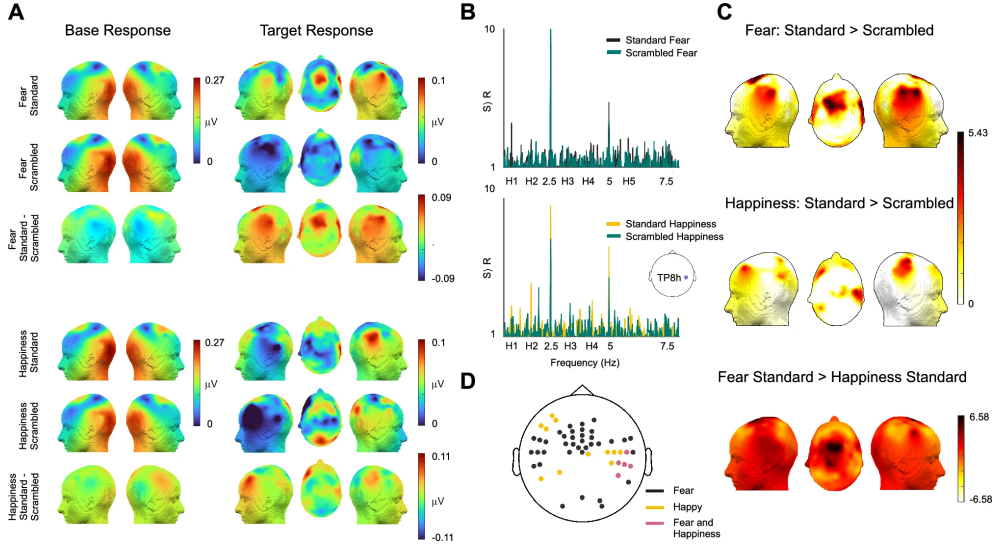


Figure 2. 2 EEG Results - Emotion Selective Response: A Sum of baseline subtracted amplitudes are represented as topographies. The base response appears similar across all conditions and types. The target response to Intact Fear is localized to bilateral temporal and central areas while response to Intact Happiness is elicited in the right temporal and left frontal region. Higher responses observed for intact stimuli sequences than scrambled stimuli sequences. B Signal to noise ratio (SNR) at channel TP8h for both emotion conditions to visualize the response across frequencies. H1 and Hn refer to the first harmonic of target frequency i.e. 0.83 Hz and higher harmonics, respectively. C Scalp regions eliciting significant emotion-selective responses Fear is significant at bilateral temporal and central regions; Happiness at right temporal and left frontal areas. Channels in central and bilateral temporal regions are more selective to Fear than Happiness while no channels are more selective to Happiness. D Distinct, yet overlapping channels selective to Fear and Happiness

We found five significant overlapping channels across Fear and Happiness conditions over right temporal electrodes, that may indicate elicitation from regions common to emotion processing, regardless of the category: P6, TP8h, TP8, CP6 and T8h (Figure 2.2d). For visualization, the signal-to-noise ratio (SNR) at the right temporal channel TP8h (significant in both Fear and Happiness conditions) shows higher SNR at the harmonics of interest for intact in comparison to scrambled (Figure 2.2b). However, we also observed different clusters of significant electrodes across conditions. To investigate further whether some channels were more selective to Fear than Happiness or vice versa, we subtracted the FFT of Happiness intact from Fear intact and computed the z-score at the bins of interest. Since we previously selected different number of harmonics to quantify Fear (5 harmonics) and Happiness (4 harmonics) responses, we equalized the number of harmonics to the maximum (i.e. 5 harmonics) since adding responses at non-significant harmonics is not detrimental for the calculation of the response (i.e., adding zeros, Retter et al., 2018). The two-sided FDR corrected p-values from the z-test revealed 38 significant channels that were more selective to Fear than Happiness (Figure 2c) mostly

clustered at and around the central region (Fear intact—Happiness intact > 0: maximum at FFC1: $z = 6.4061$, $p = 7.4645 \times 10^{-11}$). We did not find any channels significantly more selective to Happiness than Fear.

Furthermore, we also quantified each participant's response to Fear and Happiness by averaging the summed baseline subtracted amplitude across 42 and 16 significant channels, respectively. One-tailed t-tests revealed that the contrast Intact and Scrambled is significantly superior to 0 (Intact—scrambled > 0: Fear: mean = 0.0485 μ V, S.D. = 0.0638, $t(23) = 3.7237$, $p = 5.5720 \times 10^{-4}$, Cohen's $d = 0.7602$; Happiness: mean = 0.0361 μ V, S.D. = 0.0621, $t(23) = 2.8480$, $p = 0.0045$, Cohen's $d = 0.5813$, all p-values are FDR corrected). Individual baseline subtracted amplitudes are plotted in Figure S2.1 in supplementary material.

2.3.1.2 Frequency Domain: Base Response

Base Response 2 consecutive significant harmonics for Fear (2.5, 5 Hz) and 3 for Happiness were identified (2.5, 5, 7.5 Hz). A similar topography was observed visually at the general rate of presentation of sounds across intact and scrambled sequences (Figure 2.2a). However, z-scores revealed differences between intact and scrambled types for both Fear and Happiness emotion conditions (Intact—Scrambled > 0: Fear: 1 channel at FC6: $z = 5.2978$, $p = 5.8611 \times 10^{-8}$; Happiness: 38 channels. maximum at FC6: $z = 6.4441$, $p = 5.8128 \times 10^{-11}$, p-values are FDR corrected).

2.3.1.3 Time Domain

We compared the grand averaged trials of intact and scrambled targets with a time point-by-time point 1-tailed t-test separately for each emotion condition. The analysis revealed two significant windows for the Fear condition (Figure 2.3): between 135 and 176 ms (Intact > Scrambled, peak at Cz at 152 ms: mean = 0.6780 μ V, SD = 0.7978, $t(23) = 4.1631$, $p = 0.0012$, Cohen's $d = 1.0257$, p-values are FDR corrected) and between 311 and 398 ms (Intact > Scrambled, peak at Cz at 339 ms: mean = 0.8361 μ V, SD = 0.8528, $t(23) = 4.6508$, $p = 8.8496 \times 10^{-4}$, Cohen's $d = 0.9345$, p-values are FDR corrected). These differences expressed over central, frontal and parietal electrodes. Happiness intact trials elicited a stronger response than their scrambled version across right temporal channels in only one late time window—between 319 and 356 ms (Intact > Scrambled, peak at CP6 at 337 ms: mean = 0.5764 μ V, SD = 0.5628, $t(23) = 5.0178$, $p = 0.0013$, Cohen's $d = 1.1390$, p-values are FDR corrected) expressing over right temporal channels. Thus, in addition to different topographic representations, we also found different temporal windows of categorization of Fear and Happiness, where responses to Fear occurred earlier than Happiness. Lastly, four channels in the right temporal region: P6, TP8h, CP6 and T8h showed greater responses for intact sequences than scrambled for both

Fear and Happiness conditions, consistent with the results of the frequency domain analysis.

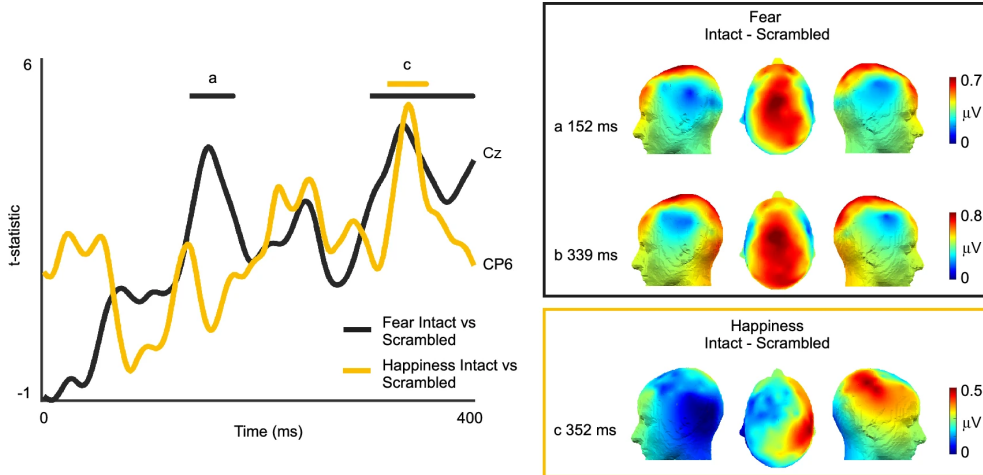


Figure 2. 3 EEG time domain results: t-statistic of Fear intact>scrambled and Happiness intact> scrambled plotted with respect to time. The representation of Fear on contrasting with scrambled sequences, evolves early (peaking at 152 ms) and later (339 ms) across fronto-central-parietal channels; only one significant window for happiness was found peaking at 352 ms in the right temporal-parietal region. The topographies represent the EEG amplitude at the peak time points

2.3.2 Cochlear Model and Envelope

After preparing the sequences of affective sounds for each subject, the FFT spectrum of the envelope of sequences of both emotion conditions (Fear and Happiness) and types (Frequency Intact and Scrambled) were averaged separately. Then, the FFT magnitude at the harmonics were summed (defined from the empirical EEG results: 5 harmonics for Fear, 4 for Happiness) and contrasted with a 1-tailed t-test suggesting similar temporal structure for both type of sequences (Intact > Scrambled: Fear: $t(23) = -0.3718$, $p = 0.6432$, Cohen's $d = -0.1759$; Happiness: $t(23) = -0.7062$, $p = 0.75644$, Cohen's $d = -0.1442$, p-values are FDR corrected across participants; Figure 2.1e). Similarly, we computed a 1-tailed t-test between the summed FFT magnitudes of harmonics of interest of the simulated cochlear response to the two types of sequences (Intact > Scrambled: Fear: $t(23) = -6.9465$, $p = 0.9999$, Cohen's $d = -1.4180$; Happiness: $t(23) = -11.7324$, $p = 0.9999$, Cohen's

$d = -2.3949$) indicating similar processing of acoustic cues at the cochlear level for both type of sequences and each emotion condition (Figure 2.1f).

2.3.3 Experiment 2 - Behavior

We conducted a separate behavioral experiment with the same participants who took part in the EEG experiment to validate: (a) if participants could categorize a short non-verbal affective vocalization played in a sequence (Sequence task), b) the unintelligibility of the scrambled sounds (Isolation task). In the Sequence task, all participants identified the presence of a target emotion amongst a stream of other emotional sounds well above chance, for both Fear (one tailed t-test $d' > 0$: d' values mean = 1.3866, SD = 0.5116, $t(20) = 12.4212$, $p = 7.37 \times 10^{-11}$, Cohen's $d = 2.7103$) and Happiness (one tailed t-test $d' > 0$: d' values mean = 2.2399, SD = 0.8638, $t(20) = 11.882$, $p = 1.61 \times 10^{-10}$, Cohen's $d = 2.5931$; Figure 2.4a).

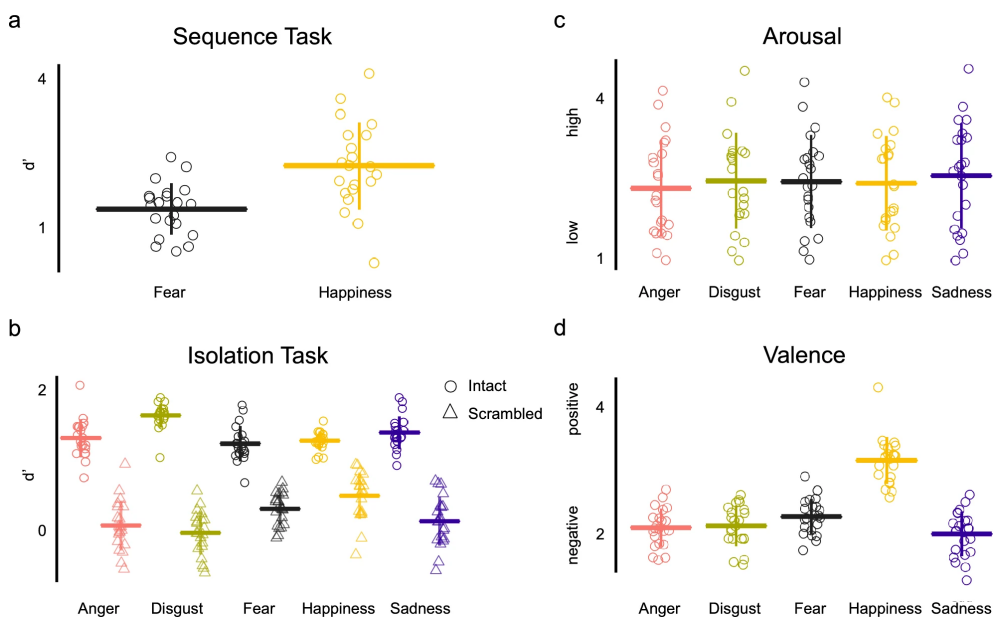


Figure 2. 4 Behavior Results: A Sequence task: Participants could identify the presence of an exemplar of the target emotion category amongst a short sequence of sounds well above chance level. B Isolation task: Intact sounds were categorized to the appropriate emotion category, with worse performance for scrambled sounds, thus verifying the disruption in unintelligibility. C Similar arousal ratings across all emotion categories suggest no involvement of arousal in the EEG responses. D Happiness was the only 'positive' emotion amongst other categories. Greater values indicate positive valence

The following results in the isolation task are also tabulated in Table S2.2 and S2.3 in supplementary material. All intact stimuli were correctly categorized above chance level (Figure 2.4b: one tail t-test $d' > 0$: Anger: d' values mean = 1.3214, SD = 0.2711,

$t(20) = 22.3376$, $p = 1.33 \times 10^{-15}$, Cohen's $d = 4.8742$; Disgust: d' values mean = 1.6475, $SD = 0.1768$, $t(20) = 42.7111$, $p = 0$, Cohen's $d = 9.3184$; Fear: d' values mean = 1.2428, $SD = 0.2511$, $t(20) = 22.6799$, $p = 9.99 \times 10^{-16}$, Cohen's $d = 4.9494$; Happiness: d' values mean = 1.2828, $SD = 0.1309$, $t(20) = 44.9205$, $p = 0$, Cohen's $d = 9.7998$; Sadness: d' values mean = 1.4006, $SD = 0.2331$, $t(20) = 44.9205$, $p = 0$, Cohen's $d = 6.0086$, all p -values are FDR corrected). For scrambled stimuli, the categorization of three categories: Anger, Disgust and Sadness vocalizations were at chance level (one tail t -test $d' > 0$: Anger: d' values mean = 0.0665, $SD = 0.3587$, $t(20) = 0.8501$, $p = 0.4053$, Cohen's $d = 0.1854$; Disgust: d' values mean = -0.0387, $SD = 0.3132$, $t(20) = -0.5670$, $p = 0.577$, Cohen's $d = -0.1236$; Sadness: d' values mean = 0.1284, $SD = 0.3523$, $t(20) = 1.6704$, $p = 0.1104$, Cohen's $d = 0.3645$, all p -values are FDR corrected). However, d' of scrambled versions of Fear and Happiness were found to be significantly higher than chance (one tail t -test $d' > 0$: Fear: d' values mean = 0.3071, $SD = 0.2332$, $t(20) = 6.0349$, $p = 6.71 \times 10^{-6}$, Cohen's $d = 1.3169$; Happiness: d' values mean = 0.4956, $SD = 0.3341$, $t(20) = 6.7958$, $p = 1.31 \times 10^{-6}$, Cohen's $d = 1.4834$, all p -values are FDR corrected). Two explanations could be given for this result. First, there were a higher number of Fear and Happiness stimuli than the stimuli of other emotion categories, since we chose the number of stimuli based on equal presentation of each sound in a sequence (see methods Sect. "Sequences and Procedure"). Therefore, higher d' values for Happy and Fearful vocalization could indicate an effect of training. Second, it is possible that the scrambled sounds of Fear and Happiness categories carried some amount of acoustic information which made them slightly recognizable. However, contrasting d' of intact sounds with their frequency scrambled versions indicated that the intact vocalizations carried significantly more information than their scrambled version (two-tailed t -test Intact > Scrambled: Fear: $SD = 0.2996$, $t(20) = 14.306$, $p = 5.74 \times 10^{-12}$, Cohen's $d = 3.12$; Happiness: $SD = 0.3841$, $t(20) = 9.3916$, $p = 8.99 \times 10^{-9}$, Cohen's $d = 2.04$, all p -values are FDR corrected).

We further analyzed the arousal and valence ratings of the intact sounds. We observed similar arousal levels across all the discrete emotion categories (ANOVA $F = 0.15$, $p = 0.9612$, $\eta^2 = 0.5848$; Figure 2.4c) and an effect of emotion on the valence ratings (ANOVA $F = 28.15$, $p = 6.8228 \times 10^{-16}$, $\eta^2 = 17.7659$; Figure 2.4d). As expected, post hoc t -tests revealed that valence ratings for Happiness were significantly different than other categories (Happiness vs Anger: $p = 2.98 \times 10^{-8}$; Happiness vs Disgust: $p = 1.98 \times 10^{-8}$; Happiness vs Fear: $p = 1.81 \times 10^{-8}$; Happiness vs Sadness: $p = 5.96 \times 10^{-8}$, all p -values are FDR corrected).

2.4 Discussion

In our study, we demonstrated how combining electroencephalographic recordings (EEG) in humans with a frequency-tagging paradigm provides a robust and objective

measure of a categorical brain response to short bursts of affective vocalizations. To disentangle the potential contributions of low-level auditory processing of affective vocalizations to representations of a higher order emotion categorization process, we implemented a careful selection of affective vocalizations to match their spectro-temporal properties across emotion categories. In particular, we first selected stimuli with comparable spectral features: spectral center of gravity, harmonicity-to-noise ratio (HNR) and pitch, such that no emotion category was spectrally different than the others (Figure 2.1a). These sounds could be accurately classified to a discrete category of emotion: anger, disgust, fear, happiness and sadness, validated with a behavioral experiment (Figure 2.1b). In addition to the periodic sequence created with intact sounds, we introduced a sound sequence with identical periodic constraints like the intact sequences but with frequency-scrambled sounds such that their spectrotemporal structure is similar but their intelligibility and harmonicity are disrupted (Barbero et al. 2021; Dormal et al. 2018; Figure 2.1c,d,e). It is important to note that the difference in EEG responses between intact and scrambled sequences was unlikely to solely rely on differences in the harmonicity of our sounds since the intact sounds had comparable HNR across categories that were all presented in the same intact sequence (Figure 2.1a). To ensure that emotion-selective EEG responses could not be explained by different temporal envelopes across emotion categories, we compared the FFT magnitude of the sequences' envelopes of frequency-intact and frequency-scrambled sequences to find no significant difference at the frequency bins of interest (Figure 2.1e). As a final validation, we also used Gammatone filters to simulate the cochlear response to intact and scrambled sequences and showed no differences at frequency bins of interest, at the level of early peripheral auditory processing (Figure 2.1f). Thus, any difference in the EEG response to intact versus scrambled sequence cannot be simply explained by a cochlear simulation of acoustic response, nor by their spectral or envelope structure. Altogether, our stimuli selection procedure and control analyses assert that the observed categorical response to specific emotion expression categories is at least partially independent from low-level acoustic features and, therefore, likely reflects a higher-level categorization process.

We relied on a frequency-tagging technique to isolate the responses to emotion vocalizations objectively, with a high signal-to-noise ratio and automatically at specific known frequencies without having the participants to overtly respond to the sounds, thus avoiding the decisional and attentional processes to contaminate the EEG response (Levy et al. 2003). Unlike other ERP studies where emotion responses are compared to responses of neutral vocalizations causing confounds due to differences in acoustic features and in terms of arousal and valence, the method presented here provides a direct approach to obtain vocal emotion responses to discrete categories by including various heterogeneous exemplars of

each category. Although the frequency of presentation of sounds of the target emotion category (0.83 Hz) lies at the lower end of the frequency EEG spectrum which is susceptible to noise (Luck 2014), we acknowledged the trade-off between the frequency of presentation of target emotion category and presenting stimuli of sufficient length to allow correct emotion categorization (Falagiarda and Collignon 2019).

We observed a higher emotion-selective response to the intact sequences than the scrambled version at the target frequency and its harmonics (Figure 2.2a, b). This could only occur if the brain was able to discriminate the target emotion from other emotion categories as well as generalize all sounds of the target emotion to one common category (Barbero et al. 2021). Nevertheless, the target response for the scrambled condition alone was significantly greater than 0. Such significant response in the scrambled condition could either be explained by some low-level properties of the sounds of an emotional category (as seen in the envelope or cochlear simulation) or that the response arises from some residual intelligibility that was retained, as indicated by the behavior experiment showing that the categorization of scrambled Fear and Happiness vocalizations were above chance level, but significantly lower than the categorization accuracy of the intact sounds (Figure 2.4b). Thus, it can be concluded that the significantly stronger responses to intact affective vocalizations when compared to scrambled sequences (Figure 2.2c, d), in conjunction with controlling of low-level features of the sounds indicate that the emotionselective EEG response is at least partially independent of low-level acoustic features and characterize a higher-order categorization process. We also observed a robust response and a similar scalp topography at the general rate of presentation of all stimuli for both intact and scrambled sequences. However, we found a higher response to intact sequences than scrambled that could be accounted for by the overlap of the first harmonic of the general rate of presentation and third harmonic of the target response. Additionally, the higher response can be explained by a perceptual advantage of the intact sounds that may trigger enhanced attention to the stimulation.

We tested two types of sequences with either Fear or Happiness as the target emotion category repeated periodically at every third position (unknown to the participant). They expressed different topographies with part of their response spanning across few common right temporal channels suggesting the involvement of distinct yet overlapping neural substrates in emotion processing (Hamann 2012; Johnstone et al. 2006; Mauchand and Zhang 2022; Phan et al. 2002; Phillips et al. 1998). The distribution of the response to the Fear category suggests the involvement of fronto-central and temporal areas, in line with neuroimaging studies indicating large scale networks involved in the processing of fearful expressions (Kober et al. 2008; Zhou et al. 2021). Further, the different scalp topographies found

for Fear and Happiness suggest different generators might be involved in the processing of the two different emotion expressions. In fact, previous research in visual emotion discrimination using Fast Periodic Visual Stimulation (FPVS) found a differential topography of the responses to distinct facial emotion expressions (Dzhelyova et al. 2017; Leleu et al. 2018; Poncet et al. 2019). Further studies are required to validate the claim by exploring the EEG responses to other emotion categories such as anger, disgust and sadness. Although the arousal ratings of all discrete emotions were similar in our study, a possible reason for different topographies may be due to the difference in valence between Fear and Happiness (Figure 2.4d; Happiness has a positive valence). However, valence does not seem to be a strong driver since we observed a strong target response when Fear is the target category and Fear was not deviant in terms of valence when compared to the other emotion categories presented in the sequence.

We also investigated the time-course of the vocal categorization process. Both emotions elicited a significant response 300 ms post stimulus presentation (Intact > Scrambled), in line with electrophysiological studies that observed the late positive component (LPP) to emotional utterances (Frühholz et al. 2011; Jessen and Kotz 2011). Additionally, we observed the categorical response to Fear as early as 135–175 ms post stimulus presentation. While previous literature has reported P200 time window to be modulated by emotional interjections and non-verbal vocalizations (Charest et al. 2009; Jessen and Kotz 2011; Sauter and Eimer 2010; Schirmer et al. 2013), it is debated whether these modulations are linked to the categorical nature of emotion expressions or to differences in arousal (Paulmann et al. 2013). But due to comparable affective dimensions (valence and arousal) of Fear with other categories in our stimuli sequence (Figure 2.4c, d), we speculate that the early ERP evoked by various vocalizations depicting Fear could be an early marker of its categorization, supporting recent evidence of the brain's ability to represent discrete categories as early as < 200 ms (Giordano et al. 2021). In fact, faster responses to fear may be possible due to the potential contribution of subcortical pathways (Pessoa and Adolphs 2010) and primitive circuits recognizing danger or fear which are preserved in mammals and in humans (LeDoux 2012). Further, the absence of differences between intact and scrambled sequences before 100 ms supports the idea that the EEG responses to both emotion categories are not elicited by low-level acoustic properties of the sounds, which are well known to modulate early ERP components such as the N100 (Näätänen and Picton 1987). Another interesting observation found in both frequency and time domain analysis (post 300 ms), was that the responses from four channels in the right temporal region (P6, TP8h, CP6 and T8h) were found to be significantly greater for intact sequences than scrambled for both Fear and Happiness conditions, putatively suggesting common processes for vocal emotion processing in the right temporal regions.

There are well known differences between the spectral and temporal properties between fear, happiness and other emotion categories (see Juslin and Laukka 2003; Scherer 2003). These differences are integral in disassociating one emotion category from another. For instance, studies that used morphing of affective vocalizations of various categories, hence impacting the acoustic cues of the sounds, manipulated perceived emotions (Giordano et al. 2021) or found lower reaction time to classify a morphed vocalization (Whiting et al. 2020). The debate does not lie at how these acoustic differences express over sensory regions, but at the spatio-temporal representation of categorical discrimination of emotion expressions in the human brain, i.e. the higher-order representation of emotions. However, the mental states that support mechanisms of emotion processing can be abstract and highly dimensional, rendering it difficult to disentangle which part of the processes linked to emotion discrimination are better conceptualized as categorical or dimensional as well as, which of these processes are driven by acoustic cues (see Giordano et al. 2021; Hamann 2012; Kragel and LaBar 2016; Skerry and Saxe 2015). Thus, to add to this knowledge gap, we propose a method that demonstrates categorical responses from the human brain to different discrete vocal emotions that are at least partially independent from processing of acoustic features. In the future, frequency tagging can also be used to investigate dimensional aspects of emotions such as valence, arousal, intensity etc. to unravel the interplay between categorical and felt perception of emotions and investigating the contribution of acoustic features to both processes. This makes frequency-tagging a valuable technique to study emotion categorization, suitable to use in populations that are more difficult to test (e.g., individuals with autism, infants etc.) with traditional paradigms such as ERP design studies and other neuroimaging techniques.

2.5 Supplementary Information

Table S2. 1: Z-scores for all significant channels found as per the results section 2.3.

Channel	z-score	Channel	z-score	Channel	z-score
Fear Intact > Scrambled FDR corrected					
Cz	2.4133	B30	2.6852	D1	3.5852
I1	2.2329	FC4h	2.7230	FCC1h	5.1274
Oz	2.8293	FCC2h	2.3914	FFC3h	5.0419

I2	2.5596	FCC2	4.5105	FFC3	5.4294
POI2	2.3645	FFC2	3.7994	FT7	2.7384
P10	3.3620	F2	3.0374	FT7h	3.1347
P6	2.2327	AFF4h	3.1518	FC5	2.7468
TP8	4.2504	AFFz	3.4333	FC3h	4.4714
TP8h	5.1871	Fz	2.9558	C1	3.3253
CP6	3.1889	FFCz	3.3953	C5	2.8902
T8h	4.4285	FCz	4.7923	T7h	2.4928
T8	3.4197	FFC1	4.7615	T7	3.1867
FT8	4.8867	F1	4.3491	TP7	3.9254
FT8h	5.3219	AFF1	3.5015	TP7h	4.7200
Happiness Intact > Scrambled FDR corrected					
CPP3	3.0811	CP6h	2.7744	AFF5	4.3399
C2h	2.9064	C4	3.3134	F5	2.6541
P6	2.7621	C6h	4.8022	F7	3.0511
TP8	3.3129	C6	2.9586	P7	3.5407

Chapter 2. Categorization of Auditory Emotion Expressions

TP8h	2.9090	T8h	3.6893		
CP6	3.6049	AF7	3.0840		
Fear Intact > Happiness Intact FDR corrected					
Cz	3.6662	FCC2	4.9567	FC5	2.8393
CCPz	3.3739	FFC2	6.5864	FC3	4.4699
CPz	3.4205	F2	3.5489	FC3h	3.2799
I1	2.4174	Fz	2.5007	C1	3.8339
I2	3.5345	FFCz	5.6050	C1h	2.9282
POI2	2.9950	FCz	6.5102	CCP1h	6.3831
O2	2.7675	FFC1	6.4061	C3h	2.9730
P2	2.2944	F1	5.0717	C3	2.2431
C2h	3.1041	FCC1h	4.1906	C5	2.8818
C4	2.4353	FCC1	4.8892	T7	3.4979
FT8	2.8416	FFC3h	3.2197	TP7	3.0771
FT8h	4.7199	FFC3	3.0675	P9	3.6068
FCC2h	3.9286	FT7	2.5508		

Table S2. 2

Isolation task : Statistics of scrambled sounds					
Emotion	Mean of d'	S.D. of d'	t value	p value	Cohen's d
Anger	0.0665	0.3587	0.8501	0.4053	0.1854
Disgust	-0.0387	0.3132	-0.5670	0.577	-0.1236
Fear	0.3071	0.2332	6.0349	6.71×10^{-6}	1.3169
Happiness	0.4956	0.3341	6.7958	1.31×10^{-6}	1.4834
Sadness	0.1284	0.3523	1.6704	0.1104	0.3645

Table S2. 3

Isolation task : Statistics of intact sounds					
Emotion	Mean of d'	S.D. of d'	t value	p value	Cohen's d
Anger	1.3214	0.2711	22.3376	1.33×10^{-15}	4.8742
Disgust	1.6475	0.1768	42.7111	0	9.3184
Fear	1.2428	0.2511	22.6799	9.99×10^{-16}	4.9494
Happiness	1.2828	0.1309	44.9205	0	9.7998
Sadness	1.4006	0.2331	44.9205	0	6.0086

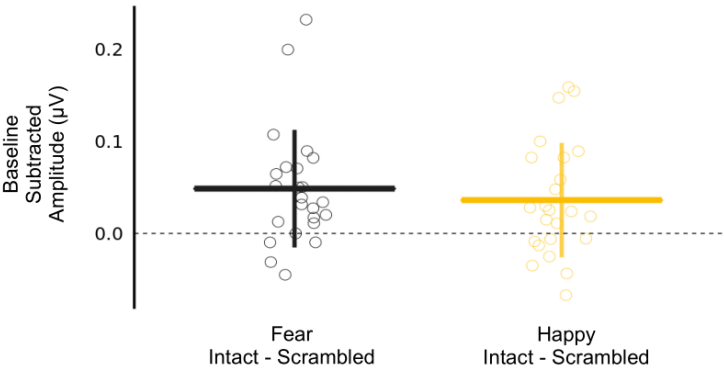


Figure S2. 1 Individual baseline subtracted amplitude

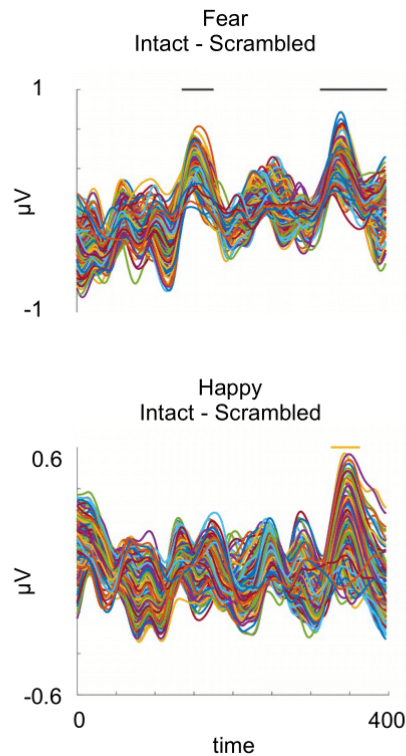


Figure S2. 2 Contrasting ERP between intact and scrambled conditions

Acknowledgements

We would like to extend our gratitude to Anna Krzyżak for her help in EEG data acquisition.

2.6 Bibliography

Banse R, Scherer KR (1996) Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol* 70(3):614

Barbero FM, Calce RP, Talwar S, Rossion B, Collignon O (2021) Fast periodic auditory stimulation reveals a robust categorical response to voices in the human brain. *eNeuro*. <https://doi.org/10.1523/ENEURO.0471-20.2021>

Belin P, Fillion-Bilodeau S, Gosselin F (2008) The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behav Res Methods* 40(2):531–539. <https://doi.org/10.3758/BRM.40.2.531>

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc: Ser B (methodol)* 57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Boersma P, Weenink D (2001) PRAAT, a system for doing phonetics by computer. *Glott International* 5:341–345

Bostanov V, Kotchoubey B (2004) Recognition of affective prosody: continuous wavelet measures of event-related brain potentials to emotional exclamations. *Psychophysiology* 41(2):259–268 <https://doi.org/10.1111/j.1469-8986.2003.00142.x>

Bottari D, Bednaya E, Dormal G, Villwock A, Dzhelyova M, Grin K, Pietrini P, Ricciardi E, Rossion B, Röder B (2020) EEG frequency-tagging demonstrates increased left hemispheric involvement and crossmodal plasticity for face processing in congenitally deaf signers. *NeuroImage* 223:117315. <https://doi.org/10.1016/j.neuroimage.2020.117315>

Brainard DH (1997) The psychophysics toolbox. *Spat vis* 10(4):433–436. <https://doi.org/10.1163/156856897X00357>

Calder AJ, Lawrence AD, Young AW (2001) Neuropsychology of fear and loathing. *Nat Rev Neurosci*. <https://doi.org/10.1038/35072584>

Charest I, Pernet CR, Rousselet GA, Quiñones I, Latinus M, Fillion-Bilodeau S, Chartrand J-P, Belin P (2009) Electrophysiological evidence for an early processing of human voices. *BMC Neurosci*. <https://doi.org/10.1186/1471-2202-10-12867> *Brain Topography* (2023) 36:854–869 13

Chen L, Cichy RM, Kaiser D (2021) Semantic scene-object consistency modulates N300/400 EEG components, but does not automatically facilitate object representations. *Cerebral Cortex*. <https://doi.org/10.1093/cercor/bhab433>

Cornew L, Carver L, Love T (2010) There's more to emotion than meets the eye: a processing bias for neutral content in the domain of emotional prosody. *Cogn Emot* 24(7):1133–1152. <https://doi.org/10.1080/02699930903247492>

Cowen AS, Keltner D (2017) Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc Natl Acad Sci* 114(38):E7900–E7909. <https://doi.org/10.1073/pnas.1702247114>

Darwin C (1872) The expression of the emotions in man and animals. John Murray. <https://doi.org/10.1037/10001-000>

Dormal G, Pelland M, Rezk M, Yakobov E, Lepore F, Collignon O (2018) Functional preference for object sounds and voices in the brain of early blind and sighted individuals. *J Cogn Neurosci* 30(1):86–106. https://doi.org/10.1162/jocn_a_01186

Dzhelyova M, Jacques C, Rossion B (2017) At a single glance: fast periodic visual stimulation uncovers the spatio-temporal dynamics of brief facial expression changes in the human brain. *Cereb Cortex* 27(8):4106–4123. <https://doi.org/10.1093/cercor/bhw223>

- Ekman P (1993) Facial expression and emotion. *Am Psychol* 48(4):384–392. <https://doi.org/10.1037/0003-066X.48.4.384>
- Ekman P (2009) Darwin's contributions to our understanding of emotional expressions. *Philoso Trans R Soc B: Biol Sci* 364(1535):3449–3451. <https://doi.org/10.1098/rstb.2009.0189>
- Ekman P, Friesen WV (1978) Facial action coding system: manual. Consulting Psychologists Press, Palo Alto
- Elfenbein HA, Ambady N (2002) On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychol Bull* 128(2):203–235. <https://doi.org/10.1037/0033-2909.128.2.203>
- Ethofer T, Van De Ville D, Scherer K, Vuilleumier P (2009) Decoding of Emotional Information in Voice-Sensitive Cortices. *Curr Biol* 19(12):1028–1033. <https://doi.org/10.1016/j.cub.2009.04.054>
- Ethofer T, Bartscherer J, Gschwind M, Kreifelts B, Wildgruber D, Vuilleumier P (2012) Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. *Cereb Cortex* 22(1):191–200. <https://doi.org/10.1093/cercor/bhr113>
- Etkin A, Egner T, Kalisch R (2011) Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends Cogn Sci* 15(2):8593. <https://doi.org/10.1016/j.tics.2010.11.004>
- Falgaudiard F, Collignon O (2019) Time-resolved discrimination of audio-visual emotion expressions. *Cortex* 119:184–194. <https://doi.org/10.1016/j.cortex.2019.04.017>
- Fecteau S, Belin P, Joanette Y, Armony JL (2007) Amygdala responses to nonlinguistic emotional vocalizations. *Neuroimage* 36(2):480–487. <https://doi.org/10.1016/j.neuroimage.2007.02.043>
- Flom R, Bahrick LE (2007) The development of infant discrimination of affect in multimodal and unimodal stimulation: the role of intersensory redundancy. *Dev Psychol* 43(1):238–252. <https://doi.org/10.1037/0012-1649.43.1.238>
- Frühholz S, Grandjean D (2013a) Multiple subregions in superior temporal cortex are differentially sensitive to vocal expressions: a quantitative meta-analysis. *Neurosci Biobehav Rev* 37(1):24–35. <https://doi.org/10.1016/j.neubiorev.2012.11.002>
- Frühholz S, Grandjean D (2013) Processing of emotional vocalizations in bilateral inferior frontal cortex. *Neurosci Biobehav Rev* 37:2847–2855. <https://doi.org/10.1016/j.neubiorev.2013.10.007>
- Frühholz S, Jellinghaus A, Herrmann M (2011) Time course of implicit processing and explicit processing of emotional faces and emotional words. *Biol Psychol* 87(2):265–274. <https://doi.org/10.1016/j.biopsycho.2011.03.008>
- Giordano BL, Whiting C, Kriegeskorte N, Kotz SA, Gross J, Belin P (2021) The Representational dynamics of perceived voice emotions evolve from categories

to dimensions. *Nat Hum Behav* 5(9):1203–1213. <https://doi.org/10.1038/s41562-021-01073-0>

Gomez MI, Hwang S-A, Sobotova L, Stark AD, May JJ (2001) A comparison of self-reported hearing loss and audiometry in a cohort of New York farmers. *J Speech Lang Hear Res* 44(6):1201–1208. [https://doi.org/10.1044/1092-4388\(2001/093\)](https://doi.org/10.1044/1092-4388(2001/093))

Grandjean D, Sander D, Pourtois G, Schwartz S, Seghier ML, Scherer KR, Vuilleumier P (2005) The voices of wrath: brain responses to angry prosody in meaningless speech. *Nat Neurosci* 8(2):145–146. <https://doi.org/10.1038/nn1392>

Hamann S (2012) Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends Cogn Sci* 16(9):458466. <https://doi.org/10.1016/j.tics.2012.07.006>

Hannula S, Bloigu R, Majamaa K, Sorri M, Mäki-Torkko E (2011) Self-reported hearing problems among older adults: prevalence and comparison to measured hearing impairment. *J Am Acad Audiol* 22(8):550–559. <https://doi.org/10.3766/jaaa.22.8.7>

Hautus MJ, Macmillan NA, Creelman CD (2021) *Detection theory: A user's guide*, 3rd edn. Routledge, New York

Izard CE, Huebner RR, Risser D, Dougherty LM (1980) The young infant's ability to produce discrete emotion expressions. *Dev Psychol* 16(2):132–140. <https://doi.org/10.1037/0012-1649.16.2.132>

Jessen S, Kotz SA (2011) The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *Neuroimage* 58(2):665–674. <https://doi.org/10.1016/j.neuroimage.2011.06.035>

Jessen S, Obleser J, Kotz SA (2012) How bodies and voices interact in early emotion perception. *PIOS one* 7(4):e36070. <https://doi.org/10.1371/journal.pone.0036070>

Johnstone T, Van Reekum C, Oakes T, Davidson R (2006) The voice of emotion: An fMRI study of neural responses to angry and happy vocal expressions. *Soc Cogn Affect Neurosci* 1:242–249. <https://doi.org/10.1093/scan/nsi027>

Juslin PN, Laukka P (2003) Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol Bull* 129(5):770–814. <https://doi.org/10.1037/0033-2909.129.5.770>

Kiely KM, Gopinath B, Mitchell P, Browning CJ, Anstey KJ (2012) Evaluating a dichotomized measure of self-reported hearing loss against gold standard audiometry: prevalence estimates and age bias in a pooled national data set. *J Aging Health* 24(3):439–458. <https://doi.org/10.1177/0898264311425088>

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C (2007) What's new in psychtoolbox-3. *Perception* 36(14):1–16

Kober H, Barrett LF, Joseph J, Bliss-Moreau E, Lindquist K, Wager TD (2008) Functional grouping and cortical-subcortical interactions in emotion: a meta-

- analysis of neuroimaging studies. *Neuroimage* 42(2):998–1031. <https://doi.org/10.1016/j.neuroimage.2008.03.059>
- Kotz SA, Kalberlah C, Bahlmann J, Friederici AD, Haynes J-D (2013) Predicting vocal emotion expressions from the human brain. *Hum Brain Mapp* 34(8):1971–1981. <https://doi.org/10.1002/hbm.22041>
- Kragel PA, LaBar KS (2016) Decoding the nature of emotion in the brain. *Trends Cogn Sci* 20(6):444–455. <https://doi.org/10.1016/j.tics.2016.03.011>
- LeDoux J (2012) Rethinking the emotional brain. *Neuron* 73(4):653–676. <https://doi.org/10.1016/j.neuron.2012.02.004>
- Leleu A, Dzhelyova M, Rossion B, Brochard R, Durand K, Schaal B, Baudouin J-Y (2018) Tuning functions for automatic detection of brief changes of facial expression in the human brain. *Neuroimage* 179:235–251. <https://doi.org/10.1016/j.neuroimage.2018.06.048>
- Levy DA, Granot R, Bentin S (2003) Neural sensitivity to human voices: ERP evidence of task and attentional influences. *Psychophysiology* 40(2):291–305. <https://doi.org/10.1111/1469-8986.00031>
- Luck SJ (2014) *An Introduction to the event-related potential technique*, 2nd edn. A Bradford Book, Denver
- Makeig, S., Bell, A., Jung, T.-P., & Sejnowski, T. J. (1995). Independent Component Analysis of Electroencephalographic Data. *Advances in Neural Information Processing Systems*, 8. <https://proceedings.neurips.cc/paper/1995/hash/754dda4b1ba34c6fa89716b85d68532b-Abstract.html>
- Mauchand M, Zhang S (2022) Disentangling emotional signals in the brain: an ALE meta-analysis of vocal affect perception. *Cogn Affect Behav Neurosci*. <https://doi.org/10.3758/s13415-022-01030-y>
- Mehler J, Bertoncini J, Barriere M, Jassik-Gerschenfeld D (1978) Infant recognition of mother's voice. *Perception* 7(5):491–497. <https://doi.org/10.1068/p070491>
- Näätänen R, Picton T (1987) The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24(4):375–425. <https://doi.org/10.1111/j.1469-8986.1987.tb00311.x>
- Norcia AM, Appelbaum LG, Ales JM, Cottareau BR, Rossion B (2015) The steady-state visual evoked potential in vision research: a review. *J vis* 15(6):4. <https://doi.org/10.1167/15.6.4>
- Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011:156869. <https://doi.org/10.1155/2011/156869>

- Paulmann S, Bleichner M, Kotz S (2013) Valence, arousal, and task effects in emotional prosody processing. *Front Psychol.* <https://doi.org/10.3389/fpsyg.2013.00345>
- Pell MD, Kotz SA (2011) On the time course of vocal emotion recognition. *PLoS one* 6(11):e27256. <https://doi.org/10.1371/journal.pone.0027256>
- Pell MD, Rothermich K, Liu P, Paulmann S, Sethi S, Rigoulot S (2015) Preferential decoding of emotion from human nonlinguistic vocalizations versus speech prosody. *Biol Psychol* 111:14–25. <https://doi.org/10.1016/j.biopsycho.2015.08.008>
- Pelli DG (1997) The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spat vis* 10(4):437–442. <https://doi.org/10.1163/156856897X00366>
- Pessoa L, Adolphs R (2010) Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nat Rev Neurosci* 11(11):773–783. <https://doi.org/10.1038/nrn2920>
- Phan KL, Wager T, Taylor SF, Liberzon I (2002) Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage* 16(2):331–348. <https://doi.org/10.1006/nimg.2002.1087>
- Phillips ML, Young AW, Scott SK, Calder AJ, Andrew C, Giampietro V, Williams SC, Bullmore ET, Brammer M, Gray JA (1998) Neural responses to facial and vocal expressions of fear and disgust. *Proc R Soc B: Biol Sci* 265(1408):1809–1817
- Poncet F, Baudouin J-Y, Dzhelyova MP, Rossion B, Leleu A (2019) Rapid and automatic discrimination between facial expressions in the human brain. *Neuropsychologia* 129:47–55. <https://doi.org/10.1016/j.neuropsychologia.2019.03.006>
- Poncet F, Leleu A, Rekow D, Damon F, Dzhelyova M, Schaal B, Durand K, Faivre L, Rossion B, Baudouin J-Y (2022) A neural marker of rapid discrimination of facial expression in 3.5- and 7-month-old infants. *Front Neurosci.* <https://doi.org/10.3389/fnins.2022.901013>
- Regan D (1989) Human brain electrophysiology: Evoked potentials and evoked magnetic fields in science and medicine. Elsevier, Amsterdam Retter TL, Rossion B (2016) Uncovering the neural magnitude and spatio-temporal dynamics of natural image categorization in a fast visual stream. *Neuropsychologia* 91:9–28. <https://doi.org/10.1016/j.neuropsychologia.2016.07.028>
- Retter TL, Jiang F, Webster M, Rossion B (2018) Dissociable effects of inter-stimulus interval and presentation duration on rapid face categorization. *Vis Res* 145:11–20. <https://doi.org/10.1016/j.visres.2018.02.009>
- Retter TL, Rossion B, Schiltz C (2021) Harmonic amplitude summation for frequency-tagging analysis. *J Cogn Neurosci* 33(11):122. https://doi.org/10.1162/jocn_a_01763

Rossion B, Torfs K, Jacques C, Liu-Shuang J (2015) Fast periodic presentation of natural images reveals a robust face-selective electrophysiological response in the human brain. *J Vis.* [https:// doi.org/10.1167/15.1.18](https://doi.org/10.1167/15.1.18)

Rossion B, Retter TL, Liu-Shuang J (2020) Understanding human individuation of unfamiliar faces with oddball fast periodic visual stimulation and electroencephalography. *Eur J Neurosci* 52(10):4283–4344. <https://doi.org/10.1111/ejn.14865>

RStudio Team (2020) RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. <http://www.rstudio.com>

Salvia E, Bestelmeyer PEG, Kotz SA, Rousselet GA, Pernet CR, Gross J, Belin P (2014) Single-subject analyses of magnetoencephalographic evoked responses to the acoustic properties of affective non-verbal vocalizations. *Front Neurosci.* <https://doi.org/10.3389/fnins.2014.00422>

Sauter DA, Eimer M (2010) Rapid detection of emotion from human vocalizations. *J Cogn Neurosci* 22(3):474–481. <https://doi.org/10.1162/jocn.2009.21215>

Sauter DA, Eisner F, Ekman P, Scott SK (2010) Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc Natl Acad Sci USA* 107(6):2408–2412. <https://doi.org/10.1073/pnas.0908239106>

Scherer KR (2003) Vocal communication of emotion: a review of research paradigms. *Speech Commun* 40:227–256. [https://doi.org/10.1016/S0167-6393\(02\)00084-5](https://doi.org/10.1016/S0167-6393(02)00084-5)

Scherer KR, Banse R, Wallbott HG (2001) Emotion inferences from vocal expression correlate across languages and cultures. *J Cross Cult Psychol* 32(1):76–92

Schirmer A, Kotz SA (2006) Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends Cogn Sci* 10(1):24–30. <https://doi.org/10.1016/j.tics.2005.11.009>

Schirmer A, Chen C-B, Ching A, Tan L, Hong RY (2013) Vocal emotions influence verbal memory: neural correlates and interindividual differences. *Cogn Affect Behav Neurosci* 13(1):8093. <https://doi.org/10.3758/s13415-012-0132-8>

Skerry AE, Saxe R (2015) Neural representations of emotion are organized around abstract event features. *Curr Biol* 25(15):1945–1954. <https://doi.org/10.1016/j.cub.2015.06.009>

Slaney M (1994) Technical Report #1998–010 Interval Research Corporation malcolm@interval.com. 52. Stanley HM (1984) A study of fear as primitive emotion *Psychol Rev* 1(3):241–256. <https://doi.org/10.1037/h0066176>

Talavage TM, Sereno MI, Melcher JR, Ledden PJ, Rosen BR, Dale AM (2004) Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity. *J Neurophysiol* 91(3):1282–1296. <https://doi.org/10.1152/jn.01125.2002>

Chapter 2. Categorization of Auditory Emotion Expressions

Tanner WP Jr, Swets JA (1954) A decision-making theory of visual detection. *Psychol Rev* 61(6):401–409. <https://doi.org/10.1037/h0058700>

Volfart A, Rice GE, Lambon Ralph MA, Rossion B (2021) Implicit, automatic semantic word categorisation in the left occipitotemporal cortex as revealed by fast periodic visual stimulation. *NeuroImage* 238:118228. <https://doi.org/10.1016/j.neuroimage.2021.118228>

Vytal K, Hamann S (2010) Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *J Cogn Neurosci* 22(12):2864–2885. <https://doi.org/10.1162/jocn.2009.2136>

Waller BM, Micheletta J (2013) Facial expression in nonhuman animals. *Emot Rev* 5(1):54–59. <https://doi.org/10.1177/1754073912451503>

Waller BM, Julle-Daniere E, Micheletta J (2020) Measuring the evolution of facial ‘expression’ using multi-species FACS. *Neurosci Biobehav Rev* 113:1–11. <https://doi.org/10.1016/j.neubiorev.2020.02.031>

Whiting C, Kotz S, Gross J, Giordano BL, Belin P (2020) The perception of caricatured emotion in voice. *Cognition* 200:104249

Wiethoff S, Wildgruber D, Grodd W, Ethofer T (2009) Response and habituation of the amygdala during processing of emotional prosody. *NeuroReport* 20(15):1356–1360. <https://doi.org/10.1097/WNR.0b013e328330eb83>

Zhao C, Schiessl I, Wan MW, Chronaki G, Abel KM (2021) Development of the neural processing of vocal emotion during the first year of life. *Child Neuropsychol* 27(3):333–350. <https://doi.org/10.1080/09297049.2020.1853090>

Zhou F, Zhao W, Qi Z, Geng Y, Yao S, Kendrick KM, Wager TD, Becker B (2021) A distributed fMRI-based signature for the subjective experience of fear. *Nat Commun*. <https://doi.org/10.1038/s41467-021-26977-3>

CHAPTER 3

Temporal Dynamics of Natural Sound Representation in the Infant Brain

Siddharth Talwar ¹, Roberta P. Calce ¹, Laura Bourgaux ², Stefania Mattioni ³, Adélaïde de Heering ², Olivier Collignon ^{1,4}

¹ Institute for Research in Psychology (IPSY) & Neuroscience (IoNS), Louvain Bionics, University of Louvain (UCLouvain), Louvain, Belgium

² Unité de Recherche en Neurosciences Cognitives (UNESCOG), ULB Neuroscience Institute (UNI), Center for Research in Cognition & Neurosciences (CRCN), Université libre de Bruxelles (ULB), Brussels, Belgium

³ Department of Experimental Psychology; UGent, Gent, Belgium

⁴ School of Health Sciences, HES-SO Valais-Wallis, The Sense Innovation and Research Center, Lausanne and Sion, Switzerland

3.1 Introduction

Giving meaning to an extremely rich and dynamic sensory world around us is a fundamental challenge, particularly during development. Infants need to progressively learn to group vast information that surrounds them into coherent and meaningful perceptual classes. This process, called perceptual categorization (Mervis & Rosch, 1981), is fundamental to avoid perceiving each entity uniquely, leaving the observer overwhelmed by the sheer diversity of what they experience in their environment.

In human adults, it was shown that the visual ventral stream – located in the occipitotemporal cortex - contains regions that are selective for specific categories of high ecological relevance such as faces (Kanwisher et al., 1997), bodies (Downing et al., 2001), places (Epstein & Kanwisher, 1998), and words (Cohen et al., 2000). Importantly, each of these category-selective regions interacts with a unique large-scale network of brain areas relevant to each category of interest (Beeck et al., 2008; Konkle & Oliva, 2012; Mattioni et al., 2020; Osher et al., 2016; Peelen & Downing, 2017). Furthermore, multivariate pattern analysis (MVPA) applied to brain recordings moved the field beyond selectivity mapping by highlighting the nature of represented information in the human brain (Haxby et al., 2001; Kriegeskorte et al., 2008). For example, these analyses revealed that early visual regions primarily extract low-level features of visual objects early in time, while downstream regions in the ventral occipito-temporal stream represent object categories later in time, largely independently of their low-level properties (Bracci & Op De Beeck, 2016; Carlson et al., 2013; Cichy et al., 2014, 2019; Iamshchinina et al., 2022; Kaiser et al., 2016; Peelen & Caramazza, 2012; Proklova et al., 2016). However, these studies investigated predominantly the adult brain.

In the developing brain, behavioral studies using a preferential looking paradigm (Aslin, 2007; Fantz, 1964; Quinn & Eimas, 1996) showed evidence for visual category processing (Rakison & Yermolayeva, 2010) and learning (Mareschal & Quinn, 2001) in the first years of life. Infants' looking behavior, when compared with the neural object representation in the adults' visual cortex showed progressive recruitment across the visual cortex with age (Spriet et al., 2022). Specifically, infants exhibited a general animacy categorization between 4 and 10 months old, transitioning to a fine-grained, adult-like object representation between 10 and 19 months old (Spriet et al., 2022). Brain recording studies identified similar precursors of category-specific brain response in infants (Buiatti et al., 2019; de Heering & Rossion, 2015; Peykarjou, 2022). These findings suggest early neural specialization for visual object categories (Deen et al., 2017), despite categorical representations having a significant lag in their brain time course compared to the adult brain (Xie et al., 2022).

While the processing of sound categories in the human brain remains far less understood than in vision (Iamshchinina et al., 2022), it has been proposed that similarly to the visual cortex, the auditory system processes sounds in a hierarchical fashion, with primary auditory regions extracting low-level acoustic information early in time, while higher-order regions (e.g., middle and superior temporal sulci and gyri) implement a more abstract representation of sound categories at later stages (Giordano et al., 2013; Leaver & Rauschecker, 2010; Russ et al., 2007). However, the extent to which categorical auditory representations are independent of their characteristic acoustic features remains debated (Barbero et al., 2021; Giordano et al., 2023; Ogg et al., 2020). Furthermore, little is known about the emergence of sound representation during development. Several studies showed evidence of selective brain response to socially relevant auditory stimuli such as human voices (Calce et al., 2024), speech classes (Gennari et al., 2021; Kuhl, 1983) and also auditory sense numerosity (Gennari et al., 2021) in the first months of life (Dehaene-Lambertz, 2000; Dehaene-Lambertz et al., 2002). However, by treating categories and abilities as separated entities, these studies lack a comprehensive mapping of sound categorization in the infants' brains. What is more, investigating the role of low-level features through the implementation of control conditions may not encompass the entirety of features inherent in the investigated sound (Barbero et al., 2021; Calce et al., 2024). The challenge lies in the difficulty of controlling for all features within each condition, hindering a comprehensive understanding of the low-level contribution. Therefore, whether and when different acoustic and categorical features of sounds are represented in the infant's brain remains elusive.

In the present study, we aimed to address the above mentioned issues by examining the temporal dynamics of auditory categorization of multiple animate and inanimate

object classes in the infant's brain. For that purpose, we used time-resolved MVPA applied to EEG data (Cichy et al., 2014; Grootswagers et al., 2017) collected in a group of infants and adults to investigate the time course of auditory object representations in development. Preliminary data show a substantial lag in the time-course of auditory object representation in infants compared to the adult brain. However, although immature, the developing brain implements sound representation early in development. Importantly, using Representational Similarity Analysis (RSA), we found categorical representation in infants which are independent of the spectral-temporal processing of the sounds, with a unique preference for non-human mammalian sounds and human speech sounds. Both infants and adults were found to exhibit sound representations; however, these representations were not yet synchronized in infants, as they unfolded differently over time. Future stages of this study will include a larger sample of infants of different ages to identify the time course at which a reliable and stable object categorization appears.

3.2 Methods

3.2.1 Stimuli

For the current experiment, we employed 24 natural sounds previously used in (Mattioni et al., 2020), Figure 3.1), categorized into two superordinate groups: animate and inanimate. Each sound could be further classified into one of the four semantic categories: animals, humans, big objects/places and manipulable objects, with six unique sounds per category. Each of these categories included two sub-categories, resulting in a total of eight sub-categories: birds and mammals within animals; human verbalizations and non-verbalizations within humans; big mechanical objects and environmental scenes within big objects/places, graspable objects and tools within manipulable objects, with three unique sounds per subcategory. Each sound had a duration of 1 second and was presented through a mono-channel (44,100 Hz sampling rate). The root mean square values of the stimuli were normalized. To prevent clicking, 10 ms ramps were applied at the beginning and end of each sound. Stimuli can be accessed at https://github.com/sid0309/CatinTime_Stimuli.git

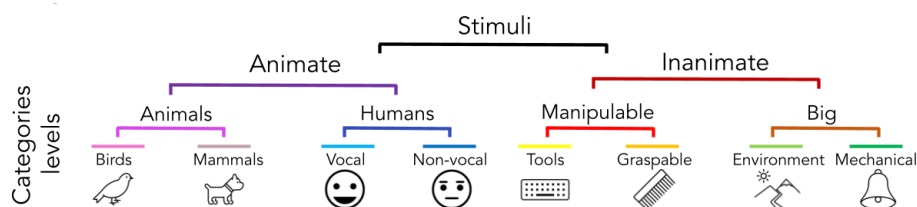


Figure 3. 1 Stimuli set of natural sounds (n=24) arranged in a categorical hierarchy - 8, 4 and 2.

3.2.2 Participants

Two groups of participants took part in this study: 4–7-month-old infants and adults. Thirty-four infants took part in the study, of which 12 were excluded due to an insufficient number of trials post artifact rejection (or exemplars of a sound) ($n = 9$) or artifacts in the continuous recording ($n = 3$), yielding a final sample analyzed of 22 infant participants (mean age = 41.66 years, S.D. = 12.38, range: 145 to 221 days). All participants were born at term (i.e., > 37 gestational weeks), and had an APGAR score of at least 9 at birth, and their caregivers reported no auditory or neurological disorders. The infant sample was tested at the BabyLab (LulLABy) of the Free University of Brussels. Caregivers were recruited through flyers in local nurseries, gave written consent after being fully informed about the aim and procedure of the study, and received reimbursement for their transport cost as well as a toy for infants. The experiment was conducted according to the Declaration of Helsinki and under the approval of the ethical committee of the Hospital Erasme (Project 2021-445).

The adult sample comprised 18 participants (12 females, mean age = 41.66 years, S.D. = 12.38) with no history of hearing or neurological disorders. Participants gave written consent after being fully informed about the aim and procedure of the study and they received monetary compensation. The experiment took place at the University of Louvain and was conducted according to the Declaration of Helsinki and under the approval of the ethical committee of the University of Louvain (Project 2016-25).

3.2.3 Data Acquisition

After electrode-cap placement, infants were installed on their caregiver's lap, in a light- and sound-attenuated room. The seat was positioned at a 60 cm distance in front of a monitor behind which the speakers were centered. Sounds were presented at an intensity of 60 dB, measured with a decibel meter at the distance of the seat from the speakers. Caregivers were asked not to interact with the infants unless they manifested distress. Additionally, they were instructed to wear noise-cancelling headphones to prevent inadvertently influencing the babies' responses to sounds. Infants were continuously monitored using a camera placed on top of the monitor and non-periodic visual distractors were used when needed to reorient attention toward the speakers.

All experiment codes were implemented on Psychtoolbox and MATLAB. In infants, sounds were presented in 7-12 runs. Each run consisted of 1 repetition of each sound (24 sounds per run). The order of presentation within each repetition was

randomized. A run was a priori rejected and not considered for analyses in case of excessive movement or manifest distress before the end of the run. The experimenter continued launching the runs until the first sign of tiredness, discomfort, or caregiver's demand.

In adults, before the EEG experiment, all sounds were presented in a randomized order to the participants to familiarize them with the stimuli and ensure correct recognition. During the EEG experiment, sounds were presented in 15 runs. Each run consisted of 3 repetitions where 1 repetition refers to the presentation of all 24 sounds resulting in a total of 72 sounds per run. Thus, each sound was presented once before any sound was repeated. The order of presentation within each repetition was randomized. Participants performed a one-back recognition task, pressing a button if they heard two consecutive identical sounds. These 6 catch trial sounds were additionally inserted pseudo-randomly in each run (2 within each repetition), where the catch trial sound was the same as the previous sound. The inter-trial interval (time between end of previous trial and beginning of the next) was set to 1.5 seconds, plus a jitter computed randomly from a Gaussian distribution with a mean of 0 and a range of ± 0.25 seconds. Each run lasted for approximately 3.5 minutes. Participants were blindfolded during the experiment to ensure the eyes remained closed and avoid ocular artifacts in the EEG trace. Participants performed the experiment seated in a dimly lit room at approximately 1 m from the speakers. The sound intensity was kept at a comfortable level of approximately 60dB.

EEG was acquired with a Biosemi ActiveTwo System (<https://www.biosemi.com/products.htm>) using 32 channel Ag/AgCl electrodes for infants and 128 channels for adults. Recording sites included the standard 10-20 system locations as well as intermediate positions (channel coordinates can be found at: <https://www.biosemi.com/headcap.htm>). EEG was acquired at a sampling rate of 512 Hz. The magnitude of offset of all electrodes that are referenced to common mode sense (CMS) and driven right leg (DRL) was kept below ± 25 mV.

3.2.4 Analysis

3.2.4.1 EEG Preprocessing

Infants

Infant data was preprocessed using Automatic Pipeline for Infants Continuous EEG (APICE, Fló et al., 2022) which is a fully automated, flexible and modular pipeline that provides standardization for infant data preprocessing. In conjunction, we used the Fieldtrip toolbox (Oostenveld et al., 2013) and custom scripts running on MATLAB 2016b. Infant data was downsampled to 256 Hz and a band-pass filter from

0.5 Hz to 40 Hz was applied on continuous EEG data. The following analysis was performed using the APICE pipeline adapted from (Fló et al., 2022; Gennari et al., 2023).

APICE relies on adaptive thresholds to find bad samples, rather than absolute thresholds to account for inter-individual variability. Thresholds were set independently for each subject and each of the 32 electrodes based on the distribution of various measures throughout the entire recording. The threshold was defined as the median $\pm$ $n \cdot IQ$, where IQ represents the interquartile range of the distribution. A series of algorithms identified artifacts on the basis of: the voltage amplitude and its first derivative; the variance across a 500ms-long moving time window; the fast running average and the deviation between the fast and the slow running averages within a 500ms-long sliding time window. Two additional algorithms identified whether the power within the 0 - 10 Hz band was excessively low or within 20 - 40 Hz excessively high relative to the total power; and whether the voltage amplitude displayed by each sensor at a given time point was disproportionate relative to that recorded by the other sensors at the same instant.

Artifact detection was conducted on the continuous recording in an iterative fashion (4 loops in total). At each loop, previously identified bad samples were kept aside for the subsequent artifact detection steps. Algorithms were applied twice in order to identify very short signal disruptions (80ms max), corresponding to heart beats or jumps. These short segments were corrected by estimating their principal components (PCA) and removing the first n components determining 90% of the variance. This operation was followed by a high-pass filter (0.5 Hz) to eliminate possible drifts that could have been created by this local correction. These detection algorithms were applied to a maximum of 2 times. At the end of these iterations, we obtained a rejection matrix of the same size as the EEG recording indicating “bad” timesamples for each electrode.

Data was further epoched from -500 ms to 2000 ms in reference to the sound onset. Then, we used the rejection matrix to mark, in each epoch, the time points containing prominent artifacts (bad times) and channels that did not function properly (bad channels). Specifically, bad times were periods longer than 80ms with a percentage of rejected channels superior to 30% or beyond $2IQ$ from the 3rd quartile of the distribution of the percentage of rejected channels across time. Similarly, bad channels were the ones not working properly for more than 30% of the time points composing the epoch or with a percentage of bad samples that went beyond $2 \cdot IQ$ from the 3rd quartile of the distribution of the percentage of rejected samples across channels. With this step we extended in time and space the detected bad samples based on the assumption that when a portion of channels/times are detected as

“bad”, the temporal/spatial neighbors are likely to be affected by the artifact as well. Bad channels and long rejected segments of a given electrode were corrected using spherical splines interpolation only if at least 50% of the neighboring channels were intact. Epochs were discarded if more than 25% of their samples contained artifacts or if more than 2.5% of their channels were marked as bad. Since in infants ongoing activity is particularly large and can hinder the extraction of the evoked potential, epochs were also discarded based on their Euclidean distance from the average, i.e. when their mean or maximum distance from the average response was an outlier in the distribution ($> 3\text{rd quartile} + 1.5 \times \text{IQ}$). Next, we applied baseline subtraction and visually inspected the trials. Few trials still presenting obvious artifacts were rejected.

Adults

In adults, we applied a simpler preprocessing pipeline. A fourth order Butterworth band-pass filter from 0.1 Hz to 100 Hz was applied on continuous EEG data. Additionally, a fourth order Butterworth band-stop filter at 50 Hz was implemented to remove power-line noise. Continuous EEG data was segmented into 1080 trials (24 sounds X 15 runs X 3 repetitions) of 2.5 seconds length, including 0.5 seconds before stimulus onset and 2 seconds post onset (i.e., 1 seconds of sound presentation and 1 second post-stimulus). Baseline subtraction was applied on the extracted trials using the mean pre-stimulus activity as the baseline window. Then, the trials were down-sampled to 256 Hz to facilitate data storage and handling, providing a temporal resolution of 3.9062 milliseconds. A semi-automatic artifact rejection pipeline was employed: trials with absolute amplitude exceeding 250 μV were deleted and trials with a maximum absolute amplitude between 150 μV and 250 μV were visually inspected and potentially discarded. Lastly, all trials were re-referenced to common average referencing.

3.2.4.2 Time resolved sound decoding from EE

All analyses were conducted in MATLAB 2017b. Decoding analyses were conducted using the LibSVM library (Chang & Lin, 2011).

3.2.4.2.1 Decoding Across All Channels

The preprocessed trials were first divided into their corresponding sounds ($n=24$) making sure that there were at least of 40 trials per sound for any adult participant and at least 6 per sound for infants. To improve signal-to-noise ratio, 3 pseudo-trials per sound were computed by averaging trials together. Further analysis was performed on these 72 pseudo-trials (24 sounds x 3 pseudo-trials) in both groups. A timepoint-by-timepoint binary decoding of each comparison (276 binary comparisons across 24 sounds) was implemented using a 3-fold cross validation with a ‘leave one trial out’ strategy, i.e., training on four trials and testing on remaining

two. Support Vector Machines (SVM) were used for training and classification, as they are generally better at handling high-dimensional feature sets (Grootswagers et al., 2017). Area under the curve (AUC) was computed as a metric to quantify decoding performance, providing a measure of both specificity and sensitivity. The AUC values from all folds were then averaged, yielding a decoding metric for each time point and comparison. The entire process from creating pseudo-trials to computing the binary decoding AUC at each time-point was repeated 100 times, to reduce the arbitrariness in creating pseudo-trials and ensure a reliable decoding AUC.

3.2.4.2.2 Sensor-Space Searchlight

Sensor-space searchlight analysis was performed to find the channels contributing to the decoding AUC as a function on time (Kaiser et al., 2016; Kriegeskorte et al., 2008). Using the same procedure as decoding across all channels (section 2.3.2.1), we computed decoding AUC for each channel separately. The seed channel itself and 5 or 3 neighboring channels in adults or infants respectively, underwent timepoint-by-timepoint binary decoding across 276 comparisons. This provided a decoding AUC for each channel, comparison and timepoint. Finally, the AUC were averaged across all comparisons and were illustrated as topoplots averaged within different time-windows in Figure 3.2b.

3.2.4.3 Representation Similarity Analysis

RSA allows us to abstract the neural responses from the multivariate space and other properties of stimuli from their native space, and find shared representation between neural geometries and other representational models across time. The essence of RSA is the dissimilarity matrices (DSM) which consists of distances between pair-wise conditions. Thus, the comparison of neural DSMs and external representational models can reveal how information unfolds over time and across spatial regions. In our study, we used this tool to investigate the unfolding of different neural representations of sounds through time in infants and adults.

We created sensor-level EEG dissimilarity matrices (EEG DSM 24x24) at every time-point (~4 ms) using sensor space decoding AUC across all comparisons. These neural models were smoothed by averaging using a moving time window of length of 5 bins at each time-point (2 bins on either side of the central time bin, 5 bins ~20 ms). DSMs were also computed at the sensor level using the decoding AUC derived from sensor searchlight decoding analysis. Since the DSMs are symmetric matrices with a diagonal of zeros, we used only the lower triangle, excluding the diagonal, for all subsequent analyses. Participant specific neural models were correlated with the external representational models as explained below. Correlations and partial

correlations at each time point were computed at each time point using Spearman's rank correlation.

Three groups of external representational models were used to characterize the sounds: acoustic features, a priori categorical models and models created from a deep neural network.

The acoustic model was computed from the modulation transfer function using NSL tools (MTF, (Chi et al., 2005), which is inspired from the psycho-acoustical and neuro-physiological observations in the early and central auditory systems. The model is composed of two parts – the early acoustic transformation through a band of 128 ERB filters mimicking the processing in the cochlea and the midbrain, and the later cortical stage that simulates processing in the primary auditory cortex. For our analysis, we considered only the cortical simulation part in which the MTF provided a 4-dimensional representation of the sound's spectral and modulation features including: temporal window (ms), frequency (Hz), modulation-rate (up vs. down) and scale (Ω). The modulation spectrum was computed using 2D modulation selective filters tuned to default temporal modulation frequencies $w = 4, 8, 16, 32$ Hz and spectral modulation frequencies $\Omega = 0.25, 0.5, 1, 2, 4, 8$ cycles / octave. Finally, the multidimensional structure was averaged across time to obtain a total of 6144 features for each sound. Additionally, other acoustic properties such as center of gravity (COG), harmonicity-to-noise ratio (HNR) and pitch were computed using Praat ((Boersma & Weenink, 2001). Cosine distance between each pairwise sound representation was computed, providing dissimilarity measures in 24 x 24 dimension for all acoustic models (24 sounds in total).

Categorical models were created based on the a priori memberships of sounds at each level of the categorical hierarchy (2, 4, 8 categories, Figure 3.1). Thus, the distance between each comparison of sounds was set to 1 if they belonged to different categories or 0 if they had the same category membership.

To create representational models based on different layers of a DNN, we chose YAMNet, a feed-forward convolutional deep neural network developed by Tensorflow and pretrained on the Google Audioset corpus. The selection of this specific DNN was based on results from a previous study (Giordano et al., 2023) showing that YAMNet is the best model for predicting responses from superior temporal regions, outperforming other acoustic models and other deep neural networks. By providing an auditory stimuli as an input (sampled every 0.96), YAMNet provides a prediction from 521 audio events from the AudioSet ontology. The Mobilenet_v1 architecture includes an input layer and a standard convolutional layer, followed by 13 depthwise separable convolutional layers. This provides a 3 x 2

dimension of 1024 kernels which are averaged and put through a single logistic layer of 521 audio ratings. Sound embeddings from all 14 layers were considered for analysis. Scripts for implementation in Python and MATLAB were adapted from. For all DNN layers, cosine distance between each pairwise sound representation was computed, providing dissimilarity measures in 24 x 24 dimension.

To reveal the consistency of neural representations within each group, each participant's neural DSM was correlated with the mean of their respective group for each time-point combination, while taking out that particular participant from the computation of the group mean (individual infant with mean of infants and individual adult with mean of adults). A diagonal from bottom-left to top-right corner reflects temporal synchrony between the two estimates. This diagonal also reflects the lower bound of the noise ceiling in each group, representing a threshold above which correlations with any external model indicate that the model successfully predicts the variance in the neural data (Nili et al., 2014). For intergroup comparison, individual adult EEG DSMs were correlated with the mean of the infant group (individual adults with the mean of infants). We averaged the correlations of each adult participant and subsequently performed statistics of the estimate. We chose this to obtain a better estimate of neural representation in infants due to their poor SNR.

3.2.4.4 Statistical Analysis

One-sample sign-permutation tests were performed to find the time-points when decoding AUC or Spearman's rank correlation were significantly higher than chance level or 0 respectively for each group. AUC values were subtracted by chance level (0.5) before sign flipping. For one-sample test, at each time-point, the individual values were multiplied with either +1 or -1 (random sign flip) and then the group average was computed. 5000 sign-flip permutations were computed to create a null distribution and a p-value was determined by calculating the proportion of values in the null distribution higher than the observed empirical mean.

Subsequently, clusterwise inference (Nichols & Holmes, 2002) in 1-D (sensor space decoding and correlations), 2-D (sensor space searchlight decoding, within-group and inter-group correlations) were employed to correct for multiple comparisons by converting the null distribution to p-maps across all time points or source points and determining the largest cluster thresholded at $p < 0.05$. All significant empirical time-clusters that were bigger than the largest cluster found in the null distribution survived the correction.

Bootstrap tests were implemented to estimate confidence intervals (95%) of the temporal onset of the first significant cluster, offset of the last significant cluster and peak latencies. We created 1000 bootstrapped samples by sampling the participants

with replacement. The same statistic as above was applied to each bootstrap which resulted in an empirical distribution of onset, offset and peak latencies on which the confidence intervals were determined.

3.3 Results

3.3.1 Sound Decoding

We performed multivariate pattern decoding at the sensor level between pairwise sounds (sound 1 vs sound 2, sound 1 vs sound 3 etc) leading to 276 binary comparisons in infants and adults yielding a decoding accuracy at each time point. Significance at each timepoint was computed using a sign-permutation test followed

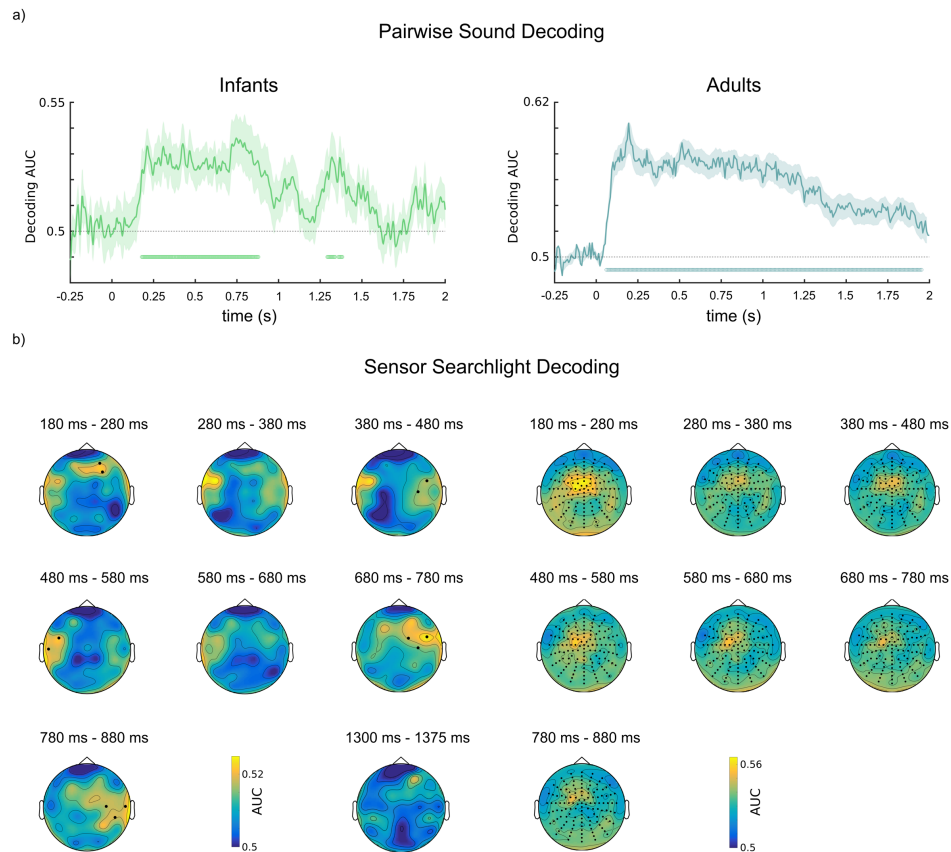


Figure 3. 2 Decoding in infants and adults. a) Grand average of pairwise sound decoding implemented at each timepoint, i.e., mean of 276 binary comparisons and participants (right tail sign-permutation test, $p < 0.05$, cluster based correction). While the curve in adults shows a sharp peak and gradual decline, infants exhibit a slow rising curve and broad peaks with maximum AUC (area under the curve) at a higher latency than adults. A gradual decline of the classification curve is followed by another rise post sound

offset only in infants. b) Sensor searchlight decoding reveals channels contributing to sound decoding as illustrated in bold dots. They were computed within time windows of 100 ms beginning from the onset of significant AUC observed in infants (right tail sign permutation test, $p < 0.05$, cluster based correction).

by a cluster based correction to address multiple comparisons. The following results present the peak latency, onset, and offset of significant time points for both groups. 95% confidence intervals were calculated from bootstrap tests, presented in square parentheses only for adults. The following reported latencies are calculated in reference to the sound onset. Infants exhibited significant decoding accuracy ranging from 179 ms to 878 ms, peaking at 746 ms (Figure 3.2a). Additionally another cluster was found post sound offset from 1296 ms to 1375 ms. In adult participants, a higher and different classification curve emerged sharply at 59 ms [41-72 ms], peaking at 175 ms [163-951 ms] and ending at 1949 ms [1902-1992 ms].

Sensor searchlight decoding in infants revealed the channels contributing maximally to the decoding curve. In infants, temporal and frontal channels contained maximum sound relevant information at different time-windows (Figure 3.2b). For adults, most channels exhibited significant sound decoding over the same time span, with the highest area under the curve (AUC) values localized to fronto-central channels.

3.3.2 Representation Similarity Analysis (RSA)

To disentangle the contribution of acoustic processing and categorical representation in infants and adult EEG, we employed RSA to correlate neural dissimilarity matrices (DSM) with external representation models. In infants, acoustic representation simulated by the modulation transfer function (MTF) emerged later in time in comparison to adults (Infants: onset = 800 ms, peak = 933 ms, offset = 960 ms; Adults: onset = 85 ms [74-156 ms], peak = 187 ms [93-242 ms], offset = 214 ms [113-1062 ms], Figure 3.3a). To investigate further the emergence of acoustic representation in reference with the unravelling of the sound presentation over time, we computed MTF within consecutive 100 ms windows and consequently compared each of them with the neural models. Interestingly, acoustic processing in infants emerged only after the presentation of the sounds in contrast to adults who displayed acoustic representation even before the relevant portion of the sound was fully presented (Figure 3.3b). This indicates feedback connections in adults that facilitate acoustic processing, as well as the lack of the same in infants who show more feedforward processing in their response to stimuli yet with more processing delays.

To find unique categorical representation independent of the spectral-temporal properties of the sounds, we partialled out the MTF model (computed using the whole time of the sound) from the correlations of the 8-categorical model and neural DSMs. Similar to the pattern observed in the late latency of the emergence of

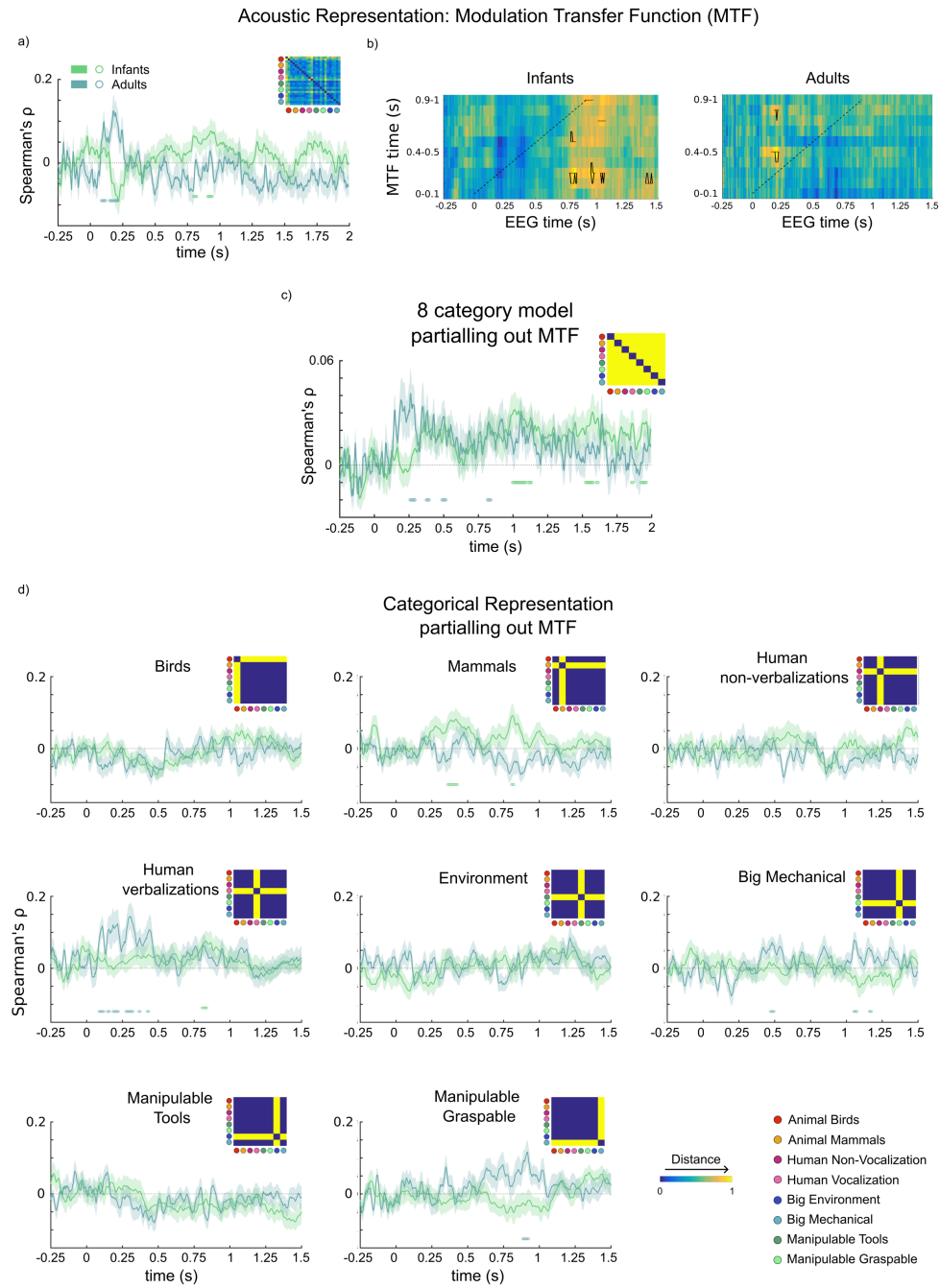


Figure 3. 3 Representation Similarity Analysis a) Comparison of time-varying EEG and acoustic model (Modulation Transfer Function, MTF) computed from the entire duration of the sounds. Delayed and

Chapter 3. Temporal Dynamics of Sound Representation in Infants and Adults

weaker acoustic representation of sounds emerge in infants compared to adults. b) Correlation of time-varying EEG and MTF models computed in consecutive 100 ms windows. The dotted line represents the linear progression of sound presentation in relation to the time of EEG activity. Significant correlation to the right of the dotted line reflects responses appearing later in time to previously heard portions of sounds, while activity to the left indicates emergence of acoustic representation before those portions of sounds are presented. In adults, early brain activity encodes acoustic features of upcoming sounds, whereas in infants, acoustic representation emerges only after the sound has been presented. Significant clusters are shown in continuous black lines (right tailed sign permutation test, $p < 0.05$, 2-D cluster wise correction). c) Categorical representation after partialling out the acoustic model emerged at sound offset in infants and earlier in adults (right tailed sign permutation test, $p < 0.05$, 1-D cluster wise correction). d) Individual categorical representations in both groups. Mammals and human verbalizations (HV) are uniquely represented in infant EEG, while HV, big mechanical and manipulable graspable tools are uniquely represented in adults (right tailed sign permutation test, $p < 0.05$, 1-D cluster wise correction).

representation, we found late categorical representation in infants compared to adults emerging post sound offset (Infants: onset = 1000 ms, peak = 1074 ms, offset = 1968 ms; Adults: onset = 183 ms [49-408 ms], peak = 269 ms [148-1636 ms], offset = 1148 ms [519-1988 ms], Figure 3.3c). These results show the temporal dynamics of categorical selectivity in infants to natural sounds for the first time. Interestingly, the categorical representations in both groups peaked soon after their respective acoustic representations, suggesting a presence of hierarchical auditory processing in both infants and adults.

We also investigated finer categorical representations in infants with sub-categorical models- birds, mammal, human non-verbalization, human verbalizations, big environmental, big mechanical, manipulable tools and graspable tools (Figure 3.3d). These models assume that the target subcategory is distinctly unique while treating all other subcategories as identical and disregarding the relationships among the non-target subcategories. Similar to above, we partialled out the acoustic representation from the correlations between sub-categorical models and neural models. In infants, we found maximum correlation with the non-human mammal model which was absent in adults (Infants: onset = 363 ms, peak = 414 ms offset = 863 ms). Strikingly, this strong representation of mammals emerged much earlier than the acoustic processing in infants indicating the presence of a neural ensemble that is selective to non-human mammal sounds and is independent of the spectral-temporal properties of the sounds. Conversely, adults exhibited maximum selectivity to human verbalizations (HV) which emerged later in infants (Infants: onset: 820 ms peak: 832 ms offset: 847 ms; Adults: onset: 89 ms peak: 183 ms offset: 445 ms). No unique representation of human non-verbalizations was found in either populations. Interestingly, the emergence of categorical selectivity for HV in both groups occurred during the same latency as the emergence of their respective acoustic representations, suggesting the link between spectral-temporal cues of speech-like sounds and HV selectivity. Additionally in adults, we also found unique categorical representations of big mechanical sounds (Adults: onset: 472 ms peak: 1054 ms

offset: 1074 ms) and graspable tools (Adults: onset: 886 ms peak: 918 ms offset: 933 ms).

In summary, the brain of infants aged 4-7 months exhibit a general categorical representation post acoustic processing, but also show a pronounced and early predilection for non-human mammal sounds, a preference that is independent from the spectra-temporal modulations of the sounds and diminishes to non-existence in adulthood. Concurrently, they possibly develop neural mechanisms that subserve processing of speech-like sounds. Showing early selectivity to speech sounds, these processes establish a foundation for the acquisition of adult-like speech perception and early language skills.

3.3.3 Deep Neural Networks (DNN)

DNNs allow us to investigate complex auditory representations that could explain neurophysiological processes aiding auditory categorization. We used YAMNet, a pre-trained feed-forward convolutional network consisting of 14 layers, that performs classification of sound events based on 521 sound labels (architecture illustrated in Chapter 4, Figure 4.4a). We passed our stimuli through the DNN, extracted the embeddings of each layer, modeled and correlated them with the neural models.

We hypothesized that auditory complexity would increase over time, as reflected by the progressive emergence of early to late DNN layers in the neural models of both groups. This expectation was based on the finding that MTF uniquely predicted the variance of early DNN layers better than the 8-categorical model, while the inverse was true for the late DNN layers (Figure 3.4a). Peaks of representations of different layers can also be viewed in Table 3.1 and Figure 3.4b. Amongst early layers, we found layers 1 and 4 to emerge early in infant EEG (*infants-layer 1*: onset: 250 ms, peak: 218 ms, offset: 265 ms; *infants-layer 4*: onset: 93 ms, peak: 113 ms, offset: 128 ms). Infants' neural models also shared representation with intermediate to late layers 7, 8, 10, 11, 12 later in time (*infants-layer 7*: onset: 882 ms, peak: 1281 ms, offset: 1359 ms; *infants-layer 8*: onset: 886 ms, peak: 929 ms, offset: 1328 ms; *infants-layer 10*: onset: 1109 ms, peak: 1253 ms, offset: 1890 ms; *infants-layer 11*: onset: 1878 ms, peak: 1878 ms, offset: 1890 ms; *infants-layer 12*: onset: 1210 ms, peak: 1160 ms, offset: 1253 ms). Layers 2 to 6 were also represented in a large cluster post sound offset (*infants-layer 2*: onset: 1328 ms, peak: 1363 ms, offset: 1496 ms; *infants-layer 3*: onset: 1308 ms, peak: 1363 ms, offset: 1480 ms; *infants-layer 4*: onset: 1269 ms, peak: 1359 ms, offset: 1519 ms; *infants-layer 5*: onset: 1128 ms, peak: 1363 ms, offset: 1402 ms; *infants-layer 6*: onset: 1250 ms, peak: 1339 ms, offset: 1402 ms).

Chapter 3. Temporal Dynamics of Sound Representation in Infants and Adults

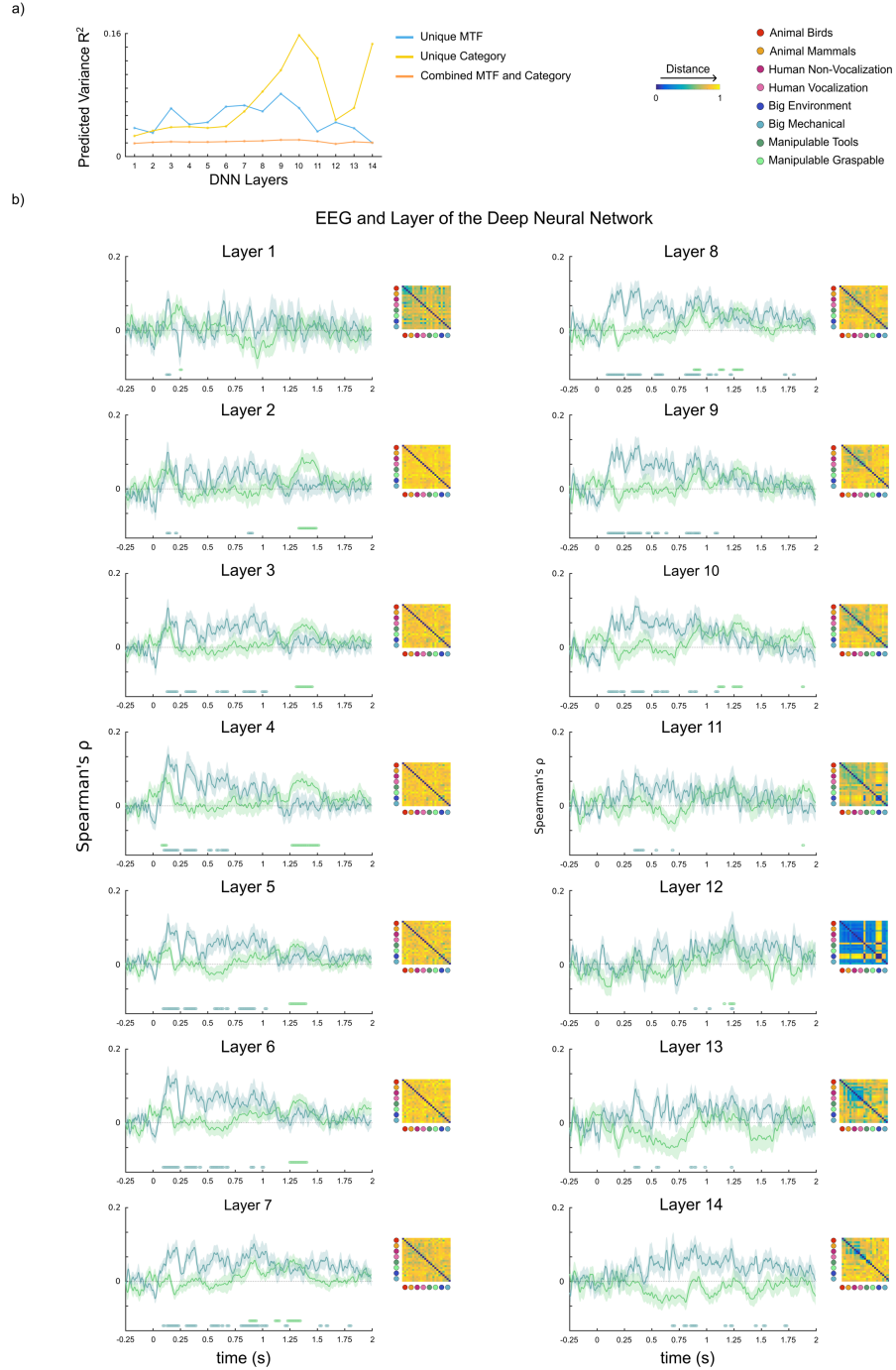


Figure 3. 4 Deep Neural Network a) Unique variance of the acoustic MTF model and the 8-categorical

model predicting different layers of the DNN. b) Comparison of time-varying EEG and 14 layers of the DNN which progressively increase in complexity. The DSMs show the representation of each layer.

All DNN layers were represented in adults, exhibiting a hierarchical pattern with the early layers peaking earlier in time and the late layers peaking later (*adults-layer 1*: onset: 125 ms, peak: 140 ms, offset: 160 ms; *adults-layer 2*: onset: 128 ms, peak: 132 ms, offset: 894 ms; *adults-layer 3*: onset: 125 ms, peak: 136 ms, offset: 1046 ms; *adults-layer 4*: onset: 101 ms, peak: 140 ms, offset: 1011 ms; *adults-layer 5*: onset: 93 ms, peak: 136 ms, offset: 1043 ms; *adults-layer 6*: onset: 89 ms, peak: 140 ms, offset: 1011 ms; *adults-layer 7*: onset: 125 ms, peak: 921 ms, offset: 1816 ms; *adults-layer 8*: onset: 93 ms, peak: 378 ms, offset: 1730 ms; *adults-layer 9*: onset: 97 ms, peak: 378 ms, offset: 1016 ms; *adults-layer 10*: onset: 105 ms, peak: 351 ms, offset: 1109 ms; *adults-layer 11*: onset: 343 ms, peak: 351 ms, offset: 699 ms; *adults-layer 12*: onset: 89 ms, peak: 1234 ms, offset: 1808 ms; *adults-layer 13*: onset: 343 ms, peak: 351 ms, offset: 996 ms; *adults-layer 14*: onset: 687 ms, peak: 695 ms, offset: 1734 ms).

In adults, the latencies of acoustic processing representations (MTF) and early DNN layers aligned, similar in the case of categorical representation and late DNN layers aligning later in time. However, in infants, the stages of auditory processing appear to be only partially temporally synchronized compared to adults. Although early layers 1 and 4 emerged early in infant EEG, the representation of layers 2 to 6 peaked higher in a late cluster at ~350 ms post sound offset. This suggests that certain auditory processes in infants are still developing, as its weak representation indeed emerged earlier in time, followed by the recurrence of the same auditory representation later in time. Additionally, infants displayed a capacity to encode more complex auditory representations denoted by the intermediate-late DNN layers, however no significant representations of the last two DNN layers in the EEG indicate a lack of higher-order structure.

MTF predicted early DNN layer representations better than the categorical model, however a significant portion of variance remains unexplained (>0.95). This suggests that early DNN layers encode an alternate and complex auditory representation that may extend beyond spectral-temporal cues aiding auditory categorization.

Chapter 3. Temporal Dynamics of Sound Representation in Infants and Adults

Decoding analysis	Infants Peak (ms)	Adults Peak (ms)
Pairwise (sound vs sound)	746	175
RSA		
Modulation Transfer Function (MTF)	933	187
Deep Neural Network		
Layer 1	218	140
Layer 2	1363	132
Layer 3	1363	136
Layer 4	1359	140
Layer 5	1363	136
Layer 6	1339	140
Layer 7	1281	921
Layer 8	929	378
Layer 9	-	378
Layer 10	1253	351
Layer 11	1878	351
Layer 12	1160	1234
Layer 13	-	351

Layer 14	-	695
Categorical model		
8 - category partialling out MTF	1074	269

Table 3. 1. Peak latency of all external representational models in EEG

3.3.4 Shared Neural Representations Between Infants and Adults

We investigated whether representations related to auditory object processing and categorization are consistently shared within groups and across infants and adults. Due to non-linearities and different trajectories of auditory processing in both groups, direct identification of shared representations is difficult. Therefore, we correlated the neural models at each timepoint and participant in the adult population with the group mean of neural models of infants at every timepoint. A perfect linear synchrony in the temporal unfolding of auditory processing could be visualized as a diagonal from sound onset time (0 ms) until some time around or after sound offset (bottom left to top right corner).

Within the infant group, comparisons revealed no diagonal suggesting a lack of shared auditory representation (Figure 3.4a). Nevertheless, some significant small clusters were found along the diagonal at around 750-850 ms coinciding with the peak of sound decoding and emergence of the acoustic MTF model. Additionally, off diagonal clusters were found on the comparison of neural models between 750 ms to 1350 ms and 1700 ms to 200 ms. In the case of adults, a clear diagonal was found showing shared auditory representations within the group (Figure 3.4b). The diffusion of the significant timepoints away from the diagonal reflects the generalizability of the neural models at neighbouring timepoints. The diagonal itself also signifies the noise ceiling in each group (Figure 3.4d). In infants, the noise ceiling was only significant during 800 ms to 832 ms, while the significance ranged from 54 ms to 1097 in adults. This suggests that while the adult auditory representations synchronize well within their own group almost throughout the duration of the sound, infants predominantly exhibit temporally non-linear representations which converge at a latency of 800 ms.

Across-group correlations revealed shared auditory representations between infants and adults as most significant clusters were observed between 100 ms to 500 ms in infants and 100 ms to 2000 ms in adults (Figure 3.4c). Although the absence of a diagonal reflects non-existence of one-to-one mapping, large-scale temporal

dynamics are still shared between the two populations at off-diagonal timepoints. Overall, these findings suggest that auditory processes in infants are non-linear and require more time to develop feedforward processes that are typically observed in adults, especially in response to sounds early in time.

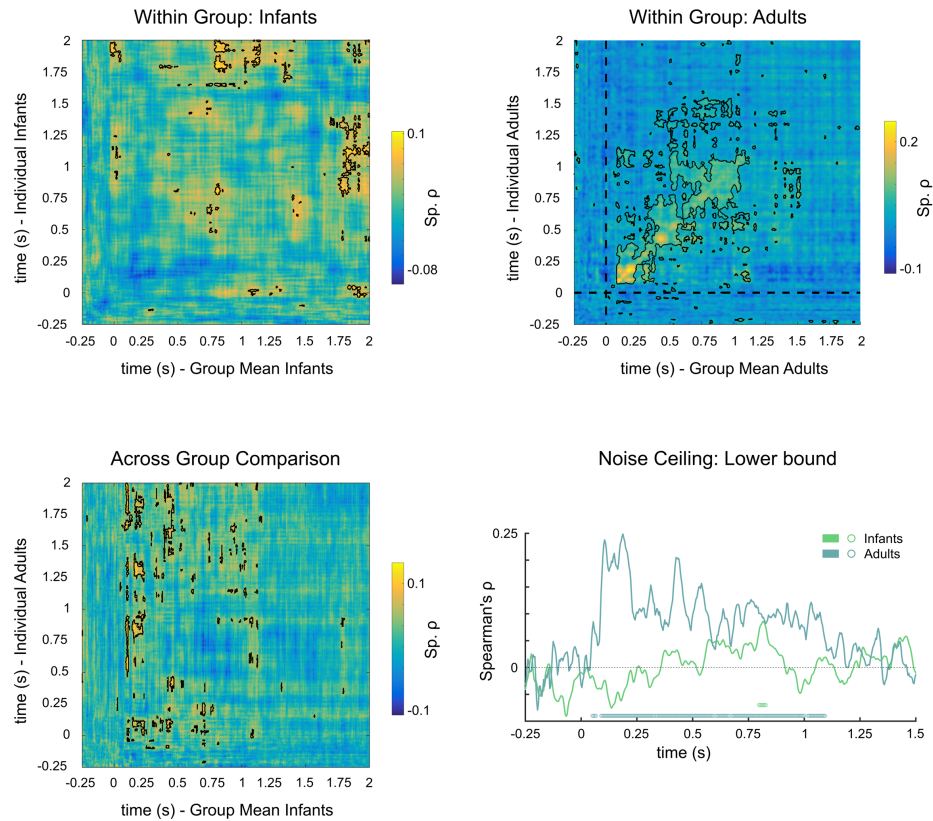


Figure 3. 5 Time generalization averaged across participants. a): Correlations between EEG DSMs within the infant group at all combinations of timepoints predominantly reveal nonlinear dynamics. b) Within the adult group, auditory representations are more consistent as depicted by the diagonal. c) Comparison of auditory representations between infants and adults. d) Lower bound of the noise ceiling of adults and infants. Black outlines denote significant clusters (right tailed sign permutation test, $p < 0.05$, cluster based correction).

3.4 Discussion

In this study, we explored and compared the large-scale temporal dynamics of auditory processing and categorization in infants and adults. While the findings of infants presented are preliminary and will be further refined with a larger sample of infants, they already provide significant insights into auditory object representation in the infant brain.

The first notable difference between the two groups was observed in the AUC curve of individual sound decoding, its peak and latency. In adults, the early sharp increase of the decoding curve aligns with the established notion that responses to auditory features emerge within the first 100 ms post sound onset (Giordano et al., 2013; Ogg et al., 2020). In comparison to the sharp rise and narrow classification peak in adults, infants exhibited a slow rise, broad peaks and a gradual decline in their decoding curve until sound offset. Similar dynamics were also found in infants categorizing visual objects, independently of the specific category presented (Xie et al., 2022). Strikingly, only the decoding curve of infants showed another rise after sound offset, with a latency overlapping that of the representation of early Deep Neural Network (DNN) layers indicating recurrent sound representations which are absent in adults (see layer 4 in Fig 3.4b).

The late emergence of the peak in infants compared to adults suggests an immature processing of sound objects that could be partially explained by neurophysiological differences between infants and adults. For instance, myelination progresses temporally and spatially between 3 and 11 months (Deoni et al., 2011) and adult-like synaptic connectivity is reached in some brain regions at 15 months of age (Huttenlocher & Dabholkar, 1997). Another possible explanation for the lower classification curve in infants is a poorer signal-to-noise ratio (SNR) compared to adults, along with a smaller number of trials available for analysis for infants. We cannot also exclude the possibility of differences between age groups. We chose not to equalize the SNR in adults to infants to maximize the likelihood of identifying similarities between the two populations while also preserving the true characterization of auditory processing in adults. Although low SNR presents challenges in determining whether a higher decoding AUC is driven by noise or the actual signal, these difficulties are mitigated by focusing on peak interpretation, as peak size, but not latency, depends on SNR (Xie et al., 2022).

The acoustic representation of sounds in infants was also found to have a higher latency. We selected the modulation transfer function (MTF, Chi et al., 2005) to model spectral-temporal processing in the primary auditory cortex, which emerged early in adults but later in infants. In adults, the latency of MTF representation found in our study aligns with other studies investigating acoustic processing (Lowe et al., 2021; Ogg et al., 2020). Here, we provided evidence of feedback processes in adults that facilitate acoustic representations earlier in time before the full presentation of the sounds, leading to a faster and more efficient sound categorization process. A potential reason for this could also be the repetitive presentation of sounds in the experiment causing habituation of the sounds. These feedback processes were found to be absent in infants as the acoustic representation only emerged after the presentation of the sound. Interestingly, the latency of acoustic representation of infants (800 ms to 960 ms, peak = 0.0755) overlapped with the latency of the peak

of the lower bound of the noise ceiling (800 - 832, peak = 0.085). This suggests that the most consistent auditory representation in infants may be linked to the spectral-temporal processing of sounds, as it can explain most of the variance during this duration.

Importantly, we showed that in addition to adults, infants encode categorical representations to natural sounds using the 8-categorical model. This finding is particularly noteworthy, especially when considering that prior studies failed to detect categorical information in infants for different categories, and at different ages (Grossmann, 2010; Spriet et al., 2022). However, we did not find any significant categorical representation of abstracted categories in infants using either the 4-category model (animals, humans, big and objects) or the 2-category model (animate vs. inanimate; Supplementary Figure S3.1). In contrast, these abstracted representations were observed in adults. These results demonstrate the presence of finer categorical representation in infants, yet they also highlight the absence of conceptual abstraction in the developing brain.

One of the striking findings of this study was the unique and early representation of mammalian sounds in infants, which was absent in adults. We did not find any selectivity for birds or for all animal sounds, reflecting the infant brain's unique selectivity to non-human mammalian sounds. Object categorization in 3-4 month old infants has been previously found to be supported by the vocalization of blue-eyed black lemur (*Eulemur Macaco Flavifrons*, Ferry et al., 2013). Additionally, our results indicated a delayed yet significant emergence of the human verbalization (HV) category in infants, compared to its much earlier emergence in adults. No unique categorical representation was found for human non-verbalizations, or for the abstracted category including all human sounds in infants. This finding echoes a previous study where infants matched speech sounds to human but not non-verbal sounds (Vouloumanos et al., 2009). This also aligns with the finding of enhanced responses to infant directed speech sounds compared to silence and backward speech in neonates (Peña et al., 2003), suggesting an early sensitivity to certain phonetic formats which are lost in backward speech. Infants are known to exhibit sensitivity to temporal modulations relevant for phoneme perception (Telkemeyer et al., 2009), when compared with tones (Ferry et al., 2010).

It has been suggested that at 3-6 months of age, cognitive development and early language acquisition are engaged by a broader set of natural sounds including non-human primate vocalizations and human speech sounds, although no preference has been reported of the former over the latter (Carr et al., 2021; Ferry et al., 2013; Perszyk & Waxman, 2016; Vouloumanos et al., 2010). However, after 6 months, non-human primate vocalizations no longer confer this cognitive advantage that persists for speech (Carr et al., 2021). While supporting these studies, our results

provide further insights into the temporal dynamics of these categorical selectivities. We found different temporal trajectories of the categorical representations of non-human mammals and HV in infants, suggesting different neural processes underlying these predilections. Interestingly, HV representation emerged during the same time as the acoustic processing. Although this categorical selectivity remained significant after partialling out the MTF model, the correlation weakened, suggesting that this representation was only partially independent of the sounds' spectral and temporal properties (Supplementary Figure S3.2). This aligns with the results of fast periodic auditory stimulation (FPAS) implemented in infants where in addition to the standard sequence eliciting voice selective response, the frequency scrambled sequences also evoked a target response, suggesting differences in the processing of low-level features between voices and objects (see Introduction section 4 or Calce et al., 2024). Although the target stimuli set in Calce et al., 2024 consisted of both human verbal and non-verbal sounds, by separating them as target stimuli in different sequences, it would be interesting to find differences between the two groups in future FPAS studies. Also in adults, our results found reduction in correlations with the HV model after controlling for the MTF model, though they remained significant. This suggests that the development of acoustic processing may at least partially support speech processing and consequently language skills. Conversely, the mammalian representation emerged earlier than the acoustic representation, indicating alternate auditory mechanisms subserving this selectivity. Future research and usage of biologically realistic auditory models could potentially reveal auditory processes in infants that cannot be simply explained by processing of spectral-temporal features of the sounds.

To this end, we used a sound-to-event DNN with 14 progressively complex convolutional layers to predict the neural models at each timepoint in both groups. In adults, the layers provided evidence for the hierarchical processing of auditory objects (Arnott et al., 2004; Bizley & Cohen, 2013), since early layers peaked early in time and later layers later in time. This aligns with the reports of spatially organized hierarchies for auditory object processing in adults, in which the early DNN layers could predict responses from primary auditory regions and later layers represented the belt and parabelt areas, suggesting that the auditory system is a layered organization that supports both sensory and conceptual auditory tasks (Kell et al., 2018; Rupp et al., 2024). However, these layers revealed non-linearities in the temporal order of auditory processing in infants. Notably, the early layers exhibited recurrent processes that unfolded following the emergence of intermediate layers. Even though auditory representations in infants were observed in a format of low to intermediate complexity, the non-linear emergence of these representations do not match the fully developed dynamics of adults.

The non-linearities in infant auditory representation and the variability in their timing

across participants may have contributed to the ‘off-diagonal’ shared auditory representations observed between adults and infants, as well as within the infant group. Additionally, the poor signal-to-noise ratio, a common issue in infant EEG, may contribute to the apparent lack of synchronized auditory processing across infant participants. Nevertheless, using a multivariate pattern approach (MVPA) in conjunction with representational similarity analysis (RSA), instead of univariate methods implemented in previous studies, proves to constitute a novel framework to investigate auditory processes in infants. MVPA allows us to acquire more sensitivity as it relies on the relationships between the features (channels, in our case) and may not directly align with univariate results. In conclusion, this novel approach and usage of biologically realistic models have the potential to reveal mechanisms subserving auditory processing and categorization, as well as to confirm theories on topics related to auditory and cognitive development such as speech processing, object learning and knowledge systems.

3.5 Supplementary Figures

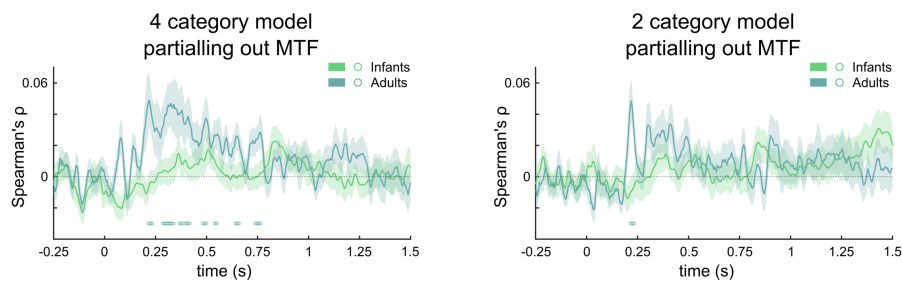


Figure S3. 1 Comparison of the 4- and 2-category model with time-varying EEG after partialling out the acoustic model reveals more abstracted categorical representation in adults, but not in infants.

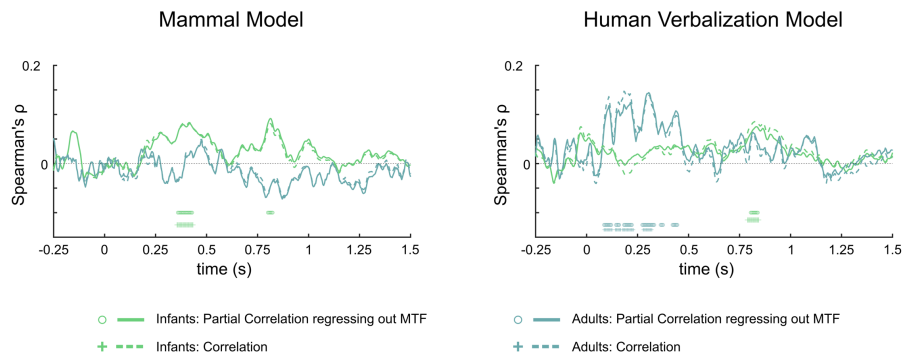


Figure S3. 2 Correlations and partial correlations between EEG and individual models. The early representation of the mammal category in infants remains unaffected by partialling out the acoustics, however the representation of human verbalization reduces after controlling for the MTF model in both infants and adults, yet still remaining significant.

3.6 Bibliography

- Arnott, S. R., Binns, M. A., Grady, C. L., & Alain, C. (2004). Assessing the auditory dual-pathway model in humans. *NeuroImage*, 22(1), 401–408. <https://doi.org/10.1016/j.neuroimage.2004.01.014>
- Aslin, R. N. (2007). What's in a look? *Developmental Science*, 10(1), 48–53. <https://doi.org/10.1111/j.1467-7687.2007.00563.x>
- Barbero, F. M., Calce, R. P., Talwar, S., Rossion, B., & Collignon, O. (2021). Fast Periodic Auditory Stimulation Reveals a Robust Categorical Response to Voices in the Human Brain. *eNeuro*, 8(3). <https://doi.org/10.1523/ENEURO.0471-20.2021>
- Beeck, H. P. O. de, Torfs, K., & Wagemans, J. (2008). Perceived Shape Similarity among Unfamiliar Objects and the Organization of the Human Object Vision Pathway. *Journal of Neuroscience*, 28(40), 10111–10123. <https://doi.org/10.1523/JNEUROSCI.2511-08.2008>
- Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10), Article 10. <https://doi.org/10.1038/nrn3565>
- Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glott International*, 5, 341–345.
- Bracci, S., & Op De Beeck, H. (2016). Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. *The Journal of Neuroscience*, 36(2), 432–444. <https://doi.org/10.1523/JNEUROSCI.2314-15.2016>
- Buiatti, M., Di Giorgio, E., Piazza, M., Polloni, C., Menna, G., Taddei, F.,

- Baldo, E., & Vallortigara, G. (2019). Cortical route for facelike pattern processing in human newborns. *Proceedings of the National Academy of Sciences*, 116(10), 4625–4630. <https://doi.org/10.1073/pnas.1812419116>
- Calce, R. P., Rekow, D., Barbero, F. M., Kiseleva, A., Talwar, S., Leleu, A., & Collignon, O. (2024). Voice categorization in the four-month-old human brain. *Current Biology: CB*, 34(1), 46–55.e4. <https://doi.org/10.1016/j.cub.2023.11.042>
- Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10), 1–1. <https://doi.org/10.1167/13.10.1>
- Carr, K. W., Perszyk, D. R., Norton, E. S., Voss, J. L., Poeppel, D., & Waxman, S. R. (2021). Developmental changes in auditory-evoked neural activity underlie infants' links between language and cognition. *Developmental Science*, 24(6), e13121. <https://doi.org/10.1111/desc.13121>
- Chang, C.-C., & Lin, C.-J. (2011). A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 1–27.
- Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2), 887–906. <https://doi.org/10.1121/1.1945807>
- Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194, 12–24. <https://doi.org/10.1016/j.neuroimage.2019.03.031>
- Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3), Article 3. <https://doi.org/10.1038/nn.3635>
- Cohen, L., Dehaene, S., Naccache, L., Lehéicy, S., Dehaene-Lambertz, G., Hénaff, M. A., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain: A Journal of Neurology*, 123 (Pt 2), 291–307. <https://doi.org/10.1093/brain/123.2.291>
- Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, 85(4), 1330–1345. <https://doi.org/10.1111/cdev.12193>
- de Heering, A., & Rossion, B. (2015). Rapid categorization of natural face images in the infant right hemisphere. *eLife*, 4, e06564. <https://doi.org/10.7554/eLife.06564>
- Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., Kanwisher, N., & Saxe, R. (2017). Organization of high-level visual cortex in human infants. *Nature Communications*, 8(1), 13995. <https://doi.org/10.1038/ncomms13995>
- Dehaene-Lambertz, G. (2000). Cerebral specialization for speech and non-

speech stimuli in infants. *Journal of Cognitive Neuroscience*, 12(3), 449–460. <https://doi.org/10.1162/089892900562264>

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional Neuroimaging of Speech Perception in Infants. *Science*, 298(5600), 2013–2015. <https://doi.org/10.1126/science.1077066>

Deoni, S. C. L., Mercure, E., Blasi, A., Gasston, D., Thomson, A., Johnson, M., Williams, S. C. R., & Murphy, D. G. M. (2011). Mapping Infant Brain Myelination with Magnetic Resonance Imaging. *Journal of Neuroscience*, 31(2), 784–791. <https://doi.org/10.1523/JNEUROSCI.2106-10.2011>

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science (New York, N.Y.)*, 293(5539), 2470–2473. <https://doi.org/10.1126/science.1063414>

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598–601. <https://doi.org/10.1038/33402>

Fantz, R. L. (1964). Visual experience in infants: Decreased attention familiar patterns relative to novel ones. *Science*, 146(Whole No. 3644), 668–670. <https://doi.org/10.1126/science.146.3644.668>

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-Month-Old Infants: An Advantage of Words Over Tones. *Child Development*, 81(2), 472–479. <https://doi.org/10.1111/j.1467-8624.2009.01408.x>

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(38), 15231–15235. <https://doi.org/10.1073/pnas.1221166110>

Fló, A., Gennari, G., Benjamin, L., & Dehaene-Lambertz, G. (2022). Automated Pipeline for Infants Continuous EEG (APICE): A flexible pipeline for developmental cognitive studies. *Developmental Cognitive Neuroscience*, 54, 101077. <https://doi.org/10.1016/j.dcn.2022.101077>

Gennari, G., Dehaene, S., Valera, C., & Dehaene-Lambertz, G. (2023). Spontaneous supra-modal encoding of number in the infant brain. *Current Biology*, 33(10), 1906–1915.e6. <https://doi.org/10.1016/j.cub.2023.03.062>

Gennari, G., Marti, S., Palu, M., Fló, A., & Dehaene-Lambertz, G. (2021). Orthogonal neural codes for speech in the infant brain. *Proceedings of the National Academy of Sciences of the United States of America*, 118(31), e2020410118. <https://doi.org/10.1073/pnas.2020410118>

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26(4), 664–672. <https://doi.org/10.1038/s41593-023-01285-9>

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract Encoding of Auditory Objects in Cortical Activity Patterns. *Cerebral*

Cortex, 23(9), 2025–2037. <https://doi.org/10.1093/cercor/bhs162>

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, 29(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience*, 28(2), 219–236. <https://doi.org/10.3233/RNN-2010-0499>

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, 293(5539), 2425–2430. <https://doi.org/10.1126/science.1063736>

Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *The Journal of Comparative Neurology*, 387(2), 167–178. [https://doi.org/10.1002/\(sici\)1096-9861\(19971020\)387:2<167::aid-cne1>3.0.co;2-z](https://doi.org/10.1002/(sici)1096-9861(19971020)387:2<167::aid-cne1>3.0.co;2-z)

Iamshchinina, P., Karapetian, A., Kaiser, D., & Cichy, R. M. (2022). Resolving the time course of visual and auditory object categorization. 10.

Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016). Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of Neurophysiology*, 115(4), 2246–2250. <https://doi.org/10.1152/jn.01074.2015>

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 17(11), 4302–4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997>

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), 630–644.e16. <https://doi.org/10.1016/j.neuron.2018.03.044>

Konkle, T., & Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, 74(6), 1114–1124. <https://doi.org/10.1016/j.neuron.2012.04.036>

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. <https://doi.org/10.3389/neuro.06.004.2008>

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior & Development*, 6(3), 263–285. [https://doi.org/10.1016/S0163-6383\(83\)80036-8](https://doi.org/10.1016/S0163-6383(83)80036-8)

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object

- Category. *Journal of Neuroscience*, 30(22), 7604–7612. <https://doi.org/10.1523/JNEUROSCI.0296-10.2010>
- Lowe, M. X., Mohsenzadeh, Y., Lahner, B., Charest, I., Oliva, A., & Teng, S. (2021). Cochlea to categories: The spatiotemporal dynamics of semantic auditory representations. *Cognitive Neuropsychology*, 38(7–8), 468–489. <https://doi.org/10.1080/02643294.2022.2085085>
- Mareschal, D., & Quinn, P. C. (2001). Categorization in infancy. *Trends in Cognitive Sciences*, 5(10), 443–450. [https://doi.org/10.1016/s1364-6613\(00\)01752-6](https://doi.org/10.1016/s1364-6613(00)01752-6)
- Mattioni, S., Rezk, M., Battal, C., Bottini, R., Cuculiza Mendoza, K. E., Oosterhof, N. N., & Collignon, O. (2020). Categorical representation from sound and sight in the ventral occipito-temporal cortex of sighted and blind. *eLife*, 9, e50732. <https://doi.org/10.7554/eLife.50732>
- Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32, 89–115. <https://doi.org/10.1146/annurev.ps.32.020181.000513>
- Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1), 1–25. <https://doi.org/10.1002/hbm.1058>
- Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology*, 10(4), e1003553. <https://doi.org/10.1371/journal.pcbi.1003553>
- Ogg, M., Carlson, T. A., & Slevc, L. R. (2020). The Rapid Emergence of Auditory Object Representations in Cortex Reflect Central Acoustic Attributes. *Journal of Cognitive Neuroscience*, 32(1), 111–123. https://doi.org/10.1162/jocn_a_01472
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2022). *FieldTrip* (Version ") [MATLAB]. <https://github.com/fieldtrip/fieldtrip> (Original work published 2013)
- Osher, D. E., Saxe, R. R., Koldewyn, K., Gabrieli, J. D. E., Kanwisher, N., & Saygin, Z. M. (2016). Structural Connectivity Fingerprints Predict Cortical Selectivity for Multiple Visual Categories across Cortex. *Cerebral Cortex*, 26(4), 1668–1683. <https://doi.org/10.1093/cercor/bhu303>
- Peelen, M. V., & Caramazza, A. (2012). Conceptual object representations in human anterior temporal cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 32(45), 15728–15736. <https://doi.org/10.1523/JNEUROSCI.1953-12.2012>
- Peelen, M. V., & Downing, P. E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 105, 177–183. <https://doi.org/10.1016/j.neuropsychologia.2017.03.033>
- Peña, M., Maki, A., Kovačić, D., Dehaene-Lambertz, G., Koizumi, H.,

Bouquet, F., & Mehler, J. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences*, 100(20), 11702–11705. <https://doi.org/10.1073/pnas.1934290100>

Perszyk, D. R., & Waxman, S. R. (2016). Listening to the calls of the wild: The role of experience in linking language and cognition in young infants. *Cognition*, 153, 175–181. <https://doi.org/10.1016/j.cognition.2016.05.004>

Peykarjou, S. (2022). Frequency tagging with infants: The visual oddball paradigm. *Frontiers in Psychology*, 13. <https://doi.org/10.3389/fpsyg.2022.1015611>

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate-Inanimate Distinction. *Journal of Cognitive Neuroscience*, 28(5), 680–692. https://doi.org/10.1162/jocn_a_00924

Quinn, P. C., & Eimas, P. D. (1996). Perceptual organization and categorization in young infants. In *Advances in infancy research*, Vol. 10 (pp. 1–36). Ablex Publishing.

Rakison, D. H., & Yermolayeva, Y. (2010). Infant categorization. *WIREs Cognitive Science*, 1(6), 894–905. <https://doi.org/10.1002/wcs.81>

Rupp, K. M., Hect, J. L., Harford, E. E., Holt, L. L., Ghuman, A. S., & Abel, T. J. (2024). A hierarchy of processing complexity and timescales for natural sounds in human auditory cortex. *bioRxiv: The Preprint Server for Biology*, 2024.05.24.595822. <https://doi.org/10.1101/2024.05.24.595822>

Russ, B. E., Lee, Y.-S., & Cohen, Y. E. (2007). Neural and behavioral correlates of auditory categorization. *Hearing Research*, 229(1), 204–212. <https://doi.org/10.1016/j.heares.2006.10.010>

Spriet, C., Abassi, E., Hochmann, J.-R., & Papeo, L. (2022). Visual object categorization in infancy. *Proceedings of the National Academy of Sciences*, 119(8), e2105866119. <https://doi.org/10.1073/pnas.2105866119>

Telkemeyer, S., Rossi, S., Koch, S. P., Nierhaus, T., Steinbrink, J., Poeppel, D., Obrig, H., & Wartenburger, I. (2009). Sensitivity of Newborn Auditory Cortex to the Temporal Structure of Sounds. *The Journal of Neuroscience*, 29(47), 14726–14733. <https://doi.org/10.1523/JNEUROSCI.1246-09.2009>

Vouloumanos, A., Druhen, M. J., Hauser, M. D., & Huizink, A. T. (2009). Five-month-old infants' identification of the sources of vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, 106(44), 18867–18872. <https://doi.org/10.1073/pnas.0906049106>

Vouloumanos, A., Hauser, M. D., Werker, J. F., & Martin, A. (2010). The Tuning of Human Neonates' Preference for Speech. *Child Development*, 81(2), 517–527. <https://doi.org/10.1111/j.1467-8624.2009.01412.x>

Xie, S., Hoehl, S., Moeskops, M., Kayhan, E., Kliesch, C., Turtleton, B., Köster, M., & Cichy, R. M. (2022). Visual category representations in the infant brain. *Current Biology*, 32(24), 5422–5432.e6. <https://doi.org/10.1016/j.cub.2022.11.016>

CHAPTER 4

Temporal Dynamics of Sound Representation in Sighted and Blind

Siddharth Talwar ¹, Stefania Mattioni ², Eléonore Giraudet ¹, Roberta P. Calce ¹, Francesca M. Barbero ¹, Olivier Collignon ^{1,3}

¹ Institute for Research in Psychology (IPSY) & Neuroscience (IoNS), Louvain Bionics, University of Louvain (UCLouvain), Louvain, Belgium

² Department of Experimental Psychology; UGent, Gent, Belgium

³ School of Health Sciences, HES-SO Valais-Wallis, The Sense Innovation and Research Center, Lausanne and Sion, Switzerland

Abstract

Early blindness triggers a reorganization in brain networks supporting sound processing. How the representational dynamics of auditory discrimination are impacted by visual deprivation remains however unexplored. Using time-resolved multivariate pattern analyses and representational similarity analyses (RSA) applied to electroencephalic recordings and acoustic, behavior, categorical and deep neural networks (DNN) models, we reveal how blindness impacts the cascade of acoustic to semantic coding of natural sounds. Decoding analyses revealed a protracted enhancement of sound representation in congenitally blind (EB) compared to sighted controls (SC), beginning at 152 ms after sound onset and involving the occipital cortex for the entire duration of the enhancement. Temporal generalization showed prolonged sequential auditory processing stabilizing later in time in EB. Models of early acoustic processes explained brain representation early in time with no differences between the two groups. Correlations between brain activity and early to intermediate DNN layers were enhanced in EB around 200 ms likely representing modulations in intermediate acoustic processing. The categorical and behavioral representation of sounds was also enhanced in EB, peaking at 574 ms, predominantly mapping onto occipito-temporal regions. These results reveal a specific pattern of reorganization in brain networks coding for sounds in EB, providing unique insights on the dynamics of functional reorganization in visual deprivation.

4.1 Introduction

A key approach to understanding how sensory experiences shape brain functions is to study individuals with “extreme” atypical sensory experience (Blakemore & Cooper, 1970), like people born without one sense (Rakic, 1988). Congenital or early blindness (EB) has long served as a crucial model system to study whether and how brain functions can be (re)shaped by (the lack of) sensory experience (Sadato et al., 1996).

Building on the accounts that blind people show enhanced acoustic abilities (Gougoux et al., 2004), some studies indicated enhanced responses from temporal brain regions to sounds in EB (Elbert et al., 2002; Gougoux et al., 2004; Manjunath et al., 1998; Naveen et al., 1998; Pascual-Leone & Torres, 1993; Rauschecker, 2002; Röder et al., 2002). However, conflicting results reported reduced activation or sound representation in the same brain area (Bedny et al., 2015; Burton et al., 2002; Mattioni et al., 2022; Pietrini et al., 2004; Ricciardi et al., 2009; Stevens & Weaver, 2009; Striem-Amit et al., 2012; Wallmeier et al., 2015). Moreover, studies investigating the encoding of acoustic properties in the temporal cortex of EB also yielded mixed results. While some studies pointed to the presence of enhanced low-level acoustical features in the occipital cortex of EB (Elbert et al., 2002; Gougoux et al., 2004), other studies did not find any evidence for low-level sound representation in the occipital cortex of EB (Mattioni et al., 2020, 2022).

In contrast, the most reliable expression of the consequence of EB is arguably the enhanced occipital response to sounds (Bavelier & Neville, 2002). For instance, transiently interfering occipital functions with transcranial magnetic stimulation (TMS) induces impaired auditory processing in EB (Collignon et al., 2007, 2009). But what stages of sound processing are encoded in the reorganized occipital cortex of blind people? Interestingly, it has been consistently demonstrated that the occipital involvement in sound processing in the blind show categorical selectivity similar to the one observed for visual processing in sighted people (Dormal et al., 2018; Mattioni et al., 2020, 2022; Peelen et al., 2014; Striem-Amit & Amedi, 2014; van den Hurk et al., 2017; Vetter et al., 2020; Wang et al., 2015). Similarly, posterior oscillatory responses to sounds in blind individuals were found to be similar to visual stimulation in sighted people (Schepers et al., 2012). It has been suggested that part of the categorical coding of sounds could anchor on the typical coding of visual categories in the ventral occipito-temporal cortex (vOTC, Pietrini et al., 2004). This suggests that the higher-order visual areas may preserve their computational role for categorical coding while extending their function to auditory modality in blind individuals (Heimler et al., 2014; Mattioni et al., 2020).

All these findings highlight that blindness causes alterations in the auditory networks responsible for processing and categorizing sounds. Yet it remains elusive how blindness impacts the dynamics of sound processing, i.e., from extracting acoustic features in the early auditory system to the later emergence of categorical representation abstracted from the low-level features in higher-order brain networks (Giordano et al., 2013). In other words, it remains unclear how sighted and blind individuals differ in parsing sound processing into a structured sequence of operations that convert acoustic signals into meaning. This is partially due to a lack

of detailed characterization of the temporal evolution of auditory processes in blind individuals, especially in the context of natural sound processing.

Indeed a consistent endeavor in auditory neuroscience is to map the temporally evolving representations of the acoustic-to-semantic transformation of auditory signals (Giordano et al., 2013). For instance, previous studies have shown that Modulation Transfer Function (MTF) provides a spectral-temporal representation of sounds that can reliably predict responses in the primary auditory cortex (Chi et al., 2005; Santoro et al., 2014). In secondary auditory areas, Deep Neural Networks (DNN) have recently been shown to outperform most auditory models including MTF and others based on pitch, loudness, harmonicity among others (Giordano et al., 2023; A. J. E. Kell et al., 2018). The layers of DNNs progressively encode increasingly abstract representations of natural sounds that may explain and bridge early acoustic processing in primary auditory regions with the semantic and categorical representations. Lastly, semantic embeddings and categorical coding of natural sounds have been found in the inferior temporal regions and the VOTC, even in sighted people (Lowe et al., 2021; Mattioni et al., 2020). Therefore, the comparison of computational models of sound representation with time-resolved electrophysiological brain recordings can provide a framework to elucidate the propagation of auditory functions in EB, from their sensory encoding to their abstract meaning, further revealing the stages of sound processing which are altered by EB.

Here, we compared the dynamics of sound representation between EB and sighted control (SC) individuals using multivariate pattern decoding and representational similarity analyses (RSA, Kriegeskorte et al., 2008; Mattioni, 2024) applied to high-density electroencephalographic recordings (EEG). Using a diverse set of natural sounds belonging to multiple categories, we concomitantly probed alteration in the brain encoding of low-level acoustical (e.g. MTF) and high-level (e.g. categorical) features of the afferent auditory input. Additionally, we directly compared the brain's response of the two groups with the representation of different DNN layers (YAMNet), which can emulate how auditory information is sequentially extracted in the human brain (Giordano et al., 2023). We found that blindness does not impact early acoustic encoding of sounds in the temporal cortex but enhances the representation of intermediate and higher-level stages of sound processing later in occipital-temporal regions.

4.2 Methods

4.2.1 Stimuli

We employed 24 natural sounds adapted from Mattioni et al., (2020) and categorized into two superordinate groups (animate and inanimate), four ordinate groups

(animals, humans, big objects/places and manipulable objects), and eight subordinate categories (birds and mammals within animals; human verbalizations and non-verbalizations within humans; big mechanical objects and environmental scenes within big objects/places, graspable objects and tools within manipulable objects), with three unique sounds per subcategory (Figure 4.1). Each stimulus had a duration of 1 second, and the root mean square (RMS) values of the stimuli were normalized. To prevent clicking, 10 ms ramps were applied at the beginning and end of each sound. Stimuli can be accessed at https://github.com/sid0309/CatinTime_Stimuli.git

4.2.2 Participants

Eighteen early blind (EB) and eighteen sighted controls (SC) participated in the experiment and were matched for gender and age (12 males in both groups, Age: EB mean = 41.38, S.D. = 10.42, SC mean = 41.66, S.D. = 12.38, two-sample t-test, $t_{(23)} = 0.058$, $p = 0.9541$). No history of hearing or neurological disorders were reported in both groups. Participants gave written consent after being fully informed about the aim and procedure of the study and they received monetary compensation. The experiment took place at the University of Louvain and was conducted according to the Declaration of Helsinki and under the approval of the ethical committee of the University of Louvain (Project 2016-25).

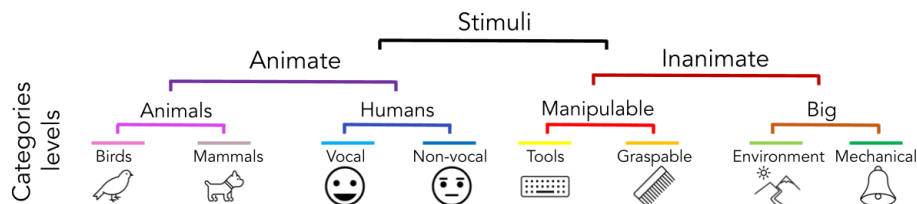


Figure 4. 1 Stimuli set used in the experiment ($n = 24$) that can hierarchically be divided into 2-, 4- and 8-categories.

4.2.3 Data Acquisition

4.2.3.1 EEG Experiment

Before the EEG experiment, all sounds were presented in a randomized order to the participants to familiarize them with the stimuli and ensure correct recognition. During the EEG experiment, sounds were presented in 15 runs. Each run consisted of 3 pseudo-blocks unbeknown to the participant, containing the presentation of all 24 sounds resulting in a total of 72 sounds per run. Thus, each sound was presented once before any sound was repeated and twice before any sound was heard for the third time. The order of presentation within each pseudo-block was randomized. Participants performed a one-back recognition task, pressing a button if they heard

two consecutive identical sounds. These 6 catch trial sounds were additionally inserted pseudo-randomly in each run (2 within each pseudo-block), where the catch trial sound was the same as the previous sound. The inter-trial interval (time between end of previous trial and beginning of the next) was set to 1.5 seconds, plus a jitter computed randomly from a Gaussian distribution with a mean of 0 and a range of ± 0.25 seconds. Each run lasted for approximately 3.5 minutes. Participants from both groups were blindfolded during the experiment to ensure the eyes remained closed and minimise ocular artifacts in the EEG recording. All experiment codes were implemented on Psychtoolbox and MATLAB 2016b.

Participants performed the experiment seated in a dimly lit room at approximately 1 m from the speakers. The sound intensity was kept at a comfortable level of approximately 60dB. EEG was acquired with a Biosemi ActiveTwo System (<https://www.biosemi.com/products.htm>) using 128 channel Ag/AgCl electrodes. Recording sites included the standard 10-20 system locations as well as intermediate positions (channel coordinates can be found at: <https://www.biosemi.com/headcap.htm>). EEG was acquired at a sampling rate of 1024 Hz. The magnitude of offset of all electrodes that are referenced to common mode sense (CMS) and driven right leg (DRL) was kept below ± 25 mV. After the EEG experiment, electrodes and fiducials positions were digitized by capturing 3D images of the participants using Structure Sensor Pro (<https://structure.io/structure-sensor>) connected to an 8th generation iPad (Homölle & Oostenveld, 2019). Electrodes and fiducials positions were then later used for co-registration with the individual MRI T1-image.

4.2.3.2 Behavioral Experiment

After the EEG session, the participants completed a behavioral experiment in which they rated the pairwise similarity between the stimuli. For each of the 276 possible pairs of sounds, they rated how similar the 2 sounds were on a scale from 1 (most dissimilar) to 7 (most similar).

4.2.3.3 MRI Structural Images

High-resolution anatomical images were acquired within a different project that involved some of the participants reported in this study (Barilari et al, in prep.) using an MP2RAGE sequence (Marques et al., 2010). To create a T1-weighted uniform (UNI) image, two whole-brain images with different inversion time (TI1, TI2) were acquired and then combined (repetition time, TR = 4300 ms; echo time, TE = 2.27 ms; echo spacing = 7.2 ms; repetition time excitation = 4.54 ms; slices per slab = 192 slices; voxel size = 0.75 mm (isotropic), matrix size = 192 x 300 x 320 and flip angle = 4°, 4°, GRAPPA = 3, TI1 = 1000 ms and TI2 = 3200 ms). Scans of 9 early blind participants and 8 sighted controls out of the total of 36 participants were

acquired. For the rest of the participants, analysis was performed on a common MNI template.

4.2.4 Analysis

4.2.4.1 EEG Preprocessing

Data was preprocessed using Fieldtrip toolbox (Oostenveld et al., 2011) and custom scripts running on MATLAB 2016b. A fourth order Butterworth band-pass filter from 0.1 Hz to 100 Hz was applied on continuous EEG data. Additionally, a fourth order Butterworth band-stop filter at 50 Hz was implemented to remove power-line noise. Continuous EEG data was segmented into 1080 trials (24 sounds X 15 runs X 3 repetitions) of 2.5 seconds length, including 0.5 seconds before stimulus onset and 2 seconds post onset (i.e., 1 seconds of sound presentation and 1 second post-stimulus). Baseline subtraction was applied on the extracted trials using the mean pre-stimulus activity as the baseline window. Then, the trials were down-sampled to 256 Hz to facilitate data storage and handling, providing a temporal resolution of 3.9062 milliseconds. A semi-automatic artifact rejection pipeline was employed: trials with absolute amplitude exceeding 250 μV were deleted and trials with a maximum absolute amplitude between 150 μV and 250 μV were visually inspected and potentially discarded. Lastly, all trials were re-referenced to common average referencing.

4.2.4.2 Time Resolved Sound Decoding

All analyses were conducted in MATLAB 2017b. Decoding analyses were conducted using the LibSVM library (Chang & Lin, 2011). Source-time series was computed using Fieldtrip toolbox (Oostenveld et al., 2011).

4.2.4.2.1 Decoding Across All EEG channels

The preprocessed trials were first divided into their corresponding sounds ($n=24$) such that there were at least 40 trials per sound for any participant. To improve signal-to-noise ratio, 3 sets of 13 trials were chosen randomly and averaged separately to form 3 pseudo-trials per sound. Further analysis was performed on these 72 pseudo-trials (24 sounds x 3 pseudo-trials). A timepoint-by-timepoint binary decoding of each comparison (276 binary comparisons across 24 sounds) was implemented using a 3-fold cross validation with a 'leave one trial out' strategy, i.e., training on four trials and testing on remaining two. Support Vector Machines (SVM) were used for training and classification, as they are generally better at handling high-dimensional feature sets (Grootswagers et al., 2017). Area under the curve (AUC) was computed as a metric to quantify decoding performance, providing a measure of both specificity and sensitivity. The AUC values from all folds were then

averaged, yielding a decoding metric for each time point and comparison. The entire process from creating pseudo-trials to computing the binary decoding AUC at each time-point was repeated 100 times, to reduce the arbitrariness in creating pseudo-trials and ensure a reliable decoding AUC. For temporal generalization, we performed training and testing trials at all combinations of time points with the same procedure as above (King & Dehaene, 2014).

4.2.4.2.2 Sensor-Space Searchlight

Sensor-space searchlight analysis was performed to find the channels contributing to the decoding AUC as a function of time (Kaiser et al., 2016; Kriegeskorte et al., 2006). Using the same procedure as decoding across all channels (section 2.3.2.1), we computed decoding AUC for each channel separately. The seed channel itself and 5 neighboring channels underwent timepoint-by-timepoint binary decoding across 276 comparisons. This provided a decoding AUC for each channel, comparison and timepoint. Finally, the AUC were averaged within different time-windows and were illustrated as topoplots in Figure 4.2b. The time-windows were chosen based on the significant clusters of difference found between the two populations from the decoding analysis across all channels.

4.2.4.2.3 Source-Space Searchlight

To localize brain regions contributing to the decoding observed in the sensor space analyses, we performed analyses in source space. For each participant, a head model was created by reslicing and segmenting individual T1 images, which included the brain, skull and scalp volumes (boundary element method, BEM). To obtain accurate electrode placements, individual co-registration was implemented by first identifying the fiducial locations (nasion, left and right preauricular) in the T1 images and then matching them with the participant-specific fiducials and channel locations. A common MNI template and Biosemi channel location template were used for the participants who were unavailable or incompatible for MRI. Lastly, source points were placed 10 mm apart in the brain of each subject.

For each participant, a spatial filter was calculated at each source point using Linear Constraint Minimum Variance (LCMV, Van Veen et al., 1997). Preprocessed EEG data was then applied to the spatial filters to reconstruct the time series at each source point. The resulting source-time series was averaged within time windows of 20 ms to reduce computational load. Finally, a searchlight decoding analysis was performed on the simulated time-series using a 15 mm sphere at each source point, such that the source points within the sphere were considered as features for the classifier (as opposed to channels as in the sensor level analysis). This yielded a decoding AUC for each source point, comparison, time window and participant. The

classification procedure was the same as the analysis performed at the sensor level. Post statistics, significant clusters (see section 2.3.4) were interpolated onto the Desikan-Killiany atlas to identify the locations (Desikan et al., 2006).

4.2.4.3 Representation Similarity Analysis

RSA allows us to abstract the neural responses from the multivariate space and find shared representation between neural geometries and representational models of sounds across time (Kriegeskorte et al., 2008). The essence of RSA is the dissimilarity matrices (DSM) which consists of distances between pair-wise conditions. Thus, the comparison of neural DSMs and external representational models can reveal how information unfolds over time and across spatial regions (Cichy et al., 2014). In our study, we used RSA to investigate the unfolding of different neural representations of sounds through time in early blind people in comparison to sighted controls.

We created sensor-level EEG dissimilarity matrices at every time-point (~4 ms) using sensor space decoding AUC across all pairwise sound comparisons. These neural models were smoothed by averaging using a moving time window of 5 bins length at each time-point (2 bins on either side of the central time bin, 5 bins ~20 ms). DSMs were also computed at the source level using the decoding AUC derived from source searchlight decoding analysis. Each source point was represented by a 24x24 DSM at each time window (length of 20 ms). Since the DSMs are symmetric matrices with a diagonal of zeros, we used only the lower triangle, excluding the diagonal, for all analyses.

Four groups of external representational models were used to characterize the sounds: acoustic features, a priori categorical models, behavior ratings based on perceived similarity of sound pairs and models created from a deep neural network.

The acoustic model was computed from the modulation transfer function using NSL tools (MTF, Chi et al., 2005), which is inspired from the psycho-acoustical and neuro-physiological observations in the early and central auditory systems. The model is composed of two parts – the early acoustic transformation through a band of 128 ERB filters mimicking the processing in the cochlea and the midbrain, and the later cortical stage that simulates processing in the primary auditory cortex. For our analysis, we considered only the cortical simulation part in which the MTF provided a 4-dimensional representation of the sound's spectral and modulation features including: temporal window (ms), frequency (Hz), modulation-rate (up vs. down) and scale (Ω). The modulation spectrum was computed using 2D modulation selective filters tuned to default temporal modulation frequencies $w = 4, 8, 16, 32$ Hz and spectral modulation frequencies $\Omega = 0.25, 0.5, 1, 2, 4, 8$ cycles / octave. Finally, the

multidimensional structure was averaged across time to obtain a total of 6144 features for each sound. To further test brain representation of other low-level acoustic features, DSMs based on center of gravity (COG), harmonicity-to-noise ratio (HNR) and pitch were computed using Praat (Boersma & Weenink, 2001). Cosine distance between each pairwise sound representation was computed, providing dissimilarity measures in 24 x 24 dimension for all acoustic models (24 sounds in total).

To create the behavioral models we acquired the similarity ratings from each participant and created a DSM from the perceived responses from each pair of sounds for each individual (we flipped the scale to form dissimilarity from similarity ratings). Categorical models were created based on the a priori memberships of sounds at each level of the categorical hierarchy (2, 4, 8 categories, Figure 4.1). Thus, the distance between each comparison of sounds was set to 1 if they belonged to different categories or 0 if they had the same category membership.

Recent studies suggest that state-of-the-art convolutional ‘deep’ neural networks (DNNs) capture important aspects of human sound perception (Giordano et al., 2023; A. J. E. Kell et al., 2018). To create representational models based on different layers of a DNN, we chose YAMNet, a feed-forward convolutional deep neural network developed by Tensorflow (Abadi et al., 2015?) and pretrained on the Google Audioset corpus. The selection of this specific DNN was based on results from a previous study (Giordano et al., 2023) showing that YAMNet is the best model for predicting responses from superior temporal regions, outperforming other acoustic models and other DNN. By providing an auditory stimuli as an input (sampled every 0.96), YAMNet provides a prediction from 521 audio events from the AudioSet ontology. The Mobilenet_v1 architecture includes an input layer and a standard convolutional layer, followed by 13 depthwise separable convolutional layers. This provides a 3 x 2 dimension of 1024 kernels which are averaged and put through a single logistic layer of 521 audio ratings. Sound embeddings from all 14 layers were considered for analysis. Scripts for implementation in Python and MATLAB were adapted from. For all DNN layers, cosine distance between each pairwise sound representation was computed, providing dissimilarity measures in 24 x 24 dimension.

Participant specific neural models were correlated with external representational models. We first correlated the sensor-level neural models with the external representational models to identify the time points when these models were significantly represented in the blind. Next, we correlated the source-level neural models with the same external models during this specific time window(s) to pinpoint the brain regions that may encode these representations in the blind. Correlations

and partial correlations were computed at each time point using Spearman's rank correlation. For the behavior models, participant-specific ratings (rather than group averages) were used to correlate with their respective neural DSMs.

To reveal the consistency of neural representations within each group, each participant's neural DSM was correlated with the mean of their respective group for each time-point combination, while taking out that particular participant from the computation of the group mean (individual EB with mean EB and individual SC with mean SC). A diagonal from bottom-left to top-right corner reflects temporal synchrony between the two estimates. The diagonal also reflects the lower bound of the noise ceiling in each group, representing a threshold above which correlations with any external model indicate that the model successfully predicts the variance in the neural data (Nili et al., 2014). For intergroup comparison, each individual EEG DSM was correlated with the mean of the other group (individual EB with mean SC and individual SC with mean EB). We averaged the correlations of each EB participant with the same of its respective matched SC and subsequently performed statistics of the estimate. Multidimensional scaling was computed for a given DSM using the *cmds* command in MATLAB R2016b.

4.2.4.4 Statistical Analysis

One-sample sign-permutation tests were performed to find the time-points when decoding AUC or Spearman's rank correlation were significantly higher than chance level by first subtracting chance level (0.5) before sign flipping. For one-sample test, at each time-point, the individual values were multiplied with either +1 or -1 (random sign flip) and then the group average was computed. 5000 sign-flip permutations were computed to create a null distribution, and a p-value was determined by calculating the proportion of values in the null distribution higher than the observed empirical mean. To find significant differences in decoding and correlations between the two populations, a two-sample sign permutation test was conducted using the same procedure except that sign flipping was performed separately in both populations and that the null distribution was created by computing the difference of means of EB and SC in each iteration. Two-sample tests were also applied for source space decoding AUC and correlations, in which during a specific time window, sign-flips were computed at each source point.

Subsequently, clusterwise inference (Nichols & Holmes, 2002) in 1-D (sensor space decoding and correlations), 2-D (temporal generalization, sensor space searchlight decoding, within-group and inter-group correlations) and 3-D (source searchlight decoding and correlations) were employed to correct for multiple comparisons by converting the null distribution to p-maps across all time points or source points and determining the largest cluster thresholded at $p < 0.05$. All significant empirical time-

clusters that were bigger than the largest cluster found in the null distribution survived the correction.

Bootstrap tests were implemented to estimate confidence intervals (95%) of the temporal onset of the first significant cluster, offset of the last significant cluster and peak latencies. We created 1000 bootstrapped samples by sampling the participants with replacement. The same statistic as above was applied to each bootstrap which resulted in an empirical distribution of onset, offset and peak latencies on which the confidence intervals were determined.

4.3 Results

4.3.1 Enhanced Sound Decoding in Blind Individuals

We compared the temporal dynamics of the EEG responses to sounds in blind and sighted people. Auditory responses to different sounds were characterized by performing multivariate pairwise classifications across all 24 sounds (e.g., sound 1 vs sound 2, sound 1 vs sound 3 etc) at each time-point. Figure 4.2b represents the mean of all 276 pairwise comparisons revealing enhanced decoding AUC in early blinds (EB) in comparison to matched sighted controls (SC). The following results present the peak latency, onset, and offset of significant time points, with 95% confidence intervals, calculated from bootstrap tests, presented in square parentheses. In EB participants, significant sound decoding (AUC) began at 53 ms [33-72 ms], peaked at 183 ms [175-226 ms] and remained significant until 1996 ms post-stimulus onset [1929-1996 ms]. A similarly shaped but lower classification curve in SC participants was significant from 59 ms [41-72 ms], peaked at 175 ms [163-951 ms] and ended at 1949 ms [1902-1992 ms]. Importantly, significant differences between the two groups (EB vs. SC) started at 160 ms [59-263 ms] and continued until 1007 ms [363-1988 ms]. Within this extended period of enhanced sound classification in EB, seven distinct significant time clusters were identified (see Figure 4.2b).

Further, we localized channels and brain areas that contributed to the enhanced decoding in the blind across these seven time-clusters employing sensor searchlight decoding for channels and source searchlight decoding for brain regions (Figure 4.2c). The largest difference between the two groups was found in the first time-cluster (152-285 ms) involving anterior and posterior channels. Source analyses revealed that many brain regions displayed significant enhanced decoding in EB; in the occipital cortex: bilateral cuneus, lateral occipital cortex (LOC), lingual, pericalcarine, bilateral fusiform right parahippocampal regions; in temporal regions:

Chapter 4. Temporal Dynamics of Sound Representation in Blind and Sighted

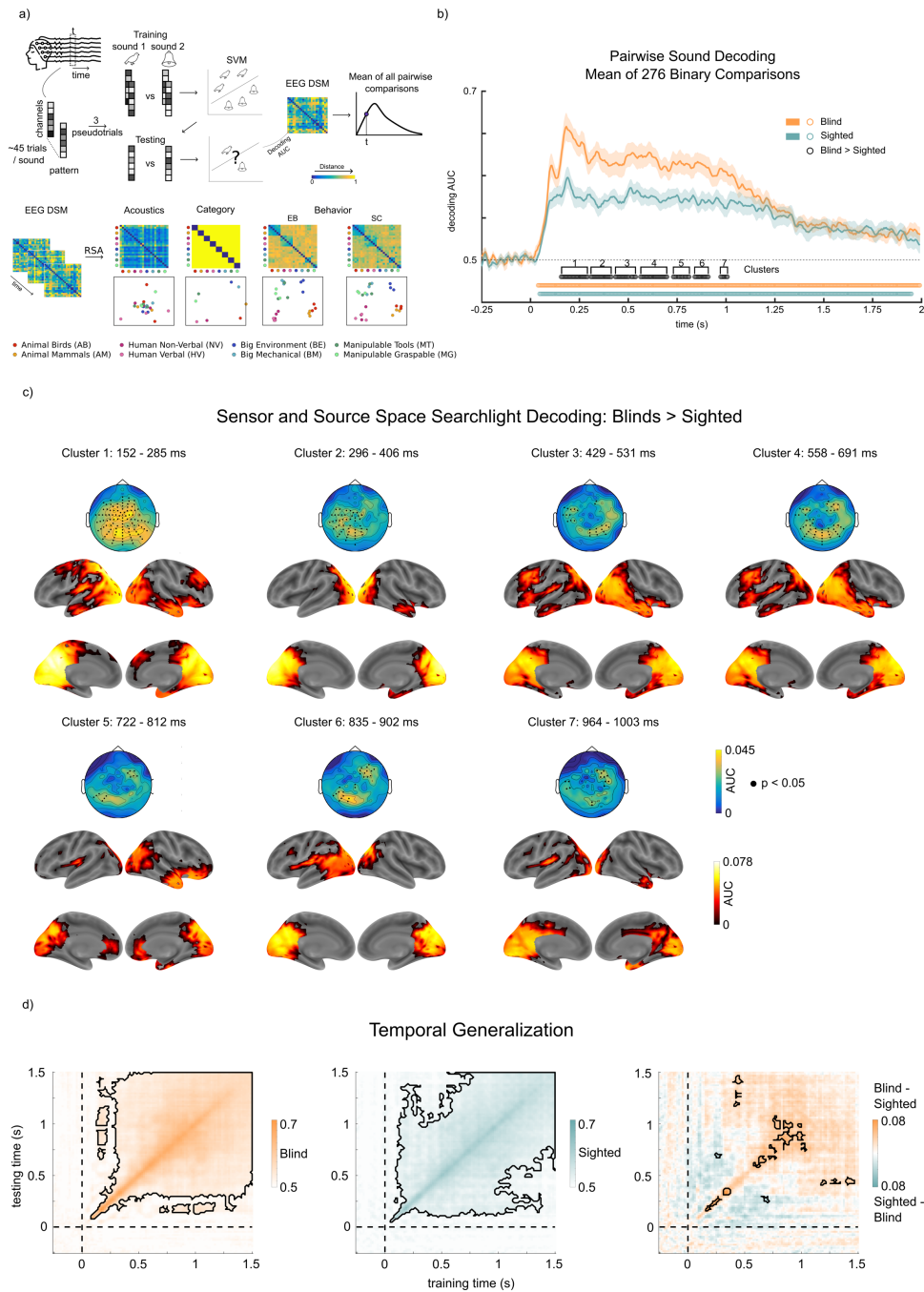


Figure 4. 2 Decoding pairwise sounds in blinds (EB) and sighted (SC). a) Pipeline illustrated for decoding analysis and representational similarity analysis. b) Enhanced decoding in EB across all sensors with the

black dots reflecting the difference, resulting in 7 visually identified time-clusters. Shaded region displays the Standard Error of Mean (SEM). c) Sensor and source space decoding show enhanced involvement of posterior channels and brain regions, respectively in EB in comparison to SC. d) Temporal Generalization reveals the dynamics of auditory processing in EB and SC. On contrasting EB and SC, a series of transient neural representations unravel for a prolonged time in EB, while in SC, the representations stabilize and diffuse early. All statistics were computed using sign permutation test ($p < 0.05$) and 1-D or 2-D cluster-wise correction.

bilateral inferior and middle temporal, right superior temporal cortex; other areas: precuneus, inferior parietal, superior frontal, insula, left cingulate, supramarginal area. The magnitude of difference diminished in other time-clusters however notably, the posterior channels and occipital brain areas such as bilateral cuneus, LOC, pericalcarine, lingual as well as precuneus displayed higher decoding AUC in all time-clusters in EB. All significant brain areas within other time clusters can be found in the Supplementary Table S4.1.

Temporal Generalization analysis allowed us to reveal how stable or dynamic neural representations are over time (King & Dehaene, 2014). We found significant decoding AUC in EB and SC commencing at 53 ms and 59 ms respectively along the diagonal (Figure 4.2d). The curves in both groups started narrowly along the diagonal indicating evolution of different representations where each time point's neural representation influenced subsequent ones. The narrow curves shortly broadened over time suggesting that the transient neural representations were transformed to more stable and diffused representations over time. By contrasting the two groups (EB - SC), we found higher decoding AUC in EB along and around the diagonal both early (~168-385 ms) and later in time (~601-1004 ms). These results overall suggest that different neural representations in the blind sequentially unfold almost until sound offset and then generalize later in time, while in the sighted the neural representations are sustained and generalize early in time. Some significant clusters were also found off-diagonal which suggest the reactivation of specific neural representations.

4.3.2 Disentangling Acoustic, Behavior and Categorical Representations

Multivariate pattern decoding analysis revealed the temporal evolution of neural patterns in response to sounds in EB and SC. However, it could not disentangle between the different dimensions of sound processing that contributed to the group differences. We relied on Representational Similarity Analysis (RSA) to find the similarities of the representational geometries of sound processing in EB and SC and to find which aspects of sound processing drive the observed decoding enhancement in EB.

Both EB and sighted control groups displayed a strong correlation with the acoustic Modulation Transfer Function (MTF) model early in time (EB: onset: 97 ms [47-101 ms], peak: 199 ms [109-218 ms], offset: 281 ms [238–1748 ms]; SC: onset: 85 ms [74-156 ms], peak: 187 ms [93-242 ms], offset: 214 ms [113-1062 ms]; Figure 4.3a). No significant differences were observed between the two groups at either sensor-level or source-level analysis. To further attest that early acoustical representation is unaltered by EB, we also explored other acoustic dimensions such as Center of Gravity, Harmonicity-to-noise ratio and Pitch with EEG at the sensor level (Supplementary Figure S4.1). While they also emerged early in time in both groups, the three acoustic models also did not reveal any differences in acoustic representation between EB and SC.

We explored how similarly both populations perceive the sounds that were presented to them. DSMs created from individual participants' similarity ratings within a group were correlated with the mean of the other group, yielding 18 correlation values for each group. No significant difference was found between the perceptual ratings of both groups (two-sample sign-permutation test, $p = 0.1454$). We further correlated participant specific EEG and their respective behavior model partialling out the acoustic MTF model (EB: onset: 86 ms [31-92 ms], peak: 617 ms [164-1546 ms], offset: 1632 ms [1595-1992 ms]; SC: onset: 289 ms [234-685 ms], peak: 293 ms [203-1218 ms], offset: 550 ms [535-1916 ms]; Figure 4.3b). Significant differences between the two populations were observed later in time (EB>SC: onset: 609 ms [160-1476], offset: 625 ms [193-1762 ms]). Left temporal regions such as banks-STS, middle and inferior temporal and the fusiform area were found to code for this enhancement in behavior representation of EB compared to SC.

To obtain a categorical representation independent of acoustic properties in both populations, we computed the correlation between the neural models and different category models while partialling out the acoustic MTF model. While SC displayed similar correlation peaks and curves for 8-, 4- and 2-categories models, EB showed a similar trend for 2-categories animacy model *but not* for 8- and 4-categories models (*8-categories EB*: onset: 222 ms [195-292 ms], peak: 480 ms and 578 ms [308-621 ms], offset: 1636 ms [1558-1988 ms]; *8-categories SC*: onset: 183 ms [49-408 ms], peak: 269 ms [148-1636 ms], offset: 1148 ms [519-1988 ms], Figure 4.3c; *4-categories EB*: onset: 89 ms [59-1543 ms], peak: 562 ms, [207-613 ms], offset: 1730 ms [802-1988 ms]; *4-categories SC*: onset: 218 ms [50-394 ms], peak: 218 ms [207-761 ms], offset: 769 ms [574-1874 ms], Figure 4.3d; *2-categories animacy EB*: onset: 210 ms [55-300 ms], peak: 222 ms [210-691 ms], offset: 543 ms [406-1578 ms]; *2-categories animacy SC*: onset: 218 ms [191-394 ms], peak: 218 ms [207-935 ms], offset: 414 ms [226-1976 ms], Figure 4.3e). We found an enhanced categorical representation in EB later in time only with 8- and 4-category models (*8-categories*

EB>SC: onset: 574 ms [234-619 ms], offset: 621 ms [328-1960 ms], Figure 4.3c; 4-categories *EB>SC*: onset: 574 ms [78-1326 ms], offset: 699 ms [621-1960 ms], Figure 4.3d). No difference was found between the groups in the case of the 2-category model. For 8-categories model, source-space RSA attributed these enhancements to bilateral cuneus, inferior and superior parietal, isthmus cingulate, precuneus and right banks-STS, transverse, superior, middle and inferior temporal, supramarginal, insula, parahippocampal, fusiform, lateral occipital, lingual, pericalcarine, paracentral, posterior cingulate regions. Similarly for 4-categories model, we found bilateral banks-STS, transverse, superior, middle and inferior temporal, supramarginal, insula, parahippocampal, fusiform, cuneus, lingual, lateral occipital, pericalcarine, inferior and superior parietal, precuneus, precentral, isthmus cingulate, left caudal middle frontal and right paracentral, postcentral and posterior cingulate regions.

Although EB showed a higher correlation with the 8- and 4-categories model later in time, surprisingly both groups coded similarly for animacy early in time (humans, animals vs objects, big sounds). Why do EB display enhanced categorical sensitivity to 8- and 4- categories but not to 2 (animate vs non animate)? Do they further organize these categories not only based on animacy but also according to an additional conceptual framework? We investigated the qualitative structure of category encoding in EB at ~600 ms by rearranging the sound categories to form another 2-category model. We defined this model as the ‘frequency of interaction’ model (referring it to as the interaction model in this chapter) as it consists of 2 classes, one which EB could frequently interact with (humans, manipulable and graspable objects) and the second where instances of interactions could be limited (mammals and birds, big environmental and mechanical sounds, see discussion for more). Interestingly, we found the 2-category interaction model to correlate with the neural model of EB peaking at a similar latency compared to 8- and 4-categories models (2-categories interaction in EB: 2 categories interaction in EB: onset: 168 ms [130-388 ms], peak: 574 ms [546-617 ms], offset: 1608 ms [910-1928 ms]; 2 categories interaction in SC: onset: 320 ms [195-542 ms], peak: 531 ms [203-855 ms], offset: 656 ms [330-1886 ms]; Figure 4.3f). This model could also best explain the enhancement in EB later in time (*EB>SC*: onset: 546 ms [93-605 ms] , offset: 714 ms [703-1282 ms]). The brain areas representing this enhancement were bilateral inferior and middle temporal, parahippocampal, fusiform, cuneus, lateral occipital, lingual, pericalcarine, inferior and superior parietal, paracentral, postcentral, precentral, precuneus, posterior and isthmus cingulate, and right banks-STS and supramarginal regions.

Chapter 4. Temporal Dynamics of Sound Representation in Blind and Sighted

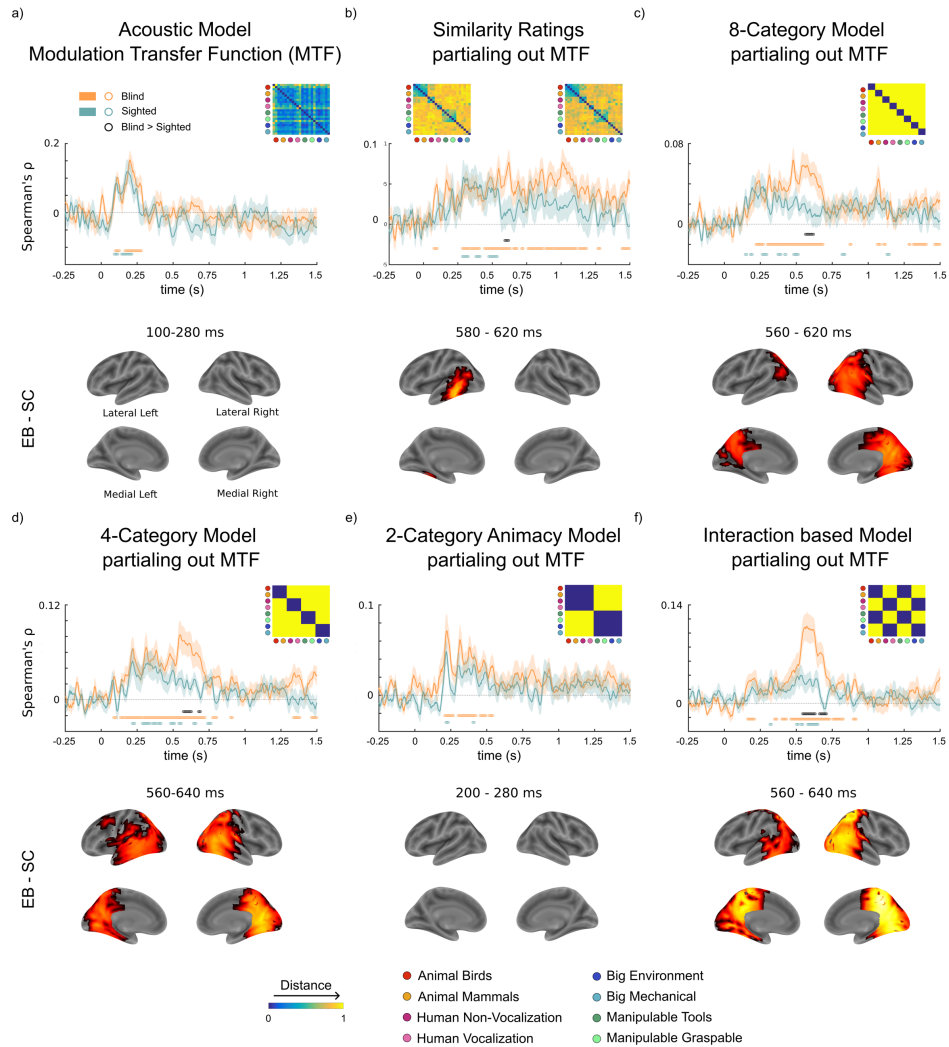


Figure 4. 3 Representation Similarity Analysis performed both at sensor and source levels. a) Comparison of time-varying EEG models and acoustic model- Modulation Transfer Function (MTF) showed no difference between early blinds (EB) and sighted controls (SC). b) Neural models of all participants were correlated with their respective behavior model. An enhancement in EB emerged later in time which can be attributed to the left inferior temporal regions. c and d) Representation of the 8- and 4-category models were enhanced in EB at a similar latency as that of the behavior model in occipital, temporal and parietal regions. e) 2-category animacy model emerged early and similarly in both groups. f) Rearrangement of categories revealed that instead of animate vs inanimate, EB exhibit an enhanced categorical representation to human and object sounds vs animal and big mechanical and environmental sounds. We refer to this as the 2-category interaction model, hypothesizing that EB may naturally cluster frequently encountered sounds (humans and objects), developing a more refined internal representation of their sources, compared to sounds with which they have lesser experience and a weaker internal representation. The 2-category interaction model best explains neural responses occurring ~600 ms. The

acoustic model was partialled out from the correlations of the neural models and all categorical and behavior models. All statistics were performed with the sign-permutation test ($p < 0.05$) with cluster-wise correction.

Indeed, the enhanced representation of individual behavior and 8-, 4-categorical models peaking around the same time (~ 600 ms) in EB is corroborated due to high collinearity between the two external models (EB perceptual model - 8 categorical model Sp.rho , mean = 0.3492, S.D. = 0.0875; SC perceptual model - 8 categorical model Sp.rho within SC, mean = 0.2738, SD = 0.1234). Furthermore, the categorical and behavior models were found to be orthogonal to the acoustic MTF model, emphasizing the absence of acoustic contribution to the enhanced behavioral and categorical representation in EB (Sp.rho EB perceptual model - MTF, mean = -0.0519, S.D. = 0.0666, one sample sign permutation test $p = 0.2774$; Sp.rho SC perceptual model - MTF, mean = 0.0854, SD = 0.0134, one sample sign permutation test $p = 0.9760$).

These results collectively indicate that blindness does not affect early acoustic processing but it modulates the temporal emergence and enhancement of perceptual and categorical representation later in time, which are primarily encoded in occipital, temporal and parietal regions. This enhanced categorical encoding in EB is qualitatively unique, where they cluster humans and objects together, and animals with big sounds.

4.3.3 Deep Neural Network - YAMNet

We found enhanced sound decoding in EB than SC that can be accounted for by enhanced behavioral and categorical representation around ~600 ms. However, similar acoustic representation in the neural data of both groups does not explain the higher decoding AUC in EB earlier in time. If not the acoustics, what aspect of sound processing promoted the decoding enhancement in EB early in time? To investigate an alternate format of sound representation and account for the missing variance in our data, we employed a deep neural network (DNN). The layers of the DNN allow us to capture the transformation and complexity of auditory representations that may explain neurophysiological events occurring during auditory processing and categorization. We used YAMNet, a pre-trained feed-forward convolutional network consisting of 14 layers, that performs classification of sound events based on 521 sound labels (Figure 4.4a). We passed our stimuli through the DNN, extracted the embeddings of each layer, modeled and correlated them with the neural models. We found shared representations between the neural DSM and multiple DNN layers at various time points in EB and SC (Supplementary Figure S4.2). Interestingly, 4 layers expressed differences between the two populations (layers 2 to 5, Figure 4.4b). Following results show the onset, offset and peaks with confidence intervals of these 4 layers. The peaks are also tabulated in Table 4.1.

In EB Layer 2: onset: 86 ms [39-179 ms], peak: 187 ms [152-316 ms] and 1035 ms [820-1050 ms], offset: 1261 ms [1082-1988 ms]; Layer 3: onset: 89 ms [35-179 ms], peak: 187 ms [52-316 ms] and 1035 ms [820-1050 ms], offset: 1617 ms [1082-1988 ms]; Layer 4: onset: 89 ms [27-101 ms], peak: 191 ms [171-328 ms] and 1035 ms [738-1035 ms], offset: 1609 ms [1082-1980 ms]; Layer 5: onset: 124 ms [35-126 ms], peak: 144 ms [128-661 ms], offset: 1593 ms [527-1980 ms].

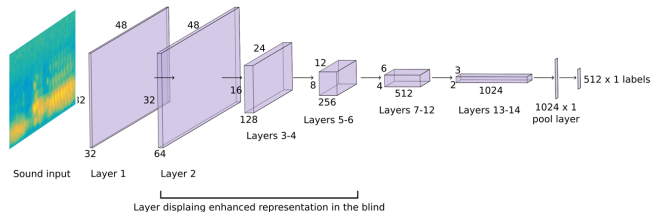
In SC Layer 2: onset: 125 ms [39-832 ms], peak: 132 ms [125-359 ms], offset: 914 ms [144-1958 ms]; Layer 3: onset: 128 ms [42-308 ms], peak: 137 ms [128-953 ms], offset: 1046 ms [69-1988 ms]; Layer 4: onset: 105 ms [35-126 ms], peak: 144 ms [128-671 ms], offset: 1109 ms [527-1955 ms]; Layer 5: onset: 62 ms [37-101 ms], peak: 191 ms [171-328 ms] and 1050 ms [738-1035 ms], offset: 1972 [1082-1988 ms].

In case of EB>SC Layer 2: cluster 1 onset: 179 ms [0-246 ms], cluster 1 offset: 203 ms [0-280 ms], cluster 2 onset: 1003 ms [500-1035 ms], cluster 2 offset: 1070 ms [500-1972 ms]; Layer 3: cluster 1 onset: 183 ms [0-242 ms], cluster 1 offset: 199 ms [0-449 ms], cluster 2 onset: 1031 ms [500-1035 ms], cluster 2 offset: 1066 ms [500-1964 ms]; Layer 4: cluster 1 onset: 179 ms [0-250 ms], cluster 1 offset: 203 ms [0-472 ms], cluster 2 onset: 1027 ms [500-1127 ms], cluster 2 offset: 1066 ms [500-1972 ms]; Layer 5: cluster 1 onset: 191 ms [0-250 ms], cluster 1 offset: 203 ms [0-472 ms], cluster 2 onset: 1043 ms [500-1047 ms], cluster 2 offset: 1160 ms [500-1976 ms].

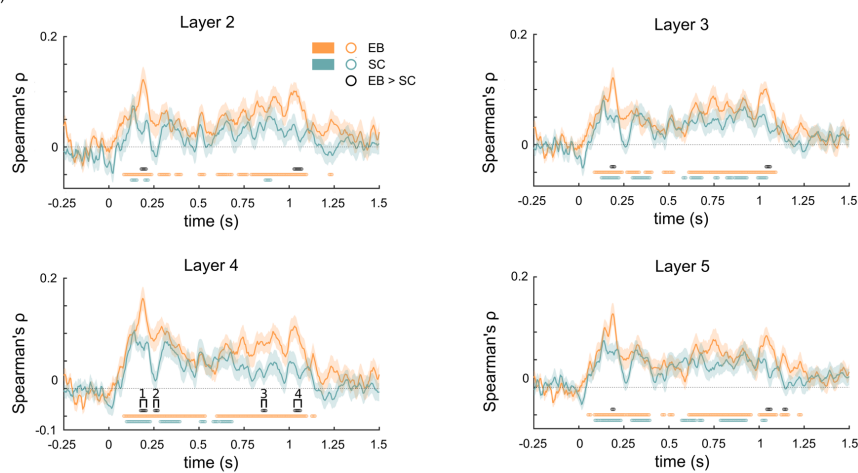
We illustrated the brain regions sharing the representation of the fourth layer since this layer correlated maximally within each group, as well as revealed the maximum difference between both groups (Figure 4.4c). EB displayed enhanced representation of this DNN layer early in time (180-200 ms) underlying widespread brain areas localized to bilateral superior parietal regions and precuneus, and right cuneus, entorhinal, fusiform, superior, middle and inferior temporal, temporal pole, insula, lateral occipital, pericalcarine, postcentral, precentral, rostral anterior cingulate, lateral orbitofrontal, medial orbito frontal, caudal middle frontal, pars opercularis, pars triangularis, rostral middle frontal, superior frontal, caudal and rostral anterior cingulate and left inferior parietal region. A second time-cluster of enhancement emerged soon after, localized to bilateral frontal gyri and right rostral middle frontal region between 240-280 ms. The third cluster arose from anterior parts of right, middle and inferior temporal regions. A final re-emergence of the same DNN representation was observed just after sound offset involving bilateral banks-STS, middle and superior temporal, supramarginal, fusiform, parahippocampal, precuneus, lingual, pericalcarine, cuneus, inferior parietal, isthmus cingulate and

a)

Deep Neural Network: YAMNet

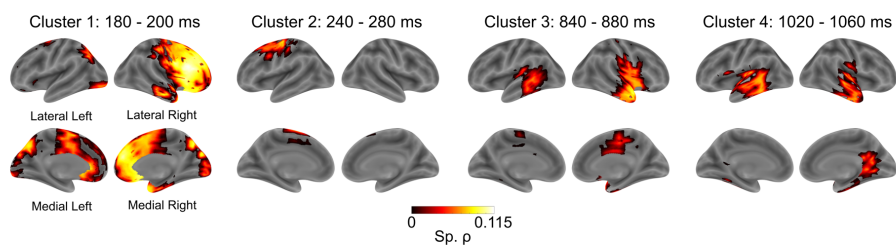


b)

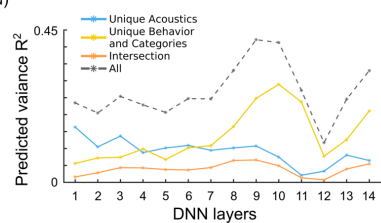


c)

EEG - YAMNet Layer-4 Blind > Sighted



d)



e)

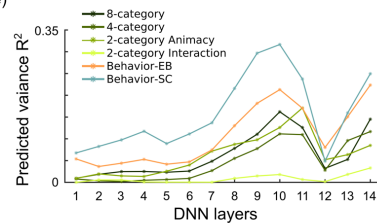


Figure 4. 4 Deep Neural Network a) A schematic architecture of the 14 layers of the neural network. b) EB exhibit maximum enhancement of representation of layers 2 to 4 early in time and close to sound offset (1 sec). c) Layer 4, showing maximum correlation with both groups and maximum difference between them, is represented predominantly in the fronto-temporal brain regions at different time clusters. d) Unique and combined predictive variance of acoustic models and higher-order categorical and behavior models predicting all 14 layers separately. e) Each DNN layer was predicted by individual higher-order models revealing that the DNN aligns best with the behavior model of SC and worst with the 2-category interaction model.

right inferior and transverse temporal, paracentral, postcentral, posterior cingulate, precentral, superior parietal and insula. The localization of peaks of all layers are illustrated in Supplementary Figure S4.3.

To gain a better understanding of the representation of the DNN layers and plausible neural processes these may represent (acoustic or categorical/behavior), we predicted the variance of each DNN layer using two classes of models: one consisting of all acoustic components such as the MTF, Center of Gravity, Harmonicity-to-Noise Ratio (HNR) and Pitch and the other consisting of the 8–categorical and group averages of behavior models of EB and SC. Using variance partitioning in conjunction with commonality analysis, we computed the unique and combined accounted variances for these 2 classes of models predicting each layer of the DNN (Figure 4.4d). As the layers evolved, a transformation from low-level acoustic properties of the sounds progressed to a higher-level representation of categories and behavior. While the acoustics uniquely explained the variance in the layers 1-3 better than categories and behavior, inversely, the latter could uniquely predict the later layers 8-14 better than the former. In the context of the enhanced representation of layers 2-5 in the neural model of EB, these early layers captured both unique acoustics and behavioral/categorical representations. However, the higher unaccounted variance in the early layers compared to the later ones suggests that layers 2-5 support an abstract and complex auditory representation that extends beyond acoustics and behavior.

Furthermore, a similar correspondence was observed in the emergence of the DNN layer representation and the time-evolving neural models of both EB and SC (Supplementary Figure S4.2). The early layers (2-5) emerged earlier in the EEG (~200 ms), while the later layers (9-11) peaked similarly around 350 ms similarly in both populations, supporting the hierarchical view of auditory circuits: acoustic-relevant DNN layers emerged earlier, while category/behavior-relevant layers emerged later in time. However, the later DNN layers did not show any enhancement in EB in comparison to SC, as opposed to representation of categorical and behavior models observed at ~600 ms in EB. Therefore, we explored the similarities between models by calculating the amount of variance that can be explained by individual categorical and behavior models for each DNN layer (Figure 4.4e). The maximum

variance in the later layers of YAMNet could be best explained by the behavior model of SC, followed by the same of EB. While the 8-, 4-, and 2-category animacy models exhibited lower yet comparable predictive performance for the later layers, the 2-category interaction model performed poorly. This suggests that the auditory categorization mechanisms of YAMNet are more closely modeled after those observed in sighted individuals than those associated with blindness.

In summary, we found that despite early spectral-temporal processing being preserved in EB as evidenced by RSA results with MTF and other acoustical models (Pitch, HNR, CoG), blindness causes an enhanced representation of early to intermediate DNN layers, which cannot be completely and uniquely explained by acoustics, categorical or behavior and are represented in large scale auditory networks.

Analysis	EB Peak [95% CI] (ms)	SC Peak [95% CI] (ms)	EB - SC duration (ms)	Fig no.
Pairwise decoding (sound vs sound)	183 [175-226]	175 [163-951]	160 [59-263] to 1007 [363-1988]	4.2b
RSA : Acoustic				
Modulation Transfer Function	199 [109-218]	187 [93-242]	None	4.3a
RSA : Deep Neural Network				4.4b
Layer 2	187 [152-316], 1035 [820-1050]	132 [125-359]	179 [0-246] to 203 [0-280], 1003 [500-1035] to 1070 [500-1972]	
Layer 3	187 [52-316], 1035 [820-1050]	137 [128-953]	183 [0-242] to 199 [0-449], 1031 [500-1035] to 1066 [500-1964]	

Layer 4	191 [171-328], 1035 [738-1035]	144 [128-671]	179 [0-250] to 203 [0-472], 1027 [500-1127] to 1066 [500-1972]	
Layer 5	191[171-328], 1050 [738-1035]	144 [128-661]	191 [0-250] to 203 [0-472], 1043 [500-1047] to 1160 [500-1976]	
Categorical and Behavior				
8 - category partialling out MTF	480 and 578 [308-621]	269 [148-1636]	574 [234-619] to 621 [328-1960]	4.3c
4 - category partialling out MTF	562 [207-613]	218 [207-761]	574 [78-1326] to 699 [621-1960]	4.3d
2 - category animacy	222 [210-691]	218 [207-935]	-	4.3e
2 - category interaction	574 [546-617]	531 [203-855]	546 [93-605] to 714 [703-1282]	4.3f
Behavior partialling out MTF	617 [164-1546]	293 [203-1218]	609 [160-1476] to 625 [193-1762]	4.3b

Table 4. 1. Peak and 95% confidence intervals for early blind (EB), sighted controls (SC) and EB - SC in decoding and RSA.

4.3.5 Shared Representations Within and Across Populations

We investigated the similarities and differences in the temporal dynamics of auditory processing in EB and SC. We compared the two neural models at each combination of time points. A perfect linear synchrony in the temporal unfolding of auditory processing could be visualized as a diagonal from sound onset time (0 ms) until around or after sound offset (bottom left to top right corner). We first compared within

group correlations. Correlations were highest along the diagonal albeit higher for EB than SC, peaking at 179 ms (Sp rho = 0.45) and 175 ms (Sp rho = 0.25), respectively (one-tail sign permutation test $p = 0.0158$; Figure 4.5 a) and b). The correlations gradually decreased as the distance from the diagonal increased in both groups. Interestingly, early blind participants had a broader and wider within-group correlation outline than in the sighted control group. However, EB also demonstrated a higher noise ceiling than SC which suggests that between-participants neural responses were more consistent in EB than in the SC group. A re-emergence of early auditory representation was also noted in EB, where group mean neural patterns from 144 ms to 230 ms correlated significantly with individual patterns at 777 ms to 1296 ms.

Further, we compared the neural models between the two groups. Figure 4.5c displays across-group temporal correlations between EB and SC with big clusters extending along the diagonal from 62 ms to 1192 ms in EB and 74 ms to 1105 ms in SC, peaking at 218 ms in EB and 210 ms in SC. Additionally, off-diagonal clusters were also noted. A significant cluster was also found ranging from 460 ms to 714 ms in EB and 890 ms to 1046 ms in SC. This indicates a non-linear synchrony of auditory representation between the two groups highlighting the neural processes occurring earlier in EB than SC. Another big cluster was found from 1070 ms to 1148 ms in EB and 89 to 218 ms in SC, indicating the early auditory representation in SC re-emerges in EB post sound offset.

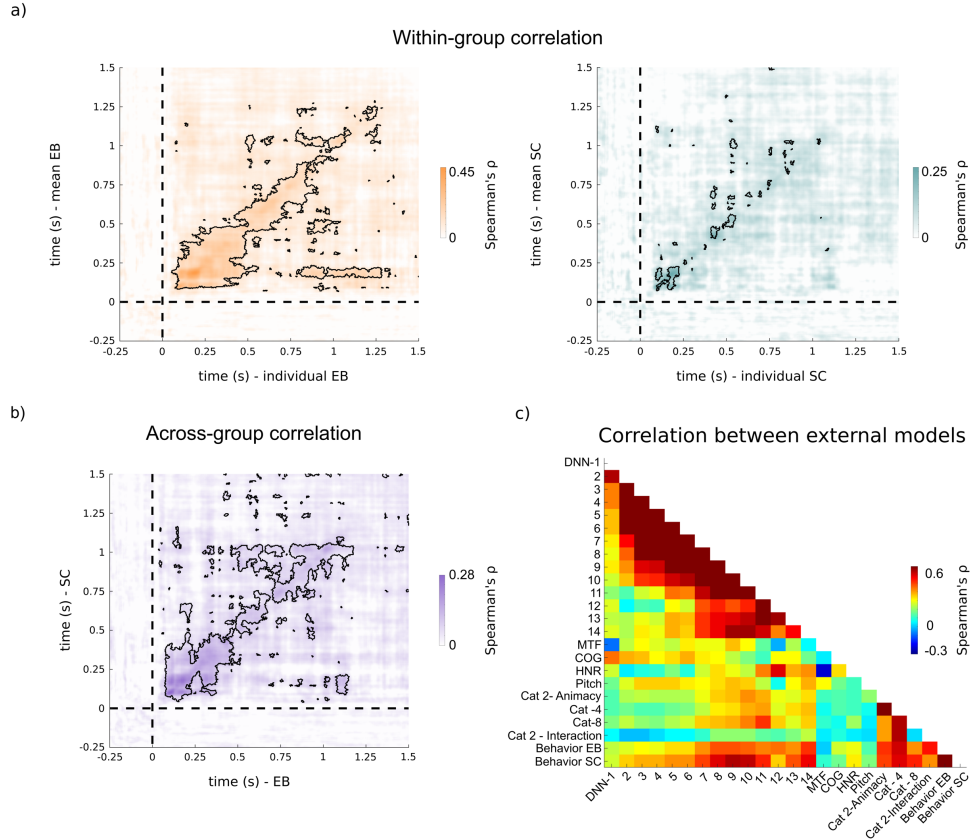


Figure 4. 5 Within and across-group comparisons a) and b) Each participant was correlated with the mean of its respective group computed by taking the concerned participant out of each computation. Clusters marked by black outline along the diagonal show unraveling of consistent neural representations in EB and SC (sign permutation test, $p < 0.01$, cluster-wise correction) c) Shared representation between EB and SC (sign permutation test, $p < 0.05$, cluster-wise correction). d) Correlations between all models used in the study.

4.4 Discussion

We investigated the representation of natural sounds through model-based analyses to provide a comprehensive picture of how sound processing -from acoustic to semantic- unfolds dynamically in the brain of the blind and sighted people. Using whole scalp EEG, we showed that the neural patterns in response to 1 second natural sounds yield higher decoding AUC (area under the curve) in EB compared to SC over a protracted period of time ranging from 152 to 1003 ms post sound onset (Figure 4.2b). Previous neuroimaging studies have also decoded natural sounds in the early visual cortex of blind individuals (Mattioni et al., 2020; Vetter et al., 2020); however, the exact timing and representational content of the enhanced involvement of these visual regions in sound processing remained unknown. By testing at which

moment a specific sound becomes a unique decodable entity in brain activity, we showed that EB exhibit an enhanced representation of discrete natural sounds over an extended period of time when compared to SC. Additionally, we found that the enhanced decoding in EB was contributed by posterior channels and occipital regions from 152 ms till the entire duration of the sounds. The temporal generalization analysis (King & Dehaene, 2014) further revealed transient neural representations (as shown by the diagonal) unfolding for a prolonged time before stabilizing in EB compared to SC. This extended enhanced decoding suggests effects across multiple successive auditory processing stages. A potential reason for this could be the extended involvement of the occipital cortex in EB that may facilitate a sequence of unique auditory representations, unfolding one after another and delaying the stabilization of neural patterns. While the visual cortex of blind individuals has been shown to synchronize more effectively with speech and language tasks compared to SC (Abboud et al., 2019; Striem-Amit et al., 2012; Van Ackeren et al., 2018), our findings demonstrate that the involvement may extend to non-speech sounds as well, lasting throughout the entire duration of the auditory stimuli. In contrast in SC, although we found non-significant results, the trends showed that the neural patterns underlying auditory processes appear to stabilize and diffuse much earlier than EB. Overall, these decoding results provide evidence for significant, quantitative and temporally extended variations in the brain activity supporting sound discrimination in EB.

However, characterizing the temporal trajectory offers limited insights on the stages of auditory processes reorganized in EB. To elucidate which aspects of sound processing contributed to the enhancement observed in EB, we used RSA to compare brain representations to that of acoustic, categorical, behavioral and convolutional neural networks (CNNs) trained to categorize sounds. Our RSA results suggest an “embedded” representational structure in both groups where neural representation is first organized according to acoustic and later to categorical and behavioral relations (Giordano et al., 2013; Heer et al., 2017; Lowe et al., 2021; Ogg et al., 2020). We first focused on isolating low-level acoustic processes of sounds that are typically well represented in the auditory regions and are known to create low-level bias for higher order categorical responses (Leaver & Rauschecker, 2010; Moerel et al., 2012; Santoro et al., 2014). We modeled the early acoustic processing using the modulation transfer function (MTF) model that provides a spectro-temporal representation of the sounds simulating the responses from primary auditory cortex as demonstrated in fMRI studies (Chi et al., 2005; Giordano et al., 2023; Santoro et al., 2014). The MTF model correlated early in time with EEG responses and similarly in EB and SC (Figure 4.3a). This suggests that the early decoding enhancement in EB cannot be explained by spectral-temporal processing. Unaltered acoustic brain representation in EB was also confirmed by other acoustic models based on center

of gravity, pitch and harmonicity-to-noise ratio of sounds (see Supplementary Figure S4.1). These results resolve a long debate by unequivocally showing that despite the tendency to rely more on natural sounds in their daily life, early blindness does not trigger a significant alteration onto how the brain encodes the low-level acoustic properties of sounds (Mattioni et al., 2022). These results may echo studies showing no behavioral differences between blind and sighted in the discrimination of low-level sound properties (Mattioni et al., 2020, 2022).

A previous study revealed the temporal unfolding of cochlear and categorical representations using fMRI and MEG fusion in sighted adults, showing that cochlear representations peaked at 127 ms in primary and secondary auditory regions while semantic representation peaked later at 218 ms in VOTC and lateral occipital region (Lowe et al., 2021). Aligning with those results, the peak emergence of acoustic and categorical / behavior representation in SC in our study was found similarly at 187 and 250-300 ms respectively. The same representations began at similar latencies in EB; however, we found enhanced categorical and behavior representations in EB that peaked later in time (~600 ms), originating from occipital-temporal cortex and inferior parietal sources. These enhanced categorical representations were independent of the spectral-temporal modulations of the sounds as those acoustic features were regressed out (Figure 4.3b-d). Previous studies have found enhanced categorical encoding of natural sounds in the ventral occipital temporal regions in blind people (Mattioni et al., 2020; van den Hurk et al., 2017), however the timing and precise features leading to this enhanced sound encoding was not fully characterized. Interestingly, other studies found differentiated activity in the blind during this latency in occipital regions - one showed high gamma activity elicited by natural sounds peaking at around 500 ms in the occipital regions of blind people (Schepers et al., 2012) and another found semantic enhancement in blind individuals beginning from 580 ms arising from occipital sources (Abboud et al., 2019), suggesting a potential overlap of neural circuits for categorical representation of linguistic and natural sounds in blind individuals (Handjaras et al., 2016; Mattioni et al., 2020; Peelen & Downing, 2017). Another possible explanation for the observed categorical enhancement at this particular latency could be found from studies investigating event related potentials (ERPs). Although the N400 (negative peak 400 ms post onset of the stimulus) is well known to index semantic congruency and memory retrieval in sighted and blind populations performing linguistic tasks (Röder et al., 2000), as well as in braille (Glyn et al., 2015), the late ERP can also be elicited in the left hemisphere in response to meaningful environmental sounds (see Kutas & Federmeier, 2011). In fact, we found the maximum univariate difference between the EEG amplitudes of the two groups in response to all sounds between 300 to 700 ms from left lateralized temporal-central-occipital channels (Supplementary Figure S4.4). Despite the fact that there was no semantic task in this study (participants

performed a one-back task), the latency of the enhanced categorical representation in EB lies close to the trough of the N400 (Categorical peak = 574 ms, N400 peak = 538 ms). Thus, it is plausible that the top-down semantic memory retrieval function of the N400 (Kutas & Federmeier, 2011) could extend to include the retrieval of category-specific conceptual information for natural sounds, whose representation is enhanced in blind individuals, particularly in occipital regions (Bedny et al., 2011).

Furthermore, we found that the categorical enhancement in EB qualitatively clusters human and object sounds together, and animals with big sounds. In fact, objects and big sounds yielded high decoding AUC at ~600 ms when the classifier was trained on human and animal sounds respectively, and vice versa (Supplementary Figure 4.5). Thus, neural responses to sounds of diverse acoustic properties converged to an enhanced higher-order categorical encoding at ~600 ms in EB, validating that the categorical representation is indeed independent of acoustic properties of the sounds. But why specifically this combination of categories? One hypothesis concerns the frequency of interaction with the sources producing the sounds. EB may frequently interact with humans and objects more than animals, environmental and big mechanical objects. For eg., the tactile exposure of humans and objects could be higher than other categories in EB, providing a rich internal representation of these sources of sounds. Complementing this, the lack of visual representation of concepts such as animals and environment have been demonstrated (Tian et al., 2001) that could prove to be a principle for categorizing sounds. Additionally, these combinations of categories could relate to auditory scene analysis, in which some sounds may be segregated or grouped based on context and occurrence (Campus et al., 2019). For instance, an object sound is usually accompanied with a human voice; animals could be better associated with environmental sounds. Moreover, these intricate relationships between these categories could be coded differently in EB due to affordance, i.e. the interaction of an individual's capabilities with the surrounding environment (Gibson, 1950). Conversely in our study, the behavior similarity ratings of EB and SC were similar, suggesting that there are no differences in the behavior inferences about the sounds between both groups. However, the ratings also reflect conceptual knowledge which can be imbibed through experience and/or language (Campbell & Bergelson, 2022). Thus, specific models of concept learning in visually deprived individuals, whether driven by experience or language, could provide deeper insights into the neural representation of both sounds and language. This approach may help unravel the mechanisms guiding concept encoding in early blindness.

While our decoding approach indicates that differences in auditory processing emerge in EB from early (152 ms) to late (~1 s) processing windows, our theoretically driven RSA modeling approach only explained part of these processing

windows. Specifically, while higher categorical and behavioral representations were found to be enhanced during late time periods (~600 ms) in EB using this computational approach, it still remained to be shown which specific representations are critical in differentiating EB from SC in earlier time windows. In this context, we decided to rely on DNNs that have served as plausible models predicting neurophysiological mechanisms guiding behavior (Cichy et al., 2016; A. J. Kell & McDermott, 2019). Thus, to investigate alternate representation of auditory processing, we employed a Deep Neural Network (DNN) called YAMNet which performs categorization of a sample of sound to pre-existing labels (Sound-to-Event DNN), processing features of gradually higher complexity and abstraction along its layers (Cichy et al., 2016; Giordano et al., 2023; A. J. E. Kell et al., 2018; Rupp et al., 2024). Layers of the DNN were represented in the neural models of both groups, but compared to SC, we found that the neural models of EB showed greater representation of layers 2 to 5 of the DNN distributed over fronto-temporal regions, especially early in time at ~200 ms and post sound offset between 1020-1070 ms. These results indicate that the representation of early layers of an object-trained auditory DNN matched the representations of EB more closely than those of SC and explained the enhancement early in time. Additionally, both low-level acoustic and higher-order categorical and behavior models predicted the representation of early layers to an extent, however there remained a high unaccounted variance (~0.75) for the early layers as well. This suggests that these layers capture auditory representations of acoustic processing and categorical/behavioral models to some extent, while also reflecting greater complexity and abstraction beyond theoretical models (Figure 4.4d).

Although the hierarchical architecture of DNNs have been inspired from research in vision, they have also supported the hypothesis that the auditory networks are more hierarchically organized than distributed. For instance, spatially organized hierarchies for auditory object processing were reported in which the early layers predicted responses from primary auditory regions and later layers represented belt and parabelt areas, suggesting that the auditory system is a layered organization that supports both sensory and conceptual auditory tasks (A. J. E. Kell et al., 2018; Rupp et al., 2024). From a temporal perspective, this aligns with the emergence of representations of early layers at ~200 ms and that of the later layers (8-11) at ~350 ms in both populations, bolstering the hierarchical view of the auditory networks (Supplementary Figure 4.2). Moreover, the enhanced representation of early DNN layers in EB was attributed to fronto-temporal brain regions, indicating that blindness causes modulations in large-scale auditory networks. Thus, while the progressive DNN layers illuminated the hierarchy of auditory processing, each individual layer revealed representations shared by distributed auditory-related brain regions, supporting parallel computation for audition.

An unexpected finding in our results was that the representations of later DNN layers were not enhanced in EB, despite the enhanced representation of the 8-category and behavior model compared to SC. This is because we found that YAMNet's sound categorization mechanisms align most closely with the behavior patterns of SC, followed closely by EB, while the 2-category interaction model, which could best predict the neural responses at ~600 ms in EB, performed the worst (Figure 4.4e). Consequently, the later layers did not share any similarity with the enhanced categorical representation in EB, indicating that the DNN's categorization mechanisms differ from those in EB.

In our last analysis, we compared the neural models of EB and SC and found a strong diagonal till almost sound offset indicating the correspondence of neural activity between the two groups. In addition, a non-linear emergence of representation was also noted from large significant off-diagonal clusters. Particularly we found representation in EB at the offset similar to that of itself and SC at ~200 ms indicating a re-emergence of early auditory processing after the sound offset, which was absent in the sighted. Interestingly, the early DNN layers representations were also significantly higher at sound offset in EB. This re-emergence could potentially indicate feedback from higher auditory circuits. Another large off-diagonal cluster displayed similar representations between neural processes of SC at offset and that of EB from 460 to 714 ms. While the 2-category interaction model explains the neural representation of EB during this time window, the same external model does not correlate with the neural representation of SC at offset. This suggests unaccounted variance in both neural models of EB and SC. Therefore, future exploration of more nuanced models could better explain the variance of the neural models and reveal more accurate temporal dynamics in both populations.

In summary, while early acoustic processing remained unaffected in EB, alternate auditory representations were modulated primarily in the fronto-temporal networks early in time as shown by the initial layers of the DNN. Importantly, EB showed enhanced categorical coding later in time which is abstracted from the acoustic features of the sounds. Together these results reveal a specific pattern of reorganization in the brain networks coding for sounds in EB, providing unique insights on the dynamics of functional reorganization in visual deprivation.

4.5 Supplementary Figures

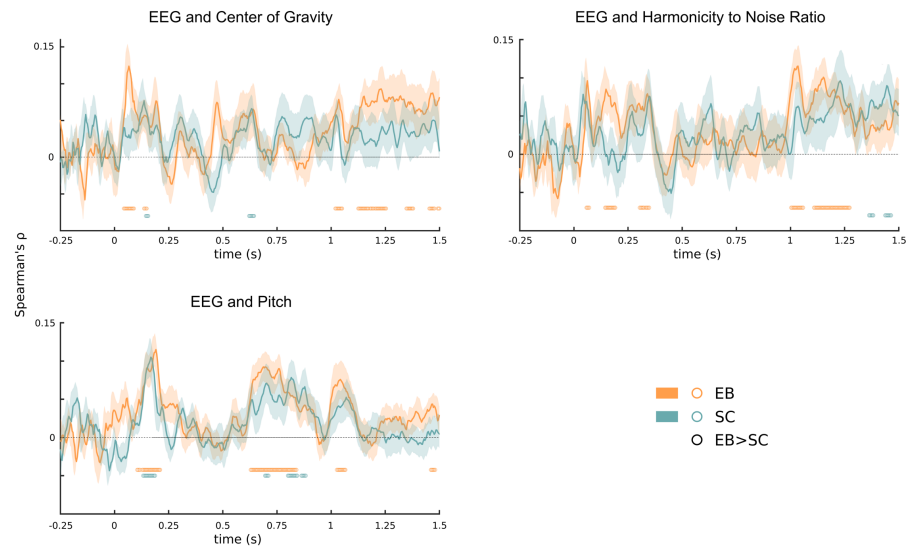


Figure S4. 1 RSA with acoustic models: Correlation of time-varying EEG models and acoustic models.

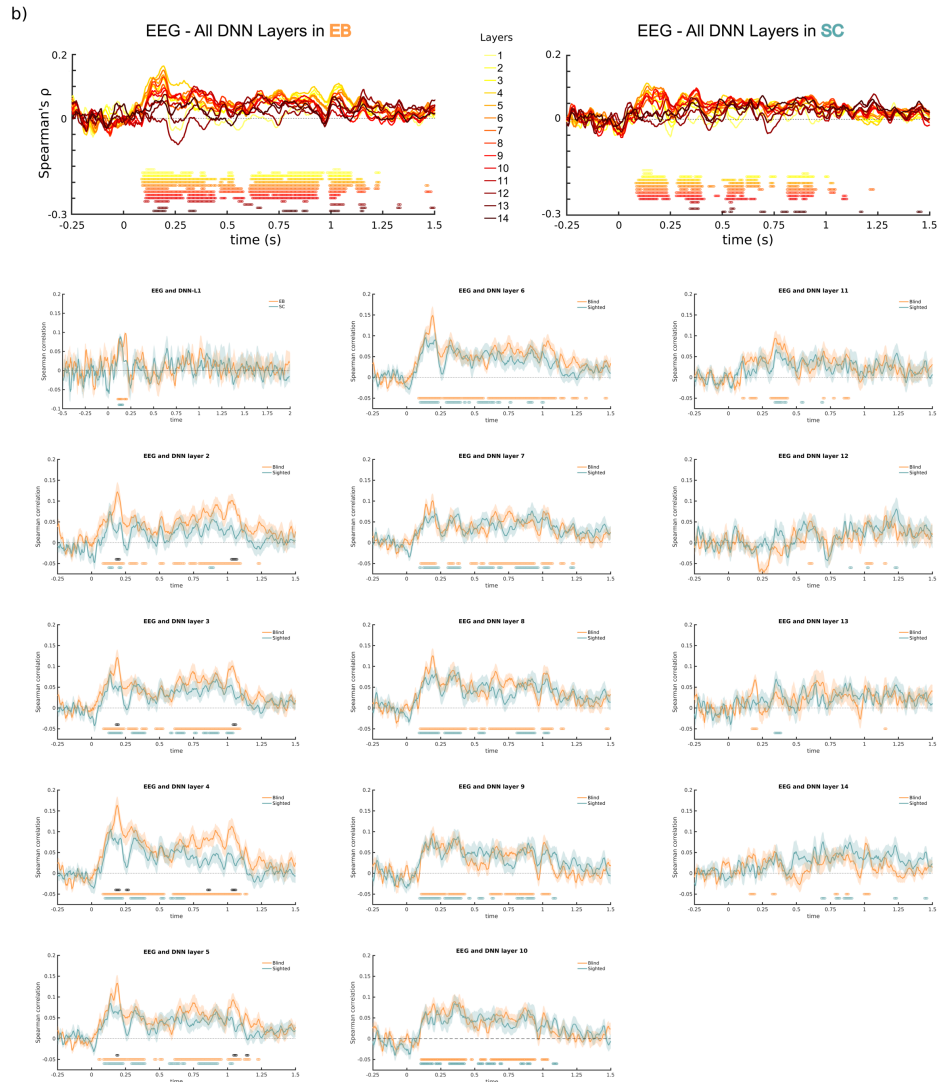


Figure S4. 2 RSA between EEG and all DNN layers. Correlation of EEG and all 14 layers of the deep neural network (DNN)

Chapter 4. Temporal Dynamics of Sound Representation in Blind and Sighted

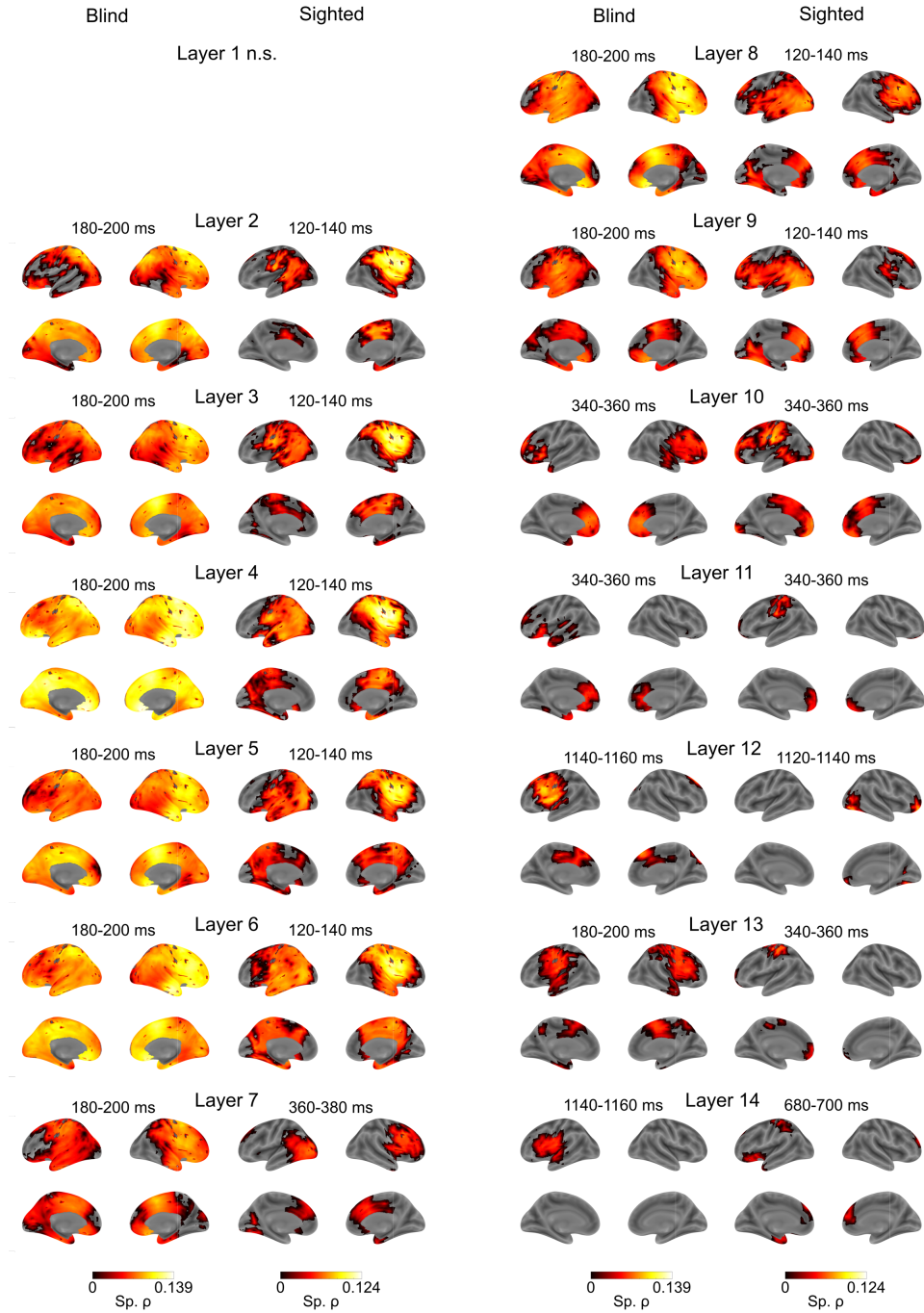


Figure S4. 3 Source space RSA with all DNN layers for blind and sighted separately. The time windows correspond to the peak correlation found at the sensor level (see Figure S4. 2)

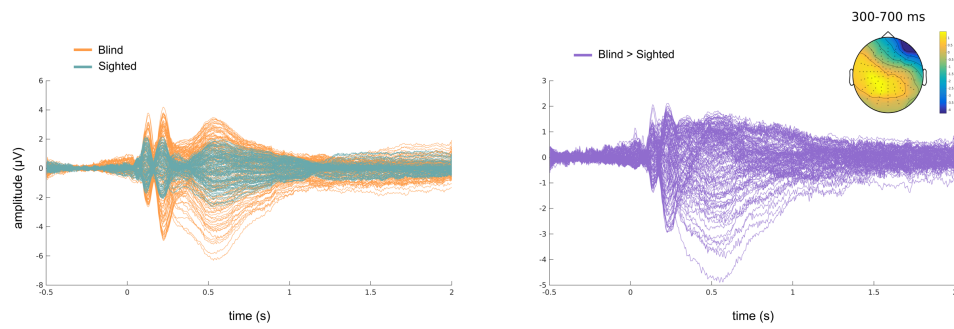


Figure S4. 4 Event Related Potentials (average re-referenced). Enhanced univariate responses found in EB. The enhanced categorical representation overlaps the N400 ERP in EB compared to SC.

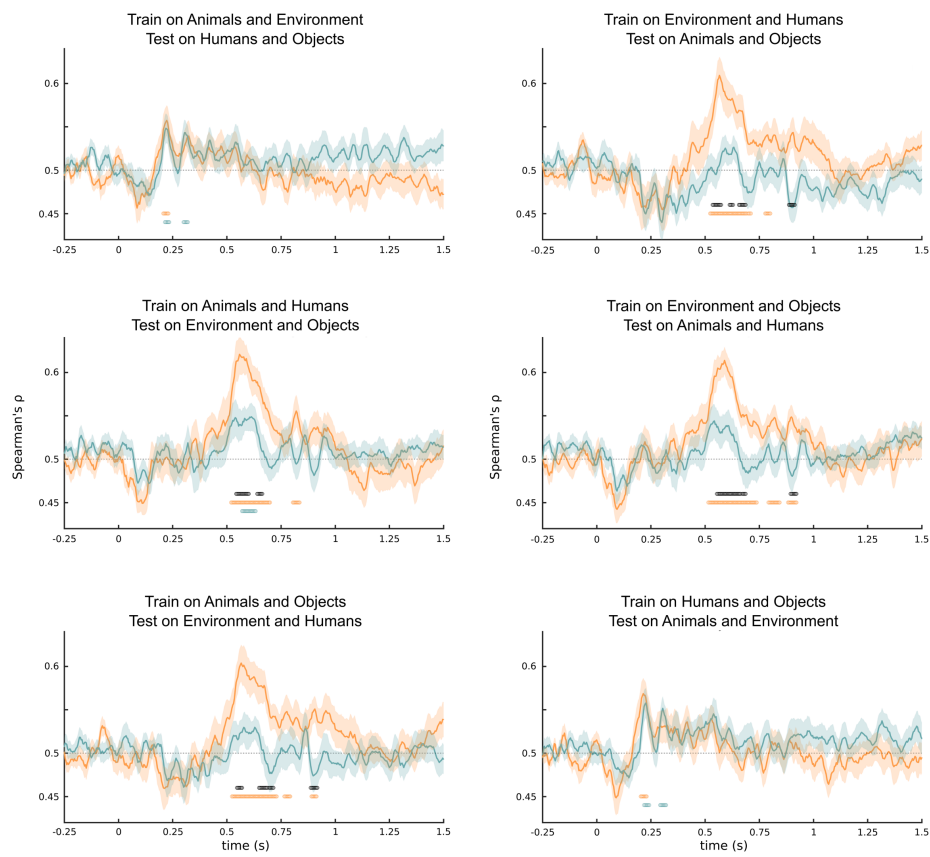


Figure S4. 5 Cross Category Decoding. Training performed on 2 out of 4 categories and testing on the remaining 2. Humans and Objects or/and environment (big) and animals share similar neural patterns that may drive the enhanced categorical coding in blind individuals.

4.6 Bibliography

Abboud, S., Engemann, D. A., & Cohen, L. (2019). *Semantic coding in the occipital cortex of early blind individuals* (p. 539437). bioRxiv. <https://doi.org/10.1101/539437>

Bavelier, D., & Neville, H. J. (2002). Cross-modal plasticity: Where and how? *Nature Reviews Neuroscience*, 3(6), 443–452. <https://doi.org/10.1038/nrn848>

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, 108(11), 4429–4434. <https://doi.org/10.1073/pnas.1014818108>

Bedny, M., Richardson, H., & Saxe, R. (2015). “Visual” Cortex Responds to Spoken Language in Blind Children. *Journal of Neuroscience*, 35(33), 11674–11681. <https://doi.org/10.1523/JNEUROSCI.0634-15.2015>

Blakemore, C., & Cooper, G. F. (1970). Development of the brain depends on the visual environment. *Nature*, 228(5270), 477–478. <https://doi.org/10.1038/228477a0>

Burton, H., Snyder, A. Z., Diamond, J. B., & Raichle, M. E. (2002). Adaptive Changes in Early and Late Blind: A fMRI Study of Verb Generation to Heard Nouns. *Journal of Neurophysiology*. <https://doi.org/10.1152/jn.00129.2002>

Campbell, E. E., & Bergelson, E. (2022). Making sense of sensory language: Acquisition of sensory knowledge by individuals with congenital sensory impairments. *Neuropsychologia*, 174, 108320. <https://doi.org/10.1016/j.neuropsychologia.2022.108320>

Campus, C., Sandini, G., Amadeo, M. B., & Gori, M. (2019). Stronger responses in the visual cortex of sighted compared to blind individuals during auditory space representation. *Scientific Reports*, 9, 1935. <https://doi.org/10.1038/s41598-018-37821-y>

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2), 887–906. <https://doi.org/10.1121/1.1945807>

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1), 27755. <https://doi.org/10.1038/srep27755>

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3), Article 3. <https://doi.org/10.1038/nn.3635>

- Collignon, O., Davare, M., Olivier, E., & De Volder, A. G. (2009). Reorganisation of the Right Occipito-Parietal Stream for Auditory Spatial Processing in Early Blind Humans. A Transcranial Magnetic Stimulation Study. *Brain Topography*, 21(3), 232–240. <https://doi.org/10.1007/s10548-009-0075-8>
- Collignon, O., Lassonde, M., Lepore, F., Bastien, D., & Veraart, C. (2007). Functional cerebral reorganization for auditory spatial processing and auditory substitution of vision in early blind subjects. *Cerebral Cortex (New York, N.Y.: 1991)*, 17(2), 457–465. <https://doi.org/10.1093/cercor/bhj162>
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3), 968–980. <https://doi.org/10.1016/j.neuroimage.2006.01.021>
- Dormal, G., Pelland, M., Rezk, M., Yakobov, E., Lepore, F., & Collignon, O. (2018). Functional Preference for Object Sounds and Voices in the Brain of Early Blind and Sighted Individuals. *Journal of Cognitive Neuroscience*, 30(1), 86–106. https://doi.org/10.1162/jocn_a_01186
- Elbert, T., Sterr, A., Rockstroh, B., Pantev, C., Müller, M. M., & Taub, E. (2002). Expansion of the Tonotopic Area in the Auditory Cortex of the Blind. *The Journal of Neuroscience*, 22(22), 9941–9944. <https://doi.org/10.1523/JNEUROSCI.22-22-09941.2002>
- Gibson, J. J. (1950). *The perception of the visual world* (pp. xii, 242). Houghton Mifflin.
- Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26(4), 664–672. <https://doi.org/10.1038/s41593-023-01285-9>
- Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract Encoding of Auditory Objects in Cortical Activity Patterns. *Cerebral Cortex*, 23(9), 2025–2037. <https://doi.org/10.1093/cercor/bhs162>
- Glyn, V., Lim, V. K., Hamm, J. P., Mathur, A., & Hughes, B. (2015). Behavioural and Electrophysiological effects related to semantic violations during braille reading. *Neuropsychologia*, 77, 298–312. <https://doi.org/10.1016/j.neuropsychologia.2015.09.008>
- Gougoux, F., Lepore, F., Lassonde, M., Voss, P., Zatorre, R. J., & Belin, P. (2004). Pitch discrimination in the early blind. *Nature*, 430(6997), 309–309. <https://doi.org/10.1038/430309a>
- Handjaras, G., Ricciardi, E., Leo, A., Lenci, A., Cecchetti, L., Cosottini, M., Marotta, G., & Pietrini, P. (2016). How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. *NeuroImage*, 135, 232–242. <https://doi.org/10.1016/j.neuroimage.2016.04.063>

- Heer, W. A. de, Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *Journal of Neuroscience*, 37(27), 6539–6557. <https://doi.org/10.1523/JNEUROSCI.3267-16.2017>
- Heimler, B., Weisz, N., & Collignon, O. (2014). Revisiting the adaptive and maladaptive effects of crossmodal plasticity. *Neuroscience*, 283, 44–63. <https://doi.org/10.1016/j.neuroscience.2014.08.003>
- Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), 630–644.e16. <https://doi.org/10.1016/j.neuron.2018.03.044>
- Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, 55, 121–132. <https://doi.org/10.1016/j.conb.2019.02.003>
- King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4), 203–210. <https://doi.org/10.1016/j.tics.2014.01.002>
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. <https://doi.org/10.3389/neuro.06.004.2008>
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>
- Leaver, A. M., & Rauschecker, J. P. (2010). Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. *Journal of Neuroscience*, 30(22), 7604–7612. <https://doi.org/10.1523/JNEUROSCI.0296-10.2010>
- Lowe, M. X., Mohsenzadeh, Y., Lahner, B., Charest, I., Oliva, A., & Teng, S. (2021). Cochlea to categories: The spatiotemporal dynamics of semantic auditory representations. *Cognitive Neuropsychology*, 38(7–8), 468–489. <https://doi.org/10.1080/02643294.2022.2085085>
- Manjunath, N. K., Srinivas, R., Nirmal, K. S., Nagendra, H. R., Kumar, A., & Telles, S. (1998). Shorter Latencies of Components of Middle Latency Auditory Evoked Potentials in Congenitally Blind Compared to Normal Sighted Subjects. *International Journal of Neuroscience*, 95(3–4), 173–181. <https://doi.org/10.3109/00207459809003339>
- Marques, J. P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P.-F., & Gruetter, R. (2010). MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *NeuroImage*, 49(2), 1271–

1281. <https://doi.org/10.1016/j.neuroimage.2009.10.002>

Mattioni, S. (2024). Exploring brain representations through the lens of similarity structures. *Nature Reviews Psychology*, 3(8), 506–506. <https://doi.org/10.1038/s44159-024-00335-8>

Mattioni, S., Rezk, M., Battal, C., Bottini, R., Cuculiza Mendoza, K. E., Oosterhof, N. N., & Collignon, O. (2020). Categorical representation from sound and sight in the ventral occipito-temporal cortex of sighted and blind. *eLife*, 9, e50732. <https://doi.org/10.7554/eLife.50732>

Mattioni, S., Rezk, M., Battal, C., Vadlamudi, J., & Collignon, O. (2022). Impact of blindness onset on the representation of sound categories in occipital and temporal cortices. *eLife*, 11, e79370. <https://doi.org/10.7554/eLife.79370>

Moerel, M., Martino, F. D., & Formisano, E. (2012). Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. *Journal of Neuroscience*, 32(41), 14205–14216. <https://doi.org/10.1523/JNEUROSCI.1388-12.2012>

Naveen, K. V., Srinivas, R., Nirmala, K. S., Nagarathna, R., Nagendra, H. R., & Telles, S. (1998). Differences between Congenitally Blind and Normally Sighted Subjects in the P1 Component of Middle Latency Auditory Evoked Potentials. *Perceptual and Motor Skills*, 86(3_suppl), 1192–1194. <https://doi.org/10.2466/pms.1998.86.3c.1192>

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1), 1–25. <https://doi.org/10.1002/hbm.1058>

Ogg, M., Carlson, T. A., & Slevc, L. R. (2020). The Rapid Emergence of Auditory Object Representations in Cortex Reflect Central Acoustic Attributes. *Journal of Cognitive Neuroscience*, 32(1), 111–123. https://doi.org/10.1162/jocn_a_01472

Pascual-Leone, A., & Torres, F. (1993). Plasticity of the sensorimotor cortex representation of the reading finger in Braille readers. *Brain: A Journal of Neurology*, 116 (Pt 1), 39–52. <https://doi.org/10.1093/brain/116.1.39>

Peelen, M. V., & Downing, P. E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 105, 177–183. <https://doi.org/10.1016/j.neuropsychologia.2017.03.033>

Peelen, M. V., He, C., Han, Z., Caramazza, A., & Bi, Y. (2014). Nonvisual and Visual Object Shape Representations in Occipitotemporal Cortex: Evidence from Congenitally Blind and Sighted Adults. *Journal of Neuroscience*, 34(1), 163–170. <https://doi.org/10.1523/JNEUROSCI.1114-13.2014>

Pietrini, P., Furey, M. L., Ricciardi, E., Gobbini, M. I., Wu, W.-H. C., Cohen, L., Guazzelli, M., & Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 101(15), 5658–5663.

<https://doi.org/10.1073/pnas.0400707101>

Rakic, P. (1988). Specification of cerebral cortical areas. *Science (New York, N.Y.)*, 241(4862), 170–176. <https://doi.org/10.1126/science.3291116>

Rauschecker, J. P. (2002). Cortical map plasticity in animals and humans. In *Progress in Brain Research* (Vol. 138, pp. 73–88). Elsevier. [https://doi.org/10.1016/S0079-6123\(02\)38072-5](https://doi.org/10.1016/S0079-6123(02)38072-5)

Ricciardi, E., Bonino, D., Sani, L., Vecchi, T., Guazzelli, M., Haxby, J. V., Fadiga, L., & Pietrini, P. (2009). Do We Really Need Vision? How Blind People “See” the Actions of Others. *Journal of Neuroscience*, 29(31), 9719–9724. <https://doi.org/10.1523/JNEUROSCI.0274-09.2009>

Röder, B., Rösler, F., & Neville, H. J. (2000). Event-related potentials during auditory language processing in congenitally blind and sighted people. *Neuropsychologia*, 38(11), 1482–1502. [https://doi.org/10.1016/S0028-3932\(00\)00057-9](https://doi.org/10.1016/S0028-3932(00)00057-9)

Röder, B., Stock, O., Bien, S., Neville, H., & Rösler, F. (2002). Speech processing activates visual cortex in congenitally blind humans. *The European Journal of Neuroscience*, 16(5), 930–936. <https://doi.org/10.1046/j.1460-9568.2002.02147.x>

Rupp, K. M., Hect, J. L., Harford, E. E., Holt, L. L., Ghuman, A. S., & Abel, T. J. (2024). A hierarchy of processing complexity and timescales for natural sounds in human auditory cortex. *bioRxiv: The Preprint Server for Biology*, 2024.05.24.595822. <https://doi.org/10.1101/2024.05.24.595822>

Sadato, N., Pascual-Leone, A., Grafman, J., Ibañez, V., Deiber, M.-P., Dold, G., & Hallett, M. (1996). Activation of the primary visual cortex by Braille reading in blind subjects. *Nature*, 380(6574), 526–528.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*, 10(1), e1003412. <https://doi.org/10.1371/journal.pcbi.1003412>

Schepers, I. M., Hipp, J. F., Schneider, T. R., Röder, B., & Engel, A. K. (2012). Functionally specific oscillatory activity correlates between visual and auditory cortex in the blind. *Brain*, 135(3), 922–934. <https://doi.org/10.1093/brain/aws014>

Stevens, A. A., & Weaver, K. E. (2009). Functional characteristics of auditory cortex in the blind. *Behavioural Brain Research*, 196(1), 134–138. <https://doi.org/10.1016/j.bbr.2008.07.041>

Striem-Amit, E., & Amedi, A. (2014). Visual cortex extrastriate body-selective area activation in congenitally blind people “seeing” by using sounds. *Current Biology: CB*, 24(6), 687–692. <https://doi.org/10.1016/j.cub.2014.02.010>

Striem-Amit, E., Cohen, L., Dehaene, S., & Amedi, A. (2012). Reading with Sounds: Sensory Substitution Selectively Activates the Visual Word Form Area in

- the Blind. *Neuron*, 76(3), 640–652. <https://doi.org/10.1016/j.neuron.2012.08.026>
- Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional Specialization in Rhesus Monkey Auditory Cortex. *Science*, 292(5515), 290–293. <https://doi.org/10.1126/science.1058911>
- Van Ackeren, M. J., Barbero, F. M., Mattioni, S., Bottini, R., & Collignon, O. (2018). Neuronal populations in the occipital cortex of the blind synchronize to the temporal dynamics of speech. *eLife*, 7, e31640. <https://doi.org/10.7554/eLife.31640>
- van den Hurk, J., Van Baelen, M., & Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences of the United States of America*, 114(22), E4501–E4510. <https://doi.org/10.1073/pnas.1612862114>
- Van Veen, B. D., Van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, 44(9), 867–880. *IEEE Transactions on Biomedical Engineering*. <https://doi.org/10.1109/10.623056>
- Vetter, P., Bola, Ł., Reich, L., Bennett, M., Muckli, L., & Amedi, A. (2020). Decoding Natural Sounds in Early “Visual” Cortex of Congenitally Blind Individuals. *Current Biology*, 30(15), 3039–3044.e2. <https://doi.org/10.1016/j.cub.2020.05.071>
- Wallmeier, L., Kish, D., Wiegrecbe, L., & Flanagan, V. L. (2015). Aural localization of silent objects by active human biosonar: Neural representations of virtual echo-acoustic space. *European Journal of Neuroscience*, 41(5), 533–545. <https://doi.org/10.1111/ejn.12843>
- Wang, X., Peelen, M. V., Han, Z., He, C., Caramazza, A., & Bi, Y. (2015). How Visual Is the Visual Cortex? Comparing Connectional and Functional Fingerprints between Congenitally Blind and Sighted Individuals. *Journal of Neuroscience*, 35(36), 12545–12559. <https://doi.org/10.1523/JNEUROSCI.3914-14.2015>

Chapter 4. Temporal Dynamics of Sound Representation in Blind and Sighted

Chapter 5

DISCUSSION

In this thesis, I presented three chapters investigating auditory emotion categorization in neurotypical population (Chapter 2), development of auditory processing and categorical representation in infants (Chapter 3) and reorganization of auditory functions in congenitally blind individuals (Chapter 4). Due to the similarities in the experiment design and analysis techniques between Chapters 2 and 3, the following discussion will be divided into two parts- first concerning the findings of emotion categorization and the second will discuss auditory representations found in all populations from Chapter 3 and 4.

5.1 Auditory Emotion Categories in Neurotypical Population

Voices carry a wealth of information such as emotion, identity, gender, age, physical size and health status (Belin et al., 2004, 2011; Frühholz, S & Belin, P., 2024; Klasen et al., 2012; Pernet & Belin, 2012; Whiteside, 1998; Zäske et al., 2013). We demonstrated that the brain is able to discriminate different non-verbal vocalizations and generalize them to a common emotion category (Talwar et al., 2023). These neural processes occurred automatically since the participants performed another unrelated task. By selecting stimuli from different categories that were acoustically similar and shown in our simulation to produce comparable responses within the cochlea, along with incorporating a scrambled control condition, we could determine that the observed categorical response to emotion (Fear and Happiness) is at least partially independent of the sounds' acoustic properties. Furthermore, the different topographic responses to Fear and Happiness indicate distinct yet overlapping regions subserving categorical representations of emotion categories. The following sections explore additional key points in more detail.

Is the emotion category response completely independent from the acoustic properties of the sounds?

There is an ongoing debate about whether the categorical representation of emotions in the human brain can be explained by the low-level acoustic features of sounds. On one hand, evidence indicates the existence of the emotion voice area that is preferentially selective to emotional vocalizations (Ethofer et al., 2009, 2012). On the other hand, the differential neural patterns to emotion categories could arise

due to response bias to spectral-temporal cues that are characteristic to an emotion category (Banse & Scherer, 1996).

To isolate the categorical response to emotions, we selected stimuli with acoustic features that can well represent non-verbal sounds, ensuring they were matched for center of gravity (COG), harmonicity-to-noise ratio (HNR), and pitch. However, there are many other acoustic features that were not considered but could provide a better explanation of neural acoustic processing. Some properties that are known to modulate neural responses include *fundamental frequency* (Allen et al., 2017; Giordano et al., 2013; Patterson et al., 2002), *spectro-temporal modulation* (S. Norman-Haignere et al., 2015; Ogg et al., 2020; Schönwiesner & Zatorre, 2009), *spectral flatness* (Alluri et al., 2012), *spectral centroid* (Allen et al., 2017; Alluri et al., 2012; Giordano et al., 2013), *spectral envelope* (Ogg et al., 2019; Warren et al., 2005), *loudness* (Giordano et al., 2013; Langers et al., 2007), *noisiness* (Giordano et al., 2013; Lewis et al., 2009), *attack time* (Menon et al., 2002) and *decay* (Aramaki et al., 2011). Therefore, it may be possible that the categorical neural response to emotion could be at least partially explained by these acoustic features. However, since many of these properties are highly collinear, accurately characterizing the acoustic profile of a sound becomes complex. A potential solution could be using models inspired by psycho-acoustical and neurophysiological findings in early and central stages of the auditory system, such as the modulation transfer function (MTF, Chi et al., 2005) that can simulate acoustic processing in the primary auditory cortex of humans (Santoro et al., 2014). We passed our sounds through the MTF model that generated a rich acoustic representation of the affective vocalization used in the study (see details of representation in Chapter 3 or 4 section 2, methods). After building a dissimilarity matrix, we illustrated the multidimensional scaling of the DSM, showing that the spectral-temporal modulations of the sounds across categories were fairly similar (Figure 5.1). This further bolsters our claim that the categorical representation of emotion categories are at least partially independent of acoustic properties of the sounds.

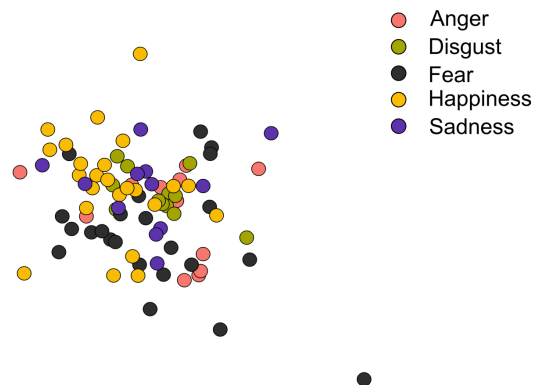


Figure 5. 1 Multidimensional scaling of acoustic representation of affective vocalizations used in the study across five emotion categories.

In an attempt to overcome the overwhelming number of acoustic properties influencing category perception, we introduced a control condition ‘scrambled condition’, where the frequency bins of the sounds were scrambled such that the frequency content was similar while the intelligibility of the sounds was disrupted (Dormal et al., 2018a). Although the emotion selective responses in the standard condition were significantly higher than the target response in the scrambled condition, the significant target response in the scrambled sequence can be attributed to some residual intelligibility (see the isolation task of the behavior experiment in Chapter 2) or acoustic features shared by sounds within an emotion category. Future studies should consider using other scrambling techniques to preserve the acoustic representation as much as possible while disrupting intelligibility. For instance, we used consecutive 200 Hz windows within which the frequency bins were scrambled (Dormal et al., 2018). Since the human processing of frequency is logarithmic and not linear (e.g., there is more perceptual difference between 100 Hz and 200 Hz tones opposed to 1100 and 1200 Hz), windows of logarithmically increasing lengths can be applied to make the spectral representation even more similar to the original sounds while disrupting the recognition of the sounds.

Furthermore, we simulated the cochlear response finding no differences between the responses to our selected sounds across different categories in both standard and scrambled sequences. However, it has also been reported that subcortical structures may initially process acoustics before relaying the processed output to the auditory cortical network. For instance, spectral and harmonic processing occurs in the cochlear nucleus (Rhode & Greenberg, 1994; Young & Davis, 2002), the medial geniculate body refines sounds by encoding speech-relevant temporal and spectral

patterns (Mihai et al., 2019). Future studies should aim to simulate different subcortical regions while carefully considering additional acoustic features, consequently regressing them out to isolate higher-order categorical selectivity to auditory emotion.

Different neural substrates for different emotion categories

In our study, we found different topographical responses to Fear and Happiness which could be due to the different categorical representation or different valence. Using frequency tagging, future studies can possibly investigate responses for other emotion categories which can validate whether different brain regions are recruited for other emotions. Nevertheless, a plethora of literature supports the hypothesis that post acoustic processing, different emotion categories are represented by distinct cortical pathways. (Fecteau et al., 2007; Grandjean et al., 2005; Hamann, 2012; Johnstone et al., 2006; Kotz et al., 2003; Mauchand & Zhang, 2022; Phan et al., 2002; Phillips et al., 1998). On one hand, many regions are common between these studies such as the anterior and superior temporal gyrus (STS), amygdala, inferior frontal gyrus, ventromedial prefrontal cortex amongst others forming a fronto-temporo-striatal network for emotion evaluation, categorization and appreciation (Frühholz & Grandjean, 2013; Kotz et al., 2013; Petrides, 2005; Schirmer et al., 2004). On the other hand, these brain areas are not consistently activated across conditions and studies casting doubt on whether these regions represent categorical information or other dimensions of emotion processing. The best example to demonstrate this is the amygdala, which was popularly known to be activated by fearful vocalizations (LeDoux, 1996; Adolphs et al., 1995), was later found to respond to arousal and emotion salience regardless of the emotion category presented (Phan et al., 2002; Vuilleumier, 2005).

The same way multiple features can be extracted from voice, emotions need to be represented as multidimensional entities carrying categorical, salience and affective information. It is easier to extract the emotion category of affective vocalizations, as validated by cross-cultural studies (Juslin & Laukka, 2003; Sauter et al., 2010; Sauter & Eimer, 2010; Scherer et al., 2001), but they can vary in terms of other dimensions (such as valence or arousal) causing inconsistent reports of brain regions facilitating emotion processing. A recent study did disentangle the time varying categorical and affective representations (valence & arousal) of auditory emotions by linking it with behavior ratings of emotions (Giordano et al., 2021). They compared the time-varying neural models created from MEG responses to affective vocalizations constrained from the brain regions of interest, and external representational models of categories and dimensional features, while partialling out the acoustics characterized by the temporal-spectral modulations of the sounds. Categories were found to emerge in middle STG early at 157 ms, while the dimensions dominated

later at 717 ms in posterior STG (Figure 5.2). This study reconciled the two opposing views by showing the interplay of cerebral networks with different representational dynamics emphasizing categories or dimensions.

In future, more theories need to be investigated on how humans actually conceptualize emotions. In addition to using advanced electrophysiological analysis methods such as frequency tagging or multivariate pattern analysis to extract neural patterns across emotions, further attention should be given to the stimuli itself and its multidimensional nature.

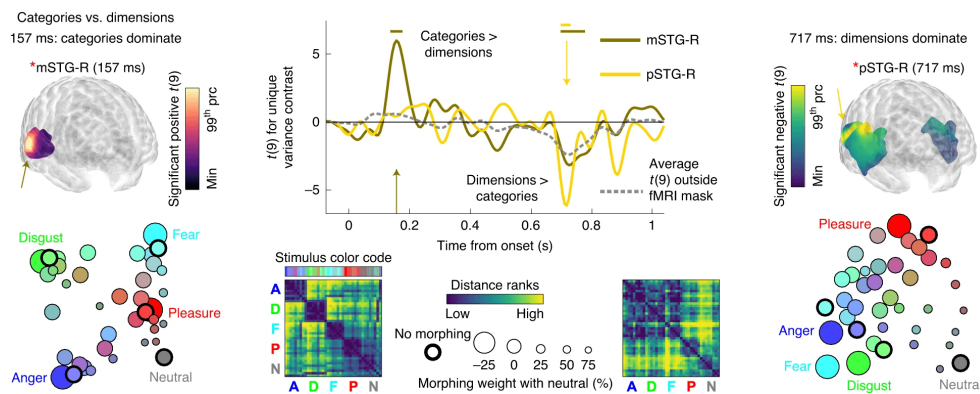


Figure 5. 2 Emotion categorical and dimensions representations in MEG, adapted from *Giordano et al., 2021*. Categories emerge earlier in time in the right middle superior temporal gyrus, while affective dimensions dominate in the right posterior middle gyrus later in time. The neural dissimilarity matrices and the multidimensional scaling highlight how sounds used in the study are arranged at the peaks of categorical and dimensional representations.

Fast Periodic Auditory Stimulation (FPAS)

We used the FPAS (Barbero et al., 2021) to investigate emotion selective responses in the adult human brain. FPAS is adapted from fast periodic visual stimulation (FPVS, Rossion et al., 2015) which has been used previously to investigate high-level responses such as faces, emotion and numerosity (Dzhelyova et al., 2017; Lochy et al., 2015; Marlair et al., 2024; Retter & Rossion, 2016; Rossion et al., 2015; Volfart et al., 2021). FPAS allows us to isolate the response at the predefined target frequency, indicating that the brain can not only *discriminate* sounds presented at the target frequency from other sounds, but also *generalize* across all target sounds to one common category. If the generalization across all target sounds is high, the periodicity in the neural signal is maintained and a response can be observed at the target frequency; however if there is no generalization between the target stimuli, then the aperiodicity will break leading to no response at the target frequency.

Opposed to other classical ERP studies, this method does not rely on subtraction of the target response from the baseline or control condition, thus, providing a *direct* measurement of the response. By predefining the presentation frequency, FPAS offers an *objective* approach to investigate neural responses. The method allows us to present a high number of heterogeneous stimuli, which to an extent, controls for the low-level acoustic bias in the target response. Furthermore, no overt responses are required from the participants, making this technique *automatic* and not contaminated by attention and decisional processes. Lastly, this fast experimental paradigm allows for testing populations that are typically challenging to test, such as infants and individuals with autism. Therefore, FPAS is a promising tool for investigating different conceptual information that can be extracted from a group of sounds with a high signal-to-noise ratio within a few minutes of recording (See section 4 of discussion for drawbacks).

5.2 Auditory Representations - Development and (Re)Organization

Nature endows us with an innate proto-neuroarchitecture for auditory functions at birth that are shaped by sensory experience. However experience can significantly vary, as in the case of people with congenital blindness and sighted individuals. The following sections will discuss auditory development in infants and the two trajectories of experience (with visual and without visual input) and their impact on neural representations of sounds processing and categorization. Note: The neurotypical adults from Chapter 3 and sighted controls from Chapter 4 of this thesis come from the same dataset.

5.2.1 Typical Development - Infants and Neurotypical Adults

In chapter 3, a delayed yet significant acoustic representation was found in infants, followed by a categorical representation across 8 presented categories, which was independent of the spectral-temporal modulations of the sounds. Although infants exhibited consistent and special neural selectivity to non-human mammalian sounds and human verbalizations, other auditory representations of sounds show non-linear dynamics as revealed by the sound-to-event deep neural network. These results indicate that a 4 to 7 month-old infant's brain shows categorical selectivity and undergoes early auditory development, but its dynamics still do not align with those of an adult.

To what extent the neonates' brain is equipped with the genetic blueprint to develop early auditory competence without which they cannot acquire auditory perception of speech and other sounds, and what are the mechanisms reliant on exposure to sounds to shape these auditory functions?

A significant amount of evidence now indicates that neonates exhibit certain auditory capabilities. For instance, early development for speech acquisition is suggested by evidence from studies showing enhanced behavioral responses to non-human primate and human vocalizations over synthetic sounds (Vouloumanos et al., 2010), sensitivity towards phonemes (Telkemeyer et al., 2009), consonants (Cabrera & Gervain, 2020), speech (Peña et al., 2003), preference to vowels similar to how adults map them (Aldridge et al., 2001) and segregate different auditory streams (Winkler et al., 2003). After a few months of experience, neuroimaging studies in infants showed a left frontotemporal network activated for speech processing (Dehaene-Lambertz et al., 2002). Similar to the pattern observed in adults, neural activation in infants hearing their native language is already concentrated in left-hemispheric perisylvian regions rather than being broadly distributed (Dehaene-Lambertz et al., 2002, 2006). A hierarchical processing of speech sounds in their native language was also revealed in infants in perisylvian areas similar to adults (Dehaene-Lambertz et al., 2006). These findings bolster the notion that early infants possess a proto-neuroarchitecture to acquire speech and language skills which develops further with exposure to sounds to reach adult-like auditory capabilities.

It is intriguing that, despite newborns not showing a preference for human sounds over animal sounds (Vouloumanos et al., 2009, 2010), they can still distinguish between different attributes of human speech. For example, they show a greater sensitivity to adult-like speech over backward speech (Peña et al., 2003) and can even differentiate between their mother's and father's voices (Beauchemin et al., 2011; Lee & Kisilevsky, 2014). One hypothesis suggests that neonates are exquisitely poised to "crack the speech code" (Kuhl, 2010), but may have underdeveloped regions to distinguish human vocalizations from other non-human vocalizations. For instance, the well known temporal voice areas (TVA, Belin et al., 2000), that respond more to human vocal sounds than environmental or animal sounds and are proven to be independent of low-level features of the sounds, develop around 7 months in infants (Grossmann, 2010), potentially enabling infants to differentiate human sounds from others. Thus, newborn responses to speech-like stimuli may not be supported by TVA, but from specific regions in the primary auditory cortex that are tuned to certain acoustic modulations specific to speech and/or higher-order networks that are known to be activated for speech sounds (Dehaene-Lambertz et al., 2002). Our results align with this hypothesis. In Chapter 3, we found no evidence of categorical selectivity for human sounds in 4- to 7-month-old infants, including both human verbal (HV) and non-verbal sounds. This suggests that the TVA may not have been fully developed in our sample of infants. However, we found that the latency of human verbal (HV) categorical representation in infants overlapped with the latency of spectral-temporal processing of sounds, suggesting a link between the two processes. Moreover, HV categorical representation diminished yet remained significant after controlling for the acoustic model, indicating

that HV processing is only partially dependent on acoustic processing and extends beyond mere spectral-temporal modulations. This aligns with research suggesting that infants possess a hierarchical structure for processing speech sounds as found in 3 month old infants (Dehaene-Lambertz et al., 2006). Further research using source localization or fMRI can reveal auditory networks that may support HV processing in early infancy, and whether the development of TVA after 7 months of age facilitates the discrimination of voices from other natural sounds.

Additionally, different processing timescales may support human meaningless speech and non-human mammalian sound representations. Supporting this hypothesis, different temporal trajectories of neural representations of human meaningless speech sounds and non-human mammalian sounds were observed in our sample of infants, where the latter occurred even before the emergence of acoustic processing. This may indicate alternate auditory representations enabling non-human mammal sound perception, the mechanisms of which need to be investigated further.

In adults, the representation of the HV was enhanced and occurred as early as the emergence of the acoustic processing of sounds, while the unique selectivity to non-human mammal sounds disappeared. These results collectively suggest that humans, early in life, are naturally inclined to develop speech and language skills, but also have a predisposition to lose certain innate functions over time. Another similar example can be found in early infants who can distinguish all phonemes across different languages, but by around 6-12 months, they lose sensitivity to sounds not used in their native language (Kuhl et al., 1992). This is because the brain is able to discard rarely activated connections due to limited exposure to certain stimuli such as animal vocalizations and phonemes of different languages (synaptic pruning) while simultaneously specializing in recognizing frequently encountered sounds (perceptual narrowing). These processes, enabled by neuroplasticity, are maximal during the critical period i.e., the first two-three years of life enabling infants to develop not only auditory but other sensory systems and cognition (Cisneros-Franco et al., 2020). Thus, early exposure to sounds, especially human sounds in early life, becomes a crucial factor for developing speech, language skills and auditory perception in general. Accounts of feral children further validate this, as infants who are raised by animals and are not exposed to human sounds during the critical period, potentially do not develop normal language skills in their life.

Apart from speech, neonates also show differential responses to musical sounds (Virtala et al., 2013), encode numerosity by preferentially looking towards similar numerous quantities in audio and visual modalities (Izard et al., 2009) and differential responses to a deviant pitch presented amongst standards (Stefanics et al., 2009), indicating early cognitive processing and development. In fact, newborns are shown

to extract statistical patterns of sounds suggesting that auditory development is well supported by cognition (Bosseler et al., 2016; Saffran et al., 1996; Teinonen et al., 2009).

Feedforward and feedback processing

We revealed in chapter 3 that adults encode acoustic properties of the sounds even before the complete presentation of the sound. This indicates feedback processes that may facilitate faster and more efficient recognition and categorization of sounds. The temporal latency of the acoustic presentation in our results match previous research employing multivariate patterns to investigate temporal dynamics of natural sound processing (Lowe et al., 2021; Ogg et al., 2020). This also opens avenues to investigate predictive mechanisms for acoustic processing. This effect may also be found due to habituation to the sounds, as adults paid attention to the sound as instructed by the task demands. Future studies could investigate this effect and find whether this feedback process depends on feedback mechanisms and predictive coding or task demands and memory operations.

In contrast, we found more feedforward processing in infants as the acoustic representations occurred after the presentation of the relevant portion of the sound. Indeed, the infants performed fewer trials than adults, however I hypothesize that feedback connections may be missing in 4-7 month old infants. Dynamic Representational Similarity Analysis with more sophisticated time varying acoustic models and brain region constrained neural models can shed more light on feedback mechanisms and their development in infants.

Lack of categorical abstraction in infants

In Chapter 3, adults showed emergence of significant representations of individual (8-category model) and abstract categories constituted by the 4-categorical model consisting of humans, animal, big environmental and places, and objects; and the 2-categorical model of animate vs inanimate sources of sound. In contrast, infants lacked abstraction, only displaying representation of the 8-categorical model and not for 2- or 4-categories (Chapter 3, Supplementary Figure S3.1). This finding contrasts with previous research suggesting that infants encode animacy features, as differential responses to combined human and human action sounds compared to environmental and mechanical sounds were observed (Geangu et al., 2015). However, our animate category constituted a wider range of sounds including mammal and bird sounds within the animate category, indicating that animate representation in infants may only be limited to human sounds.

Furthermore, an fMRI study showed early specialization for emotion processing in infants where they found that affective vocalization evokes greater response in the temporal cortex compared to non-human sound and neutral vocalizations (Blasi et al., 2011). Our results did not align with this finding as non-human verbalization, consisting of laughing and crying sounds, did not show any unique representation in infants, nor adults. This is because the unique human non-verbal categorical model assumes all other categories to be the same, which may not be the case. In contrast, our results reveal that infants uniquely encode human verbal sounds. Thus, future research requires use of models that consist of more nuanced relationships between categories, which can disentangle how infants and adults process and associate different categories.

Non-linearity in infant neural representations

In chapter 3, we found that temporally, the acoustic representation in infants and adults was followed by categorical representation, albeit at different latencies. These results align with the hierarchical view of auditory processing found previously in primates (Tian et al., 2001) and infants (Dehaene-Lambertz et al., 2006). However, infants exhibited non-linear processing with respect to representations of the deep neural network (DNN), where some early layers (1, 4) emerged before acoustic representation, most of the early layers (2 to 6) emerged after categorical representation, and intermediate to late layers (7,8,10,12) also emerged around similar time as that of the acoustic and categorical representations. Since these layers progressively emerged in adults, the DNN revealed that the temporal dynamics underlying auditory representations in 4-7 month old infants do not completely align with those of adults.

Interestingly, layer 4 of the DNN correlated highest with early EEG both in infants and adults, yet only in infants, the same representation emerged post sound offset. These recurrent auditory activations may potentially help refine and optimize auditory processing in infants, leading to development of perception, comprehension, and communication. As discussed in chapter 3 and 4, we found that the early DNN layers, including layer 4, can be best predicted by the acoustic properties of the sound (spectral-temporal properties, center of gravity, harmonicity, pitch) compared to behavior and categorical models, however most of the variance is unaccounted for (~ 0.75). One interpretation may be that these early layers represent intermediate-stage auditory processes that may bridge the link between acoustic and higher order semantic and categorical representations (Giordano et al., 2013, 2023). Further research on biologically plausible models could uncover alternative auditory representations (Kell & McDermott, 2019) that explain the missing variance, shedding further light on these auditory processes.

Finally, variability and nonlinearity of auditory representations in infants were also demonstrated by within group correlations in infants, indicating a lack of common sequence of neural processes that synchronize temporally (Chapter 3, Figure 3.5). Future research on different age groups can reveal the age when more stable auditory representations emerge, as observed in adults. This is exacerbated by a relatively bigger age range of 4-7 months, low signal-to-noise ratio and limited data which are discussed in the following section.

Methodological Considerations

Although most research reports milestones of auditory abilities from birth to adulthood, the neural mechanisms underlying these abilities are far less understood. This is because, despite studies utilizing near-infrared spectroscopy (NIRS), EEG, MEG, and, in rare cases, fMRI to identify neural substrates supporting these theories, they often lack certain essential methodological approaches. For example, many infant studies use fewer stimuli, which may not be enough to ensure the results are generalizable. They may also have small sample sizes and a wider age range due to limited access to infants, or rely on less effective artifact rejection methods and univariate analysis. These factors reduce sensitivity, which is especially important since infant neural data typically have low signal amplitude and high variability. Apart from testing constraints, there is a need for better methodological approaches that can find patterns within the data that univariate methods cannot.

As demonstrated in chapter 3, we applied multivariate pattern analysis on EEG of 4-7 month old infants to characterize the temporal dynamics of auditory processing of natural sounds. For this purpose we created pseudo trials to improve the SNR by compromising the number of trials used for cross-validation (Grootswagers et al., 2017). The process of making pseudo trials and decoding was performed in 100 iterations to remove the arbitrariness of pooling specific trials together. Since classifiers can be sensitive in finding patterns to separate classes, our decoding pipeline reduced the possibility of observing significant decoding due to noise. For this reason we discarded 12 babies who had insufficient data (less than 6 trials per stimulus type) or displayed more noise. Furthermore, in chapter 3 we emphasized on the peaks of representations rather than the onset and offsets as SNR can affect the latter and peak size, but not the latency of the peak (Xie et al., 2022).

We also cannot exclude the possibility that some differences may have emerged due to different age groups. Previous studies have found differences in neural signatures comparing 4 and 7 month old infants (Carr et al., 2021; Grossmann, 2010). In the near future, our study can be expanded by including a larger sample of infants within

narrower age ranges. This would enable a more detailed analysis of neural responses by age group, helping to uncover developmental differences in auditory representations.

There are several important considerations for future work related to this project. For instance, due to differences in signal-to-noise ratio (SNR) between the two groups, we did not conduct direct statistical comparisons. A logical next step could involve equalizing the SNR between infants and adults to allow statistical comparison. The application of multivariate noise normalization techniques may also help reduce noise, particularly in infant data (Guggenmos et al., 2018). Another potential strategy for addressing variability in infant responses would be to model subjects as random factors. However, direct comparisons between infants and adults remain challenging, as adults engaged in an active task during the experiment. Future studies should carefully consider cognitive and task demands when designing cross-group comparisons, ensuring that experimental conditions are as closely matched as possible. Despite these limitations, the current findings offer compelling evidence for the presence of previously unmapped representations, offering new insights into the emergence of categorization and cognitive development in infancy.

5.2.2 Reorganization - Blind and Sighted

“How does the brain of blind individuals adapt and reorganize the sensory deprived occipital cortex to process and categorize auditory objects, and what neural mechanisms drive these adaptive processes?”

Enhanced behavior performances of early blind (EB) people compared to sighted controls (SC) in feature based and spatial auditory tasks are suggested to be linked to plasticity driven functional and structural modulations of cortical areas due to absence of visual input (Bavelier & Neville, 2002; Heimler et al., 2014; Sabourin et al., 2022). As mentioned in the introduction, numerous neuroimaging studies have shown that the visual cortex is engaged in non-visual functions, primarily those related to cognition (Bedny, 2017; Bedny et al., 2011; Collignon et al., 2007, 2009; Dormal et al., 2016, 2018b; Peelen et al., 2013; Pietrini et al., 2004). However, fewer studies have explored categorical representations of natural sounds in the posterior cortex of the blind. In this study, we characterized the large-scale temporal dynamics of auditory processing and categorization in early blind (EB) and sighted controls (SC), an investigation which has not been conducted before. Although most related topics were addressed in the discussion of Chapter 4, a few additional points are discussed further in response to the question above.

Factors Behind Enhanced Decoding in Blind

Although both groups displayed similar shaped classification curves, the decoding AUC was significantly higher in EB than SC where the difference originated mostly from occipital sources throughout the duration of the sounds. Additionally, auditory responses were more consistent in EB resulting in a higher noise ceiling. This trend was also found in Mattioni et al., 2020, where they found a higher noise ceiling in VOTC for EB than SC. Moreover, the univariate EEG responses of EB were higher (Chapter 4, Supplementary Figure S4.4), which may enhance the separability between sounds or categories, as classification algorithms are particularly sensitive to variability in neural signals (King & Dehaene, 2014). The involvement of the visual cortex could indeed lead to higher univariate responses in EB, however factors such as synchrony of neural firing and orientation of alternate neural generators could also cause larger univariate responses.

Despite the decoding enhancement in EB, the latencies of the peaks were similar in both groups. In fact, most auditory representations aligned temporally between the two groups (Chapter 4, Figure 4.5). However, differences between EB and SC emerged in terms of stabilization or generalization of neural signals, where the neural patterns of the latter could be generalized earlier in time, the former showed unraveling of a chain of unique neural processes, consequently stabilizing later in time. This indicates that the involvement of the visual cortex may extend auditory processing in time in EB showing significant decoding throughout the duration of the sound.

Theta Oscillations Integrate Auditory Information at ~200 ms

The decoding peak in both groups was found at ~200 ms post sound onset, aligning with other studies reporting similar latencies (Lowe et al., 2021; S. V. Norman-Haignere et al., 2022; Ogg et al., 2020). This latency also overlaps the ERP P200 which is known to process acoustic properties (Caclin et al., 2006; Giard et al., 1995; Näätänen & Picton, 1987; Roberts et al., 2000; Rosburg, 2003), and is a distinct ERP for speech stimuli (Charest et al., 2009). This is supported by the fact that the early DNN layers and the spectral-temporal acoustic model (Modulation Transfer Function, MTF) peak around the same time in blind. 200 ms is also known to be a merging latency for auditory responses from shorter integration windows (<200 ms) that process acoustic properties of the sounds and a longer integration window (>200 ms) that may represent categorical information (S. V. Norman-Haignere et al., 2022). Thus, in our study, the decoding analysis along with RSA results for MTF and the early DNN layer models signify that the different auditory representations of a mixture of complexity reaches their peak at approximately 200 ms.

Investigating oscillatory activity can reveal mechanisms facilitating neural processes

such as attention modulation, neuronal communication, memory consolidation or especially in our case, auditory feature binding (Gourévitch et al., 2020). In addition to the results we performed a supplementary time-frequency decoding analysis. First, we decomposed the EEG trials to 45 frequency bins placed logarithmically between 2 and 30 Hz and consequently, pseudo trials were computed for every frequency bin separately. Using the same analysis pipeline we conducted in Chapter 3 and 4 (see section 4.2.4.2 and 3.2.4.2), we performed a pairwise time-frequency decoding analysis which yielded a decoding AUC for each timepoint, frequency bin, binary comparison and participant. This further decomposes the signal to show which frequency bins between 2-30 Hz contain auditory related information and when. We found that the theta band exhibits neural responses that can differentiate sounds in both EB and SC (Figure 5.3). Additionally, frequencies in alpha and early beta bands were also significant separately in EB and SC but showed lower decoding AUC compared to the early theta band. This suggests that theta oscillations get phase-locked to sound onset and carry the maximum information differentiating the sounds. On contrasting the two populations, we found enhanced theta activity peaking at 6.733 Hz at 171 ms. Thus, the enhanced activity in theta band subserves the enhanced early decoding in EB compared to SC. As we found high univariate ERP for EB, we can predict that they may have an enhanced power in the theta range. Collectively, we can conclude that the theta band is modulated due to visual deprivation causing enhanced auditory responses in EB which peaks at ~200 ms and may facilitate synchronizing different auditory information in the human brain.

Which auditory processes do theta oscillations facilitate? We created neural DSMs specific to the theta band (3-7 Hz) and correlated them with the MTF model and the DNN layers. We found that while MTF representations emerged early and similarly in EB and SC, the DNN layer 4 representation (best performing layer in infants, SC and EB early in time) was enhanced in EB compared to SC. It could be concluded that the low-intermediate level auditory representations are mediated by theta oscillations in both groups. Visual deprivation modulates the theta activity and thereby the representation of specific auditory representations in EB. Future studies should investigate the neural substrates underlying theta activity in both groups and explore the relationship between the development of theta oscillations in infants and auditory development.

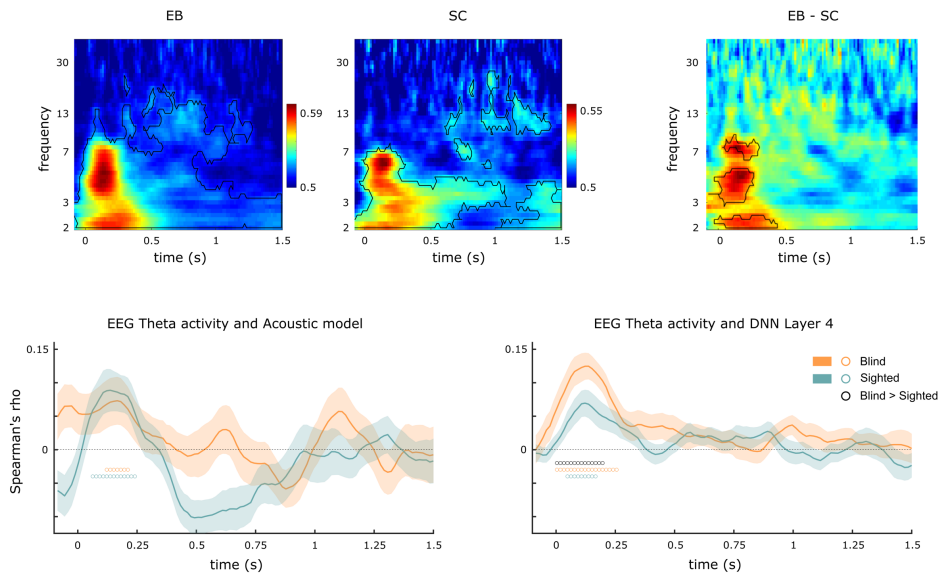


Figure 5. 3 Enhanced decoding AUC in the theta band in EB than SC. Theta oscillations carry different auditory information unique to different sounds; however, visual deprivation increases the information carried in the theta band. The enhanced representation in EB does not involve early spectral-temporal modulations of the sounds, but includes auditory representations that are of low-intermediate complexity as demonstrated by the DNN Layer 4.

Late Enhanced Abstracted Categorical Representation in Blind

EB showed enhanced categorical and behaviour representation at ~600 ms post sound onset compared to SC using the 8-, 4-categorical model and similarity ratings, respectively. Although these external representation models were highly collinear, the temporal dynamics of EEG were different between EB and SC, suggesting different neural processes subserving categorization and behavior. In fact, this late categorical representation could be best explained using the ‘interaction’ based model which clustered humans and tools vs animals and big mechanical and environmental sounds differently (Chapter 4, Figure 4.3f). Despite distinct acoustic properties of these combinations, the neural patterns of EB abstracted away to form a categorical representation later in time, supported by the cross-category decoding analysis (human, tools vs big, animals: Chapter 4, Supplementary Figure S4.5). The neural sources of this representation were localized to the occipital-temporal and parietal regions. The question persists - why these combinations of categories? Is it due to the closer association of human and human made object sounds, and/or due to low exposure and internal representation of animal and big environmental and mechanical sounds? (See more in the discussion of Chapter 4)

Although many studies have found categorical representations to sounds in temporal-occipital brain regions of EB (Handjaras et al., 2016; Mattioni et al., 2020; Peelen et al., 2013, 2014; van den Hurk et al., 2017), the categorical pattern found in our study, unexpectedly, has not been reported in previous studies. Indeed we used the same sounds from Mattioni et al., 2020, 2022, however the qualitative results did not align with these studies potentially due to the limited range of time of significance of the categorical representation (~100-200 ms) in our results and the low temporal sampling of the fMRI. Alternatively, smaller regions of interest (ROI) or alternate brain regions could be investigated to find common trends between EEG and fMRI responses. Future research should focus on testing the same participants with both modalities (fMRI + EEG) and fuse them to characterize the spatio-temporal dynamics of auditory object processing (see Lowe et al., 2021).

Nevertheless, two interesting questions emerge - firstly, could this late enhanced categorical representation in EB be invariant to different formats of presentation within the same modality (e.g., sounds and spoken words)? Secondly, could this representation be consistent across different sensory modalities (e.g. sounds and tactile objects)? While semantic or categorical selectivity has been demonstrated for words and sounds in blind people respectively, typically recruiting visual cortices for their function (Abboud et al., 2019; Bedny et al., 2011; Handjaras et al., 2016; Mattioni et al., 2020; Peelen & Downing, 2017; van den Hurk et al., 2017), an investigation is warranted to find common representations across formats within the same modality in the same blind participants. Despite evidence pointing to the dissociation of verbal and non-verbal networks in sighted (Thierry & Price, 2006), how blindness may impact the dynamics of these networks extending to occipital regions is yet to be known. Supporting evidence from vision for common conceptual neural substrates across images and written words in sighted individuals (Devereux et al., 2013; Simanova et al., 2014), along with the ability to disentangle perceptual and contextual aspects of visual processing through multivariate approach (Peelen & Caramazza, 2012), suggest the possibility of finding shared conceptual representations between sounds and words in both EB and SC. Although linguistic processes tend to involve more top-down processing and sounds are more bottom-up perceptual (Martin et al., 2018), exploring these dynamics could reveal how conceptual representations are coded in EB across different presentation formats.

Furthermore, it would be interesting to find if the categorical representations evoked by sounds in EB overlap with the same in tactile modality. As discussed in Chapter 4, a potential factor for the qualitative clustering of categories in EB could be the tactile experience of the participants with these objects (thus called the 'interaction' model). Evidence supporting cross-modal interactions comes from a study finding increased gamma activity in the visual cortex of blind people during incongruence between sounds and tactile objects (Schepers et al., 2012). Interestingly, the latency

of enhanced gamma activity peaked around 400-600 ms post stimulus onset (Schepers et al., 2012). If we discover a similar categorical structure in EB using tactile stimuli and can successfully cross-decode across modalities, it would suggest that certain posterior brain regions in the blind represent supramodal representations. This means that the occipital cortex in EB may exhibit higher-order representations that abstract away from their respective sensory properties, processing similar amodal conceptual information. Our cross-decoding results showing the emergence of categorical information abstracted from the acoustic properties support this hypothesis, at least in the context of sound processing (Chapter 4, Supplementary Figure S4.5). Alternatively, similar auditory and tactile patterns in the occipital cortex of EB may indicate that some regions are multimodal in nature i.e., different neurons in a small brain region can be activated by distinct modalities. Overall, further research on these lines would highlight the effects of cross-modal interactions on supramodal representation of external objects in blind people (Cecchetti, Kupers, et al., 2016).

Human vocalizations, especially speech sounds, are known to elicit enhanced responses and are also encoded in the occipital cortex in EB (Dormal et al., 2018b; Van Ackeren et al., 2018). In fact, it was shown that the human sounds could explain most of the variance in the responses from temporal and occipital ROIs in both early and late blind people (Mattioni et al., 2022). Similar trends were reflected in our results where the human model correlated maximally with EEG in both EB and SC (Figure 5.4). This reinforces the fact that human sounds elicit unique responses which are not only restricted to the temporal areas (Gougoux et al., 2009; temporal voice area TVA, Belin et al., 2000;), but also extend to the occipital regions in EB. Therefore, future studies investigating natural sound processing should control for the fact that voices are indeed distinctly represented in the human brain and analyse other artificial sounds separately.

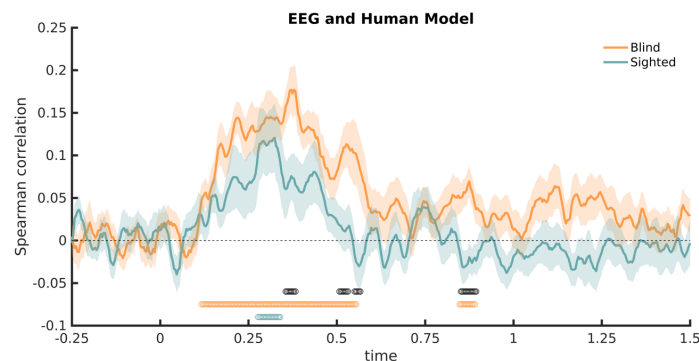


Figure 5. 4 Representation of human model with time-varying EEG in EB and SC after partialling out the acoustic model. Human sounds elicit enhanced correlation in EB spanning a long range of time.

The mechanisms that lead to the development of such categorical representations require further future research. In EB, visual cortices are known to be functionally coupled with fronto-parietal networks involving these neural substrates in language, numerical processing and executive functions (Bedny et al., 2011; Liu et al., 2007; Striem-Amit et al., 2012; Watkins et al., 2012). On the other hand, heteromodal connections between the visual and auditory cortex could facilitate engagement of both cortices during auditory tasks in EB (Collignon et al., 2013). This is also supported by the finding of similar sound representations in the occipital and temporal cortices of EB during sound presentation (Mattioni et al., 2022). Indeed, all these findings highlight that visual deprivation impacts these connections leading to the functional recruitment of the visual cortex for non-visual tasks, however whether they are recruited for perceptual and/or cognitive processing is not very clear. On one hand language related tasks were found to activate the occipital cortex in EB, with more complex words/sentences showing higher activation (Bedny et al., 2011), suggesting that the occipital cortex undergoes profound functional reorganization to perform cognitive functions (Saccone et al., 2024). On the other hand, brain regions such as the VOTC is known to preserve its innate computation prototype for semantic processing but switches to non-visual modality like sounds (Mattioni et al., 2020; van den Hurk et al., 2017) and words (Handjaras et al., 2016; Peelen & Downing, 2017). Our study does not resolve this debate, yet investigating time scales of different auditory processes and categorical representations using more nuanced models can help future studies to find the ground truth.

Lastly, future studies could compare temporal dynamics of auditory categorization in both early and late blind people. Since both groups have been found to exhibit similar categorical structure in the occipital cortex (Mattioni et al., 2022), the exploration of the temporal dynamics subversing sound categorization may reveal more nuanced differences between the two populations. Future studies may also reveal information about visual imagery, as late blind individuals may experience visual imagery of auditory objects, which have been known to evoke patterns similar to those associated with actually seeing the same object (Reddy et al., 2010). Depending on the age of acquiring blindness and its impact on auditory functions, would early visual exposure enable late blind people to conceptualize auditory objects more similar to early blind or sighted controls?

Support for Cortico-Cortical Connections in Blind

The propagation of auditory signals to the visual cortex can essentially be explained by either or a combination of two pathways, thalamo-cortical and cortico-cortical connections. While previous research has deemed the former improbable due to degeneration of the lateral geniculate nucleus (LGN) in EB (Cecchetti, Ricciardi, et

al., 2016), yet an MEG study found this pathway mediating tactile information to the occipital cortex of EB (Müller et al., 2019). On the other hand, other studies claim the role of cortico-cortical connection in mediating responses in the visual cortex of EB for spatial tasks (Collignon et al., 2007), tactile tasks (Ptito et al., 2008; Wittenberg et al., 2004) and sounds (Klinge et al., 2010). One perspective to find the relevant pathway is to investigate the timing of responses in the visual cortex (Müller et al., 2019). Thalamo-cortical routes are faster in conduction due to monosynaptic connections and consist of fewer synaptic relays than cortico-cortical pathways. Thus, If visual cortex activation in response to sounds occurs within 50 ms of sound onset, it likely follows a thalamo-cortical pathway; if the activation occurs later, it is more likely mediated by cortico-cortical connections.

Although we reported significant sound decoding enhancement in EB from 160 ms post sound onset (Chapter 4, Figure 4.2), sounds could be decoded from the visual cortex as early as 80-100 ms, as demonstrated by source-space decoding (Figure 5.5). This supports that the sound propagates using cortico-cortical connections in EB. This is further supported by the findings of connections between primary sensory cortices from tract tracing and imaging studies in humans (Eckert et al., 2008) and animals (Ghazanfar et al., 2005; Kayser et al., 2005).

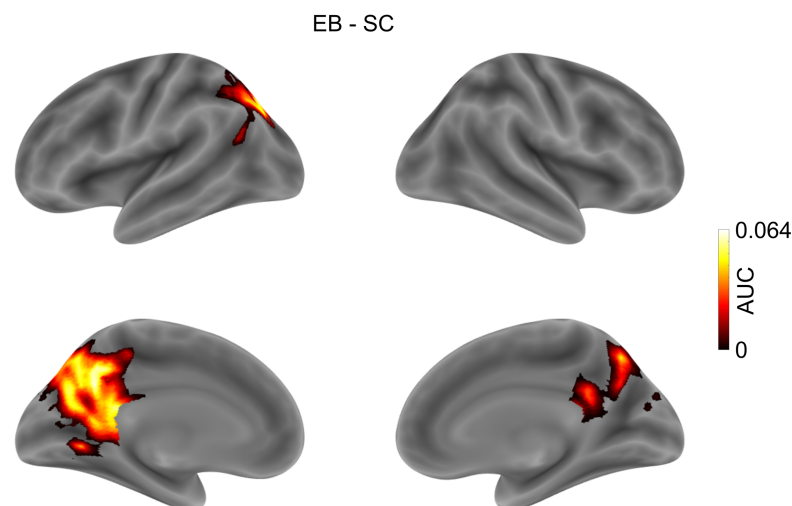


Figure 5. 5 Pairwise sound decoding (Early Blind - Sighted Control) found earliest between 80 - 100 ms. The signal to the visual cortex is thus potentially mediated by cortico-cortical connections rather than thalamo-cortical.

5.2.3 Atypical Development - Infants and Blinds

In this thesis, we did not compare the auditory representations between infants and blind, which can characterize atypical auditory development, predominantly due to the fact that the early exposure to vision may already develop visual related brain regions in infants. As this would not occur in the case of congenitally blind individuals, the ideal comparison for blind adults should be with infants with congenital blindness. However, some interesting patterns were found between the two groups that can potentially provide insights about the mechanisms of auditory processing in the two populations.

Comparison of Neural Models

We correlated the neural DSMs between the two groups at each combination of timepoints (Figure 5.6). We found large significant off-diagonal clusters from 200 ms to 1100 ms in infants and 1200 ms to 2000 ms in early blind individuals (EB). The maximum correlation was higher compared to infant-adult/sighted control (SC), maybe due to higher consistent auditory responses found in blinds than sighted controls (SC) / adults (see Chapter 4, Figure 4.5). An interesting pattern found in infant-blind and infant-adult comparisons was that the peak correlation was found around 200 ms in infants and ~1350 ms in both adults and EB. These results reflect the underdeveloped auditory development in 4-7 month old infants, highlighting the slow maturation of auditory processing. Nevertheless, using this approach, it would be interesting to find at what age do these representations align in a diagonal form as observed for EB-SC or within group comparisons of EB and SC.

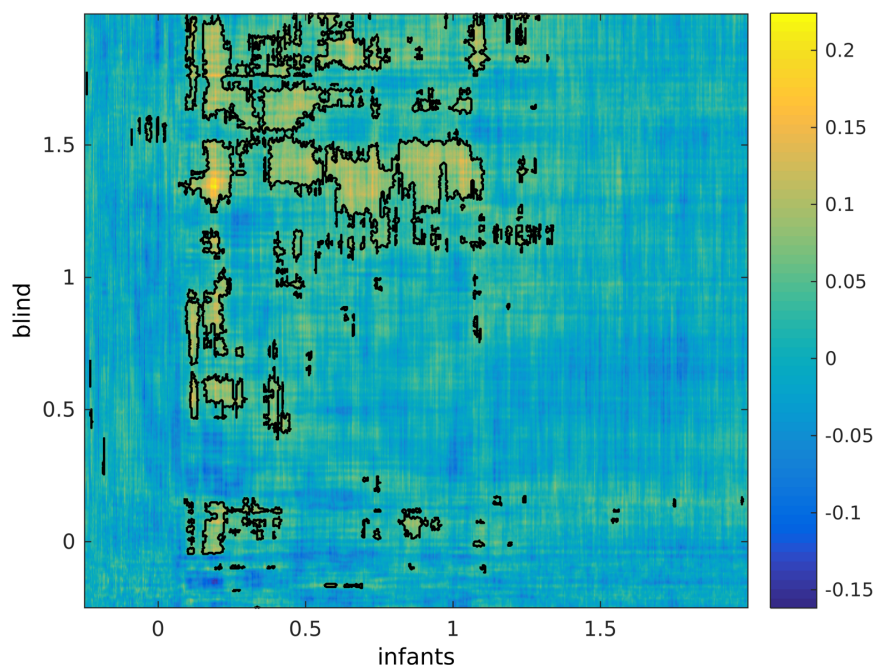


Figure 5. 6 Correlations between neural DSM of infants and early blind adults at all timepoint combinations.

Recurrent Early Deep Neural Network Representations

In EB and infants, we found a recurrent auditory representation characterized by layers number 2 to 5 of the DNN (YAMNet), which emerged post sound offset (in EB ~1050 ms, in infants ~1350 ms), significantly more than the SC/adult group (Figure 5.7). We speculate that this recurrent representation in infants, which is absent in SC, is related to refining and optimizing auditory processes, leading to development of perception, comprehension, and communication. Why visual deprivation causes preservation of this auditory representation in blind adults is indeed a mystery.

Interestingly, the same representation is enhanced in the early EEG of EB compared to SC. This suggests that visual exposure plays a crucial role in shaping auditory development, while its absence may leave certain auditory processes enhanced or underdeveloped. In the development of infants, new connections are formed and lesser used connections disappear (synaptic pruning), and this result may indicate preservation of processes due to no visual experience. In summary, according to our results, visual deprivation causes an intermediate-stage auditory representation to be enhanced early in EEG, which are facilitated by theta oscillations (see section 2.2

of discussion), while preserving its recurrence post sound offset, which is not linked to oscillatory activity, throughout the course of development.

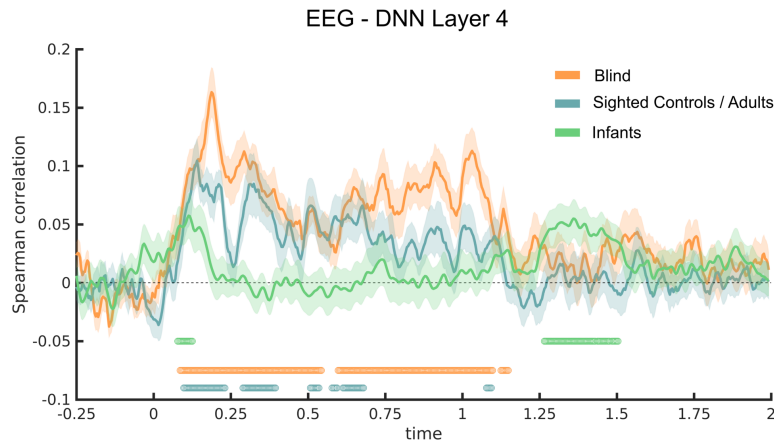


Figure 5. 7 Representation of DNN Layer 4 in time-varying EEG of all three populations. The representation peaks early in time in all groups, but emerges again after sound offset in only EB and infants.

5.4 Linking Results with Theories of Categorization

In this thesis, we found quantitatively enhanced and qualitatively different categorical representation in the blind compared to sighted control, signifying the impact of sensory experience on auditory categorization. But how these categories are embedded, or rather conceptualized in the human brain demands future research. One of the most pursued theories are statistical theories of categorization that suggest that individuals form conceptual categories based on the probabilistic structure of sensory experiences, allowing for flexible and adaptive classification of stimuli. Under the umbrella of the statistical models, the prototype theory assumes that every category has a prototype stored in memory (Rosch 1973) suggesting that the closest match of the incoming stimulus with different prototypes may evoke a particular category. The exemplar theory goes beyond the prototype theory, stating that categorization is based on all the previously categorized exemplars across all categories. Both theories indicate that defining a specific categorical space relies on the accumulation of sensory attributes of an object, which can differ in blind and sighted people. Therefore, investigating how visually rich versus visually non-rich objects are represented in blind individuals presents a compelling direction for future research, potentially revealing how categorical space is embedded in humans. For instance, consider the concept of wind, an entity primarily experienced through auditory and tactile modalities and not typically defined by visual characteristics. Would such a stimulus evoke comparable categorical representations in the brains of blind and sighted individuals? In contrast, an object like a guitar, which can be

heard, touched, and seen, may yield divergent neural representations across these groups, given the differing sensory inputs available to construct its conceptual category. Furthermore, whether the boundaries between categorical spaces are diffused or sharp is not known. Usage of morphed stimuli could help to extract these boundaries.

Moreover, infants have been found to extract statistical patterns of sounds (Bosseler et al., 2016; Saffran et al., 1996; Teinonen et al., 2009), suggesting that definition of categorical spaces may occur early in infancy. Interestingly, most evidence for statistical learning comes from studies using speech or speech-like sounds, indicating that a developed neural architecture, attuned to specific sound features, is necessary for such learning to occur. Although the stimuli set used in Chapter 3 for infants and Chapter 4 for blind people were similar, future research should focus on sounds which are encountered statistically more by infants. For instance, infants may not listen to many exemplars of church bells (part of our stimulus set), however the exposure to various sounds of toys would be higher. Thus, stimuli selection can be performed more carefully in future. Furthermore, future research with biologically plausible computational models could also shed light on the dependency of the development of sound categories on the maturation of brain regions computing for the low-level properties of the sounds.

5.5 Deep Neural Network (DNN): YAMNet

This is the first known EEG study that employed a specific sound-to-event DNN - YAMNet, which has been previously used in only two reported fMRI studies investigating auditory representations of natural sounds. Firstly, Giordano and colleagues investigated if acoustic, YAMNet or semantic models can best predict 7T fMRI responses from the Heschl's Gyrus (HG) and posterior superior temporal gyrus (pSTG, Giordano et al., 2023). They found HG could be well predicted by both acoustic and DNN layers however only the DNNs, which partly shared representations with both acoustic and semantic models, best predicted the responses from pSTS above the noise ceiling. Specifically, they found that layer 8 consisted of the most amount of information related to both HG and pSTG.

To illustrate a comparison, we extracted the sensor level neural models from the SC/Adult group at peaks of 140 ms and 316 ms corresponding to the first two peaks that consistently emerged by correlating the EEG and DNN layers (see Chapter 4, Supplementary Figure S4.2). Then, I illustrated the correlation of the group mean neural DSMs at these peaks and the DNN layers (Figure 5.8c). We found that while EEG at 140 ms is best explained by layer 4, the neural DSM at 316 ms can be best predicted by layer 8. This suggests firstly, the hierarchical nature of auditory processing as earlier layers correlated higher with the early neural DSM and intermediate-late layers with the late neural DSM. Secondly, layers 7-8 exhibit an

Chapter 5. Discussion

inflection point where an intermediate scale of auditory complexity is represented later in time. Lastly, layer 8 which maximally shares the same representation of HG and pSTG in Giordano et al., 2023 could potentially peak at ~300 ms as per these results.

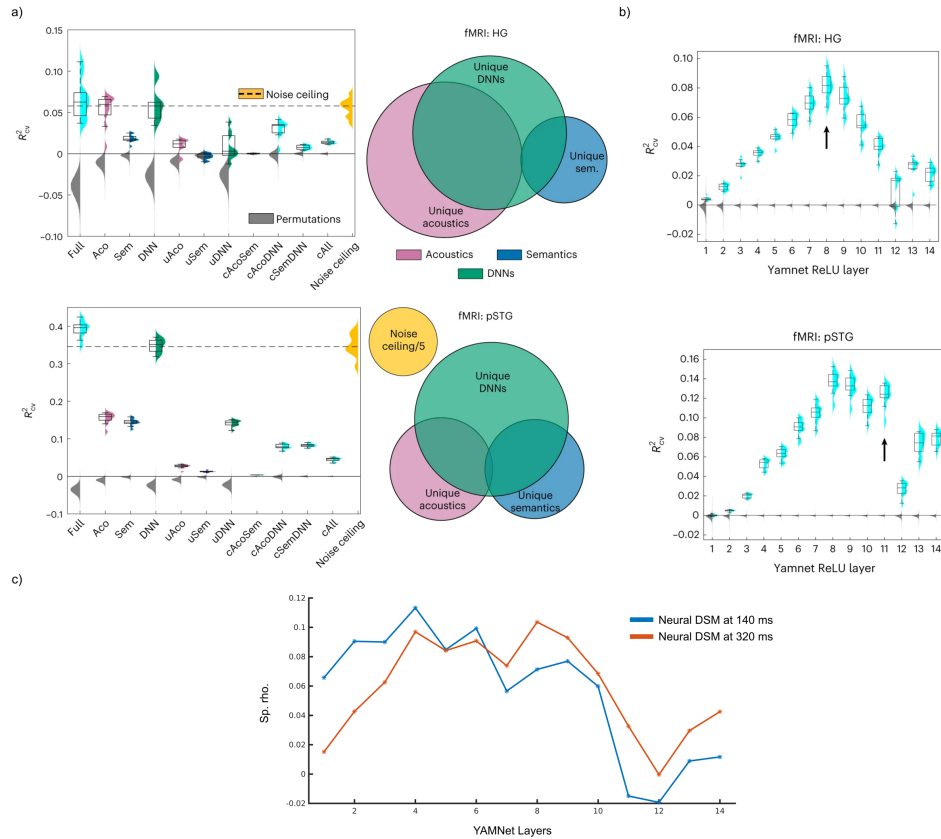


Figure 5.8 a) and b) Adapted from Giordano et al, 2023. a) Unique and combined variances of all models used in the study - acoustic, DNN and semantic models. DNNs can predict the neural responses from Heschl's Gyrus (HG) and posterior superior temporal gyrus (pSTG) well up to noise ceiling. b) Layer 8 shows higher predictive variance amongst all layers in both regions of interest. c) Results from Chapter 4 demonstrate progressively complex layers emerge in EEG with time. Layer 4 correlated maximally with early neural model, while layer 8 with late neural model.

The second study implemented intracerebral recordings at different positions in the auditory cortex and built encoding models (Rupp et al., 2024). These models were then predicted by DNN layer representations. They found that the complexity of a cortical site (belt to parabelt region) correlated with the depth of the DNN. The complexity increased from posteromedial-anterolateral direction in core and lateral belt, and posterior-to-anterior and dorsal-to-ventral dimensions in the parabelt,

indicating the hierarchies present within the primary auditory cortex (PAC). Using RSA with large-scale high density EEG of SC/Adult group and external auditory models, we demonstrated that the hierarchical auditory processing is not only restricted to the PAC, but also extends to the fronto-temporal auditory network (Figure 5.9). MTF can be localized close to the PAC early in time at 80-100 ms (first peak of MTF, see Chapter 4, Figure 4.3a), then in left frontal areas later at 180-200 ms (second and highest peak). At 140-260 ms, we found DNN layer 4 representation emerging in the right temporo-frontal network. Lastly, layer 8 was found in superior and medial frontal regions representing more cognitive processing. Statistics were performed with sign permutation test, $p < 0.01$, cluster-wise correction.

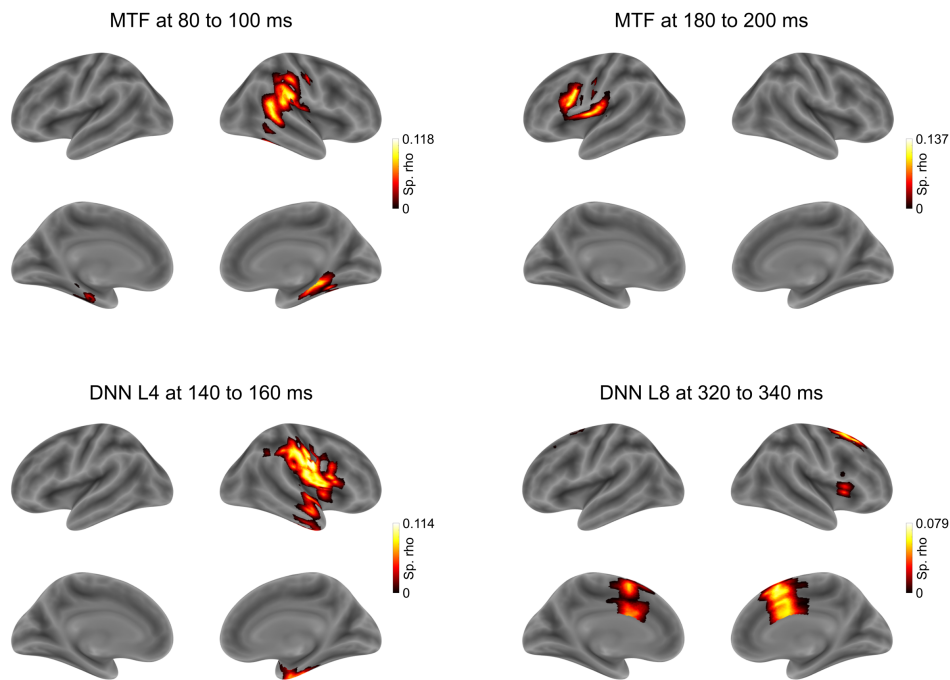


Figure 5. 9 The representation of auditory models in space and time. MTF emerges earliest in sensory regions and propagates to frontal networks. DNN Layer 4 is represented in the fronto-temporal network early in time and Layer 8 in superior temporal regions later in time. Statistics performed with sign permutation test ($p < 0.01$), cluster-wise correction.

Future research should focus on fusing fMRI and EEG together to reveal spatial and temporal dynamics of auditory representations (Cichy et al., 2019; Lowe et al., 2021). Furthermore, other DNNs such as VGGish, Kell's network can also be used to characterize auditory representations depending on the task and complexity. The development of DNNs with more biologically plausible representations in the future

could reveal detailed auditory hierarchy leading to behavior. While most results support the hierarchical point of view for auditory processing, advanced recurrent DNN models may also shed light on parallel processing which involve both feedforward and feedback across different brain regions in the auditory network.

5.6 Frequency Tagging vs Multivariate Pattern Analysis

We implemented two electrophysiological methods in this thesis. Both methods have their advantages, but some considerations for future research are listed here. Frequency tagging technique is limited to testing one higher-order selectivity at a time e.g., the target stimuli in the first study belonged to one emotion category. While stimuli from other emotions were also presented, it did not allow us to characterize their responses. Due to this limitation, combined with testing time constraints, we were unable to include other emotion categories such as anger, disgust, or sadness. If we had found responses to those emotion categories, that would have bolstered our claim of automatic emotion categorization, as well as, highlighted that valence does not contribute to the emotion selective response. Thus, FPAS is a fast, automatic, direct and objective tool if a single higher-order representation needs to be isolated across different stimuli, otherwise classical ERP style experimental paradigm in conjunction with multivariate pattern analysis could be a better approach to characterize responses of all presented stimuli. However, the latter requires more testing time and more repetitions of the same stimuli leading to the presentation of a lesser number of heterogeneous stimuli. For these reasons, we could present 84 affective vocalizations within 1 hour of testing for the first study, while acquiring optimum number of trials for 24 natural sounds in chapter 3 required 1.25 hours of testing in adults.

Additionally, FPAS is blind to the temporal onset of neural responses, but is dependent on their periodicity. Thus, if the question relates to the timing of neural processes, timepoint-by-timepoint MVPA can be implemented on time-varying epochs. Lastly, these techniques could also be combined in future. For instance, one study used frequency tagging to acquire neural responses to different rhythm sequences and used MVPA in conjunction with Representation Similarity Analysis (RSA, Kriegeskorte et al., 2008), to investigate the categorical perception of rhythmic structures (Barbero et al., 2025).

Lastly, RSA allowed us to abstract the neural and other external models (such as the acoustic model) from their native space to similarity space for comparison. A consideration for this method that consequently affects interpretation are the metrics of distances used to build the dissimilarity matrices (DSM). In this thesis, we used the decoding AUC to quantify the distances, however the possible gradation of AUC for each instance could only be 0, 0.5 and 1, due to using 2 instances for testing

during the cross-validation step. This may result in loss of information as the distances are discretized (Walther et al., 2016). An alternate method is to use continuous distance measures such as euclidean and cosine distances. The latter was used to create the external representational models in chapter 3 and 4. These continuous distances provide unbiased distance estimates and more replicable DSMs (Guggenmos et al., 2018). Future research should consider different distance metrics for implementing RSA.

5.7 Bibliography

Abboud, S., Engemann, D. A., & Cohen, L. (2019). *Semantic coding in the occipital cortex of early blind individuals* (p. 539437). bioRxiv. <https://doi.org/10.1101/539437>

Aldridge, M. A., Stillman, R. D., & Bower, T. G. R. (2001). Newborn categorization of vowel-like sounds. *Developmental Science*, 4(2), 220–232. <https://doi.org/10.1111/1467-7687.00167>

Allen, E. J., Burton, P. C., Olfman, C. A., & Oxenham, A. J. (2017). Representations of Pitch and Timbre Variation in Human Auditory Cortex. *Journal of Neuroscience*, 37(5), 1284–1293. <https://doi.org/10.1523/JNEUROSCI.2336-16.2016>

Alluri, V., Toivainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, 59(4), 3677–3689. <https://doi.org/10.1016/j.neuroimage.2011.11.019>

Aramaki, M., Besson, M., Kronland-Martinet, R., & Ystad, S. (2011). Controlling the Perceived Material in an Impact Sound Synthesizer. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2), 301–314. *IEEE Transactions on Audio, Speech, and Language Processing*. <https://doi.org/10.1109/TASL.2010.2047755>

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3), 614.

Barbero, F. M., Calce, R. P., Talwar, S., Rossion, B., & Collignon, O. (2021). Fast Periodic Auditory Stimulation Reveals a Robust Categorical Response to Voices in the Human Brain. *eNeuro*, 8(3). <https://doi.org/10.1523/ENEURO.0471-20.2021>

Barbero, F. M., Talwar, S., Calce, R. P., Rossion, B., & Collignon, O. (2025). Intermodulation frequencies reveal common neural assemblies integrating facial and vocal fearful expressions. *Cortex*, 184, 19–31. <https://doi.org/10.1016/j.cortex.2024.12.008>

Bavelier, D., & Neville, H. J. (2002). Cross-modal plasticity: Where and how? *Nature Reviews Neuroscience*, 3(6), 443–452. <https://doi.org/10.1038/nrn848>

Bedny, M. (2017). Evidence from Blindness for a Cognitively Pluripotent

Cortex. *Trends in Cognitive Sciences*, 21(9), 637–648.
<https://doi.org/10.1016/j.tics.2017.06.003>

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, 108(11), 4429–4434.
<https://doi.org/10.1073/pnas.1014818108>

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding Voice Perception. *British Journal of Psychology*, 102(4), 711–725.
<https://doi.org/10.1111/j.2044-8295.2011.02041.x>

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3), 129–135.
<https://doi.org/10.1016/j.tics.2004.01.008>

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767), 309–312.
<https://doi.org/10.1038/35002078>

Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., Deeley, Q., Barker, G. J., Renvall, V., Deoni, S., Gasston, D., Williams, S. C. R., Johnson, M. H., Simmons, A., & Murphy, D. G. M. (2011). Early Specialization for Voice and Emotion Processing in the Infant Brain. *Current Biology*, 21(14), 1220–1224.
<https://doi.org/10.1016/j.cub.2011.06.009>

Borghesani, V., Pedregosa, F., Buiatti, M., Amadon, A., Eger, E., & Piazza, M. (2016). Word meaning in the ventral visual path: A perceptual to conceptual gradient of semantic coding. *NeuroImage*, 143, 128–140.
<https://doi.org/10.1016/j.neuroimage.2016.08.068>

Bosseler, A. N., Teinonen, T., Tervaniemi, M., & Huotilainen, M. (2016). Infant Directed Speech Enhances Statistical Learning in Newborn Infants: An ERP Study. *PLOS ONE*, 11(9), e0162177. <https://doi.org/10.1371/journal.pone.0162177>

Cabrera, L., & Gervain, J. (2020). Speech perception at birth: The brain encodes fast and slow temporal information. *Science Advances*, 6(30), eaba7830.
<https://doi.org/10.1126/sciadv.aba7830>

Caclin, A., Brattico, E., Tervaniemi, M., Näätänen, R., Morlet, D., Giard, M.-H., & McAdams, S. (2006). Separate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience*, 18(12), 1959–1972.
<https://doi.org/10.1162/jocn.2006.18.12.1959>

Carr, K. W., Perszyk, D. R., Norton, E. S., Voss, J. L., Poeppel, D., & Waxman, S. R. (2021). Developmental changes in auditory-evoked neural activity underlie infants' links between language and cognition. *Developmental Science*, 24(6), e13121. <https://doi.org/10.1111/desc.13121>

Cecchetti, L., Kupers, R., Ptito, M., Pietrini, P., & Ricciardi, E. (2016). Are Supramodality and Cross-Modal Plasticity the Yin and Yang of Brain Development? From Blindness to Rehabilitation. *Frontiers in Systems Neuroscience*, 10.

<https://doi.org/10.3389/fnsys.2016.00089>

Cecchetti, L., Ricciardi, E., Handjaras, G., Kupers, R., Ptito, M., & Pietrini, P. (2016). Congenital blindness affects diencephalic but not mesencephalic structures in the human brain. *Brain Structure and Function*, 221(3), 1465–1480. <https://doi.org/10.1007/s00429-014-0984-5>

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.-P., & Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, 10. <https://doi.org/10.1186/1471-2202-10-127>

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2), 887–906. <https://doi.org/10.1121/1.1945807>

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194, 12–24. <https://doi.org/10.1016/j.neuroimage.2019.03.031>

Cisneros-Franco, J. M., Voss, P., Thomas, M. E., & de Villers-Sidani, E. (2020). Chapter 8—Critical periods of brain development. In A. Gallagher, C. Bulteau, D. Cohen, & J. L. Michaud (Eds.), *Handbook of Clinical Neurology* (Vol. 173, pp. 75–88). Elsevier. <https://doi.org/10.1016/B978-0-444-64150-2.00009-5>

Collignon, O., Dormal, G., Albouy, G., Vandewalle, G., Voss, P., Phillips, C., & Lepore, F. (2013). Impact of blindness onset on the functional organization and the connectivity of the occipital cortex. *Brain*, 136(9), 2769–2783. <https://doi.org/10.1093/brain/awt176>

Collignon, O., Lassonde, M., Lepore, F., Bastien, D., & Veraart, C. (2007). Functional cerebral reorganization for auditory spatial processing and auditory substitution of vision in early blind subjects. *Cerebral Cortex (New York, N.Y.: 1991)*, 17(2), 457–465. <https://doi.org/10.1093/cercor/bhj162>

Collignon, O., Voss, P., Lassonde, M., & Lepore, F. (2009). Cross-modal plasticity for the spatial processing of sounds in visually deprived subjects. *Experimental Brain Research*, 192(3), 343–358. <https://doi.org/10.1007/s00221-008-1553-z>

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional Neuroimaging of Speech Perception in Infants. *Science*, 298(5600), 2013–2015. <https://doi.org/10.1126/science.1077066>

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., & Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Sciences*, 103(38), 14240–14245. <https://doi.org/10.1073/pnas.0606302103>

Dormal, G., Pelland, M., Rezk, M., Yakobov, E., Lepore, F., & Collignon, O.

(2018a). Functional Preference for Object Sounds and Voices in the Brain of Early Blind and Sighted Individuals. *Journal of Cognitive Neuroscience*, 30(1), 86–106. https://doi.org/10.1162/jocn_a_01186

Dormal, G., Pelland, M., Rezk, M., Yakobov, E., Lepore, F., & Collignon, O. (2018b). Functional Preference for Object Sounds and Voices in the Brain of Early Blind and Sighted Individuals. *Journal of Cognitive Neuroscience*, 30(1), 86–106. https://doi.org/10.1162/jocn_a_01186

Dormal, G., Rezk, M., Yakobov, E., Lepore, F., & Collignon, O. (2016). Auditory motion in the sighted and blind: Early visual deprivation triggers a large-scale imbalance between auditory and “visual” brain regions. *NeuroImage*, 134, 630–644. <https://doi.org/10.1016/j.neuroimage.2016.04.027>

Dzhelyova, M., Jacques, C., & Rossion, B. (2017). At a Single Glance: Fast Periodic Visual Stimulation Uncovers the Spatio-Temporal Dynamics of Brief Facial Expression Changes in the Human Brain. *Cerebral Cortex*, 27(8), 4106–4123. <https://doi.org/10.1093/cercor/bhw223>

Eckert, M. A., Kamdar, N. V., Chang, C. E., Beckmann, C. F., Greicius, M. D., & Menon, V. (2008). A cross-modal system linking primary auditory and visual cortices: Evidence from intrinsic fMRI connectivity analysis. *Human Brain Mapping*, 29(7), 848–857. <https://doi.org/10.1002/hbm.20560>

Ethofer, T., Bretscher, J., Gschwind, M., Kreifelts, B., Wildgruber, D., & Vuilleumier, P. (2012). Emotional Voice Areas: Anatomic Location, Functional Properties, and Structural Connections Revealed by Combined fMRI/DTI. *Cerebral Cortex*, 22(1), 191–200. <https://doi.org/10.1093/cercor/bhr113>

Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of Emotional Information in Voice-Sensitive Cortices. *Current Biology*, 19(12), 1028–1033. <https://doi.org/10.1016/j.cub.2009.04.054>

Fecteau, S., Belin, P., Joanette, Y., & Armony, J. L. (2007). Amygdala responses to nonlinguistic emotional vocalizations. *NeuroImage*, 36(2), 480–487. <https://doi.org/10.1016/j.neuroimage.2007.02.043>

Frühholz, S., & Belin, P. (2024). Oxford Handbook of Voice Perception | Request PDF. *ResearchGate*. <https://doi.org/10.1177/0301006620938229>

Frühholz, S., & Grandjean, D. (2013). Processing of emotional vocalizations in bilateral inferior frontal cortex. *Neuroscience & Biobehavioral Reviews*, 37(10, Part 2), 2847–2855. <https://doi.org/10.1016/j.neubiorev.2013.10.007>

Geangu, E., Quadrelli, E., Lewis, J. W., Macchi Cassia, V., & Turati, C. (2015). By the sound of it. An ERP investigation of human action sound processing in 7-month-old infants. *Developmental Cognitive Neuroscience*, 12, 134–144. <https://doi.org/10.1016/j.dcn.2015.01.005>

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory Integration of Dynamic Faces and Voices in Rhesus Monkey Auditory Cortex. *The Journal of Neuroscience*, 25(20), 5004–5012.

<https://doi.org/10.1523/JNEUROSCI.0799-05.2005>

Giard, M. H., Lavikainen, J., Reinikainen, K., Perrin, F., Bertrand, O., Pernier, J., & Näätänen, R. (1995). Separate Representation of Stimulus Frequency, Intensity, and Duration in Auditory Sensory Memory: An Event-Related Potential and Dipole-Model Analysis. *Journal of Cognitive Neuroscience*, 7(2), 133–143. <https://doi.org/10.1162/jocn.1995.7.2.133>

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26(4), 664–672. <https://doi.org/10.1038/s41593-023-01285-9>

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract Encoding of Auditory Objects in Cortical Activity Patterns. *Cerebral Cortex*, 23(9), 2025–2037. <https://doi.org/10.1093/cercor/bhs162>

Giordano, B. L., Whiting, C., Kriegeskorte, N., Kotz, S. A., Gross, J., & Belin, P. (2021). The Representational Dynamics of Perceived Voice Emotions Evolve from Categories to Dimensions. *Nature Human Behaviour*, 5(9), 1203–1213. <https://doi.org/10.1038/s41562-021-01073-0>

Gougoux, F., Belin, P., Voss, P., Lepore, F., Lassonde, M., & Zatorre, R. J. (2009). Voice perception in blind persons: A functional magnetic resonance imaging study. *Neuropsychologia*, 47(13), 2967–2974. <https://doi.org/10.1016/j.neuropsychologia.2009.06.027>

Gourévitch, B., Martin, C., Postal, O., & Eggermont, J. J. (2020). Oscillations in the auditory system and their possible role. *Neuroscience & Biobehavioral Reviews*, 113, 507–528. <https://doi.org/10.1016/j.neubiorev.2020.03.030>

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). The voices of wrath: Brain responses to angry prosody in meaningless speech. *Nature Neuroscience*, 8(2), 145–146. <https://doi.org/10.1038/nn1392>

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, 29(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience*, 28(2), 219–236. <https://doi.org/10.3233/RNN-2010-0499>

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, 173, 434–447. <https://doi.org/10.1016/j.neuroimage.2018.02.044>

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: Controversies and consensus. *Trends in Cognitive Sciences*, 16(9), 458–466. <https://doi.org/10.1016/j.tics.2012.07.006>

Handjaras, G., Ricciardi, E., Leo, A., Lenci, A., Cecchetti, L., Cosottini, M., Marotta, G., & Pietrini, P. (2016). How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. *NeuroImage*, 135, 232–242. <https://doi.org/10.1016/j.neuroimage.2016.04.063>

Heimler, B., Weisz, N., & Collignon, O. (2014). Revisiting the adaptive and maladaptive effects of crossmodal plasticity. *Neuroscience*, 283, 44–63. <https://doi.org/10.1016/j.neuroscience.2014.08.003>

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences*, 106(25), 10382–10385. <https://doi.org/10.1073/pnas.0812142106>

Johnstone, T., Van Reekum, C., Oakes, T., & Davidson, R. (2006). The voice of emotion: An fMRI study of neural responses to angry and happy vocal expressions. *Social Cognitive and Affective Neuroscience*, 1, 242–249. <https://doi.org/10.1093/scan/nsl027>

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5), 770–814. <https://doi.org/10.1037/0033-2909.129.5.770>

Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2005). Integration of Touch and Sound in Auditory Cortex. *Neuron*, 48(2), 373–384. <https://doi.org/10.1016/j.neuron.2005.09.018>

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, 55, 121–132. <https://doi.org/10.1016/j.conb.2019.02.003>

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4), 203–210. <https://doi.org/10.1016/j.tics.2014.01.002>

Klasen, M., Chen, Y.-H., & Mathiak, K. (2012). Multisensory emotions: Perception, combination and underlying neural processes. *Reviews in the Neurosciences*, 23(4), 381–392. <https://doi.org/10.1515/revneuro-2012-0040>

Klinge, C., Eippert, F., Röder, B., & Büchel, C. (2010). Corticocortical Connections Mediate Primary Visual Cortex Responses to Auditory Stimulation in the Blind. *The Journal of Neuroscience*, 30(38), 12798–12805. <https://doi.org/10.1523/JNEUROSCI.2384-10.2010>

Kotz, S. A., Kalberlah, C., Bahlmann, J., Friederici, A. D., & Haynes, J.-D. (2013). Predicting vocal emotion expressions from the human brain. *Human Brain Mapping*, 34(8), 1971–1981. <https://doi.org/10.1002/hbm.22041>

Kotz, S. A., Meyer, M., Alter, K., Besson, M., von Cramon, D. Y., & Friederici, A. D. (2003). On the lateralization of emotional prosody: An event-related functional MR investigation. *Brain and Language*, 86(3), 366–376. [https://doi.org/10.1016/S0093-934X\(02\)00532-1](https://doi.org/10.1016/S0093-934X(02)00532-1)

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational

similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. <https://doi.org/10.3389/neuro.06.004.2008>

Kuhl, P. K. (2010). Brain Mechanisms in Early Language Acquisition. *Neuron*, 67(5), 713–727. <https://doi.org/10.1016/j.neuron.2010.08.038>

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science*, 255(5044), 606–608. <https://doi.org/10.1126/science.1736364>

Langers, D. R. M., van Dijk, P., Schoenmaker, E. S., & Backes, W. H. (2007). fMRI activation in relation to sound intensity and loudness. *NeuroImage*, 35(2), 709–718. <https://doi.org/10.1016/j.neuroimage.2006.12.013>

LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life* (p. 384). Simon & Schuster.

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., & Brefczynski-Lewis, J. A. (2009). Human Cortical Organization for Processing Vocalizations Indicates Representation of Harmonic Structure as a Signal Attribute. *The Journal of Neuroscience*, 29(7), 2283–2296. <https://doi.org/10.1523/JNEUROSCI.4145-08.2009>

Liu, Y., Yu, C., Liang, M., Li, J., Tian, L., Zhou, Y., Qin, W., Li, K., & Jiang, T. (2007). Whole brain functional connectivity in the early blind. *Brain*, 130(8), 2085–2096. <https://doi.org/10.1093/brain/awm121>

Lochy, A., Van Belle, G., & Rossion, B. (2015). A robust index of lexical representation in the left occipito-temporal cortex as evidenced by EEG responses to fast periodic visual stimulation. *Neuropsychologia*, 66, 18–31. <https://doi.org/10.1016/j.neuropsychologia.2014.11.007>

Lowe, M. X., Mohsenzadeh, Y., Lahner, B., Charest, I., Oliva, A., & Teng, S. (2021). Cochlea to categories: The spatiotemporal dynamics of semantic auditory representations. *Cognitive Neuropsychology*, 38(7–8), 468–489. <https://doi.org/10.1080/02643294.2022.2085085>

Marlair, C., Lochy, A., & Crollen, V. (2024). Frequency-tagging EEG reveals the effect of attentional focus on abstract magnitude processing. *Psychonomic Bulletin & Review*, 31(5), 2266–2274. <https://doi.org/10.3758/s13423-024-02480-w>

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife*, 7, e31873. <https://doi.org/10.7554/eLife.31873>

Mattioni, S., Rezk, M., Battal, C., Bottini, R., Cuculiza Mendoza, K. E., Oosterhof, N. N., & Collignon, O. (2020). Categorical representation from sound and sight in the ventral occipito-temporal cortex of sighted and blind. *eLife*, 9, e50732. <https://doi.org/10.7554/eLife.50732>

Mattioni, S., Rezk, M., Battal, C., Vadlamudi, J., & Collignon, O. (2022). Impact of blindness onset on the representation of sound categories in occipital and temporal cortices. *eLife*, 11, e79370. <https://doi.org/10.7554/eLife.79370>

Mauchand, M., & Zhang, S. (2022). Disentangling emotional signals in the brain: An ALE meta-analysis of vocal affect perception. *Cognitive, Affective, & Behavioral Neuroscience*. <https://doi.org/10.3758/s13415-022-01030-y>

Menon, V., Levitin, D. J., Smith, B. K., Lembke, A., Krasnow, B. D., Glazer, D., Glover, G. H., & McAdams, S. (2002). Neural Correlates of Timbre Change in Harmonic Sounds. *NeuroImage*, 17(4), 1742–1754. <https://doi.org/10.1006/nimg.2002.1295>

Mihai, P. G., Moerel, M., Martino, F. de, Trampel, R., Kiebel, S., & Kriegstein, K. von. (2019). Modulation of tonotopic ventral medial geniculate body is behaviorally relevant for speech recognition. *eLife*, 8, e44837. <https://doi.org/10.7554/eLife.44837>

Müller, F., Niso, G., Samiee, S., Ptito, M., Baillet, S., & Kupers, R. (2019). A thalamocortical pathway for fast rerouting of tactile information to occipital cortex in congenital blindness. *Nature Communications*, 10(1), 5154. <https://doi.org/10.1038/s41467-019-13173-7>

Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology*, 24(4), 375–425. <https://doi.org/10.1111/j.1469-8986.1987.tb00311.x>

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, 88(6), 1281–1296. <https://doi.org/10.1016/j.neuron.2015.11.035>

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, 6(3), 455–469. <https://doi.org/10.1038/s41562-021-01261-y>

Ogg, M., Carlson, T. A., & Slevc, L. R. (2020). The Rapid Emergence of Auditory Object Representations in Cortex Reflect Central Acoustic Attributes. *Journal of Cognitive Neuroscience*, 32(1), 111–123. https://doi.org/10.1162/jocn_a_01472

Ogg, M., Moraczewski, D., Kuchinsky, S. E., & Slevc, L. R. (2019). Separable neural representations of sound sources: Speaker identity and musical timbre. *NeuroImage*, 191, 116–126. <https://doi.org/10.1016/j.neuroimage.2019.01.075>

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, 36(4), 767–776. [https://doi.org/10.1016/s0896-6273\(02\)01060-7](https://doi.org/10.1016/s0896-6273(02)01060-7)

Peelen, M. V., Bracci, S., Lu, X., He, C., Caramazza, A., & Bi, Y. (2013). Tool Selectivity in Left Occipitotemporal Cortex Develops without Vision. *Journal of*

- Cognitive Neuroscience*, 25(8), 1225–1234. https://doi.org/10.1162/jocn_a_00411
- Peelen, M. V., & Downing, P. E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 105, 177–183. <https://doi.org/10.1016/j.neuropsychologia.2017.03.033>
- Peelen, M. V., He, C., Han, Z., Caramazza, A., & Bi, Y. (2014). Nonvisual and Visual Object Shape Representations in Occipitotemporal Cortex: Evidence from Congenitally Blind and Sighted Adults. *Journal of Neuroscience*, 34(1), 163–170. <https://doi.org/10.1523/JNEUROSCI.1114-13.2014>
- Peña, M., Maki, A., Kovačić, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., & Mehler, J. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences*, 100(20), 11702–11705. <https://doi.org/10.1073/pnas.1934290100>
- Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, 3. <https://doi.org/10.3389/fpsyg.2012.00023>
- Petrides, M. (2005). Lateral prefrontal cortex: Architectonic and functional organization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 781–795. <https://doi.org/10.1098/rstb.2005.1631>
- Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMRI. *NeuroImage*, 16(2), 331–348. <https://doi.org/10.1006/nimg.2002.1087>
- Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., Williams, S. C., Bullmore, E. T., Brammer, M., & Gray, J. A. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society B: Biological Sciences*, 265(1408), 1809–1817.
- Pietrini, P., Furey, M. L., Ricciardi, E., Gobbini, M. I., Wu, W.-H. C., Cohen, L., Guazzelli, M., & Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 101(15), 5658–5663. <https://doi.org/10.1073/pnas.0400707101>
- Ptito, M., Fumal, A., de Noordhout, A. M., Schoenen, J., Gjedde, A., & Kupers, R. (2008). TMS of the occipital cortex induces tactile sensations in the fingers of blind Braille readers. *Experimental Brain Research*, 184(2), 193–200. <https://doi.org/10.1007/s00221-007-1091-0>
- R, A., D, T., H, D., & Ar, D. (1995). Fear and the human amygdala. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 15(9). <https://doi.org/10.1523/JNEUROSCI.15-09-05879.1995>
- Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, 50(2), 818–825. <https://doi.org/10.1016/j.neuroimage.2009.11.084>
- Retter, T. L., & Rossion, B. (2016). Uncovering the neural magnitude and spatio-temporal dynamics of natural image categorization in a fast visual stream.

- Neuropsychologia*, 91, 9–28.
<https://doi.org/10.1016/j.neuropsychologia.2016.07.028>
- Roberts, T. P. L., Ferrari, P., Stufflebeam, S. M., & Poeppel, D. (2000). Latency of the Auditory Evoked Neuromagnetic Field Components: Stimulus Dependence and Insights Toward Perception. *Journal of Clinical Neurophysiology*, 17(2), 114.
- Rosburg, T. (2003). Left hemispheric dipole locations of the neuromagnetic mismatch negativity to frequency, intensity and duration deviants. *Cognitive Brain Research*, 16(1), 83–90. [https://doi.org/10.1016/S0926-6410\(02\)00222-7](https://doi.org/10.1016/S0926-6410(02)00222-7)
- Rossion, B., Torfs, K., Jacques, C., & Liu-Shuang, J. (2015). Fast periodic presentation of natural images reveals a robust face-selective electrophysiological response in the human brain. *Journal of Vision*, 15. <https://doi.org/10.1167/15.1.18>
- Rupp, K. M., Hect, J. L., Harford, E. E., Holt, L. L., Ghuman, A. S., & Abel, T. J. (2024). A hierarchy of processing complexity and timescales for natural sounds in human auditory cortex. *bioRxiv: The Preprint Server for Biology*, 2024.05.24.595822. <https://doi.org/10.1101/2024.05.24.595822>
- Sabourin, C. J., Merrikhi, Y., & Lomber, S. G. (2022). Do blind people hear better? *Trends in Cognitive Sciences*, 26(11), 999–1012. <https://doi.org/10.1016/j.tics.2022.08.016>
- Saccone, E. J., Tian, M., & Bedny, M. (2024). Developing cortex is functionally pluripotent: Evidence from blindness. *Developmental Cognitive Neuroscience*, 66, 101360. <https://doi.org/10.1016/j.dcn.2024.101360>
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294), 1926–1928.
- Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*, 10(1), e1003412. <https://doi.org/10.1371/journal.pcbi.1003412>
- Sauter, D. A., & Eimer, M. (2010). Rapid Detection of Emotion from Human Vocalizations. *Journal of Cognitive Neuroscience*, 22(3), 474–481. <https://doi.org/10.1162/jocn.2009.21215>
- Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6), 2408–2412. <https://doi.org/10.1073/pnas.0908239106>
- Schepers, I. M., Hipp, J. F., Schneider, T. R., Röder, B., & Engel, A. K. (2012). Functionally specific oscillatory activity correlates between visual and auditory cortex in the blind. *Brain*, 135(3), 922–934. <https://doi.org/10.1093/brain/aws014>
- Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural*

Psychology, 32(1), 76–92.

Schirmer, A., Zysset, S., Kotz, S. A., & Yves von Cramon, D. (2004). Gender differences in the activation of inferior frontal cortex during emotional speech perception. *NeuroImage*, 21(3), 1114–1123. <https://doi.org/10.1016/j.neuroimage.2003.10.048>

Schönwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34), 14611–14616. <https://doi.org/10.1073/pnas.0907682106>

Stefanics, G., Háden, G. P., Sziller, I., Balázs, L., Beke, A., & Winkler, I. (2009). Newborn infants process pitch intervals. *Clinical Neurophysiology*, 120(2), 304–308. <https://doi.org/10.1016/j.clinph.2008.11.020>

Striem-Amit, E., Cohen, L., Dehaene, S., & Amedi, A. (2012). Reading with Sounds: Sensory Substitution Selectively Activates the Visual Word Form Area in the Blind. *Neuron*, 76(3), 640–652. <https://doi.org/10.1016/j.neuron.2012.08.026>

Talwar, S., Barbero, F. M., Calce, R. P., & Collignon, O. (2023). Automatic Brain Categorization of Discrete Auditory Emotion Expressions. *Brain Topography*, 36(6), 854–869. <https://doi.org/10.1007/s10548-023-00983-8>

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, 10(1), 21. <https://doi.org/10.1186/1471-2202-10-21>

Telkemeyer, S., Rossi, S., Koch, S. P., Nierhaus, T., Steinbrink, J., Poeppel, D., Obrig, H., & Wartenburger, I. (2009). Sensitivity of Newborn Auditory Cortex to the Temporal Structure of Sounds. *The Journal of Neuroscience*, 29(47), 14726–14733. <https://doi.org/10.1523/JNEUROSCI.1246-09.2009>

Thierry, G., & Price, C. J. (2006). Dissociating Verbal and Nonverbal Conceptual Processing in the Human Brain. *Journal of Cognitive Neuroscience*, 18(6), 1018–1028. <https://doi.org/10.1162/jocn.2006.18.6.1018>

Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional Specialization in Rhesus Monkey Auditory Cortex. *Science*, 292(5515), 290–293. <https://doi.org/10.1126/science.1058911>

Van Ackeren, M. J., Barbero, F. M., Mattioni, S., Bottini, R., & Collignon, O. (2018). Neuronal populations in the occipital cortex of the blind synchronize to the temporal dynamics of speech. *eLife*, 7, e31640. <https://doi.org/10.7554/eLife.31640>

van den Hurk, J., Van Baelen, M., & Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences of the United States of America*, 114(22), E4501–E4510. <https://doi.org/10.1073/pnas.1612862114>

Volfart, A., Rice, G. E., Lambon Ralph, M. A., & Rossion, B. (2021). Implicit,

automatic semantic word categorisation in the left occipito-temporal cortex as revealed by fast periodic visual stimulation. *NeuroImage*, 238, 118228. <https://doi.org/10.1016/j.neuroimage.2021.118228>

Vouloumanos, A., Hauser, M. D., Werker, J. F., & Martin, A. (2010). The Tuning of Human Neonates' Preference for Speech. *Child Development*, 81(2), 517–527. <https://doi.org/10.1111/j.1467-8624.2009.01412.x>

Vuilleumier. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, 9(12). <https://doi.org/10.1016/j.tics.2005.10.011>

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137, 188–200. <https://doi.org/10.1016/j.neuroimage.2015.12.012>

Warren, J. D., Jennings, A. R., & Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *NeuroImage*, 24(4), 1052–1057. <https://doi.org/10.1016/j.neuroimage.2004.10.031>

Watkins, K. E., Cowey, A., Alexander, I., Filippini, N., Kennedy, J. M., Smith, S. M., Ragge, N., & Bridge, H. (2012). Language networks in anophthalmia: Maintained hierarchy of processing in 'visual' cortex. *Brain*, 135(5), 1566–1577. <https://doi.org/10.1093/brain/aws067>

Whiteside, S. P. (1998). The Identification of a Speaker's Sex from Synthesized Vowels. *Perceptual and Motor Skills*, 87(2), 595–600. <https://doi.org/10.2466/pms.1998.87.2.595>

Winkler, I., Kushnerenko, E., Horváth, J., Čeponienė, R., Fellman, V., Huotilainen, M., Näätänen, R., & Sussman, E. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences*, 100(20), 11812–11815. <https://doi.org/10.1073/pnas.2031891100>

Wittenberg, G. F., Werhahn, K. J., Wassermann, E. M., Herscovitch, P., & Cohen, L. G. (2004). Functional connectivity between somatosensory and visual cortex in early blind humans. *European Journal of Neuroscience*, 20(7), 1923–1927. <https://doi.org/10.1111/j.1460-9568.2004.03630.x>

Ws, R., & S, G. (1994). Encoding of amplitude modulation in the cochlear nucleus of the cat. *Journal of Neurophysiology*, 71(5). <https://doi.org/10.1152/jn.1994.71.5.1797>

Xie, S., Hoehl, S., Moeskops, M., Kayhan, E., Kliesch, C., Turtleton, B., Köster, M., & Cichy, R. M. (2022). Visual category representations in the infant brain. *Current Biology*, 32(24), 5422–5432.e6. <https://doi.org/10.1016/j.cub.2022.11.016>

Young, E. D., & Davis, K. A. (2002). Circuitry and Function of the Dorsal Cochlear Nucleus. In D. Oertel, R. R. Fay, & A. N. Popper (Eds.), *Integrative Functions in the Mammalian Auditory Pathway* (pp. 160–206). Springer. https://doi.org/10.1007/978-1-4757-3654-0_5

Zäske, R., Skuk, V. G., Kaufmann, J. M., & Schweinberger, S. R. (2013).

Perceiving vocal age and gender: An adaptation approach. *Acta Psychologica*, 144(3), 583–593. <https://doi.org/10.1016/j.actpsy.2013.09.009>