Les composantes de la parole dans la caractérisation phonétique du locuteur : étude sur la complémentarité et la redondance véhiculées des informations

Gabriele Chignoli

Université Sorbonne Nouvelle - CNRS

Ecole Dotorale 622 : Sciences du Langage
UMR7018 Laboratoire de Phonétique et Phonologie

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Phonetics:

Speech components in phonetic characterisation of speakers

a study on complementarity and redundancy of conveyed information

Submitted by:

Gabriele CHIGNOLI

gabriele.chignoli@sorbonne-nouvelle.fr

Under the supervision of:

Cédric GENDROT

Reviewer: **Damien LOLIVE** HDR, CNRS - IRISA - Université de Rennes 1
Reviewer: **Ioana VASILESCU** HDR, CNRS - LISN - Université Paris-Saclay
Examiner: **Cécile FOUGERON** DR, CNRS - LPP - Université Sorbonne Nouvelle
Examiner: **Jean-François BONASTRE** PR, CNRS - LIA - Université d'Avignon
Examiner: **Christine MEUNIER** DR, CNRS - LPL - Université Aix-Marseille
Supervisor: **Cédric GENDROT** HDR, CNRS - LPP - Université Sorbonne Nouvelle

Paris, September 15, 2022

Speech components in phonetic characterisation of speakers
a study on complementarity and redundancy of conveyed information

Abstract

The decomposition of the speech signal into phonetically meaningful units allows the analysis of between- and within- speaker variations. These are components associated with characteristics whose nature relates to the physical, psychological and social aspects of a speaker. In this thesis, we compare perceptual characterisation results with a phonetic analysis and advanced modelling techniques through Convolutional Neural Networks (CNN).

Clusterings' analysis shows that the perceptual results are coherent with those obtained by the CNN and phonetic approaches, which supports the application of these methods in Phonetics. Our results highlight that spectrograms are the most accurate speech representation for speaker identification (96 % correct answers on average). Higher formants and harmonics are more important in the characterisation of female voices. Whereas, voice quality characteristics, such as breathiness and hoarseness, play a major role in the characterisation of male speakers. The comparison between Mel Frequency Cepstral Coefficients (MFCC) and classical phonetic measurements is also examined. The MFCC are mainly linked to intensity and f0 in the characterisation of female speakers, while to the distributions of energy and low level spectral shape for male speakers.

Our findings confirm the importance of describing the within-speaker variation for a more complete understanding of between-speakers differences.

Keywords: speaker, characteristics, components, CNN, spontaneous, clustering, informedness, comparison

Les composantes de la parole dans la caractérisation phonétique du locuteur
étude sur la complémentarité et la redondance véhiculées des informations

Résumé

La décomposition du signal vocal en unités phonétiquement significatives permet d'analyser les variations inter- et intra- locuteur. Ces unités sont des composantes associées à des caractéristiques dont la nature est liée aux aspects physiques, psychologiques et sociaux d'un locuteur. Dans cette thèse, nous comparons une caractérisation perceptive, une analyse phonétique et des techniques de modélisation avancées par des réseaux de neurones à convolution (CNN).

L'analyse des clusterings montre que les résultats perceptifs sont cohérents avec ceux obtenus par les approches CNN et phonétique, ce qui soutient leurs applications en phonétique. Nos résultats mettent en évidence que les spectrogrammes sont la représentation de la parole la plus précise pour l'identification des locuteurs (96 % de bonnes réponses en moyenne). Les formants et des harmoniques plus élevés sont plus importants dans la caractérisation des voix féminines. En revanche, les caractéristiques de la qualité de la voix, telles que le souffle et la raucité, jouent un rôle majeur dans la caractérisation des voix masculines. Le lien entre les coefficients cepstraux à fréquence Mel (MFCC) et les mesures phonétiques classiques est également examiné. Les MFCC sont principalement liés à l'intensité et à f0 dans la caractérisation des voix féminines, tandis qu'aux distributions d'énergie et à la forme spectrale de bas niveau pour celle des voix masculines.

Nos résultats confirment l'importance de la description de la variation intra-locuteur pour une compréhension plus complète des différences entre locuteurs.

Mots-clés : locuteur, caractéristiques, composantes, CNN, spontané, clustering, informativité, comparaison

"I mean, this was something I made!
Something that came from me!
That was a part of me!
The only thing I ever made that was any good!"

Randy Marsh

Acknowledgements

I have learned that a PhD is both a collective and a solo experience. The latter mostly because when you face that document everyday, you are the only one who can push on through to the end, and for this to that little guy who looked at documentaries and wanted one day to be a researcher, you have made it. The collective part contribute to complete this seemingly unachievable goal, helping to maintain foolishness and hunger for it. Therefore, thanks to all the people who have shared a bit of their time with me during this journey. I want to thank everyone, for their conscious or unconscious help. Everyone who was there at the beginning, during or just at the end of this journey. Everyone who will be there for the next one. Everyone who has left earlier. Some are more obvious, some others are not cited, but it's only because of my volatile memory. In advance, I'm sorry, you and me, we know you were there.

A tutta la mia famiglia, anche se per quattro anni ho dovuto spiegare e rispiegare cosa stessi facendo o perché, mi avete sempre incoraggiato ed aiutato in tutto e per tutto.
Merci à Cédric & Cécile de m'avoir donné cette belle opportunité et d'avoir été présent tout au long de ma thèse avec vos conseils et idées.
Merci au jury me permettant de completer ce parcours.
Marie, avec tes relectures à l'autre bout du monde alors que tu étais tranquille sur ton canapé, merci beaucoup beaucoup !
Merci à Angélina, Jade et Marine, avec qui j'ai partagé le début de cette aventure et surtout le stress de la fin !
A propos du stress, merci à Amelia et Bowei d'avoir supporté tous mes doutes et questions.
Comment ne pas remercier Arthur, Francesco et Roberto, témoins de mille aventures et toujours là pour me distraire d'une journée un peu noire.
Un grand merci évidemment à Monsieur Ferragne pour toute sa disponibilité et ses conseils.
Merci à tout le LPP pour avoir accompagné ces années de travail avec plein de bonne humeur.
Merci au projet Voxcrim (ANR-17-CE39-0016) pour le financement de ma thèse. Merci à tous ses membres pour avoir permis que cette aventure soit possible en ayant durant chaque échange et réunion apporté une petite pièce à sa création.

Thanks to all my teachers because every school attended and every lesson half-heard was a small step to achieve this.
A special thanks to everyone who never believed in me, you were my best inspiration to do better.

And last but far from the least, to Fanny, whose every breath is the wind pushing my ship forward throughout this journey.

Contents

List of abbreviations

A(1-3)	Amplitude of the harmonics near the first three formants
ACC	Articulatory Cepstral Coefficient
AFCP	Association Francophone de la Communication Parlée
(A|C|D)NN	(Artificial|Convolutional|Deep) Neural Networks
AS(R|V)	Automatic Speaker (Recognition|Verification)
$\mathbf{C_{llr}}$	Log Likelihood Ratio cost validity metric
CoV	Coefficient of variation
CPP	Cepstral Peak Prominence
CSE	Cochlea-scaled entropy
CDS	Cosine Distance Scoring
dB	DeciBel
EER	Equal Error Rate
EMA	ElectroMagnetic Articulograph
ENV	Temporal Envelope
F0	Fundamental frequency
F(1-4)	First to fourth formants
FD	Formant dispersion
fMLLR	Feature space Maximum Likelihood Linear Regression
fMRI	functional magnetic resonance imaging
FVC	Forensic Voice Comparison
H1	Amplitude of the first harmonic
H2	Amplitude of the second harmonic
H4	Amplitude of the fourth harmonic
H2k	Amplitude of the harmonic nearest 2 kHz
H5k	Amplitude of the harmonic nearest 5 kHz
H1–A(1-3)	Relative amplitudes of the first harmonic and the ones near F(1-3)
H1–H2	Relative amplitudes of the first and second harmonics
H2–H4	Relative amplitudes of the second and fourth harmonics
H2k–H5k	Relative amplitudes of the harmonics near 2 kHz and 5 kHz
H4–H2k	Relative amplitudes of the fourth harmonic and the one near 2 kHz
HASR	Human Assisted Speaker Recognition
HCA	Hierarchical Clustering Analysis
HNR	Harmonics-to-Noise Ratio
(k)Hz	(kilo)Hertz
K	Cohen's Kappa
i-vectors	Intermediate representation vectors
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
(L)LR	(Log) Likelihood Ratio
LPC	Linear Prediction Coefficients

LRE	Language Recognition Evaluation
LSTM	Long-Short Term Memory
LTFD	Long-Term Formant Distribution
LTAS	Long-Term Average Spectrum
MCC	Mattheus Correlation Coefficient
MDS	Multidimensional Scaling
MFCC	Mel Frequency Cepstral Coefficient
MKD	Multivariate Kernel Density
ML	Machine Learning
MPS	Modulation Power Spectrum
NaN	Not a Number
NCCFr	Nijmegen Corpus of Casual French
NDA	Nearest-neighbour Discriminant Analysis
NIST	National Institute of Standards and Technologies
NLP	Natural Language Processing
PC(A)	Principal Component (Analysis)
(P)LDA	(Probabilistic) Linear Discriminant Analysis
PLP	Perceptual Linear Prediction
PVI	Pairwaise Variability Index
px	Pixel
(m)s	(milli)seconds
ResNet	Residual Networks
RMS	Root Mean Square
SFA	Société Française d'Acoustique
SD	Standard Deviation
SHR	Amplitude ratio between subharmonics and harmonics
SNR	Signal-to-Noise Ratio
soe	Strenght of Excitation
SRE	Speaker Recognition Evaluation
STS	Superior temporal sulcus
SVM	Support Vector Machine
TFS	Temporal Fine Structure
TF-IDF	Term Frequency–Inverse Document Frequency
TV	Vocal Tract variables
TVAs	Temporal voice areas
UBM-GMM	Gaussian Mixture Model based Universal Background Model
VPA	Vocal Profile Analysis

NB: all reported values of harmonics were corrected for the influence of formants on harmonic amplitudes and marked with a "*" in the literature, omitted in this document for reading clarity.

List of Figures

List of Tables

Chapter 1

Introduction

The speech we produce and perceive everyday is mostly spontaneous. It requires no special training, aside from the acquisition process we undergo during the first years of growth, or some specific situations such as during language classes. Moreover, efficiency of speech is a remarkable trait considering both the minimal cognitive load it requires and the wide range of information it can carry at multiple levels. During spontaneous speech production and perception we, as humans, are capable of operations such as locating hidden punctuation, coping with disfluencies or retrieving more than just words' meaning. These are just a few consequences of the highly selective and optimised intrinsic mechanisms that are activated in our brain.

Complex brain's mechanisms can be investigated through the scope of linguistics, psychology and other science. A powerful example is offered by the cocktail party effect, it represents the ability for a listener to focus only on pertinent elements from a sparse set, performing segregation of received stimuli and saliency classification. Most people show high performance in this kind of selective attention, however, as for every human ability there are exceptions, i. e. less efficient listeners [Bronkhorst, 2000]. Besides, the selective segregation of sound may as well result from the brain's ability to recognise a stimulus' pertinent characteristics, which can be known or unknown. For instance, identifying a related person calling us in a noisy crowd can be associated with the ability of segregation of known speakers' characteristics from a heterogeneous set of inputs.

"To mark something as a characteristic"[1] is a definition of to characterise. An operation where a modifier and a modified unit produce a characterised one, recognisable by the indissoluble presence of both elements. The process of characterising a voice is the object of study in this thesis. Voice characterisation implies to identify the characterised unit's modified components, which appear in a characteristic way during speech productions. These elements rely on the idea of a speaker's identity, rather his/her voice, which is far from being a static object. Furthermore, studying how the information about individual characteristics distributes with its redundancy and complementarity is fundamental to understand where the between-speaker variability is most prominent.

In order to characterise a continuously changing object, such as the voice, a two stages process is needed. First, speech signal decomposition, then, the analysis of different com-

[1] Collins Concise English Dictionary © HarperCollins Publishers, from <https://www.wordreference.com/definition/characterize>

ponents in order to associate them with the information they carry. Speech components are commonly studied by the means of phonetic measurements. How do we define which voice components to analyse? How do they interact with each other? Which factors influence their variations the most? These are some of the questions the phonetic literature on voice characterisation has tried to answer. There are others we add and try to give an answer to in this work, such as can Machine Learning and Phonetics be defined as complementary approaches? What are the implications of human perception in the speaker characterisation domains?

The main aim of this thesis is to analyse how individual characteristics can be retrieved by different speech components and their interactions, adding some knowledge about the distribution of speaker information by studying speech productions in French. In order to do this, we describe three moments of analysis in the present work, the parts that organise this document.

The present Chapter provides a general introduction to the idea of voice characterisation and its implications. Whereas the entire Part I completes these considerations with a literature review concerning voice characterisation domains and studies which had an influence on this research. These three chapters represent the moment of *fixity*. The cited works throughout this first part, as the name suggests, fixate the basic concepts for this thesis.

Part II, *instability*, represents the opposite moment of analysis in which we provide our methods and results. Before the empirical studies on phonetic measurements for speaker characterisation, a perception study is presented in Chapter 5, which aims to link the results obtained through different approaches to the human perception of speakers' characteristics, highlighting possible similarities and differences, and offering some perspectives for more in-depth analysis. Indeed, in Chapter 6 we present a set of studies on speakers characteristics through classical phonetic approaches. First, acoustic measurements and rhythmic correlates are tested showing results in line with other phonetic studies. Then, methods closer to those of Forensic Phonetics are applied in order to model dynamics of acoustic characteristics in read speech. Even though the obtained results did not meet our expectations they represent a theoretical basis for the following chapter concerning representation of speech dynamics. In Chapter 7 a new approach is presented, using deep learning techniques, through the means of Convolutional Neural Networks (CNN), we provide new methods for the representations of different speech components.

Finally, Part III of this thesis represents the moment of *unity* between the preceding parts in which our results are looked at through the established basis of the first part. Chapter 8 recapitulates the main findings and contributions of this thesis, discussing how speakers' information is distributed in various speech components for the selected French speakers. Furthermore, the merge of two approaches is discussed, showing how it is important from a Natural Language Processing (NLP) perspective to apply theories from classical Linguistics in a complementary way with modern methods. In Chapter 9, some unsolved questions and propositions for possible future research are explored.

1.1 Phonetic characterisation

Throughout the decades, linguistic research formalised concepts related to language, one common analysis method has been the establishment of dichotomies. Phonetics takes acoustic events as an object of study and decomposes it in order to analyse different aspects of speech through the means of phonetic measurements.

As [Nolan, 1983] established, a useful dichotomy in order to describe the nature of components used for speaker recognition is defining them as innate or learnt.

Innate characteristics relate to variation factors directly linked to speaker's physiology, e. g., vocal tract length. These characteristics can be seen as internal to the studied speaker whose bio-physiology can be defined as the principal influence for their variation. On the other hand, learnt characteristics are associated with paralinguistic factors like specific language habits speakers might have acquired during their lives, i. e., strictly linked to the spoken language, or context-driven influences. This opposition represents a mutually exclusive dichotomy, aiming to reduce as much as possible both biases and variation factors during a speaker recognition or voice characterisation analysis. In order to relate individual information with internal variability factors we need to identify where and how the individual characteristics appear in the speech signal.

Speech is the result of a production mechanism where different sources combine themselves to create a complex wave. By decomposing this wave, namely the speech signal, we access its components which, through the means of phonetic measurements, are commonly viewed through the scope of an acoustic-articulatory dichotomy. Searching for voice characteristics means to express the relation between phonetic measurements and the components extracted from speech productions. We consider this opposition as jointly exhaustive as every component can be analysed from both points of view in order to understand how speech mechanisms behave. Hence, from the glottal source we obtain measurements like fundamental frequency (f0) and its harmonics, integer multiples of the fundamental. From an acoustic perspective, these components represent height variation of intonation, commonly named pitch. Vocal folds oscillation speed being the articulatory correspondent. Resonances from vocal tract cavities result in formants, first to fourth are consistently used in speech studies, or anti-formants when nasal cavity steps in the production process. Acoustically, formants are defined as the reinforced harmonics or local peak in the spectrum. From the articulatory standpoint formants can be associated with mouth opening, lips and tongue configurations.

As said, decomposing the speech signal in smaller units is the first stage of a voice characterisation process. We cited only some examples of these units that we identify as speech components, a larger list is provided in Figure 1.1. The considered units are inspired by the list in [Kreiman and Stidtis, 2011], in which the Authors discussed what speech components can be characteristics for a speaker. Nine groups represent the major components which have been the object of phonetic studies but this list is not exhaustive since other aspects of speech which have still not been considered could arise through future investigations. These are mostly innate-internal characteristics, except for the prosodic ones which heavily correlate with the spoken language and the set *Other* which includes e. g. dialectal features, hence learnt-external.

These sets give a first idea of the extent of the study on speech components in order to

relate them to the information they can carry. Having speech components which rely on different aspects of the speech production means that a different analysis of the same signal can provide different information. Studying voice characterisation means making explicit the relation between phonetic measurements and components carrying individual information in speech productions.

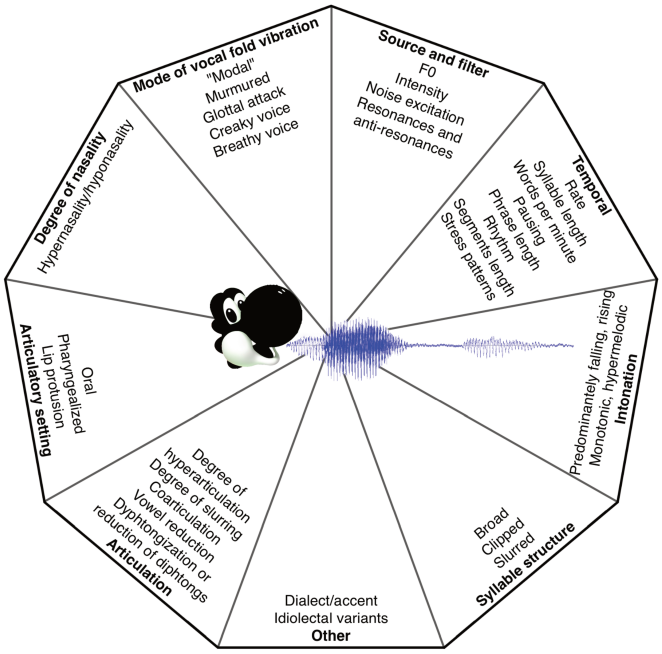


Figure 1.1: List of some known speech components that can be extracted from the speech signal, inspired by Table 4 in [Kreiman and Stidtis, 2011].

These components are not strictly independent from one another, e. g. intonation contour, syllable structure and temporal cues can be seen as representative of the Prosody, see Section 2.2, articulation and articulatory setting can be as well regrouped in a larger articulation component and so on. When considering the phonetic measurements we also observe that, as previously said, one can be used to investigate different aspects of speech depending on how the researcher makes use of it. The summary of these 9 components represent a formalisation to help list the active reference aspects of speech production.

This encourages the description of phonetic parameters interactions to a more in depth examination. We explore how complementarity and redundancy of individual information is conveyed through speech components using laboratory recordings of subjects from a similar learning context and the same communication situation, reducing influence from the external influence factors.

1.1.1 The notion of (in)variant

Speaker's identity corresponds to a large variability matrix where every characteristic can be seen not just as a single, or average, value but as a complex model of characteristic

structures. Resulting from this modelling is something close to the idea of invariant. In early research on speaker recognition, the idea of invariant was associated with the concept of voice signature. [Nolan, 1983; Barlow and Wagner, 1988] used this definition of a static object defining a category to which observations diverge. More recently, studies such as [Doddington et al., 1998; Kreiman and Stidtis, 2011; Cambier-Langeveld et al., 2014] and ours, look at this concept as something more variable, describing a matrix of distributions. When we look at the meaning of internal variability or how observations for a characteristic behave within the same speaker variability matrix, we can actually assume how far a single model can go. To have a robust model of a characteristic's distribution is an important step in order to minimise the within-speaker variability.

On the other hand, the definition of a comparison population that takes into account different distributions maximising their distances is fundamental to produce a coherent model of between-speakers variability. The way variability occurs in a group of characteristics related to a speaker identity defines the very aim of the present study. We further explore how different models of a same speech component lead to the reinterpretation of the concept of invariant. Possible invariant correlates of the speech production task in the physical space have been claimed to exist in various dimensions including articulation, acoustics and neural patterns (*Motor Theory* Liberman et al., 1967; *Acoustic Invariance Theory*, Stevens and Blumstein, 1978; Liberman and Mattingly, 1985; *Adaptive Variability Theory*, Lindblom 1988, 1990).

However, the second phase of the characterisation process is to identify where the speaker information is more prominent in the extracted components. This allows the listener to focus on a limited set of characteristics in order to create, if needed, a speaker model. Indeed, in most cases a reduced number of these factors is sufficient to explain the variability found, but it is their combination and their interactions that fully describe the sources of between-speaker variability.

An inevitable question is what really represents a speaker from a phonetic standpoint? A speaker is a person whose speech productions are taken as the object of study in a phonetic research. However, paralinguistic factors can have an important influence in the context of the study in question. Sex, stress or emotional state during a recording are characteristics as important as phonetic ones to define the "identity" of a speaker [Hansen and Bořil, 2018]. Age, e. g., is another factor of great importance when comparing speakers' phonetic characteristics. Indeed, it is the result of an unstoppable natural mechanism which influences all human characteristics, including the linguistic ones and it has been proved as a major factor of variation for multiple phonetic components. It is the object of study in [Schotz, 2007], where a large number (161) of acoustic features, divided in 7 groups, are tested on 810 female and 836 male speakers recorded over a period of 3 years. The speech material consisted of six Swedish isolated words as they contained phones which tend to convey age-related information. Segments duration, sound pressure level (SPL) range and formants were identified as the most important acoustic correlates of age. The weight of age's influence on speech components varies from both a between-speakers perspective, when considering different age groups for the same component and a within-speaker, when analysing the same speaker characteristics in a longitudinal study.

Invariance is theoretically part of what makes the perception process possible, thanks to the creation of categories with defined properties which behave in a predictable way. Concerning phonetic traits or characteristics quantified on a measurable scale in speech

signal, we look at the research of invariants not in terms of absolute measurements of a specific spectral peak but rather in terms of spectral shape. This is meant to find the minimal distance between a built model and the observed properties in the selected data. If the resulting evidence suggests that it is fully possible that two samples were spoken by the same individual, we have to take into account this evidence to model the variability matrix. Then again, the probability that their shared phonetic characteristics are not likely to be found in combination in any other individual model of the relevant population is the core of phonetic characterisation.

In [Kreiman and Stidtis, 2011] the idea of the speech carrying a large information matrix about a speaker at different levels of perception is discussed. The Authors provide a list of some of the characteristics that listeners can retrieve. We report these characteristics in Table 1.1 as they play an important role in establishing how to analyse the speech components. As for the list discussed in Figure 1.1, this is not an exhaustive set of characteristics, it still serves the purpose of understanding the basis of how far the analysis of speech can go. In our representation of the possible carried information, among other smaller modifications, we decided to change the position of two characteristics which were present in the original table from [Kreiman and Stidtis, 2011]: personal identity and meaning of the message. The first was associated with the idea of familiar voice recognition. When a listener recognises a familiar voice a perception mechanism is triggered where a series of already established characteristics are associated and compared with the heard sequence. We consider this characteristic as being a greater layer of the characterisation process, and its whole purpose, rather than just one of the possible outcomes. Recognising a familiar voice can be associated with all the other listed characteristics without the need for further analysis of a speech component.

Content of the message		Meaning of the message
Personal identity		
Physical	**Psychological**	**Social**
Age	Competence	Education
Appearance	Emotional status	Occupation
Drunk	Intelligence	Regional origin
Healthy	Lying	Role in conversation
Ethnicity	Personality	Social status
Sex	Psychatric status	
Sexual orientation	Stress level	
Smoker		
Speech disorder(s)		
+/- Teeth		
Tiredness		

Table 1.1: Some of the characteristics of the speaker that can be carried by human voice, inspired by Table 1 in [Kreiman and Stidtis, 2011].

On the other hand, the meaning of the message represents a different layer where semantic or pragmatic processes play a key role. Even if it is not the primary interest of our studies, it is important to note that whenever a speaker listens to a speech sequence the meaning and the spoken language have an influence on what the listener extracts. Following the internal-external dichotomy we have two groups which rely on the internal variability of

the speaker, physical and psychological characteristics. However since the sources of their variability rely on very different aspects of the speaker we have to consider them as two subset. The other group represents the external characteristics which only rely on factors not directly depending on the speaker.

1.2 Decoding individual signatures

Understanding the brain's activity during the decoding of stimuli is one fundamental aspect in the improvement of various domains. For instance, the use of Deep Neural Networks (DNN) to reconstruct brain's images from functional magnetic resonance imaging (fMRI) activity in [Shen et al., 2019] has shown successful results. It is shown that the DNN model can learn a direct mapping from the human brain activity during visual perception, creating images that then give consistent results in image classification tasks. This could open the way for further understanding of human mechanisms and their variability.

In a similar perspective, paralleling speaker recognition through an automatic system with human perception shows similar performance variations that can be attributed to speakers-related factors. For instance, the differences in the recognisability of speakers by automatic systems have been defined using animals in [Doddington et al., 1998; Stoll and Doddington, 2010]: *sheep* refer to speakers for whom poor performance are observed, thus whom characteristics do not show a prominent variation from the observed population; in opposition, *goats* represent the speakers with very noticeable characteristics, hence simpler to identify; *lambs* and *wolves* are the speakers whom characteristics result respectively easy to imitate and particularly successful at imitating other speakers. Thus, the ability to decode speakers' characteristics shows an important influence from the general population comparison rather than just from the isolated speaker.

An additional important consideration on voice characterisation is the ability to recognise a speaker and therefore to mark a voice as characteristic in some aspect. As mentioned above, this mechanism implies an automatic decode of the components from the heard voice. The decoded result can be either associated to a familiar voice, recalling categorical perception concepts, or to an unknown one which creates a new category within the listener. Encoding and decoding various information through vocal productions has already been studied as other species apart from humans share similar mechanisms [Ohala, 1983; Culling and Darwin, 1993; Ohala, 1994; Dediu et al., 2017]. Studies suggest that familiar voice recognition is evolutionary old [Grossmann et al., 2010; Belin and Grosbras, 2010], and emerged e. g. in amphibians [Bee and Gerhardt, 2002; Burke and Murphy, 2007], birds [Hsu et al., 2004] or primates [Rendall, 2003; Furuyama et al., 2016]. Moreover, this ability considerably precedes the evolutionary development of speech and language in human communication and cognition.

Considering examples from other animals, in bottlenose dolphins an individual signature takes the form of a specific name-like whistle which remains constant throughout the life of the animal. [King et al., 2021] gathered data during 30 years about the behaviour of a group of 14 males and tested these animals' ability to recognise allied individuals from different cooperative levels. Playback sounds of individuals' whistles show that information about the relationship between individuals is encoded in individual whistles.

Acoustical analysis was not performed, however the question remains: at what level is the information encoded? In the whistle form (the dolphin's name) or if it is in an acoustic component of the whistle (how that name is pronounced)? A different social animal tested in [Mathevon et al., 2010] shows that relationships information is present in individual vocalisation. Acoustic analysis performed on Hyenas giggling demonstrates that age and dominance appear to be related to pitch and energy distributions among the frequency spectrum. As another example, in [Mouterde et al., 2014; Elie and Theunissen, 2018] the Authors analyse zebra finches calls in order to extract individual information afterward showing that even at a long distance those signatures remain robust.

The study of how different species have developed different abilities to encode and decode individual information in vocal productions reinforces the need to better understand these mechanisms and how they work.

1.3 Prelude to the moment of *fixity*: research aims

The present chapter has served as an introduction to the basic notions involved in speaker characterisation. Furthermore, this thesis has been supported by the Agence Nationale de la Recherche project VOXCRIM (ANR-17-CE39-0016), a multilaboratory French project that focus on adding knowledge about Forensic Voice Comparison (FVC) in France, its limits and possible methodology improvements. Forensic application is not the main focus of the present work, the main contribution of this thesis rather being to bring more phonetic-driven views to the subject. In the next three chapters, a literature review unfolds in order to fixate the works and methods our investigation is based on. However, before doing this we have to provide a few more considerations which represent the research questions this thesis addresses.

The first aim is twofold, it is to provide speaker characterisation research fields with a wide overview of **speech components' interactions and their roles in characterising speakers**. Different studies on the speech components allow us to have a basic knowledge on the components themselves, their interactions, and the amount of speaker information they are capable of conveying. This analysis of speech components serves as the basis for further investigation. The **actual contribution of these components to the characterisation of speakers** is the second part of this first objective. It is strictly related to the first part as it cannot exist without the establishment of basic knowledge about the selected components. We assess the implementation of these interpretable representations in speaker recognition tasks, both individually and combined. The prior knowledge on the robustness of the components, and their link to identifiable traits of the speech production, allows to question the possibility of speech components' implementation in speaker recognition domains.

The other main objective is to assess the weight of the selected speech components in perceptual speaker characterisation. For instance, by **comparing human perception with CNN- and phonetic-based results in a clustering task**, we question the validity of our results. The possible similarity of human and non-human answers may indicate a more important weight of one or multiple components in speaker perceptual characterisation. This would invite further research on said components and allow to bring additional data to the understanding of this perceptual mechanism.

Part I

FIXITY or Literature Review

Chapter 2

Encountering speech components in phonetic literature

In this Chapter, we present the phonetic background of this thesis, showing how individual differences, or the speaker factor when applying statistical reasoning, have been studied throughout the years in Phonetics. We begin by discussing the different components on which the speaker factor has shown an influence. Both the nature of these components and the applied methods from selected studies are the object of our review.

As mentioned in the previous chapter, associating a speaker identity with a voice is an operation that can be achieved through the analysis of patterns occurring in different speech components, which are taken as characteristics of the considered speaker. Information about the speaker can be retrieved by different components at the same time. The variability matrix, which helps identify speaker characteristics and their distributions, varies from one speaker to another. In this thesis, we only focus on certain parts of the nine groups of components presented in Figure 1.1, as each part plays different roles during speech productions. As observed, phonetic measurements that are associated with a component can also be associated with another depending on what the actual object of observation is. Following this idea, we show how, throughout the years, by applying different points of views, studies have analysed the same phonetic measurements and retrieved a large number of information on individual characteristics.

Section 4.4 describes a summary of our review, presenting what aspects are retained from all the cited studies on speaker characterisation and related speech components. Our review reports on studies focusing both on large sets of phonetic measurements representing multiple components and specific ones which take in-depth analysis of the role played by a single component or measurement. This approach allows us to explain how understanding these interactions is the core of speakers' phonetic characterisation.

2.1 Source and filter

The first component we consider is named *source and filter*. During oral communication different sources operate simultaneously to create a complex wave of speech signals. Here are regrouped different variables contributing to the modulation of speech signal in the

vocal cavities. As shown in Figure 1.1, variables like f0, intensity, resonances and noise excitation are associated with this component.

Among the studies taking into account all these variables at once, we find it important to cite the works of [Keating and Kreiman, 2016; Keating et al., 2017; Lee et al., 2019], which represent, from the phonetic standpoint, a great influence on this thesis. These three studies focus on the same question, namely, the analysis of voice similarities through the scope of different phonetic measurements mainly associated with the source and filter, and voice quality components, trying to understand what roles they play in creating multiple speakers' identity matrices. The same data set is used, consisting of fifty female and fifty male speakers from the University of California, Los Angeles Speaker Variability Database [Keating et al., 2019]. All speakers are English natives and similar in age. In [Lee et al., 2019] both sexes are analysed and compared while [Keating and Kreiman, 2016] took only the female speakers as object of study and [Keating et al., 2017] only the males. The speech material is represented by 5 read Harvard sentences ([Subcommittee, 1969], Table I), for a total of six repetitions per sentence over three recording sessions on different days.

Phonetic measurements used are listed as follow: (i) **Pitch** as f0; (ii) **formant frequencies** as formants from first to fourth (F1, F2, F3, F4) and formant dispersion (FD), calculated as the average difference in frequency between each adjacent pair of formants; (iii) **Harmonic source spectral shape** as the relative amplitudes of the first and second harmonics (H1-H2) and the second and fourth harmonics (H2-H4), the spectral slopes from the fourth harmonic to the harmonic nearest 2 kHz in frequency (H4-H2k) and from the harmonic nearest 2 kHz to the harmonic nearest 5 kHz in frequency (H2k-H5k); (iv) **Inharmonic source/spectral noise Variability** cepstral peak prominence (CPP), energy, and the amplitude ratio between subharmonics and harmonics (SHR). In all three studies, moving means and coefficients of variation (CoV) values are taken as analysis measures for each sentence.

However, the statistical analysis differs between [Lee et al., 2019] and [Keating and Kreiman, 2016; Keating et al., 2017], they take two different approaches to the same question of voice similarity. In the former, a Principal Component Analysis (PCA) is used in the first study in order to show how the phonetic measurements interact with each other. This type of analysis helps the statistical description of values distributions for the different aspects measures that contribute to the source and filter component of speech. Both between and within speakers variability are studied through the explained variance and the resulting weight that the PCA assigns to each measurement.

Results obtained in the within-speaker analysis show that H2k-H5k and CPP account for the most of the variability, having the higher weights in the PCA, followed by FD and F4. This suggests that from all the studied phonetic measurements, these ones result in more stable distribution within the speaker variation matrix, but still define a limited set of elements. In addition, the fact that the first two Principal Components (PCs) account for less than half of the explained variance (32 % and 34 % respectively for female and male speakers), confirms the idea of highly variable matrices in the description of speakers characteristics.

The Authors look at the results of the variability between-speakers and find similarities between the three PCs of both sexes. Similar phonetic measurements play similar roles in describing individual differences except for H1-H2 CoV which shows a consequent weight only for men. The most important measurements are H2k-H5k, CPP and H2-H4. Indeed,

all three are present in the first PC accounting for 18 % and 20 % of variance across females and males, respectively. Formant frequencies (F4, FD, F3), correspond to the second component for women and to the third one for men, respectively, accounting for 11 % of variance in female voices and 9 % for male ones. Similarly to the second PC for females corresponding to PC3 for males, the female speakers' PC3 matches male speakers' PC2. For the two sexes the PC corresponds to spectral slope in the higher frequencies, H2k-H5k, as well as F2. This PC accounts for 10 % of variance in both female and male speakers.

All the PCA results are summarised in Figure 2.1, showing the differences of regroupments for the various phonetic measurements in the components of the statistical analysis.

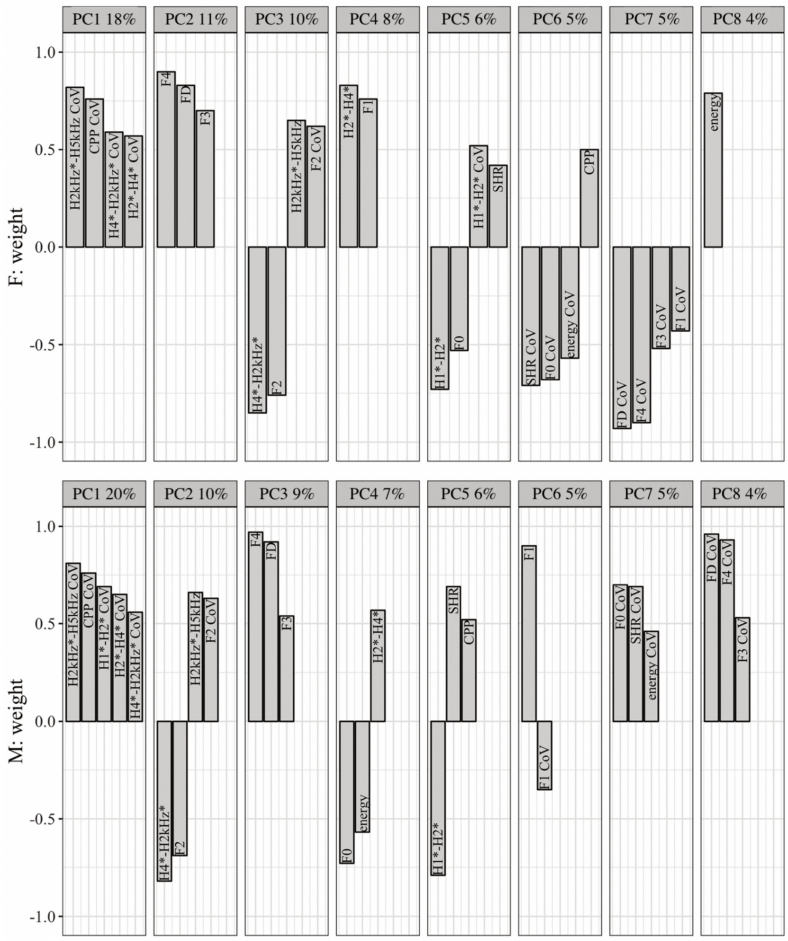


Figure 2.1: Figure 5 from [Lee et al., 2019]. Acoustic parameters emerging in 8 PCs for female speaker group (upper panel) and male speaker group (bottom panel). Variables within each PC are ordered from the highest absolute value of rotated component loadings (weight) to the lowest value. PCs variances added by Author.

The second statistical approach, shared by [Keating and Kreiman, 2016; Keating et al., 2017], involves a Multidimensional Scaling (MDS) to define voice distances in reduced subspaces and a Linear Discriminant Analysis (LDA) to classify the different speakers. In both cases, the observed measurements did not explain more than half of the between-speakers variability. However, the LDA speakers' variability modelling appears more robust than the PCA one. Regarding the classification results, men's was better with 78 %, against 68 % for women. Similarly to the PCA results, phonetic measurements have different weight in explaining the variance of both sexes with some similarities. In

these studies, the Authors compare the statistical results from both a large set of data (all 50 speakers) and smaller subsets of 5 speakers. These two approaches are interesting in order to understand if the speaker variability matrix can be influenced by the environment in which it is observed.

For female speakers, the Authors observe that f0, F4, CPP and F1 have the most weight to explain the statistical variance in the LDA over all 50 speakers. With a reduced number of speakers, i. e. 5, energy, SHR, f0 and F3 are the most important variables. Even though f0 is still present, all the other variables have changed, which demonstrate the sensitive link between speaker modelling and the number of studied speakers. An additional analysis is carried out for the whole dataset of female voices using a MDS, whereby the results correlate with both LDAs. The MDS used f0, SHR and F3 to model the dimension of its subspace and F4, H1-H2 and CPP for the second one.

In men's analysis similar measurements seem to have an important influence as f0 appears as the first one in both large and small speaker sets. Regarding the 50-speakers result the measurements having the stronger influence on the statistical models are F4, H1-H2 and SHR. In the 5-speakers comparisons, energy, H1-H2 and F4 appear after f0. Therefore, unlike female voices, reducing the number of speakers in the studied subset does not show any influence in the most important variables defining voices' acoustic subspace.

These results confirm that a single speech component, even if described by a large set of phonetic measurements, cannot account for all the needed variability to characterise a voice. Understanding the different distributions and interactions of the different component subelements is as essential as understanding the interaction between the components themselves.

2.1.1 Fundamental frequency

f0 is arguably the most cited phonetic measurement for voice characterisation. From an articulatory standpoint, it represents the frequency of vocal folds vibration which mainly correlates with the acoustic height of the produced sound. It is as well referred to as pitch, which is defined as our perception of fundamental frequency. Therefore, the term f0 describes the actual physical phenomenon, whereas pitch describes how our ears and brains interpret the signal, in terms of periodicity. Many studies have analysed f0 statistically, commonly resulting in a normal distribution. As a result, the use of mean and standard deviation (SD) values suffices when studying speakers' f0 in phonetic literature. On the other hand, the role played by f0 in speech production and in voice characterisation has been investigated through the scope of different speech habits, e. g., from mimicry to speakers' charisma. Its importance has been increasing throughout the decades. Nevertheless, as we already mentioned, a single variable cannot represent the entire highly unsteady matrix that contains a speaker's voice characteristics.

As for the study of f0 and its statistical behaviour, we find the studies and reviews carried out by [Traunmüller and Eriksson, 1995; Lindh, 2006] from Swedish read speech, and by [Hudson et al., 2007] from English, in each case read speech is used. [Traunmüller and Eriksson, 1995] studies a corpus of more than 100 male speakers from 20 to 30 years old. As multiple statistical values are studied (e. g., mean, standard deviation, range, median, etc.), all measures show a normal distribution. The mean of the mean values is 120.8 Hz.

In [Lindh, 2006], comparing the f0 of female speakers, the Authors report that mean values appear to be much higher for women while the standard deviation is approximately the same for both female and male speakers, i. e. around 15 Hz. Similar results are shown in [Hudson et al., 2007], the study of spontaneous speech by 100 male speakers, with a mean value of 105 Hz and a SD of 5 Hz. Authors noted that results from other languages such as, previously cited Swedish, and German (mean value of 115.8 Hz) appear higher due to the influence of the elicited speech. English read speech reports an average of 128 Hz, which confirms the tendency to have higher pitch values than in spontaneous speech. This phenomenon has also been reported for English, French, German and Mandarin in [Schmid et al., 2012; Gendrot et al., 2012; Swerts et al., 2014; Yuan and Liberman, 2014].

In an early work by [Boë et al., 1975], 30 male and 30 female speakers f0 distributions are analysed from French read speech. Results show that the distribution is gaussian for the given corpus, regarding separately the two sexes. The variations of f0 are low for a given speaker, on the average of 18 Hz and 20 Hz, from one record to another. The mean values of f0 are 118 Hz for men and 207 Hz for women, showing similar results to the above cited statistics in other languages. Another element highlighted by the Authors is the difference in ratio of voiced duration to the total duration of speech production which as well appear significantly higher for female speakers.

The use of f0 to differentiate men and women has been reiterated in further analysis throughout the years. [Keating and Kuo, 2012] compares 23 American English and 23 Mandarin speakers (11 men and 12 women in both languages) in multiple tasks involving both vocalisations and text reading. [Iseli et al., 2007] takes a large corpus of 335 speakers (185 males and 150 females) from different age groups, in order to understand the differences between French and English when considering the influence of sex on f0. On the other hand, [Pépiot, 2013] compares French and English speakers. Although age is demonstrated as having a great effect in decreasing f0 mean values, sex always shows a significant effect on all f0 statistical measures. This outcome is stronger for males than for females as reported by [Iseli et al., 2007]. Speaking rate, formants or duration are other studied measurements which have not shown the same consistent results and appear less robust than f0 for sex discrimination.

Concerning robustness of f0 measures, [Lindh and Eriksson, 2007] tests f0 degradation in both read and spontaneous speech through different tasks involving: the influence of emotional states; the use of different recording devices; and the influence of vocal effort through speech elicited by having the speakers communicating over various distances. The experimental language is Swedish and the material consists of recordings from 18 speakers: 6 males, 6 females and 8 children. In the first of these experiments, emotional state influence on f0 is tested asking the speakers to simulate happiness, anger and sadness. Comparing f0 from these conditions result in approximately the same values as the standard recording. Simulating different emotions results in mean, median and SD values to be affected by outliers but not to be heavily distorted. Channel distortion appears consistently when the microphone is not the recording device. Mobile phone recordings show the highest degradations due to low-band pass filters. Vocal effort influence showed that distances higher than 1.5 m resulted in an increase of the mean and median values. The Authors explain that it is actually the combined effect of an increase in the base value and in range relative to the speakers' f0 values.

As mentioned, f0 is only a part of what is called and perceived as the voice pitch. It is

important to note that, to better understand the importance of the role played by every aspect of a speech component, they need to be studied in comparison to each other and, when possible, tested through the scope of human perception.

To this effect, in order to investigate the contribution of vowel formant frequencies and f0 to sex and height identification, [Gelfer and Bennett, 2013] tests listener judgements on voices of 30 speakers of both sexes. The vowels /i/, /a/ and a read sentence are recorded and their f0 are digitally altered to five distinct values in order to represent typical f0 in the average males, the average females and ambiguous f0 ranges: 116, 145, 155, 165, and 207 Hz. The results indicate that female speakers were perceived as female even with an f0 in the typical male range, while for male speakers, gender perception was less accurate when higher than 150 Hz. The study of isolated vowels or syllables reinforces the idea that unaltered vowel formants, as well as other aspects contributing to speech production related to voice pitch, appear to be important for sex perception.

In [Eriksson and Wretling, 1997], an impersonator is instructed to reproduce three well-known Swedish public figures. The resulting f0 and formants values are compared to his standard productions along with the target ones in order to weight the influence of mimicry on these phonetic measurements. The impersonator's vowel space appears greater than the target voice in all cases he manages to make his f0 exactly the same or with a small variation of 4 Hz from the target.

The influence of speech style or speech elicitation on f0 should be considered as external influence factors on the speaker's production. In contrast, ageing can be considered as a biophysical or internal influence factor on f0 variation. However, a third type of influence has been studied in phonetic literature, namely, the interlocutor which has an interest for this thesis since the main analysis is made from a corpus of spontaneous exchanges between two speakers. A last issue we address concerning the role of f0 and its contribution to the source and filter component relates to phonetic convergence theory. As mentioned above, speakers may present characteristics not only in the colour of the produced speech, its components, but in its shape too, the lexical content. Following the findings on convergence in conversational interaction, the interacting talkers may influence one another in both the colour and shape of their speech productions. In [Pardo, 2006] phonetic convergence is tested on 12 speakers from both sexes, with recordings before and after conversational interactions where, alternatively, a subject received or gave instructions to the other about finding an objective on a map. The results suggest that the influence interlocutors exert on each other were more prominent for men as receivers than for women, in terms of perceived similarity. f0 resulted in the more variable phonetic measurement followed by F3, getting closer to the instructor values, while the other formants remained stable to their speaker "prototype".

The presented results reinforce the idea that f0 plays an important and robust role in characterising a voice but it appears to be weak against human imitations or to fail when comparing exposed interlocutors' speech. However, because of the dynamic nature of speech, pitch has to be represented by more than static values, as long as the mean of a normal distribution remains consistent. Other acoustic correlates of pitch present in the literature, such as envelope cues or spectral moments [Culling and Darwin, 1993; Ardoint and Lorenzi, 2009; Dellwo et al., 2012; Niebuhr and Skarnitzl, 2019], may be used to better understand how speaker characteristics from the source components interact.

2.1.2 Resonances

During speech production, air from the lungs is forced to pass through oral and/or nasal cavities. In this process the consequent wave is shaped by the cavities, which create resonances, and when exiting them becomes speech as we perceive it. Some of the resonances are represented as reinforced harmonics on the spectrum, which is the result of the speech signal mathematical decomposition, even though they do not always correspond to a multiple of the fundamental as harmonics do. Conventionally these are called formants and the first four (F1-4) are the most salient in normal human hearing. The passage of air through the nasal cavity generates what are called anti-formants, zeros in the spectrum, which characterise nasal segments production.

As for f0, the use of formants is limited to the study of vocalic segments. However, as mentioned hereinabove, multiple computations are present in the literature showing the extent of possible analysis application for these measurements. In [Audibert et al., 2015] five metrics for F1 and F2 are summarised to show the possible correlates between formants, and the characteristics of the studied categories, namely, French vowels from 180 speakers and different speaking styles. The metrics include the vocalic space and dispersion areas, vowel distance to the overall centroid of the speaker's vowel space and range ratios for both studied formants, designed to measure the ambitus in the respective dimensions: F1, jaw/tongue height; F2 front-back tongue dimension and/or lip rounding. As the focus of the study is on vowel and speech style discrimination, no investigation was done on the weight these metrics have in characterising between-speakers formants. However, the application of the presented metrics to study the role played by formants in speakers characterisation could help to reinforce the overall robustness of these measurements.

In Phonetics, formants have been associated to vowel discrimination studies but in recent works they have as well shown to play a role in speakers characterisation. We already mentioned that in [Eriksson and Wretling, 1997], formants are pretty robust when facing vocal imitation since the impersonator was not able to replicate the exact target vowel space while he managed to produce the same f0 values of the target. Similarly, [Yang, 1992; Eichhorn et al., 2018], report data on f0 and F1-4 for English and Korean speakers, showing that formants play an important role in characterising both sexes and different age groups. The most consistent age-related effect was a decrease in f0 for women. Concerning formants, significant differences are shown to be vowel-specific for both sexes while no significant effects of age were observed for F4. Authors explain that women's significant decrease in f0 is likely related to menopause, which is a biophysiological variable that speakers cannot control. In contrast, formant frequencies of the corner vowels slightly change throughout the decades of adult life, either due to physiological ageing which has small effects on these variables or to individuals who compensate for age-related changes in anatomy and physiology. This reinforces the idea that formants accurately represent speakers' characteristics and have an important role in describing the speaker variability matrix.

Similarly to what happens for f0, formants have been analysed using mean values in the literature. F1 and F2 have served for the classification of vowels in a subspace capable of describing basic characteristics in a language. In [de Jong et al., 2007; Nolan et al., 2011] English vowel formants are tested in read speech by 50 and 15 male speakers. The results from the first study on F1 and F2 frequencies show speaker classification rates highly

dependent from the considered vowel. This assumption is confirmed by the results from [Nolan et al., 2011] where F3 appears more stable both between and within a speaker. The influence of lexical content on speaker discrimination have been studied for French in work such as [Kahn, 2011; Ajili et al., 2016]. The first establishes a ranking for phonemic classes which are more influenced by the speaker factor. The same ranking has been confirmed in [Chignoli, 2018] and is presented in Table 2.1.

- Nasal vowels
- Nasal consonants
- Open/mid-open oral vowels
- Laterals
- Other oral vowels
- Occlusives
- Fricatives
- Approximates

Table 2.1: Ranking of phonemic classes most influenced by the speaker factor in French, established by [Kahn, 2011] and confirmed in works by [Ajili et al., 2016; Chignoli, 2018; Gendrot et al., 2020].

As shown, nasal segments are the most stable in between-speakers variability as a direct consequence of the resonances created by the passage of air through the nasal cavity. While for vowel analysis, formants are measurements capable of representing the speaker's characteristics, for consonants, other spectral measurements have been used, e. g., spectral moments [Weingartová and Volín, 2013] or Mel Frequency Cepstral Coefficients (MFCC). Oral vowel and lateral consonants are the following most stable segments, reinforcing the idea that resonances play an important role in rendering speaker characteristics.

Since the establishment of a correlation between formants and biophysical characteristics of speakers during speech production, their dynamics and temporal organisation have been the object of study in order to better understand how individual strategies are employed. Some examples are given by [McDougall, 2003, 2005, 2006; Nolan et al., 2006] where formant dynamics in English vowels or diphthongs are modelled using quadratic and cubic polynomial approximations. The results confirm that F3 is the most stable between speakers, followed by F2 slopes in diphthongs.

In addition, as reported by [Zuo and Mok, 2012], speakers are better characterised using the dynamic features of the formant transition, using polynomial linear regression to model the formant contour, rather than a small time slice. Promising results show a mean Equal Error Rate (EER) of 13 % when classifying male speakers, with the slope of F2 appearing as the most stable feature.[Kinoshita and Osanai, 2006] shows the same tendency, when comparing F1 and F2 dynamics from 20 English male speakers on the vowel /uː/ receiving nuclear stress in the hVd context 'who'd'.

[He et al., 2019] also studies formants dynamics, limited to F1, comparing positive and

negative ones, in order to better understand how individual strategies differ in representing the patterns to attain a target (the pronounced vowel) or to leave it. In this study, 16 sex-balanced speakers were considered reading 256 sentences in Zurich German which vowels' F1 dynamics were then compared. Using a Logistic regression model, negative dynamics resulted in 70 % of the explained variability. Modelling of formants dynamics and their properties is typically associated with speaker-specific articulators movement trajectories, which are, as mentioned, a combined product of inborn anatomical peculiarities of speech organs constraining their biomechanics, idiosyncratic neurological substrates regulating the motor control of articulators and individual habits acquired by speakers throughout their lifetime. Although, the opening-closing cycles of the mouth movements and the interactions with other articulators must play a non-trivial role in characterising a speaker's voice.

2.1.3 Intensity

In Figure 1.1 it has been noted that intensity is part of the source and filter component of speech since it appears as an intrinsic characteristic of produced speech resulting from a direct computation of the speech source. Throughout the decades, intensity has been shown to convey different types of information. We placed the review of this measure at the end of the source and filter and before the temporal one due to the strict correlation it has shown with both components. In a purely physic perspective, intensity is defined as the power, expressed in Watts (W), carried by sound waves per unit area (m^2) in a direction perpendicular to that area. The direct computation of intensity is obtained by the logarithmic scale of the power on area ratio and measured in dB. As for the sound amplitude, however with a doubled factor, the longer the distance from the source, the less reliable the measured intensity is. The last observation has been a major factor of controversy for the use of intensity during speech studies since the distance from the source cannot always be controlled. However, since human hearing is based on the perception of changes in both pressure and intensity levels from the sound waves, researchers have tried to obtain more reliable correlates of these measurements.

An extended discussion on intensity, on its roles in speech production, and its interactions with other phonetic measurements is done in the works by [Sorin, 1981; Künzel et al., 1995]. Although the review made by the Authors does not concern directly how intensity correlates with individual differences, it is interesting to note some of the considerations made on its perception and linguistics mechanisms. A major point of discussion is the absence of a direct articulatory correlation with intensity. Indeed, studies on vowel intensity showed that vowels of different intensity can be perceived as isophonic if they are produced with the same articulatory effort. The judgement of a listener is directly related to the physiological effort made, which is proportional to the square of the subglottal pressure. However, it has been proven that the loudness of vowels can be deduced from a physical measurement of the acoustic signal, based on the distribution of energy, especially at the level of the first and second formants. The sensation of volume therefore does not necessarily require articulatory information. It may rather be a "purely" psychoacoustic signal treatment which accounts for the effect.

The notion of speech intensity is as well associated with loudness or "volume". Indeed, the articulatory justification often evoked for the apparent confusion between pitch level

and volume is not an accurate explanation. It merely states that an increase in subglottal pressure brings a simultaneous increase in intensity and in the f0, both parameters often being correlated up until the intervention of an adjustment in vocal folds tension.

In regards to the role played in perception, [Sorin, 1981] reports that an increase in intensity brings an increase in the perceived frequency as, e. g., in French where intensity intervenes most especially at the end of an intonation group. If in French, one wants to transform a statement into a question, not only does the f0 contour have to be modified, but the intensity of the last part of the sentence must also be increased and diminished for the penultimate syllable.

In [Aubanel et al., 2018] intensity is investigated in relation to intelligibility where listeners need to identify sentences in which segments were replaced by noise following different filtering in order to impact intensity. Authors find that both are closely related, but with a marked temporal difference, namely cochlea-scaled entropy (CSE) peaks occur significantly earlier than intensity peaks in English. The disparity in temporal alignment may be associated with a small but significant advantage in capturing important speech information, since replacing high-CSE segments with noise disrupts intelligibility to a greater extent than when replacing high-intensity segments. Moreover, [el Gamal, 2015] found that intensity, at the perceptual level, is a remarkable cue for listeners to identify speaking rate, with high intensity indicating fast speaking tempo and low intensity indicating slow speaking tempo.

Other perceptive implications of intensity are discussed in works such as [Tweedy and Culling, 2014; Clark et al., 2014]. The first study demonstrates that vocal intensity increased as the background noise level increased but remained stable when Signal-to-Noise Ratio (SNR) of the interlocutor changed. The second study highlights how individuals affected by Parkinson's disease process intensity, and suggests that they may perceive speech's loudness with a deficit due to the abnormal perception of externally generated and self-generated speech intensity stimuli. These results on speech loudness perception appear to provide additional support for both sensory impairments and sensorimotor integration deficits in Parkinson's disease.

The use of intensity for speaker recognition tasks is not as common as other phonetic measurements, mainly because of the recording constraints and similar limitations. [He et al., 2015b; He and Dellwo, 2017] are some of the studies investigating the weight of intensity as a speaker characteristic. [He and Dellwo, 2017] investigates intensity dynamics using derivative and multiple intensity metrics on sentences read by German speakers showing a high statistical between-speakers variability. In [He et al., 2015b], Authors test whether the statistical results correspond to a robust modelling in a speaker classification task. The mean recognition rates were 14.2 % (duration only), 30.3 % (intensity only), and 36.9 % (duration + intensity), demonstrating yet again that statistical variability does not always result in a direct robustness of the studied measurement.

Intensity, has been mainly correlated with suprasegmental cues, such as duration, intonation or stress, but intensity fluctuations of the speech signal cannot be ascribed to a single factor. Still today, the processing precision of its linguistic functions by the auditory system needs to be understood to their full extent, especially within the natural context of continuous speech.

2.2 Prosody

Source and filter component accounts mainly for biophysical characteristics which, as the name suggests, relates to phonetic measurements of the speech sources such as glottal or noise and to the filters that through the vocal cavities modify the speech wave. However, as already mentioned hereinabove, the basic dichotomy used for speaker characteristics opposes innate/internal to learnt/external characteristics to compose the large variability matrix which phonetically defines a speaker.

In this section, we review other components which have been the subject of studies on speaker characterisation and are of interest in this thesis, following the nomenclature of Figure 1.1, **Temporal** and **Intonation** components. These have been commonly referred to under the larger definition of *prosody*. In Phonetics, the term prosody widely designates every aspect that is higher than the segmental plane, but does not refer to a single component or phonetic measurement. Hence, the convenience of regrouping studies on different prosodic aspects in a single section.

Given the acoustic-articulatory dichotomy in phonetic measurements description, an additional important aspect that has to be considered is the extent on which these measurements are investigated. As mentioned hereinabove, speech is a complex dynamic object formed by subsequent articulation of smaller units, which are the phonemes in a chosen language, whose interactions and influences on each other result in what we experience as speech units. Considering units larger than isolated segments implies the consideration of what has been called the prosodic or suprasegmental dimension. Prosody is not a single speech component, but represents a large set of different elements allowing the segmentation of continuous speech into smaller units, carrying information about the language spoken or discourse organisation, and eventually regarding speaker characteristics. Before diving into the different prosodic cues which have been studied in relation to speaker characteristics, we discuss some general aspects of prosody in phonetic literature.

Prosody involves mechanisms that extend beyond a phonemic level of speech and mostly independently of segmental features, contained in hierarchically organised linguistic units. At an acoustic level, some of these prosodic features are f0, intensity and duration. We already covered what f0 and intensity represent when studied at a lower level, but, in a prosodic perspective, their modulations are considered. Duration is influenced by the length of speech units and silent intervals.

Speech prosody has a well-investigated structure. Most theories of prosodic structure presuppose that prosodic units are organised into a hierarchy, summarised in Figure 2.2. We can see that utterance is at the highest level, followed by the intonational phrase, phonological phrase, prosodic word, (metrical) foot, and finally syllables at the lowest level. Prosodic and syntactic structures often coincide in a way that major prosodic boundaries tend to occur at major syntactic ones. The perception of prosodic boundaries has been vastly investigated and has showed, e. g. in [Lehiste et al., 1976; Nespor and Vogel, 1986; Seidl, 2007; Johnson and Seidl, 2008], that adult listeners use some prosodic cues to determine them. These include pre-boundary lengthening, pauses, pitch or other acoustic changes in the pre-boundary word.

However, infants have shown early sensitivity to prosodic cues, as many studies indicate that prosody plays an important role in early language development. Past the 28 weeks

of pregnancy, the foetus is able to hear sounds that are coming from outside low-pass filtered by the womb [Querleu et al., 1988]. Hence, only 30 % of the phonetic information manage to reach the ears of the foetus, with pitch contours and rhythmic information being amongst them. Furthermore, newborns have shown sensitivity to pitch contour [Nazzi et al., 1998b] and prosodic boundaries [Christophe et al., 2001] as well as discrimination between rhythmically different languages [Nazzi et al., 1998a; Nazzi and Ramus, 2003].

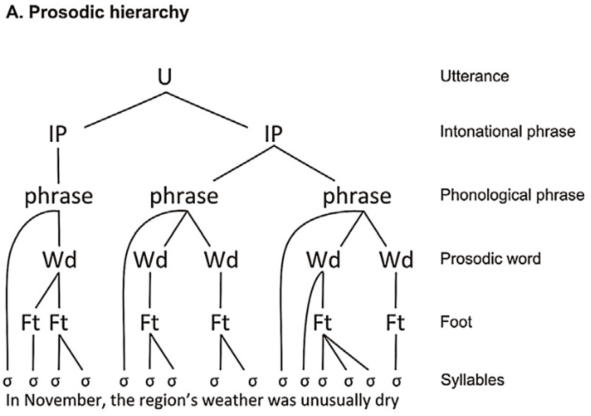


Figure 2.2: Figure 3 from [Mol et al., 2017]. The prosodic hierarchy of a sentence.

During their first year, infants have shown sensitivity to multiple prosodic cues, e. g., variations in duration [Mugitani et al., 2009; Sato et al., 2010] or in rhythm [Nazzi et al., 2000; Nazzi and Ramus, 2003], and stress patterns of their native language [Jusczyk et al., 1993; Höhle, 2009] even before the development of a large lexicon. For instance, English-learning infants show before 9 months of age that they have a preference for their native stress pattern [Jusczyk et al., 1993]. Between 4 and 9 months, tone-language-learning infants become more sensitive to native tonal patterns [Mattock and Burnham, 2006; Mattock et al., 2008; Yeung et al., 2013; Liu and Kager, 2014], while non-tone-language-learning infants gradually lose their sensitivity to those. Infants between 7 and 12 month have shown discrimination in intonational contrasts [Butler et al., 2016]. The role played by prosody since earlier development of language is consequent and can be used to explain multiple mechanisms involved in both producing and parsing the perceived speech, see as well [Morgan, 1996; Mol et al., 2017] for larger reviews.

Language discrimination covers a large part of prosodic studies including adults. As described by [Ramus and Mehler, 1998; Arvaniti, 2013], duration, rhythm, intonation or speech rate analysed in a suprasegmental perspective can explain an extremely large part of language differences. These two works used a similar approach to investigate the role played by prosodic cues and how humans use them to differentiate various languages. [Ramus and Mehler, 1998] particularly focused on syllabic rhythm which appears as an excellent, and possibly the best, prosodic cue for the discrimination of languages that are said to differ in rhythm. The main contribution of the study is to provide a methodology allowing to separate and analyse different components of prosody. In particular, transforming English and Japanese sentences following four methods: *saltanaj*, replacing fricatives with /s/, stops with /t/, liquids with /l/, nasals with /n/, all glides with /j/ and vowels with /a/; *sasasa*, replacing all consonants with /s/ and all vowels with /a/; *aaaa*, replacing all phonemes with /a/; *flat sasasa*, which is similar to *sasasa*, but with all sentences being synthesised with a constant fundamental frequency at 230 Hz. [Arvaniti,

2013] retained only the *sasasa* transformation but applied to a larger set of languages, namely Danish, Greek, Polish, Spanish and Korean.

Discrimination with impoverished stimuli is difficult and listeners take advantage of any differences they can find to help them in the task. Such differences relate primarily to speaking rate and f0 but also to localised timing differences, such as final lengthening, when speaking rate and f0 are absent. Since all these prosodic factors are ordinarily present in the speech signal and interact in perception, the present results cast doubt on the view that timing is primary and can be processed by listeners independently to all the other prosodic variables. Discrimination was possible both across and within rhythm classes, when speaking rates differed between context and test, but largely impossible once speaking rate differences were eliminated. The changes in responses associated with speech rate and f0 indicate that language discrimination arises from interactions between prosodic factors and that timing contributes to it, but on a smaller scale.

Given the vast discussions on the weight that prosodic cues play in speech production and perception, obviously authors throughout the years have as well considered the possible implications for individual differences analysis. We already cited [Barlow and Wagner, 1988] as an early work on the subject. In this study, the Authors use cues such as f0, energy or voicing modulations in order to model the patterns corresponding to each of the 5 studied speakers in English read sentences. Patterns comparison using linear prediction and clustering distances show greatly reliable results, highlighting the role played by prosody cues in characterising speakers' speech habits. Studies that have focused on prosodic cues to investigate individual differences are the object of the following sections, distinguishing between cues relying on the **temporal** component from those relying on the **intonation** component.

2.2.1 Temporal

When considering prosodic components, the relative prominence of an element over the others within the same hierarchy level and organisation are fundamental concepts, see Figure 2.2. Stress can be described as the relative prominence between syllables, i. e., one of the cues related to the prominence idea defining the temporal prosodic component. Stressed syllables are, in general, longer in duration, higher in intensity and pitch, and more distinctive in spectral quality than unstressed, or "weak", syllables. Thus, stress is correlated with multiple acoustic features that can vary in importance between languages. In addition, articulation rate and rhythmic patterns are used to define the organisation of stress and the relation between prosodic elements.

Stress is organised in a metrical structure which is part of the prosodic hierarchy. At each level only one unit is the most prominent with the others ranging at different levels. Hence, one syllable is prominent in the foot and at a higher level, one foot is prominent in the prosodic word. Another property of metrical stress in speech is its rhythmic pattern in which stressed syllables alternate with unstressed syllables. This organisation is referred to as a metrical grid in [Liberman and Prince, 1977]. Each position in a metrical grid represents a timing unit or beat. They can as well be fulfilled by silent intervals, which individual characteristics and strategies have been the subject of studies such as [Mary and Yegnanarayana, 2008; Künzel, 2013; Leemann and Kolly, 2015]. Findings in [Kolly et al., 2015] do not correlate pausing behaviour with the articulation rate, obtained by the

means of the number of automatically detected peaks in the amplitude envelope, whereas other studies have done so. In a multilingual approach, [Kolly et al., 2015] takes 16 speakers of Zurich German to read 16 sentences, and their literal translations in English and French, resulting in a total of 768 sentences containing 15 to 20 syllables. During the reading process, hesitations in the form of filled pauses were discarded and the speakers were asked to repeat the sentence, while hesitations in the form of silent pauses were not repeated. Therefore, in terms of language effects, Authors found that speakers produced the fewest and shortest pauses in their mother tongue. English and German sentences pausing values were very similar across all speakers, and the French speakers had the highest amount and the longest pauses, suggesting the influence of language typology on temporal cues. The speaker is also shown to be an important variation factor with significant between-speaker differences, in the amount as well as the duration of pauses. At the same time, measures varied little within speakers.

In a similar way, [Leemann and Kolly, 2015] performed two production experiments and one perception test, which revealed high between-speaker and low within-speaker variability concerning temporal features. In a disguised voice foreign dialect imitation experiment, the majority of speakers were not accepted as natives by listeners of the targeted dialect. The findings suggest that imitators who were accepted as native speakers have succeeded in adjusting speaking rate and global intensity peak variability, amongst other segmental and suprasegmental features, similar to the targeted dialect.

However, [Künzel, 2013] found low speaker-specific values for speakers' pausing behaviour with an important within-speaker variability when comparing read and spontaneous speech. In [el Gamal, 2015], speech rate and pauses are examined in a perception study with 60 listeners, on spontaneous Arab speech produced by both female and male speakers. Both cues show important roles in characterising speakers. Rate is another cue from the temporal component, defining the articulation speed, the manner in which speakers make use of their articulators to produce sentences/syllables/segments. Both pausing behaviour and rate can be influenced by factors external to the produced speech such as speaker's stress due to the conversational situation, the number of recording sessions, in the case of a multiple sessions study. These are just some of the examples that can determine the within speaker (ir)regularity of temporal features. [Ochi et al., 2015; Chardenon et al., 2020] studied this issue. In the first case, the ability to control speech rate with real-time visual feedback was compared between people who do and do not stutter, respectively 9 and 7 participants, with 15 Japanese sentences read aloud after repeating a played-back one. One out of the six trials, which made each of the three sessions, were accompanied by real-time visual feedback of the subject's speech rate and the target speed. No difference in speech rate between second and third sessions was shown with a significant reduction from the first one. Both groups of participants showed a high within-speaker similarity of speech rate control across the sessions with, however, a larger error in reaching the expected rate by participants who stutter.

The other cited study, [Chardenon et al., 2020], examined temporal organisation of speech in 9 French speakers in 20 recordings during 6 to 7 sessions within a one year gap. Results on pauses duration and speech rate show that representing the cues modulation and variation, both within and between the recording sessions for a speaker, is more important than the cues themselves in order to characterise the production strategies of speakers. To a similar extent, [Braun and Rosin, 2015] demonstrated that subjects exhibit distinct patterns of hesitation marker usage, concerning both the number and the

type of the marker. Whenever speech is produced, disfluencies, i. e., disruptions of the speech flow or hesitations, are bound to occur, and pausing can also manifest as hesitation markers. These can be seen as indications of verbal planning and/or monitoring of the speech signal [Shriberg, 2001]. Phonetic manifestations of hesitation have been studied through seven different markers, namely: insertions of a vowel ("uh"), a nasal ("mh"), or a sequence of vowel + nasal ("um"); initial vowel or consonant lengthening; final vowel or consonant lengthening. Phenomena of speaker-specific patterns of hesitation with a consistent lowering of f0 were observed.

Rhythm can be seen as the outcome of some phonological properties. Indeed, by paying attention to rhythm, newborns are able to discriminate between languages which have different phonological properties. There is not a single feature that defines speech rhythm and temporal organisation. The metrical characteristics of stress are useful to segment the speech signal into smaller, word-like units, as well as grouping speech sounds that belong together. Even though stress and rhythm have shown highly language dependent results as shown in language discrimination studies, these cues are investigated as well in a between-speaker perspective in studies such as [Dellwo et al., 2012, 2015; Braun and Rosin, 2015]. Authors take walking patterns in order to parallel how rhythm and temporal organisation of speech can be taken as important components to characterise speakers' oral productions. A wide set of rhythmic measures or correlates, see Table 2.2, have been studied in-depth in [Dellwo, 2010] and reviewed by [Loukina et al., 2011].

RM	Description	Type of interval	Scope	Normalisation	Reference
%V	Percentage of vocalic intervals	Ratio	Global	Yes	[Ramus et al., 1999]
Vdur/Cdur	Ratio of vowels duration to consonant duration	Ratio	Global	Yes	[Barry and Russo, 2003]
Δ V	Standard deviation of vocalic intervals	V	Global	No	[Ramus et al., 1999]
VarcoDV	Δ V/mean vocalic duration	V	Global	Yes	[Dellwo, 2006]
VnPVI	Normalized pairwise variability index (PVI) of vocalic intervals	V	Local	Yes	[Grabe and Low, 2002]
medVnPVI	VnPVI computed using median value	V	Local	Yes	[Ferragne and Pellegrino, 2004]
Δ C	Standard deviation of consonantal intervals	C	Global	No	[Ramus et al., 1999]
VarcoDC	Δ C/mean vocalic duration	C	Global	Yes	[Dellwo, 2006]
CrPVI	Raw PVI of consonantal intervals	C	Local	No	[Grabe and Low, 2002]
CnPVI	Normalized PVI of consonantal intervals	C	Local	Yes	[Grabe and Low, 2002]
medCrPVI	CrPVI computed using median value	C	Local	No	[Ferragne and Pellegrino, 2004]
PVI-CV	PVI of consonantþvowels groups	CV	Local	No	[Barry et al., 2003]
VI	Variability index of syllable duration	CV	Local	Yes	[Deterding, 2001]
YARD[a]	Variability of syllable duration	CV	Local	Yes	[Wagner and Dellwo, 2004]
nCVPVI	Normalized PVI of consonant + vowel groups	CV	Local	Yes	[Asu and Nolan, 2005]

[a]The definition for YARD is yet another rhythm determination.

Table 2.2: Table I. from [Loukina et al., 2011] summarising the used Rhythmic Measurements in that study classified by type of intervals, scope, normalisation and Reference.

In [Dellwo et al., 2015], data from Standard German and Zurich Swiss German revealed that there are strong differences between-speakers in acoustically measurable speech rhythm, even when prosodic and linguistic variability within speakers is strong. The evidence that rhythm measurements based on vocalic and consonantal intervals can vary significantly within a language as a function of speaker is highlighted in a number of works, e. g., [Dellwo et al., 2012; He et al., 2015b; Dellwo et al., 2015] on both spontaneous and read German sentences or in French [Gendrot et al., 2018]. Evidence from English speakers are discussed in [Johnson and Hollien, 1984; Wiget et al., 2010; Yoon, 2010] showing significant between-speakers variabilities for vocalic measures, temporal information derived from the amplitude envelope and both voiced and voiceless intervals of the speech signal. These and other studies [Arvaniti, 2013] have shown that rhythm measures are strongly influenced by between-speakers variations even though within-speaker prosodic and linguistic variability has little effect on them. Authors suggest that coupling these findings with articulatory measurements of speaker individual control mechanisms may

result in further understanding of individuals' rhythmic temporal characteristics.

2.2.2 Intonation

In Section 2.1.1, we described the use of f0 as a correlate of pitch and in particular its study through statistical descriptors which, at a vowel-level, can account for speaker characterisation. This is just one of the possible scopes in which f0 is studied, another is the notion of intonation. The same phonetic measurement is used to represent different components, as we discussed hereinabove, e. g., for intensity, but not in the same way. When studying pitch in general we have described how phonetic studies mainly use f0 mean values or other statistical descriptors. In order to study intonation, a representation of the f0 variations throughout the utterance needs to be used.

The functions of pitch variation in speech can be divided into two broad categories: lexical tone and intonation. In tonal languages, f0 is primarily linked to signal lexical contrasts. In stress languages, multiple features are linked to the lexical stresses, as partially discussed earlier, with intonation being the main determinant for the shape of the f0 contour. There are highly language-specific characteristics, namely, "semantic", "systemic", "realizational" and "phonotactic" distinctions in intonational structure across languages, see [Ladd et al., 1985; Vaissière, 2004]. However, all types of languages (tonal, pitch-accent, stress and boundary languages) use intonation and share intonational features at a post-lexical level.

Intonation has shown both linguistic and paralinguistic functions. Firstly, its linguistic relevance is given by pitch-meaning relationships, that can be language-specific, e. g., French informants do not classify a syllable as an accented one when it has a falling pitch movement, whereas Swedish and Dutch listeners do. In addition, the location of the onset of the pitch movement seems to have much less weight in French than in Dutch or Swedish, see [Vaissière, 2004] for further discussions.

On the other hand, the paralinguistic use of pitch can express emotional properties of the speakers and about the linguistic message, which can be interpreted on a more continuous scale. Moreover, the paralinguistic use of pitch has been argued to resemble aspects of animal communication, [Ohala, 1983, 1994]. The so-called *Frequency Code* associates high or rising pitch with smaller vocal folds, and indirectly, with a smaller size of the vocaliser. A low or falling pitch is associated with larger vocal folds and large vocalisers. Consequently, a low or falling pitch signals "big" meanings, e. g., "confident", "dominant" and "aggressive", while a high or rising pitch signals "small" meanings, e. g. "submissive", "friendly" and "vulnerable". However, human communication has evolved to be the expression of a wide range of characteristics, see Table 1.1. Languages can, nevertheless, differ in the exact implementation of the biological codes due to differences in the range of standard pitches used in a language, and the choice of specific form-function mappings to express. In addition to the interaction between different information carried by the studied speech, another important factor influencing an utterance's prosodic representation is individual intonational strategies. Research has shown that speakers should not be assumed to be homogenous even though they speak the same language.

Language-specific characteristics may have important results when looked at in a between-speaker perspective. In this sense, [Cangemi, 2009] has shown the variability of intonation

patterns between-speakers in terms of peak alignments to the stressed vowel onset or nucleus using a corpus of Neapolitan Italian recordings. In a similar way, in [Cangemi et al., 2015], speaker-specific strategies of encoding and decoding intonational contrasts, e. g., peak alignment or duration of the target words, are studied through a production task executedby German speakers and a subsequent perception study. Speaker-specific encoding of intonational contrasts is highlighted by how phonetic cues are used to encode focus structures, in terms of robustness (i. e. how many cues are employed) and partitioning (i. e. how many contrasts are expressed by a single cue). Listeners sensitivity to this variation is shown by the fact that one speaker appears to be more intelligible for one particular listener and less intelligible for another. This suggests that the interaction between phonetic cues and intonational contrasts is not linear and stable but rather relies on a highly variable individual-specific network of phonological knowledge.

These findings are discussed further by [Ouyang and Kaiser, 2015], highlighting how intonational strategies from the Standard German and Neapolitan Italian differ, as some speakers produced systematic differences in the location of the f0 peak with respect to the target syllable, while others produced systematic differences in how steep and large the f0 rise or fall was. In addition, Pisa Italian speakers differed in cue strength, those who had greater alignment differences also had greater differences in shapes, showing yet another strategy to encode informativity through intonation. In particular, [Ouyang and Kaiser, 2015] looks at American English speakers in a production study. These results demonstrate the presence of speaker-specific behaviour in prosody with speakers having individual preferences regarding the prosodic patterns of utterances and the magnitude of prosodic cues for information structure. In related works, [Andreeva et al., 2007] investigated the influence of duration, f0, intensity and vowel quality for the distinctions between narrow contrastive focus, narrow non-contrastive focus, and wide focus in German. The Authors confirm that some participants produced differences larger than others: some speakers used one parameter to a greater extent than another. Moreover, individual participants showed their own strategies in producing prosodic prominence.

Other phonetic cues are used to study intelligibility and information structure in speech such as Temporal Envelope (ENV) and Fine Structure (TFS) which have been studied in [Ardoint and Lorenzi, 2009; Søndergaard et al., 2011; Moon and Hong, 2014; He and Dellwo, 2016; Lancia et al., 2017] with regards to their individual pattern organisation for Korean, French and German speakers. The results show that both ENV and TFS information are represented in the timing of neural discharges. Therefore it conveyed important but distinct phonetic information about melody perception with the first being mostly correlated to quiet speech intelligibility but easily degraded in noise, while the latter is important for speech perception in background noise with correlation to both pitch perception and sound localisation.

Some studies use prosodic characteristics in relation to components of the between-speakers variability model such as duration, intensity, rhythmic correlates or other dynamic representations that do not exhaust our understanding of the information they carry. All of them include, more or less explicitly, the idea that a certain pattern or event is repeated over time and that these patterns can be used to characterise the outcomes of highly heterogeneous sources, such as speakers. The place of suprasegmental cues in voice characterisation and the relation with other components and the conveyed information still needs to be fully understood.

2.3 Mode of vocal fold vibration

Throughout the years there has been little agreement on how to classify voice quality or how to transcribe it as part of phonetic transcription as [Ball et al., 1995] reports. The term is often restricted to aspects of vocal fold activity but it also involves airflow features or those derived from supralaryngeal settings of the articulators. [Nolan, 1983] has used *long-term quality* as an alternative to refer to the mode of vocal fold vibration. In this section, we focus on studies that have investigated the contribution of this limited aspect of voice quality as a component for speakers' differences. We rapidly present some defining terms and studies in order to fixate basic concepts of voice quality.

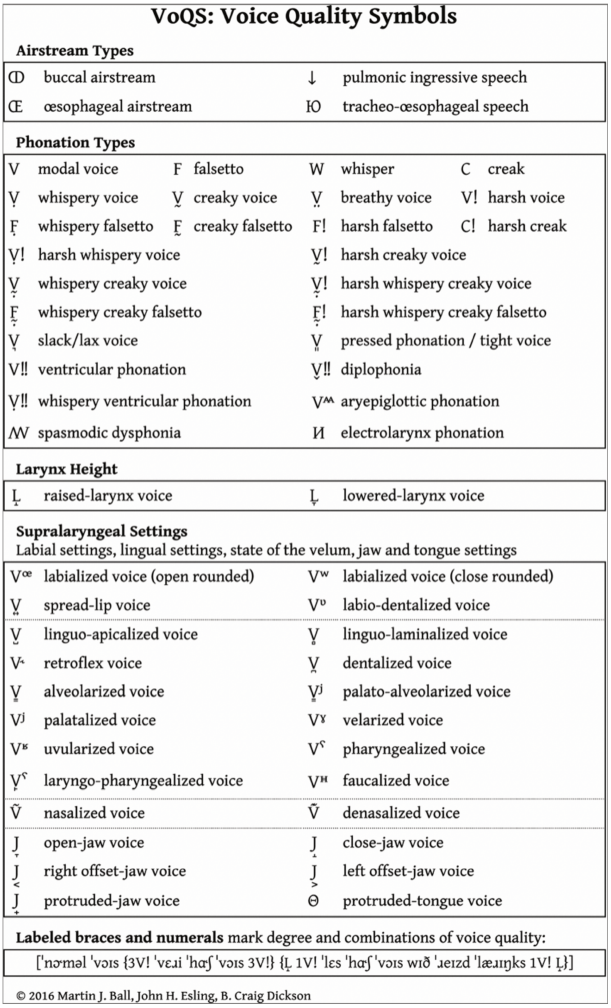
Voice quality aspects have been defined as 'light' and 'guttural' for normal speakers or 'hoarse' and 'strangled' for those with voice disorders as [Laver, 1980; Hewlett and Cohen, 1993; Esling, 1994] point out. These, together with other works in phonetics, have provided phoneticians with a widely accepted taxonomy of phonatory and articulatory settings which can be used to describe voice quality.

Figure 2.3 shows the revised version of voice quality terms and relative transcription proposed by [Ball et al., 2018]. As mentioned, it is important to note the meaning of the terms related to voice quality, as well as the use that is made of these aspects. Considering phonation types, the use of whisper does not imply that every segment is uttered using a whispered phonation. This means that, normally, expected voiced sounds are made with whisper, but expected voiceless sounds still remain voiceless. Hence, the maintenance of phonological contrast between voiced and voiceless sounds. The same would apply for most of the phonatory voice quality, [Ball et al., 1995, 2018] provides a full review.

Another important fact that has to be noted is that different voice quality processes could happen in different ways throughout an utterance. A nasalised voice does not mean that all sounds are uttered with a lowered velum. However, speakers can show a perceptually greater than normal use of nasal and nasalised articulations due to their own speaking habits or due to a cold. Similarly, defining an open-jaw voice could seem trivial, as in speech production the opening-closing movement of the jaw is more than common. Although, a more-than-normal open-jaw voice can be perceptually relevant in order to define speakers' identities from their voice. In another way, voice qualities may be discontinuous such that they can occur within certain frequency ranges, e. g., creak, or for paralinguistic reasons, e. g., whisper to mark confidentiality.

Similarly to what has been said for the prosody components, in order to define the voice quality, there is not just a singular phonetic measurement. The vast majority of these measurements are derived from studies on voice disorders and have mainly been used to assess the degree of a certain characteristic in the studied voices. In [Yumoto and Gould, 1982], sustained vowel /a/ from 41 samples with varying degrees of hoarseness and 42 normal voices are analysed through the comparison of expert evaluations on the hoarseness and HNR computations. Results show highly significant agreements, 0.85, between the expert responses and HNR values. In a similar way [Jotz et al., 2002] assessed the effectiveness of HNR values to quantify dysphonia and/or structural lesions of the vocal fold on 55 Brazilian boys of which 30 were dysphonic (3 were classified into the grade category, 5 into breathiness, 9 into roughness, and 15 into grade/breathiness). Vocal fold lesions were observed in 25 boys. The statistical analysis of the results revealed that HNR was significantly higher in boys with a structural lesion and those with dysphonia.

However, a logistic regression model showed that only the occurrence of dysphonia was explained by HNR. Nevertheless, Authors suggest that studies with larger samples are required to confirm the findings and to establish cutoff points of normality in different populations, since they obtained high false positive and false negative probabilities for HNR in these results.



VoQS: Voice Quality Symbols

Airstream Types

ʘ	buccal airstream	↓	pulmonic ingressive speech
Œ	œsophageal airstream	ʞO	tracheo-œsophageal speech

Phonation Types

V	modal voice	F	falsetto	W	whisper	C	creak
V̤	whispery voice	V̰	creaky voice	V̥	breathy voice	V!	harsh voice
F̤	whispery falsetto	F̰	creaky falsetto	F!	harsh falsetto	C!	harsh creak
V̤!	harsh whispery voice			V̰!	harsh creaky voice		
V̰̤	whispery creaky voice			V̰̤!	harsh whispery creaky voice		
F̰̤	whispery creaky falsetto			F̰̤!	harsh whispery creaky falsetto		
V̰	slack/lax voice			V̿	pressed phonation / tight voice		
V‼	ventricular phonation			V̰‼	diplophonia		
V̤‼	whispery ventricular phonation			Vᴬᴬ	aryepiglottic phonation		
ʌW	spasmodic dysphonia			И	electrolarynx phonation		

Larynx Height

V̝	raised-larynx voice	V̞	lowered-larynx voice

Supralaryngeal Settings
Labial settings, lingual settings, state of the velum, jaw and tongue settings

Vᵒᵉ	labialized voice (open rounded)	Vʷ	labialized voice (close rounded)
V̫	spread-lip voice	Vᵛ	labio-dentalized voice
V̻	linguo-apicalized voice	V̺	linguo-laminalized voice
Vˢ	retroflex voice	V̪	dentalized voice
V̼	alveolarized voice	V̟ʲ	palato-alveolarized voice
Vʲ	palatalized voice	Vˠ	velarized voice
Vᴷ	uvularized voice	Vˤ	pharyngealized voice
V̰ˤ	laryngo-pharyngealized voice	Vᴴ	faucalized voice
Ṽ	nasalized voice	Ṽ	denasalized voice
J̙	open-jaw voice	J̘	close-jaw voice
J̹	right offset-jaw voice	J̜	left offset-jaw voice
J̩	protruded-jaw voice	Θ	protruded-tongue voice

Labeled braces and numerals mark degree and combinations of voice quality:

['nɔ˞məl 'vɔɪs {3V! 'vɛ˞i 'hɑʃ 'vɔɪs 3V!} {V̰ 1V! 'lɛs 'hɑʃ 'vɔɪs wɪð 'ɹeɪzd 'læɹɪŋks 1V! V̝}]

Figure 2.3: Voice Quality Symbols chart, revised in [Ball et al., 1995].

Other measurements derived from fundamental frequency such as Jitter (the difference between consecutive periods), and Shimmer (difference between the amplitudes of consecutive periods) have shown correlation with voice disorders [Teixeira et al., 2013]. [Farrús et al., 2007; Kahn, 2011] used these measurements to differentiate between multiple speaker voices with encouraging results when combined with spectral and prosodic parameters. [Vaňková and Skarnitzl, 2014] compared between-speaker variability of formants and short-term acoustic correlates of voice quality, i. e., H1-H2, H2-H4, H1-A1, H1-A2 and H1-A3 in spontaneous and read speech by 6 Czech male speakers. The results from a LDA showed that voice quality parameters perform slightly better than formants with H1-A2 as the most reliable parameter, followed by H1-H2 and H1-A1. Authors point out how some speakers, the youngest and oldest ones, are clearly differentiated from all the others using voice quality cues and confirm the already assessed age dependency of H1-H2 and H1-A3.

The Long-Term Average Spectrum (LTAS) has been correlated as another index of dys-

phonia that has been studied in numerous variability perspectives showing a high adaptability to different factors [Harmegnies, 1992; Bele, 2006; Pettirossi et al., 2017]. It has shown language specific statistically significant differences in [Byrne and al., 1996] where read speech from 16 languages were compared. In the same study, consequent differences are shown by LTAS between female and male speakers in low-frequencies measurements. In addition, significant sex-dependent results are shown by [Mendoza et al., 1996] for 50 speakers of Spanish and for 60 Korean speakers in [Yoo et al., 2019]. CPP has been reported as an index of breathiness, [Fraile and Godino-Llorente, 2014; Keating et al., 2017], as well as a reliable between-speaker variability measure. Its robustness has been associated, to a significant extent, with the fact that there is an inverse relation between the amplitude of the first cepstral peak and the variance of amplitude, frequency and noise perturbations of the voice signal.

As shown throughout the decades, a wide set of measurements and methods has been correlated with the different cues that are identified as the voice quality. [Gold, 2014] reports the results of an international survey of forensic speaker comparison practices, which shows that voice quality ranks first among the parameters that experts have identified as most helpful for discriminating between speakers.

However, the Author discusses how a lack of consensus on the methods employed in Forensic Speaker Comparisons is present. The experts indicated that despite some individual parameters being efficient speaker discriminants, it is the combination of parameters that holds the most discriminant power. Although, the discriminant potential of phonetic measurements in combination rather than on their own is not often addressed in the forensic literature. A further description of this domain is presented in the next chapter, Section 3.1.

In a forensic voice comparison study based on voice quality, [Hughes et al., 2019], multiple voice quality cues are taken in a semi-automatic analysis in order to assess their efficacy as speaker discriminants from vowel recordings in both studio and telephone conditions. The subjects are 97 male speakers of standard southern British English. The CPP, harmonics-to-noise ratios (HNR) and a range of spectral tilt measures (H1-A1, H1-A2, H1-A3, H1-H2, H2-H4) are extracted and studied alongside f0 and MFCC. Using a standard likelihood ratio (LLR) evaluation, voice quality showed promising results with the lowest EERs at 5.8 %, outperforming formants on the same vowel-only material and an important complementarity with MFCC. However, as expected, performances were consequently degraded in the telephone recordings. Indeed, the results reveal that considerable speaker-specific information is captured by acoustic measures of laryngeal voice quality extracted from vowels under optimal conditions. In conclusion, the Authors point out that some speakers appear to be easier to characterise via voice quality than others. This means that overall performance, and the potential improvements due to these cues, are highly dependent on the studied speakers calling for caution when generalising about the speaker-discriminatory power of features or combinations of features based on overall system performances.

Phonetic measurements like HNR or the CPP have shown relations respectively to the amount of additive noise in the speech signal and the periodicity perturbations, either in amplitude, frequency or noise. Indeed, HNR has been correlated with hoarseness in voice, [Yumoto and Gould, 1982; Jotz et al., 2002; Teixeira et al., 2013], while CPP has been correlated with breathiness, [Fraile and Godino-Llorente, 2014; Keating and

Kreiman, 2016]. In addition, the SHR has also been correlated with a type of voice quality, namely, creaky voice [Keating and Kreiman, 2016; Keating et al., 2017]. Eventually these and other phonetic measurements (both correlated or completely independent from one another, such as spectral measurements or energy distribution) have been part of studies on individual characteristics through voice quality. Understanding the different strategies by which speakers modulate their voice quality during speech production is an important addition to the set of components that define the speaker matrix containing the possible variations that define their phonetic identity.

2.4 Articulatory characteristics

The last section of this phonetic literature review relates to characteristics that have not been directly investigated in this thesis. However, they are worth citing due to the coverage they have had in phonetics and their importance when considering between-speaker variability. The embodiment of speech is the basic idea of phonetic characterisation. The study of individual strategies of articulators that are used to produce intelligible speech is a question of high importance from both the acoustic and the articulatory perspective. The former can be defined as the investigation of the outcome of speech production, while the latter focuses on the manner in which this outcome is produced. We discuss hereafter some works that refer to the previously cited articulatory components: **articulation** and **articulatory setting**, see Figure 1.1. They are defined as "the overall arrangement and manoeuvring of the speech organs necessary for the facile accomplishment of natural utterance" by [Honikman, 1964].

It cannot be argued that some of the previously discussed measurements are correlated with the study of speech articulation, e. g., as the correlation between formants and lips protrusion or jaw opening. Many works have managed to integrate multiple articulatory information by the means of acoustic measurements. [Henrich et al., 2004; Bernardoni et al., 2005] show the importance of taking into account the laryngeal mechanism in which the vocal sound is produced through the open quotient, which is an indicator of laryngeal mechanism within a given voice production. In this study, spoken and sung speech (20 minutes recordings of different tasks by 18 classically trained male and female mostly professional singers) is analysed in order to explore the relationship between open quotient laryngeal mechanisms, vocal intensity and fundamental frequency. f0 and open quotient are derived from the differentiated electroglottographic signal, using the DECOM DEgg Correlation-based Open quotient Measurement method. The laryngeal mechanisms that are studied mainly derive from the different glottal configurations that male and female speakers may show: *mechanism 1* related to chest, modal, and male head register; *mechanism 2* related to falsetto for male and head register for female. The results show that open quotient depends on the laryngeal mechanisms, with lower values of open quotient found in mechanism 1, as compared to mechanism 2. The open quotient is also strongly related to vocal intensity in *mechanism 1* and to fundamental frequency in *mechanism 2*. The link between open quotient and vocal intensity mainly depends on the laryngeal mechanism. Yet, the Authors suggest that a combination of analysis, listening and measuring of open quotient and both other acoustical and EGG parameters is required to determine which laryngeal mechanism is being used.

[Kreiman and Shue, 2010] discusses the actual interest for the use of open quotient in

speakers characterisation. The increases in open quotient are assumed to cause changes in H1-H2, which in turn correspond to increases in perceived vocal breathiness. This study further examines these assumptions through the analysis of audio recordings and high-speed video images of the larynges of six phonetically, vocally healthy speakers who vary in terms of f0 and voice qualities quasi-orthogonally. Across speakers and voice qualities, open quotient, the asymmetry coefficient of the glottal area, and f0 account for an average of 74 % of the variance in H1-H2. However, analyses of individual speakers show large differences in the strategies used to produce the same intended voice qualities. Thus, H1-H2 can be predicted with good overall accuracy with its relationship to phonatory characteristics which are highly speaker dependent.

Therefore, studies, such as [French et al., 2015b], investigates individual articulatory differences by the means of vocal tract output measures: MFCC, long-term formant distributions (LTFD) and scores based on vocal profile analysis (VPA) (see Section–4.1.1) of long-term supra-laryngeal settings. Each was used to calculate a distance measure, quantifying the divergence between each pair of voice samples, and an identification score, in order to assess how well the features can classify pairs of samples. The results from spontaneous speech of 100 male speakers of Standard Southern British English show the highest error, of 12 %, for the identification score obtained from VPA features and a correlation between MFCC and LFTD while different information is carried by the third set of measurements. Therefore, the vocal tract itself provides a considerable amount of useful information to characterise individual voices. The fact that the two acoustic outputs provide similar information, while the auditory VPA provides different information for voice characterisation, suggests a potential complementary improvement in speaker discrimination power already evidenced by long term acoustic measures.

Another investigation on the complementary use and the potential for improvement of speaker/speech information representation of articulatory information and MFCC is provided by [Najnin and Banerjee, 2019]. In this case, MFCC are combined with articulatory features which the Authors present as articulatory cepstral coefficients (ACC), a representation of short-term power spectrum of an articulatory trajectory signal based on non-linear Mel-scale frequency. In order to assess the improved performance in speech recognition by adding a more in depth representation of speech production, these features are compared with classic MFCC, downsampled trajectories of different articulatory parameters from electromagnetic articulograph (EMA) and vocal tract variables (TV), representing gestures which specify the shape of the vocal tract. A total of 920 utterances from the TIMIT database, read by a male and a female native British English speakers as well as 1263 utterances from a native British English speaker in a single session (MNGU0 data set). The data used to perform phoneme classification and to evaluate the proposed features correspond to respectively 44 and 50 different phonemes. The results show that parallel use of MFCC and ACC improves recognition accuracy as compared to MFCC alone or combined with the other articulatory features. This also suggests as well that speech recognition improves significantly when classifiers are trained with a combination of acoustic and articulatory features. The same high results are obtained with cross-speaker classification, where the speech model was trained on a data set and tested on the other one, showing decreasing phoneme error rates as well as a higher robustness to noise than for the other features combinations.

Naturally, researchers that have focused primarily on articulatory investigations, and the relationships between articulators and the carried information, are vastly present in the

phonetic literature. In this sense, [Gick et al., 2013] presents a broad description of articulatory Phonetics applications and of the functional anatomy of the speech production apparatus. The Authors give an extensive overview on both the interactions between the human nervous system, respiration and the larynx as well as the actual implication of articulators in the production of specific sounds. The key concept for articulatory studies is that every speech sound results from body movements and that the movements of each articulator can be traced back and analysed in order to reconstruct the many aspects of speech sound production and perception.

Another aspect that makes articulatory studies interesting for our review, apart from the idea of embodied characteristics in speakers' productions, is the interest that these works have in representing the dynamics of articulators during speech production. As we discussed in the previous chapter, speech is a constantly changing object and the search for a good representation of these continuous changes in relation with the carried information and its influences is a fundamental challenge.

The influence of speaker-specific characteristics concerning vocal tract morphology and muscle anatomy is discussed in [Perrier and Winkler, 2015] with regards to the vowel /i/ from a reference 3D biomechanical face model compared with two subjects. Specifically, the vowel choice is justified by it being implied in the activation of muscles most likely influenced by variations in head and neck morphology. In addition, a correct acoustic realisation of /i/ requires a precise position of the tongue along the palate highlighting the different speakers patterns. Findings show that the studied muscles are important for tongue position control in vowel production. In addition, it is confirmed that speaker-specific biomechanical properties can influence the level of accuracy required for the production of given sounds. The Authors also point out findings from coarticulation strategies, highlighting that they are of great importance in order to determine speaker-specific gestures during production. Understanding coarticulation, and in general sequential organisation of gestures that speakers accomplish in order to optimise the effort for attaining the desired acoustic outcomes, is more than important to explain the underlying mechanisms. Hence, studies show that between-speaker variability is highly influenced by biomechanical factors. However, these explain only partially the existing interactions between the motor control underlying the production of the sounds and perceptual boundaries of these sounds.

In a similar way, regarding the mechanisms involving tongue muscles and vocal tract information, in [Winkler et al., 2011a,b] two biomechanical tongue models are created from a male and a female French speakers with different vocal tract morphologies regarding the relative location of the vocal tract bend and the typical morphologies found in female and male speakers. Differences concerning both the shape and steepness of the palate were considered taking results from a previous study, namely [Fuchs et al., 2008]. Analysis on these models show that, with a constant vocal tract length, the bending location appears to have an important role in the relationship between vertical and horizontal dimensions in the vocal tract and the respective degrees of freedom the tongue has within these dimensions. Results in this sense are found for back and low vowels while high front ones show a higher sensitivity to palate shape and the bend of the vocal tract location.

Findings on vocal tract geometries are reported by [Brunner et al., 2006; Fuchs et al., 2008]. These studies focus on the relationship between pharyngeal distance and smaller lower facial height, i. e., speakers with longer pharyngeal distance tend to have smaller

lower facial height and vice versa, and its influence on vowel production. Using MRI data from 9 French speakers, flesh points are defined on tongue surfaces for /ai/, /au/ and /ui/ and Euclidean distances are calculated. Results from PCA show that between-speakers tongue variability for /a/ looks similar. However, the pharynx mechanisms vary more consistently confirming the initial hypothesis that the pharynx length and facial height of speakers also has an influence on their articulatory vowel space. The speakers having long pharynx distance characteristics have smaller distances between high and low vowels, while a smaller articulatory distance between front and back vowels is shown for speakers with a short pharynx. In a similar way, [Serrurier et al., 2017] uses static MRI to quantify inter-speaker variability in vocal tract measurements of French isolated vowels and consonants from 11 French speakers, 6 male and 5 female. Average vocal tract contours coordinates, Sagittal Function (rescaled from two stable points, lip to glottis) and Phoneme-specific distributions are used in a PCA which results in the first two PCs respectively explaining 64 % and 24 % of the variance. These results confirm the assumption that speakers primarily aim towards auditory goals and adapt their articulation to the respective vocal tract boundaries.

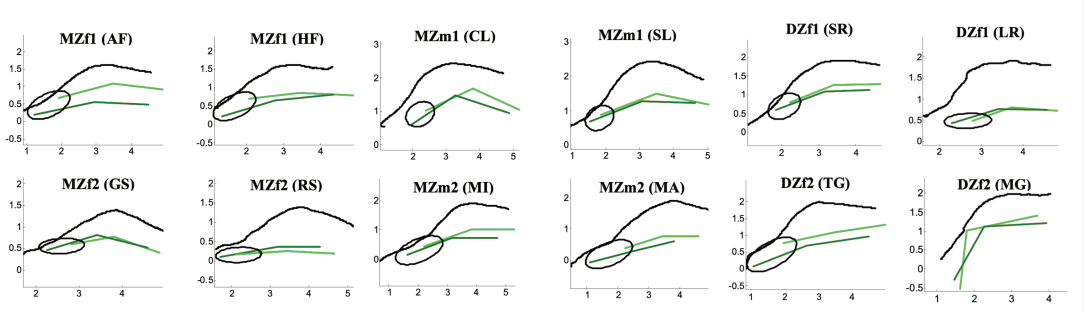


Figure 2.4: Figures 4 and 5 from [Weirich and Fuchs, 2013] combined. Mean articulatory target positions of /s/ (dark green) and /ʃ/ (light green). Different plots show different speakers. Ellipses visualise the amount of (horizontal and vertical) variation of the tongue tip. No ellipse is drawn for speaker DZf2 since she did not seem to use the tongue tip.

Articulation of consonantal contrast between /s/ and /ʃ/ is investigated by [Weirich and Fuchs, 2013] to explore the possible influence of morphological differences in contrast realisations. Results from twins and a heterogeneous group of speakers palate shapes are compared using Electromagnetic Articulography and Electropalatography. The results show that similar palate morphologies such as identical twins yield similar articulatory contrast realisations in what concerns vertical and horizontal distances of the target tongue tip positions. In addition, palatal height, anterior width of the palate, and palatal doming in the contrast region are important influence factors for the contrast realisation. [Fuchs and Toda, 2010] only considers the consonant /s/ for 24 equally distributed male and female English and German speakers. The comparison of palatal size parameters did not show consequent differences between the two sexes similarly to how it did when the language factor was considered. This highlights that other biological factors might come into play to explain the often reported male-female distinction, such as the length of the incisors. These factors may cause differences in the front cavity length and, consequently, in the acoustics. An additional report on male-female articulatory differences is presented by [Weirich, 2015] in which /i/, /a/, /u/ vowel spaces for German speakers show larger articulatory distances for male speakers, without difference in the acoustic distances between the vowels and the sexes.

This last section mostly served the purpose of completing the components description with the underlying idea that the limited number of freedom degrees available to each articulator results in recurring patterns of articulator activity and acoustic responses hypothetically determined by constraints affecting several levels of organisation of the speaker's sensorimotor system. The works we discussed show some of the interactions between physiological analyses and acoustic outcomes, obviously not exhausting the knowledge derived from this approach. Nevertheless, articulography is useful to demonstrate dialectal and language differences, e. g., in [Wieling et al., 2016] where speakers from two Dutch dialects are compared showing a structural difference in the position of the tongue. Indeed, more anterior positions of the tongue for the speakers from the southern half of the Netherlands against speakers from the northern half were found. The results may originate from a further back pronunciation of more segments for the southern dialect than the northern one, which present more frontal pronunciations. This contrasts previous findings on Dutch dialects that did not show differences in the F2 from monophthongs. However, it demonstrates that a dynamic approach using acoustic vowel information (F1 and F2) measured across multiple time points does help in uncovering regional differences, highlighting improvements in the use of articulatory data in studies investigating language variation.

This literature review chapter provides an overview of the components of speech that give rise to what we perceive as voice. The following chapters present various applications of voice characterisation, with a particular focus on on the approaches to which they are related.

Chapter 3

Beyond classical phonetics

The second chapter of this first part focuses on multiple domains that study speaker characteristics and their implication in between-speakers variability. Following the idea, stated in [Nolan, 2001; Bonastre et al., 2003; Morrison and Thompson, 2017], that there is no scientific process that allows to uniquely characterise a person's voice, speaker recognition has been linked to different approaches. The following four processes define various applications of voice recognition which influence the present work, given that our investigation is placed in a multidisciplinary perspective.

Aural (auditory) recognition is defined as the human ability to listen and recognise speakers based on their voice alone, with varying degrees of success. It is the basis of speaker recognition. As discussed in the next chapter, listeners present different levels of ability to recognise speakers, and various factors may affect the reliability of this method. Indeed, additional factors can increase the reliability of aural recognition such as familiarity with the speaker, duration or context of the audio sample, lack of background noise or vocal stress/disguise.

Spectrogram recognition involves, as the name suggests, the use of spectrographic representation to perform a speaker identification. The Voice Identification and Acoustic Analysis Subcommittee of the International Association for Identification provides methods to perform reliable and uniform spectrographic voice comparisons in a forensic investigation. Following the said standards, the known and unknown samples must contain spoken words that are comparable. An examination can only produce one of seven decisions which are 1-Identification, 2-Probable identification, 3-Possible identification, 4-Inconclusive, 5-Possible Elimination, 6-Probable Elimination, 7-Elimination. The decisions are based on aural and spectrographic comparisons of comparable words. Spectrograms have been commonly referred to as voiceprints, leading to an erroneous comparison with fingerprints. Spectrograms display speech signals in a $time \times frequency \times intensity$ representation, where the horizontal and vertical axis represent time and frequency variations respectively, while colour depth represents the intensity. They constitute useful engineering and voice analysis tools. A printed spectrogram can be seen as the only reason for the use of the term voiceprint. Further discussion on the misconception of speech as a biometric is presented in the following section.

In **forensic phonetics speaker recognition**, the speaker's information is analysed through a linguistic approach, with the expert focusing on the extraction of specific traits

or characteristics. However, in contrast with the two cited approaches, it gives a numerical estimate to how likely the examined voice sample is to be produced by a particular speaker, following the so-called likelihood ratio from the Bayesian theory. Another definition could be to estimate how many times more likely it is to observe the differences between the samples assuming that they have come from the same, rather than different, speakers [Rose, 2002].

Finally, we have the **Automatic speaker recognition (ASR)**. Its basic idea is to perform the recognition via a machine and use automatic procedures. State-of-the-art ASR techniques rely on similarity measures across acoustic parameters extracted from a set of recordings. These measures can take into account the statistical distributions for a particular speaker, the content of the message and/or information about the environment and recording medium. However, the use of a fully automated system can be challenging and is a constraint that this progression in the field is attempting to overcome in order to provide reliable responses. Furthermore, the last two recognition types are the object of discussion in the following sections.

Section 3.1 of this chapter starts with considerations about the forensic applications of Phonetics, both the methods and the studied components are presented. Some issues that have emerged throughout the years concerning the importance of strict methodology regarding the forensic application are as well discussed. Therefore, in this section, we first introduce the forensic Phonetics domain and some of its main issues, and then present some methods that characterise the forensic approach studies. Thirdly, we cover results from some studies that we considered important to mention in perspective of what is presented in Chapter 6.

In a similar way, Section 3.2 introduces the ASR approach, focusing on the basics of this domain and what the different recognition tasks provide in terms of interpretation of the between-speakers variability. Some works that cross the borders between phonetic and automatic approaches are presented as well as the National Institute of Standards and Technologies (NIST) campaigns that aim to provide a standardised evaluation of ASR systems around the world.

The last section of this Chapter focuses on the use of Artificial Neural Networks (ANN) and more specifically CNN as they cover a large part of our investigations. ANN are the core of the multidisciplinary approaches that this chapter aims to present as they are used in a continuously increasing amount of studies in various scientific domains.

As for what we discussed in the previous chapter, a summary of this chapter's review is presented in Section 4.4 analysing which aspects of the cited studies are used in this thesis. The idea of sharing methods between multiple domains in order to achieve similar objectives is fundamental in this thesis. In this sense, the present chapter provides the basis for the different approaches considered as being part of our investigation on speakers' characterisation.

3.1 Forensic speaker comparison

Even though the forensic application is not the main focus of the present work, given the strong weight of this domain on the VOXCRIM project, it is important to integrate an

overview of this approach and ideas that have contributed to our investigations. In this sense, the aim of this section is not to provide an extensive review on the forensic Phonetics domain, since more in-depth analysis have been proposed, e. g. in [Boë, 2000; Rose, 2002; Bonastre et al., 2003; Morrison, 2010; Gold, 2014], to which we refer throughout this section.

For more than thirty years in France the major actors in FVC have been the Association Francophone de la Communication Parlée (AFCP) and the Société Française d'Acoustique (SFA). These groups continue to reaffirm the idea of experts' competence, and reiterate the need to be cautious when using voice in a forensic context, e. g. [Boë et al., 1999; Boë and Bonastre, 2012]. In the courts no experts are registered as specialists in voice identification, rather experts in acoustics and vibration are recruited for legal identifications. Hence, "because of ethical concerns, it is incumbent upon any specialist to demonstrate his or her competence in speaker identification before assuming the authority of or operating as an expert"[1]. In addition to these concerns about the competence of the experts, another important aspect of the FVC reflection is the actual scientific knowledge and technological development in the field of speaker recognition. The main conclusion regarding the latter aspect (since the first investigations by both Anglo-American and French researchers, e. g. [Bolt et al., 1970; Nolan, 1983; Boë, 2000]) is the idea that despite the technological improvements and solutions for voice characterisation, at the present time there is still no scientific process that enables us to identify with absolute certainty an individual from his or her voice.

A common issue discussed by authors in the forensic domain is the difference between the biometrics, such as fingerprints or iris recognition, and voice data. The first do not show variations over the course of time and are impossible to modify, they have a significant reliability. The probability of fingerprint characteristics confusion between two individuals and the risk of a false alarm can be simply evaluated. In the case of voice data, as discussed hereinabove, we have to study an inherently complex object, for which a large part of its complexity lies in its relationship with its owner. The comparison between voices needs to be considered through what is actually being compared. When modelling the information from a voice it is important to consider how its different components relate and interact. Knowing how to interpret the ubiquitous variation in speech is necessary in order to assess the comparability of speech samples.

Voices convey a large set of information of potential forensic importance in addition to the linguistic message, e. g. the speaker's sex, emotional or health state [Bonastre et al., 2015a]. However, all of this is combined in the same channel, erroneously referred to as "voiceprint" or "voice signature" by the general public, in relation to the cited biometrics. Still, speech recordings are indirect measurements of articulatory movements rather than a trace left on a surface by contact with part of an individual's body, or a direct sample taken from it.

The speech organs give rise to instantaneous acoustic pressure variations. The parameters used to describe these variations depend on the speed of the articulation, on the pitch and loudness of voice and even on the psychological state of the speaker and the environmental stress. Within-speaker variability is intrinsically linked to the process of speech production and is another recurrent concern for the models used in forensic in-

[1]Motion adopted the 7th September 1990 by the Board of the *Groupe Communication Parlée* of the *Société Française d'Acoustique*.

vestigations of voice. The effect of parameters describing transmission, recording and the possibility that multiple voices or noises may have been superimposed must be added to these factors, as described in [Nolan, 2001]. When a speaker has been recorded over a telephone line, the characteristics of the recording device, i. e. the microphone and the telephone line, need to be considered. However, experts do not always have access to all of this data.

The possibility of imitation or disguise is another common discussion in forensic issues. Voice disguise is considered as the possible use of a whole range of techniques to distort the voice, ranging from simple spectral equalisation to the use of a vocoder or more recent techniques such as voice morphing. The resulting message may be recorded by its author under conditions that are inaccessible to the investigators. In certain inquiries, as vastly discussed in [Eriksson, 2010], the only evidence available may consist of a recording of a telephone conversation made by the suspect. Reliable speaker identification techniques may encounter an understandable interest in tackling voice disguise. In the review made by [Eriksson, 2010], we see how various types of disguise may affect both the recognition of a speaker's voice and discrimination between unfamiliar speakers. In a similar way, variations that are not artificially made are considered, namely the use of foreign language, dialect or accent that may as well influence recognition and discrimination. When voice disguise is used, we have to consider that, in most cases, the purpose is to conceal identity. If imitation is used as a disguise, it is necessary to separate how successful the actual imitation is from how well it serves the purpose of obscuring the personal characteristics of the imitator. Furthermore, the perception of a speaker's identity depends on a vast range of both naturally occurring factors like dialect and familiarity with the spoken language as well as different ways of disguising one's identity via manipulations, for example imitation and impersonation.

As considered throughout all this thesis, the large number of possible variables that intervene in the study of speech implies a consequent difficulty in experimental assessment. Moreover, gathering experimental confirmation from a sample database with specific conditions does not allow scientists and experts to generalise the results to other conditions. The researchers have to continuously propose new hypotheses and change experiments' conditions. The lack of commonly accepted methods in forensic speaker recognition is linked to this variability, and to the involvement of multiple scientific areas that focus on the different aspects of the speech object.

A typical forensic speakers comparison analysis involves at least two recordings. The first one is the criminal sample, also referred to as unknown, disputed, trace, or questioned sample [Gold, 2014], and contains the speech of an unknown individual. Other sounds can be present in this recording, associated with the crime taking place. The second recording is the suspect sample, also called known or reference, and is usually a recording of a police interview [Rose, 2002]. In the field of forensic speaker recognition, experts use a variety of methods, e. g. acoustic analysis, auditory analysis, acoustic and auditory analysis, fully ASR, or human-assisted ASR (HASR). However, a Bayesian framework is commonly used. It allows the computation of the probability of obtaining the evidence under the hypothesis that the samples came from the same person, versus the probability of obtaining the evidence under the hypothesis that two different speakers produced the criminal and suspect samples.

With respect to forensic science, the Bayes' theorem has been discussed in already cited

works such as [Rose, 2002; Kahn, 2011; Gold, 2014]. It is defined by three terms: the posterior odds, the prior odds, and the log likelihood ratio (LLR). In order to update one's beliefs, the theorem suggests that the posterior odds is equal to the prior odds multiplied by the LLR, as shown by the following equation:

$$\frac{p(H_p|E)}{p(H_d|E)} = \frac{p(H_p)}{p(H_d)} \times \frac{p(E|H_p)}{p(E|H_d)}$$

$$PosteriorOdds = PriorOdds \times LikelihoodRatio$$

Where H_p and H_d are respectively the hypothesis from the prosecution (the criminal and suspect are the same person) and defence (the criminal and suspect are not the same person), while p is the probability. The E term represents the evidence in question. Hence, each of the three terms are represented by a probability: the posterior odds is the probability of the prosecution hypothesis being correct given the evidence divided by the probability of the defence hypothesis being correct given the evidence; the prior odds corresponds to the probability of the prosecution hypothesis being correct divided by the probability of the defence hypothesis being correct; the LLR is the probability of obtaining the evidence given the prosecution hypothesis divided by the probability of obtaining the evidence given the defence hypothesis.

Numerically, the posterior odds are the multiplication of the prior odds by the LLR. However, deriving the posterior and prior odds is the responsibility of the trier(s) of fact in combination with the evidence provided by expert testimony. The prior odds represent any existing probability of the hypothesis being true prior to the consideration of new evidence being introduced. In practice, it can be problematic, because even small changes can result in significant modifications of the posterior odds. The third term of the Bayesian equation, the LLR, is a measure of the strength of the evidence [Rose, 2002; Bonastre et al., 2015b]. It is the only portion of the Bayesian framework in which a forensic expert should provide an opinion obtained from two competing probabilities. The numerator of the LLR is the probability of the prosecutor's hypothesis, while the denominator represents the probability of the defence one. In the LLR equation, when the numerator is higher than the denominator, the prosecution hypothesis is supported, while with a higher denominator, the defence is supported. The distance of the resulting LLR from 1 corresponds to the strength of evidence. It is usually converted to a logarithmic value, therefore, a positive value indicates support for the prosecution hypothesis, while a negative value indicates support for the defence.

Neither the LLR nor the prior odds on their own constitute a Bayesian probability, rather, it is the value of the posterior odds that equates to a Bayesian belief of probability. The expert opinion is presented either numerically or verbally. Various methodologies in order to obtain the results are present in the literature, in [Aitken and Lucy, 2004] different methods are studied and compared. Three main applications stand out in this comparison: the use of an LLR algorithm that can handle correlation through statistical weightings, e. g. the Multivariate Kernel Density (MKD) algorithm; Bayesian networking that accounts for correlations by considering feature distributions and variances and perform statistical weightings; and a solution proposed in the field of ASR referred to as logistic-regression fusion, which accounts for correlations in the resulting LRs and then applies statistical weightings.

Finally, a derived measure of the LLR commonly used in automatic approaches to forensic investigations is represented by the Log Likelihood Ratio cost validity metric (C_{llr}), discussed e. g. in [Morrison and Enzinger, 2016]. This is a single value summary of system performance. If the test pair is a same-speaker pair the higher the likelihood ratio the better the performance in range $[0; 1]$.

3.1.1 Results from the forensic literature

After discussing the theoretical background for forensic phonetics, hereafter we focus on some results from related studies. As noted previously, we observe that there is not just a single acoustic parameter on which the literature focuses in order to perform speakers comparison. In Table 3.1 below are listed the five most discriminant voice parameters resulting from an international forensic survey from [Gold, 2014].

1. Voice quality
2. Dialect/accent variants and vowel formants
3. Speaking tempo and fundamental frequency
4. Rhythm
5. Lexical/grammatical choices, vowel and consonant realisations, phonological processes, and fluency

Table 3.1: Top five most discriminant phonetic parameters resulting from an international survey on Forensic Speaker Comparison reported by [Gold, 2014].

Despite voice quality being considered the most discriminant parameter by forensic experts, it is less present in the literature with source and filter characteristics taking the advantage. Studies such as [Gold, 2014; Hughes et al., 2019] take into account voice quality in the forensic domain. As said in both the previous and the present chapters, they analyse multiple voice quality measurements, e. g. H1-A1, H1-A2, H1-A3, H1-H2, H2-H4. These present a strong influence by the recording conditions, which is one of the main issues affecting the forensic domain. In the same perspective, [Klug et al., 2019] compares results from CPP and HNR with the latter not resulting in significant scores and being highly influenced by the transmission channel. In the same way, [French et al., 2015a; Hughes et al., 2017; san Segundo et al., 2019] develop the idea of performing a phonetic characterisation through a vocal profile which regroups a large set of speech components the speaker has to be characterised from. The conclusion from all these studies is that voice quality cues have promising results for forensic speaker comparison, as they are capable of capturing specific traits from a speaker that are absent from another. However, since some voice quality characteristics could be completely absent in certain speakers they remain difficult to use as isolated characteristics and push researchers to increasingly adopt multiparametric approaches.

Voice quality being a recent application for forensic studies, we found that the majority of the literature from speaker comparison focuses on the source and filter's data. Both f0 and formants are examined in long-term perspectives [Asadi et al., 2018], multiple

measurements extracted throughout the analysed token. The linguistic material varies from isolated vowels and diphthongs to common whole words in order to add knowledge about a shared material to perform the comparison. Considering the high discriminant power of f0 discussed in the previous chapter, we highlight here its use in forensic studies for the analysis of voice disguise. As mentioned above in [Eriksson, 2010], f0 maintains a speaker's characteristic even when the speaker performs a voice disguise. Similar results are shown by [Zhang and Tan, 2008] on Chinese read speech. 40 male speakers were tested in a voice comparison study where they were asked to record using 10 types of disguised voice. The results show that only 5 speakers were highly misclassified, having accuracy scores under 0.50 while half of the examined speakers showed perfect scores, demonstrating the important stability of f0 even in a situation of voice disguise.

Concordant results are shown by [Künzel et al., 2004] on natural and disguised speech data from 100 German speakers recorded 5 times over a period of 7 to 9 months. Experiments are limited to estimate the performance degradation when the suspect is known to be the author of the disguised test speech. The results indicate that the three types of voice disguise, i. e. increased voice pitch (even falsetto speech), lowered voice pitch and pinching the nose while speaking, do not significantly affect the performance of the system.

The issue of voice disguise is also examined through rhythmic measurements, e. g. in [Leemann and Kolly, 2015] where dialect imitation is used to study the between- and within-speakers variation in the suprasegmental plane. Read speech from 16 German speakers, half Zurich German and Bern German is used with 3 females and 5 males per dialect. The examined measurements correspond to those presented in the previous chapter Section 2.2.1. The results reveal an important between-speakers variability while maintaining a low within-speaker variability across the disguise condition. The rate-normalised average differences between consecutive voiced interval duration ($\mathrm{nPVI_{VO}}$) shows a perfect invariance regardless of whether the speech is disguised or not. This conclusion shows complementarity between temporal and frequency domain cues for speakers characteristics. However, it indicates that speakers who performed well at imitating another dialect may have been mistaken for native speakers of that dialect because they were able to approximate the target's suprasegmental temporal features. These findings remain based on a highly controlled set of data.

The same set of temporal measurements is studied in a purely voice comparison study in [Leemann et al., 2014] where evidence is provided to assess the speaker-specific nature of suprasegmental temporal characteristics. The percentage over which speech is vocalic (% V) and voiced (% VO) revealed the strongest effects of the speaker, with % V and % VO robust to speaking style variability as well. % VO shows high robustness to channel variability. Even though % V and % VO perform well in explaining suprasegmental temporal variability between speakers the authors state that there is not yet an adequate understanding of what exactly governs variability in these two timing parameters. On one hand, speaker-specific temporal patterns may be a result of a speaker's anatomical configurations, governed by neurological motor patterns in the brain of the speaker. On the other hand, they may be caused by the speakers' acquired idiolectal way of articulation. A way to understand the role of speaker anatomy and speaker idiolect in the variability of speech temporal features is suggested in the study of twins. In particular, with identical twins, the influence of phonetic differences between the speech of two anatomically similar individuals can be examined as any difference between them normally reflects environmental effects.

In sum, this study shows relevant information for forensic cases in which there is a mismatch of speaking styles between trace and suspect material, as well as for cases where the speech signal is degraded by mobile phone transmission. Furthermore, articulation rate in read speech from 20 German speakers from both sexes is examined in [Künzel et al., 1995]. With recordings being made on telephone channels, the temporal information shows more consistent results than intensity or f0 in terms of between-speakers variability.

Formants represent the most commonly used features in forensic studies, especially through the modelling of dynamics using polynomial functions or discrete cosine transform approaches. Actual studies on twins are carried out using formants, e. g. in [san Segundo and Yang, 2019], where the two cited dynamics computations are compared as well. Vocalic sequences are taken from a corpus of 54 Spanish speakers who share different levels of familiarity, e. g. from 12 monozygotic twin pairs to 12 completely unrelated ones who are familiar to one another. The results show increasing performance with the increase of the number of coefficients for both curve computations with the highest C_{llr} at 0.15. In a similar way, making the comparison text-independent by adding vocalic sequences of different linguistic content increases the performance and demonstrates how the actual dynamic of formants remains despite the pronounced speech. However, monozygotic twin pairs deteriorate the system's performance, as well as the brother comparisons with a lower effect.

Similar sounding voices are also studied in [Rose, 1999], which investigates the nature of between- and within-speaker variation in the acoustic characteristics of the word "hello". Within-segment variation is determined with the repetition of the same word under different prosodic conditions. Six Australian males are compared with respect to f0 and patterns of the resonances below 5 kHz. The results demonstrate that between-speaker acoustic differences are not ubiquitous: 13 out of 15 pairs of speakers show significant results in the F2-F4 bandwidth. Other results on formants trajectories extracted from a whole word are discussed for the hexaphonic disyllabic Cantonese word *daihyat* (first) from 23 male speakers in [Rose and Wang, 2016]. Evaluation through C_{llr} shows best performance with lower order polynomials, with F2 requiring a cubic fit, and F1 and F3 quadratic. Fusion of formants and f0 trajectories results in considerable improvement over the individual features.

At a lower level of modelling, [Jialin and Rose, 2012] uses F2 and F3 trajectories from 18 male Cantonese speakers of the diphthong /ei/. The best C_{llr} is 0.46 with the use of both formants. The results shows a high influence of the reference population on the comparison. Further results from formants trajectories of vowels /i/, /e/ and /a/ from 64 male speakers of standard Chinese are presented in [Morrison et al., 2011]. The results show that vowel /i/ outperform its counterparts with a C_{llr} of 0.57, while including all three vowels in the comparison shows a score of 0.40. In a similar way, [Rose and Winter, 2010] investigates the first three formants trajectories from seven Australian English speakers using monophthongs extracted from words spoken in a stressed environment, showing errors below 1 %.

3.2 Automatic Speaker Recognition

ASR is the closest domain to phonetic characterisation in our analysis. The training phase of an automatic system can be compared to the within-speaker variability description while the test phase, which gives probabilities of an audio file belonging to one speaker or another, can be compared to between-speaker models. A leading technique in ASR are i-vectors (for intermediate representation vector), which are a set of features commonly modelled from MFCC and compared using a Gaussian Mixture Model based Universal Background Model (GMM-UBM) [Kahn, 2011; Bousquet et al., 2012; Sahidullah and Saha, 2012a]. Speaker characteristics statistical distributions are modelled by a descriptive approach during the train phase while the decision phase models are compared to a reference population in order to find the higher resemblance between tested speakers and their models.

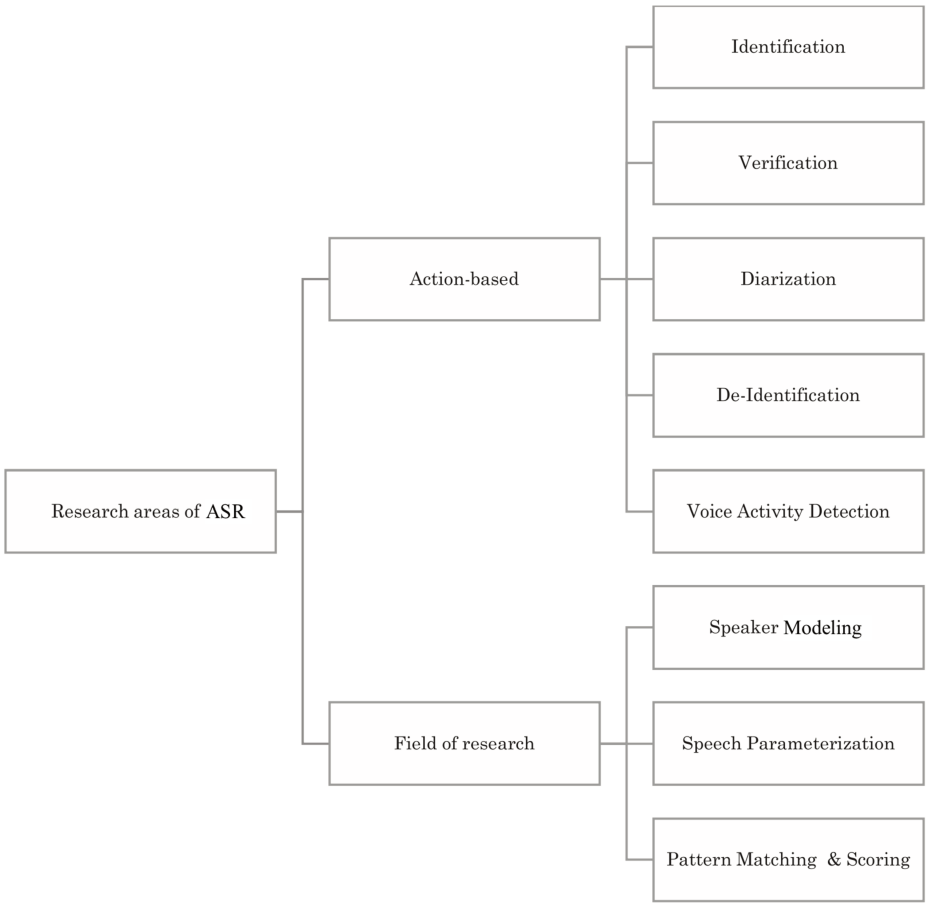


Figure 3.1: Figure 2 from [Tirumala et al., 2017]. Main areas of research for ASR.

The automatic speech processing literature traditionally distinguishes two main tasks or branches in ASR: identification (SID), where the identity of the speaker is chosen from a pool of $1 : N$ possible classifications; and verification (ASV), where the speaker's identity is tested in a binary classification, e. g. in order to confirm a claimed identity. These two tasks can both be performed via two approaches: text-dependent, where the spoken trace for the test is known or even proposed by the evaluator; and text-independent, in which the material for the test is completely open. Both identification and verification systems extract speakers' characteristics in order to assess the speaker's identity. In some cases

a system is asked not to find the identity of the speaker but to answer on whether the same speaker is talking in a chosen audio segment, this is the aim of diarisation tasks. Clustering of speakers with similar voice's characteristics is a derivation of the diarisation process that we consider in relation to perception studies in the next chapter. Figure 3.1 below shows a schematic summary of ASR tasks.

Similarly to the forensic overview in the previous section, the objective hereafter is not to give an extensive review of the ASR domain, but to point out its main constitutive aspects as more comprehensive reviews can be found e. g. in [Calliope, 1989; Bimbot et al., 2004; Kinnunen and Li, 2010]. As for many NLP problems, ASR systems working pipeline includes three stages: the selection of a linguistic input and its related extraction of features; a processing phase of the named parameters in order to create and compare speakers' models; and a decision making algorithm used in order to process the task's output.

During the enrolment or training phase, parameters extraction is performed on speech signals. A voice activity detection algorithm is usually applied in order to select the frames presenting the most energy from which the parameters are computed. Different techniques have been used in the literature, the most common representation of input parameters is through cepstral coefficients on a Mel frequency scale. After the calculation of MFCC and possibly centred and reduced, the vectors may also incorporate dynamic information, i. e. information about the way these vectors vary in time. This is classically done by using the Δ and $\Delta\Delta$ parameters, which are polynomial approximations of the first and second derivatives. Nevertheless, linear prediction approaches, i. e. Linear Prediction Coefficients (LPC), Perceptual Linear Prediction (PLP), etc., have been used as an alternative to MFCC. An extensive review on feature extraction techniques is given by [Tirumala et al., 2017].

Since [Reynolds, 1995; Reynolds et al., 2000], the GMM-UBM approach occupies a dominant position concerning the representation of the selected speakers in a D-dimensional features space. More specifically, the distribution of vectors extracted is modelled by a Gaussian mixture density, which is defined as a weighted linear combination of multiple unimodal Gaussian densities, each parameterised by a mean vector and a covariance matrix. Further computations have been integrated throughout the years in order to obtain the speakers' representations which take the name of i-vector [Dehak et al., 2009, 2011; Bousquet et al., 2012], while in recent applications of neural network based embeddings the name x-vectors have been used [Snyder et al., 2018].

In the case of a GMM-UBM system during the decision making phase a LLR computation is applied in order to compare the vectors and to give the similarity rates between speakers. The Hidden Markov Model (HMM) is the main alternative to the GMM-UBM approach for the representation of speaker characteristics. It consists of three steps: using the Viterbi algorithm by which the probability that the observed sequence is produced by the model is computed; the sequences capable of maximising this probability are selected; and a training on the model is performed aiming to adjust its parameters for maximising probability to create the best model for given training sequences. Other common ASR techniques used for the comparison between the test trace and the training models include the use of Cosine Distance Scoring (CDS), Probabilistic Linear Discriminant Analysis (PLDA) or Support Vector Machine (SVM). In the latter case, the decision is made by linearly separating the vectors through a hyperplane and afterwards calculating the

shortest distance between the test vector and those of the model.

[Reynolds, 1995] presents high performance speaker identification and verification systems based on GMM-UBM and LLR computation. The results are obtained from four databases from the English language: TIMIT, NTIMIT, Switchboard and YOHO. The first two present recordings of 10 read sentences from 630 speakers using either microphone or telephone channel. The Switchboard database provides telephone speech as well in the form of spontaneous conversations from 500 speakers collected under home/office acoustic conditions. The YOHO database was designed to support text-dependent speaker verification. It has a defined train/test scenario in which each of the 138 speakers is prompted to read a series of 24 combination-lock phrases. The recordings were collected in an office environment using a telephone handset.

The different levels of degradations and content variability found in this study allow for the examination of system performance for different task domains. Identification rates for the TIMIT and NTIMIT experiments are of 99 % and 61 %, respectively, showing the important influence of the recording support on the results. A reduced set of speakers from the Switchboard database shows an identification rate of 83 %. Concerning the verification results, expressed in EER, rates of 0.24 %, 7.19 %, 5.15 % and 0.51 % are obtained on the TIMIT, NTIMIT, Switchboard and YOHO databases, respectively.

NTIMIT and Switchboard performance shows similar trends, however the factors degrading the two databases are different. The first presents high noise levels as the main degradation, whereas hand-set variability and cross-channel echo are the major degradation factors in Switchboard. The results from the YOHO database show that very low error rates are possible for a secure access verification application, beside the overall EER of 0.51 % a false rejection rate of 0.65 % and a 0.1 % false acceptance rates are obtained. The constrained vocabulary along with the quality of the recorded speech allowed the model to focus on the speakers' characteristics without extraneous channel variabilities.

The verification experiments also demonstrated the need to select background speakers to cover the population of expected imposters. A mixed-sex experiment is also conducted, where a set of background speakers equally distributed among more or less similar speakers performs better than using only similar speakers as the background population. In the verification task, the system appears to perform better for male speakers than for female ones, the same trend is also observed in TIMIT/NTIMIT results. This study has established baseline results and discussion points for all the ASR domains. The major limiting factor is performance in transmission degradations, i. e. noise and microphone variability.

In a similar way, as mentioned above, [Bimbot et al., 2004] offers an overview of a state-of-the-art text-independent speaker verification system. Details about the training and test phases are discussed, from speech parameterisation through cepstral analysis to Gaussian mixture modelling, with modelling alternatives as well such as neural networks and SVM. Score normalisation is also explained, this represents an important step to deal with real-world data such as mismatch conditions between training and test. Through the various experiments achieved on the use of normalisation in speaker verification, the use of prior information like the handset type or speakers' sex during normalisation parameter computation is shown to be relevant to improve performance.

Alternative windowing techniques to compute MFCC are proposed in [Sahidullah and

Saha, 2012b] based on fundamental property of discrete time Fourier transform differentiation in frequency domain. The newly extracted features represent the power spectrum of the original spectrum as well as its derivative integrating phase information. 38 dimensional feature vectors are computed using 20 filters linearly spaced in Mel scale from speech frames of size 20 ms. The tested systems are shown to attain consistent performance improvements over baseline single tapered Hamming window. Speaker verification task is performed using GMM-UBM and SVM classifiers, speech data for the reference population and the test phase comes from different databases, SRE01, SRE03, SRE04, SRE06, commonly used in ASR domain and which we describe in the next section. An average of 400 utterances from more than 100 speakers for each sex are present in each of the sets, which gives a consequent evaluation database. Target models are created by adapting the means of the UBM with an established threshold's relevance factor. Speaker recognition experiments are carried out with different window functions keeping other blocks identical i. e. pre-processing, feature extraction and classification. During the score computation, the top 5 Gaussians of corresponding background models per each frame are considered.

There is a consistent performance improvement for proposed window based speaker verification for both GMM-UBM and SVM systems. In comparison with the baseline Hamming window based system, a relative decrease of EER between 0.6 % and 8 % is shown. It is also observed that performance of second order window based systems is better than first order window based systems. Furthermore, whispered speech from 24 speakers of Irish English is studied in [Mary and Yegnanarayana, 2008] using MFCC extracted on frames of size higher than 25 ms in order to capture long-term features. The results show an increase of EER around 7 % in a speaker verification task using a classic GMM-UBM but an improved robustness against noise and reverberation. In a similar way, using a SVM based system [Sahidullah and Saha, 2012a] investigate different MFCC computation showing a consistent decrease of EER, up to 22 % with a system capable of combining different representations. The linguistic material used in this study is made of 5 minutes conversation parts from the Fisher corpus, American English, divided in 3 subcorpus.

Mismatch of acoustic material in the form of whispered speech leading to increased number of identification errors is discussed as well in [Vestman et al., 2019]. The problem is approached via acoustic mismatch compensation from a feature extraction perspective. Whispered speech is intelligible, however it presents low-intensity signals, prone to extrinsic distortions. Read whispered speech from 36 English speakers, sex balanced, is used. The authors take advantage of long-term speech analysis methods that utilise slow articulatory movements in speech production through frequency domain linear prediction and time-varying linear prediction features. The experiment indicates that when tested in normal-whisper mismatched conditions, the proposed features improve speaker recognition performance by 7 to 10 % over standard MFCC in relative terms. On the other hand, performance degradation with normal-normal voice experiment is observed.

Moreover, the recognition performance for the whispered female voice is better than for the male voice, contradicting with the usual observation in speaker recognition experiments. The findings from speaker-by-speaker analysis performances show considerable accuracy differences across the speakers. This suggests that the articulatory process to produce whispered voice is highly influenced by both sex and speaker factors. Some speakers are naturally good at disguising themselves by producing close to unidentifiable whispered voices.

The influence of multiple speaking styles on speaker recognition is the object of study of works such as [Mao et al., 2020]. A corpus made of singing, humming and normal reading speech utterances from 20 male and 26 female Chinese speakers is used to test text-dependent and text-independent GMM-based and x-vector based speaker verification systems.

The results of the experiment show that the information present in humming and singing speech is more distinguishable than in normal reading speech for conventional ASV systems. Humming shows better speaker discriminant information than singing, itself being better than the normal reading. This study attempts to investigate how the multispeaking styles affect the automatic speaker verification. Furthermore, the cross-speaking style in training and test phases is also studied. Both GMM and x-vector are very vulnerable to deal with this issue as pointed out by other studies mentioned above. In addition, combining the three speaking styles significantly improves the x-vector systems' results, while no gains are shown by GMM-based systems.

Further investigations on the performance of ASR systems using different modelling and decision algorithms are present e. g. in [Nayana et al., 2017; Zhang, 2018; Jessen et al., 2019]. The latter compares two ASR systems, the first is an i-vector PLDA based on PLP parameters while the second system integrates DNN functionalities and MFCC. Both are tested in three population variants. The difference between the first and second variant lies in the number of speakers used as reference, 42 against 105. In the third variant, the background model is obtained through speakers drawn from data of a real forensic voice comparison case. In all three test phases speakers from the latter database are used.

In comparing the three variants, it was shown across the two systems that the inclusion of a background model that is dedicated to the conditions of the case leads to improved performance over the use of a default system. The difference in the size of the reference population, however, did not matter. Concerning the PLDA and DNN comparison, improved results are shown by the latter system.

The results from the PLDA system show that performance is better in Variant 2 than Variant 1, i. e. when the reference population includes fewer speakers, which is in contrast with what would be expected. However, the Authors observe that knowing that the reference population is saturated beyond 40 speakers is good news because of the difficulty to supply large amounts of case-relevant population data in forensic cases. A consistent improvement for both systems is observed in Variant 3 of the test. It determines that the use of a background model trained from case-relevant data improves performance relative to the system-internal default background model.

The differences of performance of the two versions of the ASR system show a clear pattern, as for each of the three variants, performance is improved with the integration of DNN. In a similar way, the comparison between GMM and DNN based systems is discussed in [Zhang, 2018] in an ASV test. The experiments conducted on SRE06 and SRE08 data sets, with different lengths of enrolment speech from 15 to 225 s, show that the DNN system outperforms its counterpart in most cases in terms of accuracy and discrimination metrics. In [Nayana et al., 2017] text independent ASR systems are implemented using GMM and i-vector methods with cepstral and PLP coefficients. The results show that accuracy for SID is improved when using the i-vector method with the PLDA classifier rather than with a cosine distance based classifier. Furthermore, the performance shows improvements with longer utterances. In this sense, a comprehensive overview on system

comparison and evaluation metrics is presented by [Poddar et al., 2019].

The authors attempt to incorporate supplementary information using the quality of the estimated model parameters. A class of quality measures formulated using the zero-order sufficient statistics is used during the i-vector extraction process. The proposed methods demonstrate considerable improvement in speaker recognition performance especially in short duration conditions. Similar improvements are observed over existing systems based on different duration-based quality measures.

However, even though there is a considerable improvement with the distance-based proposed quality measures, there is no consistent indication on which quality measure distance function is more appropriate. This opens up the possibility of further optimisation of distance measures for quality estimation. In this work, GMM-based i-vectors are considered, and similar investigations on quality measure can be made with DNN-based i-vector systems. Therefore, the authors state that it would be interesting to explore the general use case of the proposed quality measures where acoustic variability needs to be computed.

3.2.1 NIST Campaigns

The evaluation of i-vectors and ASR systems has been the main objective of annual campaigns from the NIST since 1996. Additional speech recognition domains are represented by Language Recognition Evaluation (LRE), Speaker Recognition Evaluation (SRE), Speaker Recognition for Biometrics, Forensics and for Investigatory purposes and HASR. Participation is open to both industry and academia in order "to drive the technology forward, to measure the state-of-the-art, and to find the most promising algorithmic approaches"[2]. The evaluation of automatic techniques and approaches is important to better represent the observed real-world. However, phonetic interpretation is fundamental in everything concerning speech data. In this section, we present results from studies that use NIST recommended data for speaker recognition, while in the next section we review studies that combine phonetic and automatic approaches.

In order to participate in the regular evaluation, participants must first complete the Conversational Telephone Speech Challenge. The evaluation rules for the last five years remain the same. Both the metrics and the linguistic material used in the NIST campaigns are established by the institution itself for each of the different campaign domains. This way, the participants have highly comparable results. A primary metric is defined by a basic cost model measure to quantify the speaker detection performance. It is defined as a weighted sum of false-reject (missed detection) and false-alarm error probabilities for a decision threshold. In addition to the primary metric, an alternative, the C_{llr}, may be used to measure how well all scores represent the likelihood ratio and that penalises for errors in score calibration.

Concerning the data, in the latest iteration of SRE21, a multimodal and multilingual corpus collected by the Linguistic Data Consortium and named WeCanTalk is used as both development and test sets. This corpus is composed of phone calls and video recordings collected outside North America, from speakers of Cantonese, English and Mandarin. Subjects made multiple calls to people in their social surroundings and recorded videos

[2]NIST SRE, from <https://sre.nist.gov>.

of themselves while performing monologues. The recorded speech is used as trials (target and non-target) with enrolment and test segments originating from different source types i. e., conversational telephone speech and audio from video, as well as cross-language trials. Similar to SRE19, a video-only track is included in the test data set as an optional additional submission. For the training sets, participants can choose both from past NIST campaigns training sets and other open access datasets.

Throughout the years, different techniques have contributed to the ASR state-of-the-art that the NIST campaigns helped to develop. One of the hallmarks of the NIST SREs has been to continuously develop the evaluations to address more challenging data and conditions as the underlying technology improves. The main issues involve channel and language mismatch with state-of-the-art systems increasing their capabilities to deal with shorter enrolment and test segments as well. The interplay between speaker characteristics and audio-visual (or multimodal) recognition is an additional challenge, as well as data confidentiality, which may continue to drive the field evaluations forward.

The works we discuss hereafter are selected from systems participating in past campaigns, and some that only used NIST SRE datasets as evaluation tools. These examples are intended to represent the standards both in terms of the score obtained and the methods used.

[Sadjadi et al., 2016] presents the IBM speaker recognition system for conversational speech for SRE16. The main advancement that contributes to this system includes the use of a nearest-neighbour discriminant analysis (NDA) approach, as opposed to LDA. The same opposition is analysed in [Khosravani et al., 2016] which shows very different results. In the former, Feature space Maximum Likelihood Linear Regression (fMLLR) features, commonly used for speech recognition, are adapted to perform channel and session compensation of i-vectors. The use of a DNN acoustic model is investigated in order to compute the frame-level alignments required in the i-vector estimation process. The results show that fMLLR features provided consistent improvements over MFCC. The DNN based UBM resulted in performance improvements with the lowest EER obtained of 0.59 %. However, error analysis of low scoring target trials reveals issues for this modelling technique in recordings presenting overlapping speech, background noise/music and signal clipping effects. Overall, NDA is more robust than LDA for multimodal data, hence more effective for channel compensation. In contrast, [Khosravani et al., 2016] presents a system based on MFCC and PLP features with a comparison between PLDA and the NDA approach. The fusion of MFCC and PLP shows a relative improvement of 10 % in terms of system accuracy metrics, compared to MFCC alone. In order to quantify the contribution of the different decision algorithms they are tested in multiple scenarios. LDA outperforms NDA in the case of PLP, however, in fusion based modelling NDA results show better performance. Overall both algorithms have better results with utterances of 10 s, which is shorter than the data set average duration of 26 s.

The primary objective of SRE16 was the development of robust speaker recognition technologies for new languages, which had much less training data than those commonly studied in the state-of-the-art of ASR. As one additional issue is the complete absence of meta-data from those languages, systems compensating for this particular evaluation are presented in [Rouvier et al., 2016; Shon and Ko, 2017] with similar investigations about decision algorithms and modelling features comparisons to those already mentioned.

[Shon and Ko, 2017] uses MFCC and a DNN based system with normalisation scores for

language and channel factors, which allows to obtain EERs between 13 % and 18 %. These results appear consistent with [Rouvier et al., 2016], where a series of systems based on the different parameter extraction and speaker modelling techniques described above are used as well as windowing and signal filtering alternatives, registering an EER of 14 % with the use of a hybrid system. The hybrid system presents both MFCC and PLP as modelling features with PLDA and DNN analysis.

The use of DNN has been increasing in recent years for the best performing systems, it is the case for e. g. [Zheng et al., 2020], the best scoring system for SRE20. It involves subsystems, including, e. g. Long-Short Term Memory (LSTM), Residual Networks (ResNet), and Visual Geometry Group (VGG) architectures. The scores are computed mainly by cosine distance and an adapted PLDA backend. For each system, calibration uses logistic regression on the SRE19 development data. The fusion score is an equal weighted average of the scores of the systems. The most important observation from these experiments is that domain adaptation of PLDA and score normalisation are not useful to reduce EER. Another interesting example from the latest SRE iteration is described in [Lee et al., 2020], which compares different systems from multiple academy participants.

The collaboration involves researchers from eight research teams across Europe and Asia. The submission is based on the fusion of top performing subsystems and subfusion systems provided by individual teams. A particular effort is made on the use of common development and validation datasets, in order to minimise inconsistencies in trial list and score file format across participants. For score fusion, a Python implementation of the BOSARIS toolkit [Brummer and de Villiers, 2013] is used for logistic regression LLR scores. When fusing the described subsystems and embeddings it emerges that, while individual systems perform good, their contribution might be little when they are integrated with many similar ones. In contrast, systems showing low individual performance integrate better thanks to their different behaviours. Moreover, some systems with little contribution have a critical impact on the remaining performance gains on progress set.

Moreover, NIST corpora has been used as an evaluation tool for ASR systems which did not actively participate in the campaigns. For instance, [Dumpala et al., 2017] presents an interesting ASR analysis through an i-vector approach on non speech sound, such as laughter and breath. An understanding of the weight of speaker-specific information carried by these variations of speech is important to build a good speaker recognition system. In this study, the i-vector system is trained only on neutral speech, from NIST SRE04-08, and its performance is evaluated on laughs. Further, the inclusion of laughter sounds during training is considered. The results demonstrate that this inclusion provides complementary speaker-specific information and an overall improved performance. The corpus for the test phases contains several hours of conversational speech recordings from 40 speakers (20 male and 20 female). Phonetic annotation is present and the laugh segments are separately labelled.

The same data sets are used in [Georgea et al., 2018], in order to perform ASR in language mismatch conditions. MFCC and relative Δ coefficients are the selected features for the GMM-UBM based system in comparison to a cosine distance features based one. Results show that reference speakers who are phonetically and geographically closer to the target ones, contribute more to the ASV performance. A reduced reference population makes the cosine based system perform better, with further gains achieved by fusing the two systems, demonstrating the complementary nature of error patterns.

Data from SRE10-11 is used as part of the investigation on feature representation by [Bousquet et al., 2012; Richardson et al., 2015]. The former investigates i-vector representation based on LDA covariance scoring compared to a Gaussian-PLDA model. Significant performance improvements are demonstrated after two iterations of Spherical Nuisance normalisation and initialising the PLDA matrices according to the new i-vector representation. Nevertheless, estimation of the PLDA parameters remains necessary as the system yields slightly worse performance when running only one iteration of the expectation-maximisation algorithm, 1.15 % EER for males and 1.75 % EER for females. Whereas in [Richardson et al., 2015], features learned a posteriori by DNN on MFCC provide significant performance gains for speaker and language recognition tasks. The unified DNN approach is shown to yield substantial performance improvements on the the 2013 Domain Adaptation Challenge speaker recognition task (55 % reduction in EER for the out-of-domain condition) and on the NIST 2011 Language Recognition Evaluation (48 % reduction in EER for the 30 s test condition).

In addition to the use of purely ASR based features in order to represent speakers' characteristics, some researchers focus on more phonetically related features that are adapted to ASR techniques. This is the case e. g. for [Reynolds et al., 2003], where SRE03 data is used as training material. A system fusion of SVM and GMM-UBM is tested through duration mismatch conditions; test material contains 2,5 minutes of speech as average. In an additional test condition, selected most common words from the studied data sets are used. Performance of 2 % EER is shown using vectors consisting of per-frame log pitch, log energy and their derivatives. The same experimental protocol is used in [Adami and Hermansky, 2003] to test a GMM-UBM system on both speech and speaker recognition. Results indicate that in both recognition tasks, the 5-frame interval estimation for features' derivatives performs better than a longer time interval, such as the 10-frame interval applied in language identification systems. The EER for the segment classes derived from the f0 and energy trajectories is respectively 14 % and 13 % relative improvement over the reported baseline. This approach uses only the information from the Δ parameters. The addition of information about segment duration provides significant improvements for the ASR results. This shows that the segment duration conveys speaker information.

In line with the findings about the discriminant power of segment duration for speaker recognition tasks, we found works such as [Shriberg et al., 2004; Kockmann et al., 2010]. In both studies, data from NIST SRE04 is used to generate a GMM-UBM based on syllable-based nonuniform extraction region features (SNERFs). These are syllable-based prosodic features relying on estimated f0, energy, and duration information. Characteristics like minimum, maximum, mean, and slope of pitch and energy trajectories, are extracted for each detected syllable in an utterance and its nucleus, as well as duration of onset, nucleus and coda of the syllable. All values are further normalised with different techniques and form several hundred features for each syllable. In both cases, the EER is less than 8 % and the approach is able to reduce the high-dimensional inputs to only about hundred dimensions, while preserving their discriminative power. A further discussion of works that combine the two approaches is presented in the following section.

3.2.2 Combining Phonetics and automatic domain

An interesting part of the field of ASR is the combination of automatic techniques and pure phonetic features. In this sense, the SNERFs described above are an important example. They quantify speakers' idiosyncratic prosodic behaviour through duration, pitch and energy features on syllable-level. A further application is described in [Shriberg et al., 2005], where speakers' characteristics are modelled at a level higher than syllable with N-grams, named SNERF-grams. They used the spontaneous English conversational telephone data from the Fisher corpus as the linguistic material on which a SVM performs speaker recognition. Results show that SNERF-grams provide significant performance gains when combined with a state-of-the-art baseline system, as well as with two highly successful long-range feature systems that capture word usage and lexically constrained duration patterns. An overall 11 % EER is obtained by the system when integrating all the features at once.

The main goal of studies such as [Shriberg et al., 2005; Ferrer et al., 2007; Ferrer and McLaren, 2018] is to use long-range features to improve the state of the art on speaker recognition. In order to do this, performance between phonetic-based systems is compared. Results show that the prosody-based system alone significantly outperforms both duration-based systems and N-gram lexical modelling, which appears as the weakest of the systems. However, duration-based systems show an important complementarity with the SNERFs. This combination is significantly better than either individual system reaching an EER of 5 %.

Further experiments examine the relative contributions of features by N-gram length and feature type. Longer N-grams perform better than the shorter ones and pitch features appear as the most useful, followed by duration and energy. For instance, the most important pitch features are those capturing pitch level, whereas the most important energy features reflect rising and falling patterns. For duration features, nucleus duration is more prominent for speaker recognition than are duration from the onset or coda of a syllable. Overall, SVM modelling of prosodic features provides consistent information for automatic speaker recognition. When studying speakers' characteristics and focusing on words and N-grams production, a major challenge is to learn which words or word groups may be more consistent for the analysis.

In this sense, important candidates are discourse-related behaviour such as filled pauses or discourse markers. Some of these elements are explored in [Kajarekar et al., 2003], where the contribution of modelling prosodic and lexical patterns is investigated on telephone conversations from the NIST SRE03 data. The comparison includes results from systems based on prosodic features derived from SNERFs, a state-of-the-art GMM-based system, and a combination of both. All features are extracted from pause-to-pause sequences in order to capture long-term patterns in speakers' productions. Results show that these features provide complementary information to both frame-level cepstral coefficients and to each other. Thus, improvement of 15 % average EER in speaker recognition performance over conventional systems is observed in particular in combination with the duration-based system.

Furthermore, [Adami et al., 2003] compares UBM-GMM based on MFCC and prosodic features, i. e. relative f0 and energy contour dynamics, for a list of words from 40 English speakers. f0-based systems produce an EER of 13.3 %, while fusion between all systems

decreases the EER to 3.7 %. These results further demonstrate that prosodic features have complementary information to standard MFCC information.

Prosodic information is not the only focus of studies that combine phonetic and automatic approaches, features representing source and filter components are highly present as well. Formant values from a large corpus of the 10 French oral vowels, more than 300000 tokens, uttered by 111 speakers are compared to estimate their speaker discrimination power in [Kahn et al., 2011]. /œ/, /ɛ/ and /a/ appear to convey more idiosyncratic information when used in a UBM-GMM system. However, no direct explanation is drawn from phonetic measures to predict performance level. In a similar way, dynamic properties of diphthong and tone trajectories in Thai from [Thaitechawat and Foulkes, 2011] show high performance. Data from English in [Kelly et al., 2016a] confirms the discriminant power of formants when integrated in a i-vector PLDA system.

Investigations on voice quality have emerged in recent years, such as in [Long et al., 2011], where a UBM-GMM based ASV system is fused with information from HNR. 76 utterances from 76 male speakers and 84 utterances from 84 female Mandarin speakers are used as training and testing material. The resulting model shows important improvements in noisy sequences, where commonly MFCC struggle. Furthermore, in [Gendrot et al., 2019], spectrograms and phonetic parameters related to voice quality are compared through a DNN system. The results from oral vowels of 45 French speaker of the ESTER corpus (a radio broadcast of prepared speech) show better classification for male (68 %) compared to female (62 %) speakers with phonetic feature, while the opposite trend is observed, however, to a lower extent, for spectrograms, 74 % and 76 %. Further comparisons show that in 22 % of the data, the network trained with spectrograms achieved successful discrimination while the training on phonetic parameters failed. The opposite was found in 10 % of the data. However, when the network trained with spectrograms failed to discriminate between some tokens, parameters related to f0 proved significant.

[Stoll and Doddington, 2010] uses jitter, shimmer, energy and LTAS in order to select similar speaker pairs that are successively passed to a UBM-GMM system. Multiple techniques are tested to assess speakers' similarity based on selected voice quality features. The largest changes in detection cost and false alarm rates for similar speaker pairs occur when speaker pairs are selected using the Euclidean distance. Even bigger differences in performance occur when speaker pairs are selected using Kullback-Liebler divergence between speaker-adapted GMM (trained on MFCC). The results confirm that integration of phonetic-based features or preprocessing of data may produce consistent improvements for ASR systems. In this perspective, phonation type classification is studied in [Chanclu et al., 2021]. The Authors propose a new workflow dedicated to voice quality, where the final objective is to provide an automatic tool to help phoneticians to improve their understanding of phonation phenomena. Its implementation may represent an important addition for the modelling of speakers through phonation patterns and, in general, their voice quality characteristics

The latter mentioned study incorporates the idea of combining knowledge from two domains in order to improve consistency of the resulting outputs. The same idea is present in works such as [Scheffer et al., 2004; Fredouille et al., 2005], investigating twins recognition and pathological voice assessment, respectively, in relation to ASR techniques. Both investigations show promising results for the adaptation of automatic methods which obtain state-of-the-art results in ASR to classical phonetic issues.

Evaluating a system performance in a speaker recognition task is an important aspect in order to assess the reliability of methods and extracted features. However, since speech is the studied object along its variations and components, the use of phonetically based parameters is increasingly common in the automatic domain. Phonetic interpretation of speech components does represent one major concern in our study. In a similar way, many researchers have used interpretable parameters that correlate with linguistic knowledge and have integrated them into ASR processes. The basic idea is to enhance the reliability of automatic results thanks to the previous knowledge acquired from the input material. Having a performing speaker recognition system but not knowing where the performance comes from is an issue that always needs to be addressed and covers an important role throughout this thesis.

3.3 Artificial Neural Networks

The last section of this chapter does not focus on a large domain in which speaker recognition is adapted, such as in the previous ones. Instead, it discusses a method that is applied to perform multiple recognition tasks. Since the advent of computer technology, modern approaches and new techniques have appeared in audio analysis and phonetic research, from automatic spectrographic computations to ANN. Results from the application of the latter on speakers' phonetic characterisation are discussed later in this thesis. ANN have been increasingly present in research, but not limited to the speech domain, throughout the decades. An extensive and comprehensive review of all the different applications of ANN is present in [Goodfellow et al., 2016].

ANN functioning is inspired by actual neurons present in animal brains, cells that are connected and exchange signals. Naturally, this comparison has to be taken lightly, in the case of ANN, the signal is a real number. A basic ANN architecture is implemented as a chain of matrix multiplications where elements from the input, or feature vectors, interact with each other by addition. Depending on the task and the neural architecture, information can be back propagated instead of going directly from the input to the output layer. Through these computing mechanisms, ANN have the ability to approximate different classes of inputs and various type of actual input objects, e. g., images, raw speech signals, text documents, number matrices [Hornik et al., 1989; Ravanelli and Bengio, 2018b; Jiang et al., 2019; Rocco et al., 2019; Crawshaw, 2020]. A separation between the train and test phase is necessary in order for the models to be constructed. Depending on the actual knowledge we have about the input material, deep learning techniques can provide meaningful and effective results.

The success registered by ANN has made researchers continuously question if the actual results reflect a deep and fundamental approximation capability or if they are the consequence of a careful response to an incidentally well stated problem. What an ANN model uses to model a real world class does not necessarily contain the same information our brain uses. Applying deep learning techniques in problems like speakers phonetic characterisation is not a difficult task, the issue is represented by providing the system with meaningful inputs. However, what a meaningful input really is and how to formalise it are questions that can be more difficult than expected to answer.

Learning is to estimate consistently the connections that perform the approximations

proven to be possible. This process involves identifying the learnable aspects of the inputs and to choose a mathematical framework that allows a formal treatment of learnability. The chosen framework should be productive enough to capture a wide variety of learning problems. Besides, establishing a determined path for learnability is demanded. However, as discussed in [Ben-David et al., 2019], this paradigm fails even in a well studied learning model since learnability cannot be characterised using the standard axioms of mathematics. Learnability does not hold a dimension that can be quantified in full generality. Hence, the choice of evaluation metrics is one of the main aspects when establishing a machine learning analysis protocol, see Section 7.1.2.

Machine learning processes are described as functions of the continuum they produce. However, their learnability is captured by the cardinality of the continuum. In other words, a designated class is learnable if and only if there are finitely many distinct cardinalities between the integers and the continuum. If and only if its approximation is provable. The main consequence is that the learnability of a family of sets F over the class of probability distributions P is undecidable. However, learning F over P may not be directly related to selected learning algorithms rather than to the definition of a learning function which can be described and deconstructed. In this perspective, the use of, e. g., attention mechanisms may provide the researcher with a look at how the input is transformed through the layers of a DNN and what part of it has been used to model a certain class.

Attention mechanisms have shown great performance in various NLP tasks, such as sentence embedding, text generation, machine translation, machine reading comprehension, etc. Nevertheless, existing attention mechanisms only describe high-level or low-level features. The lack of hierarchical mechanisms is an important issue for the improvement of learnability. [Dou and Zhang, 2018] investigates a Hierarchical Attention Mechanism based on the weighted sum of different layers of a multilevel attention mechanism. Using different attention depths to show the influence on performance, this is a first example to combine low-level features and high-level features of input sequences to output a more suitable intermediate result for decoders. The achieved results on Chinese poem generation show a nearly 6.5 % averaged improvement compared with existing machine reading comprehension models.

In a similar perspective, there are multitask models, in which the learnability problem takes a further step since multiple tasks are simultaneously learned by a shared model. Such approaches offer advantages like improved data efficiency, reduced overfitting through shared representations and fast learning by leveraging auxiliary information. It is the case for [Chen and Salman, 2011; Gresse et al., 2019], where siamese networks are used to compare fitted speakers' models and provide similarity metrics between them. However, the simultaneous learning of multiple tasks presents new design and optimisation challenges and choosing which tasks should be learned jointly is in itself a non-trivial problem [Crawshaw, 2020].

We already mentioned the implementation of DNN-based speakers' embeddings in ASR. This approach has shown consistent results for modelling speaker characteristics by the means of the so-called x-vectors. However, it requires a massive amount of training data, careful selection of network architecture and related tuning parameters. Further discussions are reported in [Garcia-Romero et al., 2019]. The Authors presented a DNN refinement approach for the DNN parameters to produce embeddings optimised for co-

sine distance scoring. Speakers in the Wild database are used for testing. It contains hand annotated speech samples from open source media in order to benchmark speaker recognition technology on single and multispeaker audio acquired across unconstrained or "wild" conditions. Results show that this approach is capable of producing embeddings that achieve record performance on said benchmark.

3.3.1 Convolutional Neural Networks

A variant of ANN used in deep learning approaches is represented by CNN, which removes the necessity of feature extraction from the inputs, allowing the use of local correlation structures such as images. Since their introduction in [LeCun and Bengio, 1995], they have become a standard method for many shape and pattern recognition tasks e. g. [Jiang et al., 2019], including image similarity e. g. [Rocco et al., 2019] and captioning e. g. [Soh, 2016]. High performance has been observed by CNN in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Krizhevsky et al., 2012; Russakovsky et al., 2015], a project aiming to evaluate algorithms for object detection and image classification at a large scale. From this challenge multiple CNN architectures have emerged, in particular the ResNet designed by [He et al., 2015a], which remains the gold-standard architecture in numerous scientific publications.

It is characterised by layers that apply residual functions directly to the layer inputs, instead of learning unreferenced ones. Residual blocks take the input, perform a rectified linear activation unit function on it and add the result to the original input. Instead of back propagating the gradient error to fit a desired underlying mapping, ResNets fit the layers through a residual mapping. This architecture has been demonstrated to be easier to optimise and gain accuracy from increased network's depth e. g. [Wightman et al., 2021]. They typically serve as the default architecture in studies, or as baseline when new architectures are proposed. In Chapter 7, we use a variant of this particular architecture, ResNet-30, applied to phonetic-based representations of speakers' characteristics.

Nevertheless, images are not the only input that have been tested for CNN, speech signals have shown promising results as well. [Palaz et al., 2015] reported an overall accuracy of 97 % for speech recognition on both clean and noisy speech from the TIMIT corpus. Results from the noised signal demonstrate the high sensitivity that CNN have in filtering the relevant information from the input. The TIMIT corpus is used to examine speaker recognition by CNN trained on raw waveform in [Ravanelli and Bengio, 2018a,b]. The results show state-of-the-art performance with the advantages of less demanding computational power compared to other ANN architectures and the ability to tune the convolutional filters to retrieve specific information directly from the signal.

Moreover, we observe that the combination of phonetic knowledge with the mentioned CNN technique brings the possibility to learn the features directly from spectrograms in order to perform speech or speaker recognition. This is the case e. g. [Ferragne et al., 2019], where /ã/ spectrograms of 45 French speakers from a radio broadcast corpus are used. Overall scores for SID are 84 % for females and 86 % for male speakers. Several versions of the same model are trained with varying low-pass filtered spectrograms showing that for certain speakers there are frequency bands providing more prominent information.

Speakers' information retrieval using CNN on spectrograms from non-speech sounds is

the object of [Zhang et al., 2017; Zhao et al., 2017]. The former study proposes a speaker recognition task on 104 speakers with cough, laugh and "Wei" (a short Chinese "Hello"). The CNN are trained on 2500 male and 2500 female speakers, with 95167 utterances randomly selected from the Fisher database. Each speaker has about 120 seconds of speech segments. The results show that there is rich speaker information within these non-speech sounds, even for coughs that may appear as less discriminant. With the proposed feature extraction approach, EER can reach 10 %-14 %, despite the short duration of tested materials, from 0.2 to 1 s. The last hidden layer of the CNN model is used to represent a feature vector in order to investigate the possible enhancement of speakers' characteristics representation. The extracted features are tested on 10 randomly selected speakers, see Figure 3.2. The Authors note that the learned features from "Wei" are reasonably discriminative for speakers, while less consistent results are shown in cough and laugh. The presence of a vocalic production in the first case is undoubtedly part of this better performance.

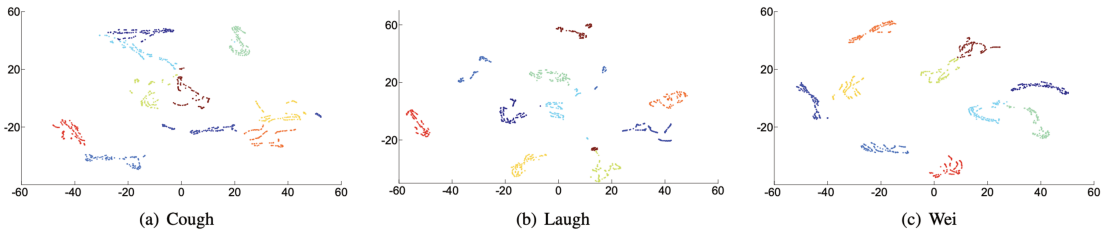


Figure 3.2: Figure 4 from [Zhang et al., 2017]. Deep speaker features on event (a) Cough (b) Laugh (c) "Wei" randomly sampled from 10 speakers. The pictures are plotted by t-SNE, with each color representing a speaker.

[Zhao et al., 2017] examines recognition potential of breath sounds in continuous speech, by the means of inhalation that occurs in order to replenish air in the lungs. Speech from radio and television news broadcast recordings of 50 English speakers is used. The results from the CNN-LTSM outperform an i-vector and SVM system, with 91 % against 71 %, confirming the high potential of the CNN framework for SID tasks.

[McLaren et al., 2014] confirms the high performance of CNN in noisy conditions, in particular when compared to conventional UBM-GMM i-vector based systems. The approach through CNN provides improvements of 26 % in miss rate, hence, considerably outperforming the traditional ASR approach. In a similar way, clustering precision is compared between a CNN and a UBM-GMM based system in [Lukic et al., 2016]. 10 sentences, 6 for training, 2 for validation, and 2 for testing from the 630 speakers of the TIMIT corpus are used, with the CNN achieving a clustering accuracy of 97 %. However, the Authors note that other works achieve better accuracy by using GMM models.

Chapter 4

Exploring voice perception

This third chapter of the moment of *fixity* introduces key elements and studies from the perception domain that have influenced this thesis. The following sections present multiple layers of the human perception in relation to speakers' characterisation with examples from the literature. Then, in Chapter 5 our perception results are presented.

In order to establish the perception background for this thesis, we begin by considering the perception principles underlying this study and the main examples from the literature we have taken as models. Once the basics of our perception analysis are discussed, we focus on how variations in specific components influenced the on listeners. Characterising a speaker's voice is an operation that can only be achieved by analysing the patterns occurring in different speech components. In this sense, understanding which component of the speech signal plays a prominent role during the listening process is a fundamental step to understand the multiple layers of speakers' characterisation.

Section 4.3 focuses on the idea of voice clustering. Methods and implications for this type of analysis are considered in our review, since our perception results are based on a very similar process. Examples from the literature show that voice clustering has been effective in particular to analyse listeners' ability to distinguish familiar from unfamiliar voices. The application of voice clustering in the analysis of pathological speech is discussed as well.

Nevertheless, when considering human perception, identification strategies may vary between individuals, semantic and pragmatic aspects can as well have an influence on the result, but some fixed elements are shared by the population. In Section 4.2, we take examples from the literature to evaluate the coherence amongst listener answers and how to reduce the gap between them during a perceptive evaluation.

In the last section, a summary of all the three *fixity*'s chapters is presented. Here, we discuss which aspects are retained from all the reviewed studies, in particular the considered methods and the components we want to focus on. In this perspective, this last section is an introduction to the second part of this thesis, the moment of instability, where we apply what we learned from the literature on speakers' characterisation to our specific cases.

The main reason for introducing perception studies and results in our literature review is not only to add yet another domain of study but rather to provide complementary findings capable of connecting all of the others. Perception is the result of the processing of an input

by the brain, which allows a subject to solve simple or complex problems. This ability has evolved to enable survival throughout the centuries and, together with behavioural abilities, it shapes neural mechanisms present in the human brain. Understanding the natural algorithms underlying a specific problem is extremely challenging, however, it may enable the formalisation of new research findings in different domains. For instance, understanding listeners' speakers characterisation through speech signal decomposition into its components and which of these are more efficient for said operation may provide useful outcomes for both classical Phonetics and NLP research.

4.1 Perception principles

As mentioned in the previous chapter, ANN are based on the idea of interconnected networks of neurons present in the brain. In particular, CNN operations take inspiration from the natural cognitive perception principle of pattern recognition. We already presented how computer vision through machine learning by the means of CNN has shown great potential and its rapidly evolving domain in Section 3.3.1.

Human vision is the result of a perception mechanism involving the eyes, which capture light using photoreceptors in the retina and transfer electric impulses to the brain. As reported in [Morgan and Wong, 2012; Yamins and DiCarlo, 2016; Schrimpf et al., 2020], works in visual systems neuroscience have shown that cortical areas encode object properties with a hierarchic organisation and tolerance generating invariant object recognition behaviour. Early visual areas capture low-level features such as edges and centre-surround patterns. In contrast, responses from the highest ventral visual areas, i. e. inferior temporal cortex, can be used to decode object categories and their significant variations present in nature. There is still a low understanding of features processing by mid-level visual areas. Nevertheless, they appear to provide intermediate computations between simple edges and complex objects, filling a pipeline of increasing receptive field sizes.

Audio perception parallels this mechanism with the ears capturing audio signal, a pressure wave propagated through a medium, converted in the cochlea and passed to the brain as an electric response. The transferred signals are decomposed by our brain. Large patterns and smaller units are extracted and associated to known or unknown categories. In linguistic fields, basic categories involved in audio and visual perception vary from simple phonemes to entire utterances, written or spoken, in a language known or unknown by the listener. In addition, as mentioned in the introduction of this thesis, other information regarding the speaker can be retrieved from the speech signal.

Research paralleling audio-visual perception has already shown that humans can extract numerous reliable information just from voices and faces. The comparison between voice and face is not always obvious, but it can be seen as more coherent when considering the underlying mechanisms of audio-visual perception, where a whole signal is decomposed in smaller units conveying information.

Furthermore, [Latinus and Belin, 2011] discussed the different treatment of voice and of intermediate levels stimuli through fMRI of a human's auditory cortex. Cerebral activity of normal subjects listening to speech sounds has shown activations along the anterior and middle parts of the superior temporal sulcus (STS). These activations remained unaltered

when the speech stimuli were played backwards. This involved the removal of most of the linguistic content, leaving unaffected the voice timbre, suggesting that these regions might be more interested in the vocal nature of the speech stimuli than in the linguistic content.

These 'temporal voice areas' (TVAs), located along the mid STS bilaterally, show sensitivity to voices whether they contain speech or not, rather than to control sounds, e. g. amplitude modulated noise. TVAs appear within a few months after birth and are also present in the brain of macaques, suggesting an early development of cerebral voice processing in evolutionary history. Despite this, little is known on the exact functional role of the TVAs, or even whether their greater response to voice implies a specific role in cerebral voice processing. For more extensive reviews see [Kreiman, 1997; Cutler et al., 2010; Belin and Grosbras, 2010; Eisner, 2015].

For instance, [Moyse, 2014] provides evidence that examining faces and voices categorisation into age ranges (18-25, 35-45 and 55-75 years) shows a high accuracy. Listeners are able to sort voices with an average percentage of correct responses of 78 % just from sustained vowels. Studies on age estimation from voices showed an impact of stimulus duration on the performance, when listening to speech sequences results increased to over 90 %. Listener groups characteristics have been shown to influence age estimation from both faces and voices. However, precise age estimation accuracy is higher from faces than from voices, implying that speech categories present a higher variability and complexity in human learning processes. As an example, the age of stimuli and the age of participants impact the performance of age estimation: younger stimuli result in better estimations than older stimuli and younger participants are more accurate than older ones. Sex also seems to impact age estimation from voices and from faces but in an opposite direction: the age of female voices is better estimated than that of male voices whereas the age of male faces is better estimated than the age of female faces.

We still need a deep understanding of what the audio signal is made of, and one of the aims of this thesis is to help describe its components interactions. The decomposition of human speech signal, which is a derivative of audio signal, enables speakers and listeners to make an inventory of discrete components. As already discussed, these components differ on the level of representation. However, their interactions create the speech production as the perceived whole from which information is extracted. Different levels of representation ranges from discrete phonemes, auditory perceptual targets specified for properties depending on prosodic, segmental and social context, to articulatory gestures, specified in terms of synchronous motor command patterns, and more, e. g., patterns involved in feedback from the articulators to higher levels of speech programming [Nooteboom and Quené, 2021]. It appears that understanding the perception of these individual parameters requires knowledge of the acoustic context in which they function.

Data from [Cutler et al., 2010] provide evidence about the influence of speakers on words and phoneme recognition. Indeed, listeners need abstract prelexical representations of speech sounds in order to deal with variation, e. g. different speakers, in the speech signal. When a particular sound is produced in an unusual way, hence there is consequent distance between the prelexical representation and the speaker's production, the use of abstract prelexical representations in decoding speech is both efficient and beneficial. Efficiency is found in knowledge about speaker's idiosyncrasies that are coded for a single sublexical representation, rather than separately for all words in the lexicon that contain

the unusual sound. On the other hand, benefits are found in the recognition of all words containing the unusual sound through the model created from the listened speaker. In this sense, it is shown that listeners asked to judge whether a word has been heard earlier recognise words spoken by the same speaker as "old" more quickly than the ones spoken by a different speaker. This suggests that speaker-specificity effects may not reflect lexical-level processing.

In addition, results in [Winkler, 2007; Gelfer and Bennett, 2013; Clark et al., 2014; Waller and Eriksson, 2016; Lee et al., 2019] suggest that understanding the process of characterising a voice involves the understanding of individual strategies used by speakers to make variation of their voices. In a sense, this means learning to recognise how patterns vary and interact with each other rather than how much realisations differ from a prototype, e. g., the average value of a phonetic measure. The large variability matrix carried by a speaker relies on related patterns appearing at different levels of speech production. A useful addition to explain these considerations is given by Gestalt theory on visual perception that follows the idea of the brain perceiving a whole as interactions of patterns more than just additions of single elements. This approach, present in [Kreiman and Stidtis, 2011; Cambier-Langeveld et al., 2014], is in line with the audio-visual parallel we already presented in this thesis. Audio and visual signals represent similar physics phenomena. As mentioned, the perception of the first is the result of interacting with an audio wave while in the latter light waves and photons allow our brain to process the visual world. In both cases different wavelengths and bandwidths are perceptively associated to different results and/or components from the whole signal.

4.1.1 Decomposition of the speech signal

Speaker's information can be present at different levels in the speech signal, i. e. conveyed by different components at the same time. The variability matrix defining speakers' characteristics and their distributions vary from speaker to speaker, however, the same components can be retrieved. Which one is efficiently identifiable by listeners, playing a role in providing the speaker's information is the question to which multiple studies have tried to answer throughout the decades. Hereafter, we discuss some examples of components, represented by phonetic measurements that have shown correlation with speakers's information such as identity, sex and age.

Variations in f0 and speech rate have been associated with consistent changes in age of the speaker in [Waller and Eriksson, 2016]. Experiments are based on read speech from a total of 36 speakers of three age groups: 20-25, 40-45 and 60-65 years old. All speakers recorded in three different conditions, one with their normal voice and two where they were instructed to sound years younger or older. Increase in f0 and rate are observed when speakers try to sound younger, while the same features are decreased when trying to sound older. The same strategy applied regardless of sex or age. Listeners attempt to identify the actual speakers' ages consistently. Whereas, in the two disguised voice conditions age estimates show 2-4 years of difference against the 15-20 speakers were instructed to. Speech rate explains more variance than f0, corroborating previous findings.

Other temporal features have shown consistency in accent discrimination, i. e. English and French accounted for German, in [Kolly et al., 2017]. Results show that listeners could identify the linguistic origin of French and English speakers based on temporal

features of these accents. In addition, listeners could also identify the accents in question in stimuli that contain strongly degraded spectral features alone. The combined presence of temporal and spectral information is thus not necessary for listeners to identify foreign accents better than chance level. However, segmental information biassed the responses. When stimuli featured uvular /r/, listeners tend to perceive a French accent, while a bias towards English accent is observed in stimuli that featured vocalised or no /r/. This and the already mentioned examples of [Ramus and Mehler, 1998; Arvaniti, 2013] show how prosody and source and filter components heavily interact in speech perception.

Furthermore, the role of speech rate and f0 in age perception is also discussed in [Winkler, 2007], through three synthesised German words. 20 listeners judged the produced words whose acoustic and temporal features corresponded to realistic variations based on a database of 23 single words spoken by 30 female and 30 male subjects. The perceived average age consistently increases with decreasing speech rate. Indeed, for female voices, all speech rates have a significant result, with the difference between slow and fast being the most significant. For the male voices there are statistically significant differences but no prominent one. The effect of different pitch levels on listeners' judgement differs regarding levels of speech rate. While for the female voices judgements regarding different pitch levels do not differ for slow and normal speech rate, a characteristic pattern can be observed for the fast speech rate. Results for mean listeners' judgement show a rise by 9.5 years from high to low pitch level. For the male voices, listeners' judgements seem to be less influenced by the pitch level in fast and middle, but more in slow speech rate. If stimulus words were spoken slowly, the high pitch level was associated with a remarkable increase in the mean listeners' judgement of a talker's age.

In a similar perspective, we already mentioned the work of [Gelfer and Bennett, 2013] in this direction, which studies the influence of f0 and formants in the physical appearance of speakers. f0 from isolated vowels /i/ and /a/ were digitally altered for 30 speakers to produce average males, average females and ambiguous value ranges. Results indicated that male speakers are less perceived as male when f0 is produced higher than 150Hz and formants present higher values as well.

Furthermore, examples from ASV with age modification through changes of source and filter characteristics are reported in [Hautamäki et al., 2017]. 60 speakers, aged from 18 to 73 years old, recorded Finnish read speech, "The Rainbow Passage" and "The North Wind and the Sun" and two TIMIT sentences in English. Speakers were required to record three sessions in total, the first using their natural voice without any intentional modification, while for the second and third they were asked to sound like an old person and a young person. 70 listeners, with only 32 of them being natives, participate in the experiment, 44 males and 26 females with similar age ranges.

Statistical analysis of source and filter characteristics show that positive changes are observed for younger voice disguise for all age groups in both female and male speakers, while in older voice disguise the change is generally lower. For a few female speakers the change is negative, a decrease of f0 in comparison to their modal voices, while for 12 of them the change is positive for the intended older voice. No statistically significant variation in formants is observed for 30 % of the speakers. However, the top variation pattern for female speakers shows a change in at least one of the formant values. More increases in mean formant differences for the young disguise condition are observed, while the old disguise had more decreases. Perceptual speaker verification and ASV results are

linked by the means of selecting easy, intermediate and difficult trials based on automatic results. The goal was to find out whether or not listeners followed the same pattern. This is confirmed: the trials considered easy for the ASV systems do not show errors by the listeners, while the ones considered difficult for the ASV systems turn out to be difficult for listeners as well. Comparison between native and non-native Finnish speakers shows similar performance.

Studying Perception of smaller units means analysing how important a component is in the categorisation of a perceived speech signal, showing that the information can be efficiently found in selected parts of the speech signal, e. g., for language discrimination. Voice disguise means to change some components of the voice in order to give a different perception to the listener. Changed component can be associated to a specific identity aspect: age, sex, native language, health aspects. When this transformation is successful the component is associated with that aspect in an intrinsic perception mechanism.

For instance, [Culling and Darwin, 1993] discussed the role of f0 and timbre in characterising speakers' voices. Formants trajectories for 12 diphthongs were derived from the formant frequencies of the steady English vowels /i/, /o/, /u/, and /ɛ/, produced by three native speakers. Formants trajectories were obtained by linear interpolation from each of the four vowel specifications to each of the other three. Diphthongs are synthesised using an implementation of the Klatt synthesiser. Subjects listen to 48 individual diphthong sounds as a practice and attempt to differentiate "continuous" (falling or rising) from "inflected" (dipping or peaking) glides for practice. Then, 264 experimental stimuli are presented twice in a random order and classified as "crossing" or “bouncing”. In the main experiment, stimuli which have the same and different timbres at the f0 intersection are compared. As predicted, subjects show no discrimination ability when the timbres are the same, whereas a highly significant ability to differentiate the two classes of stimuli is observed when stimuli have dissimilar timbres. These results are consistent with the idea that listeners use continuity of timbre to disambiguate f0 intersections.

Source and filter characteristics interactions are analysed in a speaker identification perception study in [Lavner et al., 2000]. The recordings include two short sentences in Hebrew from 20 native Hebrew speakers and from which isolated vowels have been modified in several ways: formant frequencies shifting upward or downward on a logarithmic scale or alternatively fixed to the average speaker population values; spectral envelope of the vocal tract filter shifted upward and downward on a logarithmic scale or substituted by a synthetic version, based on the formants average speaker population values; changes to values of the opening or closing quotient, without changing the fundamental frequency or the vocal tract filter; f0 changes at different rates; finally, natural voice and synthetic voice generated by the original vocal tract model used in multiple combinations. Only speakers identified in their natural voices were used in their modified variants. A total of 30 listeners participated in the experiments. It was found during an informal listening test that the changes in the closing quotient had perceptively greater effect than changes in the opening quotient. The results show that for formants, the identification percentage is related to the rate of shift of the frequency, lowering identification rate significantly more than raising. Higher formants show greater degradation in results as well, with F2 appearing as the more robust one. The same asymmetry between lowering and raising formant frequencies is also reported for modification involving shifting of spectral envelope. Average identification rate of 40 % is obtained when shifting the spectral envelope by one tone, whereas the same modification of individual formants results in significantly higher

rates. Lowest identification rates (14 %) are obtained by substituting the original vocal tract model of each speaker with a vocal tract transfer function of another whose voice was not presented in the experiment. However, replacing the glottal excitation waveform of each speaker with a parametric model, hence keeping the fundamental frequency intact, shows a significantly higher rate (44 %).

Overall, modifying the glottal excitation signal reduced the identification rate to 65 % from the original 80 % of non modified tests. These findings show that as long as the fundamental frequency was preserved, various parameters of the glottal pulse could be changed considerably without a significant perceptible degradation of identification. As for formant modification, similarly, lowering f0 reduced rates more than raising it. Authors explain this result by the natural tendency of speakers to use mainly the lower part of their vocal registers. Thus, raised frequencies are still perceived as representatives, whereas lowered ones are less perceived in the possible natural range.

The source and filter characteristics are shown as important factors contributing to speaker individuality, at least in isolated vowels. [Baumann and Belin, 2010] investigated speakers' similarity from three French vowels and found correlation between listeners' perceptual responses and a f0-formants MDS. The Authors used the three French vowels /a/, /i/, /u/ produced by 32 speakers, 16 males and 16 females. The ten listeners were presented with pairs of vowels and had to place the heard tokens based on their similarity in a two-dimensional space. Five Japanese sustained vowels by eights male speakers are used in [Matsumoto et al., 1973] providing evidence of the similarity between a f0-F1 MDS and perceptual responses. Additional features may be more important when listening to continuous speech as discussed in [Moore, 2008], where normal-hearing, partially and completely hearing loss listeners are compared. Through simple and complex tones reproduction, it is shown that cochlear hearing loss leads to a reduced ability to process TFS, thus pitch-related information.

The analysis of perception of speech components is fundamental to understand how the information is conveyed in terms of redundancy and complementarity by each of them. Phonetic studies investigating the contribution of speech components to various levels of perception have largely focused on source and filter components, highlighting the human sensitivity to pitch mechanisms and different kinds of information that can be associated with them. Evidence that complex pitch perception mechanisms are shared by humans and other species is provided by numerous studies [Schreiner and Urbas, 1986, 1988; Fontaine et al., 2013]. Findings in [Hsu et al., 2004; Vaissière, 2004; Song et al., 2016] demonstrated that the perception of pitch is highly influenced by harmonics rather than simple f0 as well as by changes in noise spectral shape and by temporal envelope. The use of modulation transfer function to represent the time varying aspects of sound shows important results, e. g. for speech intelligibility [Elliott and Theunissen, 2009; Stilp and Kluender, 2010; Aubanel et al., 2018].

Nevertheless, source and filter characteristics are not the only focus in human perception, voice quality characteristics have shown to play an important role as well. The results presented in [Kreiman and Gerratt, 2012] indicate that perception of the harmonic spectral slope and noise levels in voice are strictly related as a set of complex interactions between the shape and levels of the harmonic and inharmonic parts of the voice source. 10 listeners participate in two experiments involving synthetic voices with varying modified noise sources, which result in falling, rising and flat noise spectra: HNR, jitter, shimmer, H1-

H2 and H2-Hn (difference between H2 and higher harmonics). Listeners judged whether the stimuli were the same or different, and rated their confidence on a 5-point scale. These findings confirm that listeners' sensitivity to noise levels in voice depends in part on the shape of the higher part of the harmonic voice source spectrum. Noticeable differences for HNR vary significantly with both noise level and with harmonic spectral slope. However, it had no significant effect when harmonic slopes were steepest. In the same way, harmonic slope had no effect when HNR levels were at the highest. Similar perception changes are observed with H2–Hn, since significant changes are perceived when spectra are more steep and dependent on both HNR values and baseline H2–Hn. Listeners appear overall more sensitive to changes in H2–Hn when the noise spectrum was falling and to changes in the HNR when the noise spectrum was flat. However, in the latter case sensitivity to HNR was significantly worse when the harmonic spectrum was flattest and when noise was falling.

Furthermore, a wide variety of phonetic measurements exists describing different spectral parameters and additive noise in voice showing different levels of sensitivity in human perception. [Sundberg et al., 1988] report how the frequency characteristics of auditory feedback affect voice level, such that a low pass filtering caused speakers to raise their voice level. A stronger effect is observed for singers, overall the singer's phonation appears louder when reading normally and in noise. [Klug et al., 2019] examines two low frequency spectral slope parameters (H1-H2, H1-A1) and one additive noise parameter (CPP) in order to distinguish breathy and non-breathy voices. The results show consistency with previous studies such as [Hillenbrand et al., 1994; Keating et al., 2011], but extended the findings from elicited speech to spontaneous. Besides, no correlation with mid-to-high frequency spectral slope parameters, i. e. H2-H4, H4-H2K, H1-A2 and H1-A3, have been found to support the perception of breathiness. Results close to significance are obtained for HNR05 (p=0.097) and HNR15 (p=0.077). The linguistic material consisted of 3 minutes samples of 22 voices that, based on auditory-perceptual analysis, aim to reflect a natural mixture along the breathy/non-breathy continuum. Perceptual ratings result from a survey submitted to experts in forensic speech analysis, involved in training and research on voice quality. The between-rater consistency for these 8 voices was established using Cohen's Kappa metric. In addition, all participants regularly used the same analysis scheme - a modified VPA (VPA) - to rate voice quality in forensic caseworks.

4.1.2 Voice rating protocols

An important part of voice perception is the perception of speech disorders. Listeners are able to focus on specific sets of components. However, when voices differ substantially in quality, e. g. as in pathological voices, listeners have hints at what to listen to and which features to select in order to characterise speakers.

Different protocols have been used throughout the decades to evaluate perceptually, hence characterise, voice quality features. The first step in this perceptual evaluation is to identify if a feature is present as neutral or if it deviates sufficiently from its standard condition to be considered as non-neutral. This approach in particular supposes that quality features can be extracted from a bundle of features. However, in many cases separation of different features is more difficult since their interaction is highly intrinsic. This features' behaviour is also one of the major causes of listeners' disagreement, see

[Kreiman et al., 1993] for a review.

Protocols have been developed to rate voice quality features favouring the description of voice as patterns, comprising a large number of simpler and indivisible features. Proposals such as Shewell's Voice Skills Perceptual Profile [Shewell, 2009] aim to develop simple perceptual methods. The target in this case are voice practitioners other than speech and language therapists, such as voice teachers and singing teachers. An alternative approach is taken by the GRBAS protocol [Hirano, 1981], which consists of a Grade of Roughness, Breathiness, Asthenia Strain scale for the assessment of patients with laryngeal symptoms. When non pathological voices are considered, the voices in a set are relatively similar, as the normal voices were, listeners' strategies apparently converge on a relatively small set of perceptual features.

Vocal Profile Analysis (VPA)

	First Pass		Second Pass							
				Moderate			Extreme			
	Neutral	Non-Neutral	Setting	1	2	3	4	5	6	
A. Vocal tract features										
1.Labial			Lip rounding/protrusion							
			Lip spreading							
			Labiodentalization							
			Extensive range							
			Minimized range							
2. Mandibular			Close jaw							
			Open jaw							
			Protruded jaw							
			Extensive range							
			Minimized range							
3. Lingual tip/blade			Advanced tip/blade							
			Retracted tip/blade							
4. Lingual body			Fronted tongue body							
			Backed tongue body							
			Raised tongue body							
			Lowered tongue body							
			Extensive range							
			Minimized range							
5. Pharyngeal			Pharyngeal constriction							
			Pharyngeal expansion							
6. Velopharyngeal			Audible nasal escape							
			Nasal							
			Denasal							
7. Larynx height			Raised larynx							
			Lowered larynx							
B. Overall muscular tension										
8. Vocal tract tension			Tense vocal tract							
			Lax vocal tract							
9. Laryngeal tension			Tense larynx							
			Lax larynx							
C. Phonation features										
			Present		Scalar Degree					
					Moderate			Extreme		
	Setting		Neutral	Non-Neutral	1	2	3	4	5	6
10. Voicing type	Voice									
	Falsetto									
	Creak									
	Creaky									
11. Laryngeal frication	Whisper									
	Whispery									
12. Laryngeal irregularity	Harsh									
	Tremor									

Table adapted from Beck (2007). Shaded cells mean that the corresponding setting does not admit the specified degree(s) or label.

Figure 4.1: Appendix 1 from [san Segundo and Mompean, 2017] showing a comprehensive version of the VPA protocol for the assessment of voice quality features.

In this sense, a particular mention is needed for the above mentioned VPA. It represents one example of a perceptual assessment protocol, created in the early 1980s by John Laver and colleagues as a means to identify and rate a speaker's voice quality features. It is of common use in forensic Phonetics. Both phonatory and supra-laryngeal features are considered, making it a quite comprehensive tool. As reported by [san Segundo and Mompean, 2017] and in Figure 4.1, one of the most common versions of the protocol presents a total of 36 characteristics: 25 describe vocal tract (supra-laryngeal) features,

7 describe phonation features, and 4 describe overall muscular (laryngeal and vocal tract) tension features. Depending on the version, the VPA protocol may include extra features as well, e. g. prosody and temporal organisation. However, a comprehensive protocol often corresponds to a highly complex one as well.

[san Segundo and Mompean, 2017] suggests the idea of a simplified VPA due to four main issue when trying to analysing perceptually voice quality: the highly multidimensional nature features which may be difficult to isolate; the fact that raters can fail to agree on definitions of a voice feature; the comparison between normal and pathological ratings, with the latest which may require more complex protocols; cognitive processing constraints, which means a simpler protocol may impose fewer cognitive demands on raters. Rating voices not only implies the assessment itself but a previous process of identifying and isolating the different aspects of the stimuli. This relates to the idea of the variability matrix describing characteristics from a speaker. Even though all the elements can be present in a speech production, only part of them produce the actual vocal profile of the characteristics.

4.2 Listener reliability

In a perceptual analysis, in addition to all the variability factors already present in speech a prominent one comes to play: the listener. Further variability added by listeners may represent an issue which researchers cannot underestimate. The variability can affect both actual perceptual models of listeners, such as the expertise in speech domain. However, more trivial variations such as a lack of focus due to stress or tiredness may have a high influence on the results. For instance, performing multiple sessions for the same experiment is a common useful strategy to understand how reliable responses can be. On the other hand, having an heterogeneous group of listeners capable of representing a large variability population is also helpful to reduce control variable effects.

[Sorin, 1981] discusses the importance of a perceptual analysis, especially in terms of a heterogeneous population, in order to establish basic hypotheses and perspectives on which to examine speech features.

In this respect, [Chhabra et al., 2012] assessed the ability to discriminate between different speakers in people with schizophrenia (including 33 with and 32 without auditory hallucinations) compared to 32 healthy controls. As mentioned hereinabove, and reported by the Authors, voice discrimination may be preserved in listeners with schizophrenia, linked with deficits in processing vocal emotion, given the separation between the treatment of identity and other information present in voices. The participants rated the degree of perceived identity similarity for pairs of unfamiliar voices pronouncing three-syllable words, responses were compared with a MDS of the dissimilarity matrices and correlated with acoustic measures. A two-dimensional perceptual space was shared by both schizophrenia patients and controls, with axes corresponding to f0 and FD. Patients with schizophrenia did not differ from healthy controls in their reliance on f0 in differentiating voices, suggesting that the ability to use pitch-based cues may be relatively preserved in schizophrenia. On the other hand, patients (both with and without auditory hallucinations) made less use of FD. This suggests differences in the extraction of other information from voices by people with schizophrenia, since FD has been linked to perception of dominance, mas-

culinity, size and age in healthy individuals.

The issue on listeners reliability is studied in [Belkin et al., 1997], where auditory responses to different pitch stimuli are paralleled with odour quality perception in multiple sessions. Consistent responses are shown within the 32 subjects, 15 female and 17 male. Validation of the responses is given by 29 subjects on a further experiment. The consistency of between-subjects' responses raise the question of whether the auditory-olfactory link is merely a sensory domain or represents a more symbolic one. Similar considerations can be done considering the consistency of responses with the audio-visual association examined in the well known bouba-kiki effect [Peiffer-Smadja and Cohen, 2019]. Abstract representations studied by sound symbolism provide effective examples for the understanding of shared perceptual behaviour, with a consequent interplay of both patterns and parameters.

Voice similarity perception of normal and pathological voice pairs is presented in [Kreiman et al., 1990] with regards to listeners expertise in voice quality perception. Experts and naives differ in the characteristics their similarity judgements rely on. For pathological voices, experts' most important characteristics are correlated with f0, breathiness and H1-H2, in second place are found shimmer measures and rated roughness. Moreover, experts' ratings of pathological voices pairs similarity is explained in terms of breathiness (H1-H2), measured and rated roughness and f0. Besides, naive listeners judged the similarity of pathological voices more consistently in terms of differences in f0, jitter and roughness, breathiness playing a role only for part of the listeners. Overall, expert listeners pay more attention to breathiness and roughness than naive listeners do, the naive listeners relying more on f0.

These findings are further analysed in [Kreiman et al., 1992]. 10 experts compared to 8 naives listeners evaluate 18 male non pathological and 18 male pathological speakers' sustained /a/. Listeners' responses are compared with MDS on f0, formants, HNR, jitter, shimmer and H1-H2 in order to evaluate pertinence of these features for listeners. The expert group perceived the similarity of 18 pathological voices in terms of f0 and H1-H2, with f0 being the most important dimension in this group, 30 % of the variance for dissimilarity ratings. Individual responses varied considerably and no factor was common to all individual solutions despite f0 appearing In 9 of 10 perceptual spaces. For the naive group pathological voices are perceived in terms of f0, F1 and H1-H2. Some individuals differ from this pattern, however, differences are not as marked as for the expert group. As a group, expert listeners perceived the normal voices in terms of f0, shimmer and formant frequencies, while naive listeners relied mainly on f0. Similar within-group variations are observed, with less inter-rater reliability for experts.

Concerning perception and automatic domain comparison, as mentioned in the previous chapter, during NIST SRE a special campaign is devoted to human-assisted systems. These systems could incorporate large amounts of automatic processing with human involvement in certain key aspects, and could be based solely on human listening or somewhere in between. No restrictions are applied to the person involved in a system's decision, neither in terms of number or expertise. This is done to allow a wider perceptual response both from professionals in speech processing and naive listeners. The evaluation plan notes, however, that HASR should not be considered to be a true or representative forensic test. Indeed, many of the factors that influence speaker recognition performance and that play a role in forensic applications are controlled in the HASR test data.

In terms of evaluation, both time devoted to training/test and to pair score-decision are required. However, the decision has to be either true or false, indicating whether or not the same speaker appeared in the training and test segment. On the other hand, for the numerical score, a higher value indicates greater confidence in the fact that the two speakers were identical. Because of the smaller numbers of trials in comparison to main NIST SRE, a simple approach to scoring is adopted, with no cost functions based on miss and false alarm rates. Rather, for each system, the number of correct detections and correct rejections have to be reported. Nevertheless, no significant improvement is reported by this approach throughout the years, both in terms of system adaptation or for further understanding of listeners' perceptual behaviour. An extensive review on this particular human-machine interaction is given in [Greenberg et al., 2010].

4.3 Voice parades and clustering

The study of voice perception implies the study of both the brain with its mechanisms and the distribution of speaker information in speech components. We have reiterated the high variability brought by the latter's multiple implications. However, the multitude of underlying factors that may influence speech perception, e. g. expertise of the listener in speech domains or familiarity with the speaker, have non negligible consequences on speakers' characterisation.

Thus, many studies investigate weights of single components and rely on examining the information they convey rather than on the treatment of the recognised pattern. Even if the first approach is fundamental in the definition of characterisation, the latter process is of fundamental importance in order to provide new results on speaker characterisation. As mentioned in the previous chapter, different outcomes can be achieved by applying different methods of speaker recognition, e. g. verification, identification, diarisation. In this perspective, works such as [McDougall et al., 2015; Kelly et al., 2016b; Smith et al., 2020; O'Brien et al., 2021] examine the influence of task presentation on listeners' performance.

[Lindh, 2009] compared a total of 240 participants, from three age groups, equally distributed in target-present line-up and a target-absent line-up for nine male speakers. Both tasks showed the presence of a wolf speaker generally judged most similar to the target speaker than the actual target. Further comparisons between human listeners responses and those from an ASR GMM-UBM system are provided in [Lindh and Eriksson, 2010]. The results show a correlation between scores obtained from the automatic system and the judgements by the listeners. The latter shows more sensitive to language dependent parameters such as speaking tempo, while the former only bases its similarity spectral information.

In [O'Brien et al., 2021], two experiments are performed in order to compare listeners' SID performance across different tasks. A total of 35 French native listeners, 27 female and 8 male, participated in the first experiment, while 19 listeners participated in the second one. In the first experiment, the participants had to complete a target-lineup and a verification tasks. Target-lineup tasks are commonly used in forensic investigations. Listeners are presented with a target speech sample and a set of suspect ones in which the target speaker could be present or not, i. e. 1-out-of-N or out-of-set. See Figure–4.2 for an interface example. In the second experiment, listeners performed a clustering task.

Speech material from the PTSVOX corpus was used for both experiments consisting of utterances of 1 to 3 s from the read speech of 10 female and 10 male native French speakers.

Figure 4.2: Figure 1 from [O'Brien et al., 2021] Target-Lineup trial interface.

Overall, verification and clustering tasks show higher accuracy, with male speakers' results always higher than female results in both tasks. The target-lineup task shows an increase in false positives in comparison to the other ones, due to the possibility that the target speaker was absent from the lineup. Participants tend to give higher inset answers when presented with the out-of-set option. Application of Pearson correlation to perceptual SID performance, accuracy and task-dependent temporal-based metrics, revealed similar trends across tasks. The target-lineup task allowed participants to listen an unlimited amount of times to the speech samples. This may be another cause of the lower performance, with additional listening adding noise to the initial speaker models that the listeners created. Although, the clustering task shows promising results and potential, mainly because of its not restrictive nature allowing listeners to engage with the speech materials freely and employ different listening strategies.

In a forensic context, this raises the question of reliability of the task. Similar assumptions are presented in [Smith et al., 2020], where 92 participants, 69 female and 23 male, are tested via stimuli taken from the Dynamic Variability in Speech Database [Nolan et al., 2006] using a target-lineup task. The influence of factors such as sample duration and target presence are examined regarding identification accuracy and self-rated confidence of participants. The main objective is to address task adaptation research for voice parades in order to make them easier to conduct and to support earwitness performance, following the procedures used in England and Wales. Overall, in 1-out-of-N options, participants correctly identified the target voice with 39 % accuracy. However, when the target was absent, participants correctly rejected the parade 6 % of the time. No relationship between accuracy and confidence was observed. Consistently higher performance is observed in a sequential procedure, in which participants made a decision after listening to each voice. This findings highlight the idea of different procedures for voice parades with the potential to increase conviction rates and to reduce the risk of miscarriages of justice in cases involving voice identification.

In a similar perspective, [McDougall et al., 2015] explores the influence of telephone and studio recordings on performance in voice parades. In this experiment listeners first were familiarised with the target voice and then undertake the voice parade task. Two groups of 25 listeners in mismatched conditions and two groups with the same conditions participated. The results show that exposure to a voice recorded in studio quality followed by a studio quality voice parade led to correct identifications in 76 % of cases, while

telephone quality exposure and voice parade produced correct 64 % of identifications. The mismatch conditions of studio-telephone and telephone-studio exposure/parade gave 60 % and 32 % respectively, showing the important influence, discussed earlier for automatic systems, that transmission channels have on speakers' modelling even for human perception.

Furthermore, [Kelly et al., 2016b] shows that human SID performance correlates strongly to ASV models trained with source and filter features, i. e. f0, F1-4. Within an i-vector speaker recognition system the named features and their derivatives are used in order to identify cohorts of perceptually similar voices. A total of 43 listeners, 25 male and 18 female, participated in a voice comparison task, in which 30 comparisons presented in a random order had to be ranked from 1, very different, to 9, very similar. Besides, in [O'Brien et al., 2020] significant correlations are observed between cosine distance scores from a custom ASV system based on i-vectors and the accuracy of participants in a clustering task. French read speech from the PTSVOX corpus is used and a total of 24 listeners participated. Similar results are shown in [Gerlacha et al., 2020], comparing voice similarity estimates of ten English speakers pairs by 106 listeners with an i-vector based ASR responses. English and German native listeners' judgements are both correlated with ASR assessments, however, the latter showed a lower correlation.

The varying performance produced by the automatic system based on the demanded task has been the object of description in the previous chapter. However, as observed by [Kinnunen et al., 2000; Han et al., 2012; Lukic et al., 2016], the choice of clustering algorithms is as important as the tuning of parameter extraction and task adaptation for automatic systems, in the same way it is for human listeners.

4.3.1 Familiar voices

Task presentation, as said, is a fundamental factor to be considered when studying perceptual SID, because of the multiple mechanisms our brain is capable of. In the construction of an ASR system the choice of the algorithm and its construction provide as outcome, in most cases, an objective task response. Though, when humans are required to perform a recognition task many subjective factors may influence the task processing and its outcomes. Another important factor to be considered is the familiarity of the listeners with the voices taken as test samples and between the tested speakers.

For instance, the latter case is examined in a forensic perspective in [san Segundo et al., 2017] using 54 different speaker comparisons, 54 same speaker comparisons and 12 comparisons between monozygotic twins. Euclidean distances based on voice quality and source and filter characteristics are compared to VPA-based perceptual analysis by 15 subjects. Features have been extracted from pause fillers, which are long enough for robust feature estimation while spontaneous enough to be extracted from samples in forensic casework. The results from the acoustic analysis revealed that the differences between different and same speaker comparisons were significant in both high quality and telephone-filtered recordings, with no false rejections and limited false acceptances. In line with the literature on twin voices, mean distances for twin pairs perform as the average between the average same and different speaker comparisons. VPA is used for perceptual assessment of speaker similarity. The main purpose of this perceptual analysis is to explore the use of features combination for forensic speaker comparison, assessing

their potential via a perceptual analysis. The correlation between the two methods, even with a small number of listeners, suggests the perceptual saliency of selected features. Furthermore, individual scores show a high resemblance between related speakers both acoustically and perceptually.

Using speech rate and other temporal features such as speaking tempo, [el Gamal, 2015] demonstrate on 60 subjects that listeners with high familiarity with the target speaker show more consequent results in SID. These results confirm the findings of studies such as [Hollien et al., 1982; van Lancker et al., 1985]. The latter uses 45 famous voice stimuli presented forward and backward to 94 subjects. Listeners recognised an average of 26.6 % of the voices from 2 s forward samples presented without response alternatives. Given six choices but maintaining the duration, correct identification raises to 69.9 %. Besides, 4 s backward samples, with six choices, show 57.5 % of correct answers.

Perturbation induced by presenting spoken material backward affects speech characteristics at multiple levels, from acoustics to language identity. Despite this, pitch curves are reversed, as are the temporal structures, other kinds of information are retained, e. g. pitch range and vowel quality information. The results show that some familiar voices are nearly unrecognisable backward, while for others nearly no score degradation is observed. This suggests that information important to the recognition of one voice may be expendable in the case of another. Loss of one parameter does not impair recognisability if a voice is sufficiently distinctive on another dimension.

In [Hollien et al., 1982], some fundamental findings are reported. 2.5 minutes of read English speech from 10 speakers are used on a total of 71 listeners, divided in 3 groups: first, extremely familiar with the speakers; second, unfamiliar with the speakers but familiarised through a training phase before the actual perceptual task; finally, unfamiliar with both speakers and spoken language. The task involved three conditions, namely, normal speech, stress induced by an electrodermal response and whispered speech. Listeners from the first group show 98 % of identification in normal speech, while the middle group score at 40 % and less than 30 % for the third one. Overall, listeners whose familiarity with the tested speakers is important can be expected to identify them at very high levels of accuracy even during stress-induced speech, 75 %. Moreover, listeners show rapid adaptation to unknown speakers showing well above chance level results through a less than 10 minutes training in total. Attempted voice disguise, by the means of whispered speech here, causes confusions to members of every listeners' group. It is more pronounced for listeners who are unfamiliar with both the speaker's speech and language.

Familiarity with a voice involves more than knowledge of acoustic variability. The idea of familiar voices as unique patterns, such that a given feature may be essential for recognising one voice, but irrelevant for another is highlighted by these results. It is suggested as well that learning to recognise a voice involves learning the specific manner(s) in which that voice varies around its prototype, the mental representation of a familiar voice. Variability is essential to learn in the same way that it is essential to learn faces or other categories.

4.4 Conclusion of the literature review

Tackling a vast subject such as phonetic characterisation of speakers is not a simple task which requires investigations across multiple domains. The aim of the first part of this thesis has been to provide an overview to the different aspects covering speaker characterisation. In order to conclude this part and to give a preview on what the next part offers, in the present section we summarise the main aspects our investigation retains from the literature review.

In Chapter 2 we fixate how the phonetic literature has represented the idea of speech components by the means of phonetic measurements. Source and filter represent undoubtedly the most studied component, in particular because of its simple correlation with acoustic elements. Formants and f0 are the main focuses of studies taking on this component. Prosody and voice quality represent more heterogeneous sets into which multiple cues have been related. The first mainly studies temporal cues and changes of intonation, while the second has its focus on laryngeal and supra-laryngeal settings that influence both the harmonic and noise during speech production. Articulatory measurements are as well part of another wide domain which has shown important results concerning between-speaker analysis.

As mentioned before, this last particular set of components is not part of our further investigation since the corpora we focus on do not include any articulatory data. A comprehensive list of the components that we select for this study are listed in Table 1 of the next chapter. They aim to provide a global representation of speaker characteristics. Some of the selected phonetic measurements have already proven consistent results. However, an extensive study on all of their interactions has not been provided yet for French. In addition, a large part of the studies focused on these components used read speech, while our major focus is on spontaneous production.

Chapter 3 fixates the multidisciplinary approach that this thesis embraces. Speech-related studies do not enclose solely to a phonetic approach. While it is fundamental when studying speech data, the applications are countless, from the use of forensic experts speech knowledge in casework to automatic domains focusing on transcription. From both these domains we take inspiration in the following chapters. From the forensic domain it is mainly the focus on the representation of dynamics. The use of CNN has, however, a fundamental motivation, apart from the promising results they have shown, the fact that phoneticians have been working for decades on spectrograms to retrieve information about speech, which is done in forensic cases and also by CNN processing. The wide range of possible metrics used to evaluate systems and methods is the additional element from which we take inspiration from the automatic domain in order to find a way to further understand the data we test.

Finally, Chapter 4 fixates the importance of human perception and feedback regarding scientific findings. Reliability of human listeners is an idea that crosses multiple domains, since it is important in both automatic and perception perspectives in order to validate the phonetic information we analyse. In order to understand what machines are capable of, the comparison with human ability is fundamental. Moreover, what information can really be retrieved by humans from the speech signal and how to represent it are important questions that need to be answered in order to increase reliability of speech research. Interpretation of both results and selected components is impossible without the addition

of human-related results providing a connection between all the studied domains.

The next moment in our investigation takes all these *fixed* elements in an unstable place, which is the actual scientific investigation. The *instability* is the consequence of testing knowledge and methods through untested data in order to obtain consistent and newly informative results.

Part II

INSTABILITY or Results

Chapter 5

Perception

The second part of this thesis focuses on summarising all our results. Chapters 6 and 7 provide an overview of the study of speech components through different approaches aimed at understanding the phonetic characterisation of voices in more depth. The analyses performed in these chapters contribute to expand the basic aspects of phonetic characterisation we assess in the previous chapter. The results from the phonetic and CNN approaches highlight voice quality or formants as the most reliable speech component to characterise speakers' voices. In this chapter, we analyse voice characterisation from a perceptual standpoint. Looking at the responses of human listeners on voice clusters, we compare them with those obtained with phonetic- and CNN-driven modelling in order to study their possible differences and similarities.

What differentiates this chapter from the others is the presence of an additional variability factor, fundamental for the analysis of perception, the listener. The speaker is present as the only factor of variability in our further investigations of voice characteristics interactions. However, in a perception study, the listener represents a fundamental and important factor of variability to take into account. Human perception, especially when studying voice characterisation, can be influenced by a multitude of variables such as the listener's familiarity with the speakers, or the language being spoken as well as the task design in general. In the following introductory sections, we describe our perceptual task design and the different groups of listeners who participated. The listeners in the present study were free to refer to any voice feature that seemed relevant to them to group or differentiate the given stimuli from each other, hence they were not forced to restrict their judgements to a single feature that might provide information about the speakers.

The principal aims of this chapter are: firstly, to understand how robust are the phonetic measurements used to represent speech components when compared to human responses; secondly, this study assesses whether using a clustering task can account for a further understanding of human perception of voice characteristics. Comparing perceptual responses and clustering from speech measurements allows to improve the investigation on how individual strategies of components' modulation convey speaker information.

5.1 Methods

Table 5.1 recapitulates all the extracted phonetic measurements and the corresponding group names we use throughout the result chapters to define them. The related component is reported as well. Similar tables are present in Chapter 6 and Chapter 7 in order to remind these elements for each result chapter. We mainly used Praat [Boersma, 2001] and VoiceSauce [Shue et al., 2011] softwares to extract the selected measurements. The latter provides mainly voice quality descriptors, lacking in Praat, and offers the choice of acoustic measurements from several external programs for parameters such as fundamental frequency and formants.

Group name	**Measurements**	**Component**
Amp	Amplitudes of the harmonics near the first three formants (A1-3)	Source and filter
f0	f0 and its harmonics (H1, H2, H2k, H42k, H5k)	Source and filter
Form	First four formants (F1-4)	Source and filter
Acoust	Combination of *Amp*, *f0* and *Form*	Source and filter
Rhythm	Intensity, ENV and TFS	Temporal
Nrg	RMS, soe, Praat-based energy	Source and filter | Mode of vocal fold vibration
Ms	Four spectral moments (center of gravity, standard deviation, kurtosis, skewness)	Source and filter | Mode of vocal fold vibration
Hadiff	Differences between harmonics, and between harmonics and amplitudes (H1-A1, H1-A2, H1-A3, H1-H2, H2-H4, H2k-H5k)	Mode of vocal fold vibration
Hr	HNR at different pitch ranges (0-500 Hz HNR05, 0-1500 Hz HNR15, 0-2500 Hz HNR25, 0-3500 Hz HNR35), SHR	Mode of vocal fold vibration
Ltas	LTAS from four different frequency bandwidths between 1 and 5 kHz	Mode of vocal fold vibration
Qual	Combination of *Ltas*, CPP and *Nrg*	Mode of vocal fold vibration
MFCC	13 MFCC	?
Glob	Combination of all phonetic measurements	All
Spectros	Entire wide-band spectrograms	All

Table 5.1: Phonetic measurements groups used in Chapter 5 for the clustering experiments and related components.

Applying the same extraction method in both softwares, we obtained values every millisecond to have a precise recreation of parameters modulations for each target speaker. We took measurements on voiced consonants for parameters such as energy and included fricatives for spectral moments. Some groups use larger analysis windows, e. g., 10 ms for MFCC.

Concerning the groups of measurements: the first four formants have been obtained with VoiceSauce via Snack as in [Keating and Kreiman, 2016; Keating et al., 2017], comparing them with the extraction obtained with Praat using methods described in [McDougall and Nolan, 2007], we observed small differences in values; for intensity and CPP, Praat default settings have been used; fundamental frequency has been computed as in [Hudson et al., 2007; Nolan et al., 2011] and, like the formants, compared with the values obtained in VoiceSauce showing the same results; for all harmonics computation we used VoiceSauce; for temporal envelope (ENV) and Temporal Fine Structure (TFS) we followed the computation described in [He and Dellwo, 2016] through the Hilbert transform; energy have been computed using three different measurements, Root Mean Square (RMS) and Strength of Excitation (soe) in VoiceSauce and energy from Praat which corresponds to the integral of sound amplitude power; the Long-Time Average Spectrum peaks (LTAS) and mean values have been obtained on Praat; for the HNR and SHR computations we followed the VoiceSauce default values. In order to understand the features interactions and their weights inside each group we used modified subsets for each one during the automatic approach studies. See Chapter 7 for a more extensive analysis.

Statistical descriptions of these different groups help to understand how, from a descriptive standpoint, speakers have similar or different characteristic distributions inside their variability matrix. The comparison of results from Deep Learning approach and classic phonetic through statistical description are in accordance with the already discussed need for explicability and interpretation of somehow cryptic results from the automatic approaches. Further comparison with results from human-based studies can be an additional step to understand how and why some cues are more relevant than others in characterising a single or multiple speakers.

5.1.1 Corpora presentation

The two French corpora used throughout our studies are the Nijmegen Corpus of Casual French (NCCFr) and PTSVOX, first documented respectively in [Torreira et al., 2010] and [Chanclu et al., 2020]. They present some similarities such as the presence of Native French speakers from both sexes and similar age ranges. However, their linguistic materials are very different since NCCFr is composed of spontaneous speech recordings while PTSVOX consists of read speech. Further analysis on lexical distances between speakers in NCCFr are presented in Section 7.2.

NCCFr is composed of 46 speakers performing multiple speech tasks, including reading and a casual conversation between two of them. In our study we only consider the conversation task for 44 speakers (21 females and 23 males) since they are the only one with available annotations. The recordings have an average duration of 40 minutes performed in a single session inside a quiet chamber in the Phonetics and Phonology Laboratory of Paris and all subjects wore headset microphones in order to reduce intensity variations from head movements. Semi-automatic transcription has been made by LIMSI

laboratory [Barras et al., 2001]. Speakers are mainly aged between 18 and 27 with two females being 40 and 50 years old. They are all French Native speakers, thirty-four of them came from the Paris region, while the remaining ones came from other regions in Central and Northern France. All speakers came from the same social background and pursued similar studies. The conversation task always associated same sex speakers.

The second corpus used in our studies is the French corpus PTSVOX, aimed to recreate the main challenges of forensic voice comparison. Spontaneous and read speech are represented in this corpus and recorded in both microphone and telephone conditions. Since the PTSVOX was created during the work of this thesis, annotations were not immediately available for all the recordings. For this reason, we only considered the read subset for our experiments. It is composed of two sessions of three read passages by a total of 24 French Native speakers (12 females and 12 males). Speakers who presented incomplete sessions, e. g., they lacked one or more of the read passages or an entire recording session, have been discarded. The size of the smaller subsets we used, vary from 15 to 21 speakers depending on the protocol used. Speakers are in the same age range as NCCFr and have similar intra-corpus social background.

Throughout the results description, we use the nomenclature for speakers relative to each corpus: in the PTSVOX speakers are named using codes LG001 to LG024; while for NCCFr speakers' names present information about their sex and interlocutor, female speakers start with a F, males start with a M, interlocutors present the same number and are distinguished by a L or R as the final letter of their code, e. g. M01L, M01R and F14L, F14R.

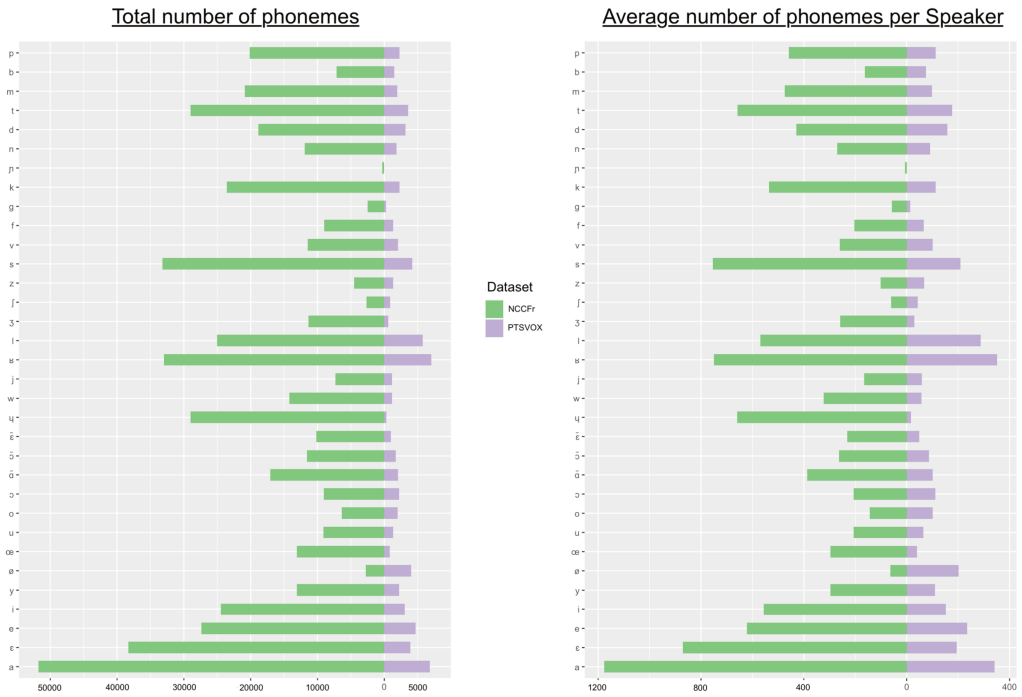


Figure 5.1: Total amount and speaker average of number of phonemes for NCCFr (in green) and PTSVOX (in blue) corpora in the studied sequences.

For both corpora, automatic segmentation has been human-validated and corrected, and present, at least, word and phoneme level annotations. Figure 5.1 shows the total number

of phonemes in both corpora and their average. Even if in our studies we do not take into consideration the phonetic content influence, it is important to know which phonemic classes are more prominent. As mentioned, speaker recognition studies have shown how different phonemic classes can be influenced by the speaker factor when comparing the amount of information they carry about speaker characteristics. The studied characteristic has an important role, some are considered to be more representative of the individual differences, indeed not every measurement is possible on every phonemic class.

In all our experiments we segmented the recordings into smaller chunks of similar duration. PTSVOX has been used in preliminary studies with sequences of multiple lengths, see next chapter's Section 7.3, and in an experiment involving comparison of the same read sentences, while for NCCFr we used 4 s sequences with a minimum of 20 phonemes each.

5.2 Perception task

In the principal perception task, listeners were asked to regroup similar speakers in a number of clusters of their choice. Figure 5.2 shows an example of the screen listeners were faced with during the experiment. The listeners could undertake the two-sessions task on their personal computer via a HTML based interface. They were asked to perform the two sessions in an interval of at least one week, in order to reduce any memory bias, a small number of them performed the two sessions the same day. There was no time limitation to complete the task, we were interested in having a reliable answer rather than a quick but unsure one.

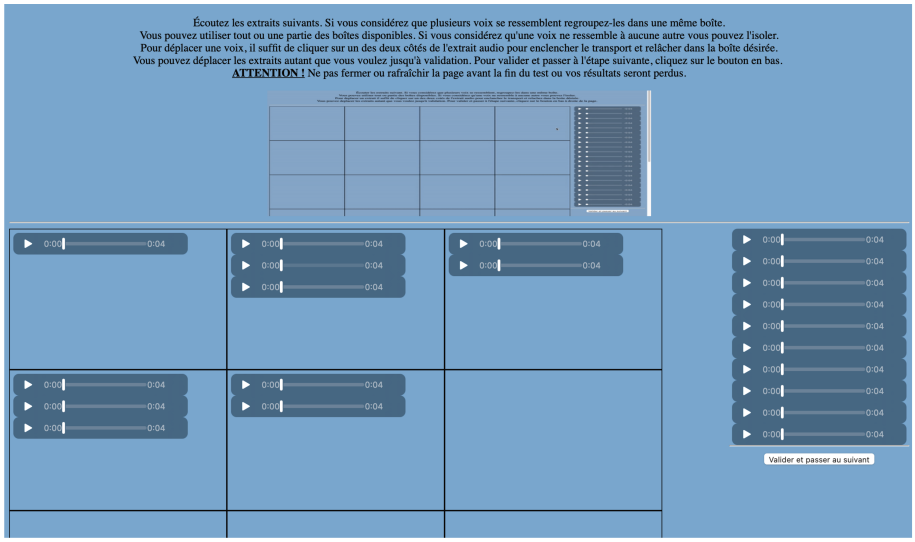


Figure 5.2: Perception task, test phase screen example.

The recordings used were the same as those used in our further experiments, see Chapter 6 and Chapter 7. We used the CNN results as a reference list in order to not have chunks that have been wrongly classified. Every NCCFr's speaker had a total of 6 chunks of 4 s which were randomly presented to the listeners and never repeated twice during the test and train phases of the task. As shown in Figure 5.2, in the upper portion of the screen listeners had the instruction and a GIF showing how to perform the task. In the actual task portion, all recordings were stored on the right, listeners could play them

indefinitely, on the left side they could find a number of empty boxes corresponding to the number of recordings. The listeners had to regroup similar voices inside the same box with no limitations on the number. Voices that were considered totally different from the others could be placed in an isolated box as well. Once they thought their answers were definitive, through a button on the left listeners could store the clustering and initialise the following phase. Stored answers could not be accessed or modified.

Concerning the enrolment of the task itself, in both sessions a *train* and a *test* phase were present. The train phase consisted of two clustering screens, one per sex, while in the test phase there was three, two of them with chunks from the same sex group. This way, every listener performed four test phases and six train phases with half of each having female voices and the other half male voices. During the train phases listeners were confronted to a reduced number of voices in order to familiarise them with both the voices and the task itself. Results from these phases were used as an additional mean to evaluate less reliable listeners in our further analysis, since two speakers had repeated chunks. Listeners that failed to regroup the same speaker in more than half of the train phases were discarded. The listeners did not receive a direct feedback on their answer.

In the following sections, we respectively report the information gathered on the listeners which serve as control variables and discuss the clustering methods and evaluations to assess the reliability of these groups and answers.

5.2.1 Participants

The participants to the perception experiment have been recruited via different means with half of them being students or researchers from the same laboratory as the Author, Laboratoire de Phonétique et Phonologie. The other half is composed by Author's family and surroundings or anonymous who responded to a Reddit post. A total of 31 listeners participated in the task, some basic information was requested in order to establish post analysis groups. We discuss here all the information and the considered control variables, an extensive reference table is reported in Appendix D. The only limitation imposed for the participation was the understanding of French. Following the results from the train phase, 4 female listeners were discarded, hence the answers from 27 listeners have been analysed in our study.

The age of the participants ranges from 21 to 77, with a mean of 27, a median of 31 and a SD of 11, 21 of them are females against 6 males. The latter's results, as for the French Non-Native listeners, have to be taken with caution and considered as possible trends or hypotheses, given the small sample we have been able to gather. None of the participants declared hearing problems. Mean duration of both sessions was 36 minutes with a minimum of 9 and a maximum of 1 hour and 59. Apart from **sex**, two other control variables were considered to compare different groups: the mother tongue; and what we named the expert status, corresponding to whether or not listeners had an education in Phonetics or Speech-Language Pathology. There are 10 French Native **Experts** in total, all females, in addition, one female expert was non-Native as were two males. The group of non-Experts, or **Naives**, is composed of 9 female and 4 male listeners. The **Natives** group is composed of 23 listeners, 19 females and 4 males. The 4 **non-Natives** all had experience in higher education in French and are 2 females, Arab and Chinese Natives speakers, and 2 males, both Chinese. Influence of non-Native speakers of French

on the results is discussed in the following analysis. Their proficiency in French and their reliability were sufficient factors to not discard them.

Crossing all the control variables, we obtain a total of 23 clusterings which, multiplied by the two tasks (female and male voices), give us 46 clusterings to compare. As we mentioned, some groups were composed of a larger number of listeners, indeed, from a scientific standpoint, we consider that smaller groups are less reliable as they are less representative. In the next session we describe the clustering methods and the evaluation metrics we used.

5.2.2 Clustering methods and evaluation

The principal aim of this chapter is to study the relation that might exist between the human answers and those obtained through the studies described in Chapters 6 and 7. As discussed in Chapter 4, clustering analysis is a common solution in order to assess similarity between voices. Looking at this process through the lens of voice characteristics, we can extend the definition to: assessing similarity of voice characteristics.

In many of the previously discussed studies on voice characterisation, authors commonly alter the speech material in order to focus the listener's attention on specific features, see Chapter 4. We took the decision to have a different approach in which listeners were free to choose what characteristics they would focus on. All answers from the same group of listeners are combined in order to create an *ideal* clustering meant to represent the perception of each group. The 46 different human clusterings are compared to those obtained by the phonetic and CNN approach, groups of measurements reported in Table 5.1. We use two types of metrics aiming to validate the clustering calculation and compare the different clusterings between them.

Taking the matrices of human, phonetic and CNN answers, we compute the Euclidean distance matrices. Afterwards, we apply a hierarchical clustering analysis (HCA) using ward criterion in order to minimise the loss of information when comparing pairs of points and creating the different clusterings. An important step in clustering validation is finding the ideal number of clusters to use. This was achieved by analysing the results of three performance metrics in R: the clustering coefficient, which is a global score on the clustering validity; the silhouette score, which computes the minimal difference inter-clusters; the gap statistic, which focuses on minimising the intra-cluster difference.

Concerning the evaluation of similarity between the obtained clusterings, we used two metrics, the **Jaccard similarity coefficient** [Levandowsky and Winter, 1971; Moulton and Jiang, 2018], a value of 1.0 corresponds to a perfect resemblance between the compared groups, and **Cohen's Kappa**. The latter is the same used when comparing multiple CNN models, here we use it to have an a priori comparison score for the different human clusterings. The Jaccard coefficient is used to measure the similarity and diversity of finite sample sets, even when different in sizes, by taking the ratio of Intersection over Union. Human clusterings are compared both intra- and inter- groups to assess both listeners and group reliability as well as inter-group variability. In both CNN and phonetic clusterings we evaluate the distance between the different speech components groups. Finally, these metrics results are also analysed by the scope of the listeners' direct feedback on their own clustering strategies.

5.3 Clusterings comparison

In the following sections we report the results of the multiple clusterings. The first section corresponds to the perceptual answers by human listeners, the second corresponds to the phonetic results, and finally the results obtained through CNN's processing. In these sections we compare the results within the 3 main clusterings, namely PHON for Phonetic clusters, CNN for the CNN and HUM for human perception. The focus of discussion in the last subsection is the similarities between all the clusters. As mentioned, Jaccard's similarity and Cohen's Kappa coefficients are used to compare the answers of the different groups we study.

5.3.1 Human clustering

Tables 5.2, 5.3 and 5.4 summarise the results obtained through the perception task by human listeners. These three tables refer to the main subgroups used to analyse the results: the first one presents responses from both sexes; the second one only the responses from female listeners; and the third one only from male listeners. As described in Section 5.2.1, the results for male listeners should be considered as less reliable because they are less numerous than women. This is confirmed by both Cohen's Kappa and clustering coefficient values. The first has always a score under 0.10 for all male listeners responses while for other listeners groups we always observe a score around 0.25 and higher. These results indicate that, when analysing the resulting matrix, the considered classifier, male listeners in our case, do slightly better than chance. The clustering coefficients are higher than 0.68 for all groups of listeners except for males where the score average is around 0.60.

	All	**Experts**	**Experts non Native**	**Naives**	**Naives Native**	**Natives**	**Non Natives**
All	1.0	0.53	0.30	0.55	0.56	0.71	0.30
Experts	0.64	1.0	**0.20**	0.56	0.60	0.54	**0.20**
Experts non Native	**0.19**	**0.24**	1.0	0.22	**0.20**	0.23	**0.87**
Naives	0.64	0.56	0.27	1.0	0.66	0.65	0.23
Naives Native	0.70	0.58	0.27	**0.89**	1.0	**0.73**	**0.19**
Natives	0.70	0.53	**0.24**	0.64	**0.72**	1.0	0.23
Non Natives	0.46	0.50	**0.24**	0.53	0.49	0.47	1.0

Table 5.2: Jaccard similarity coefficients for couples of HUM clusters, red for female speakers (top-right from the diagonal) and blue for male speakers (bottom-left). Clusters from listeners of both sexes.

5.3.1.1 Within human clustering comparison

As described earlier, expertise in Linguistics or Speech-Language Pathology and Native language are the two major control variables we consider aside from listener's sex. In all three tables (5.2, 5.3 and 5.4) the reported values result from comparison of clusterings from different listeners subgroups. When considering the female voices task, scores from both sexes listeners show that the most similar comparison is Non-Native Experts with Non-Natives with 0.87. The second most similar cluster combination involves Native Naives and all Native listeners with a score of 0.73. The least similar clusters appears to be the Experts group compared with both all Non-Natives and Non-Native Expert listeners which register 0.20. The Non-Natives groups also have low similarity with Native Naives, with a score of 0.19. Concerning the male voices perception task, we observe high similarity rates between the Native Naives group and the Naive group with a score of 0.89 as well as the Native Naives group compared with the all Native, with a score of 0.72. The least similar clusterings are those obtained by Non-Native Experts with a 0.19 score when compared to all listeners and a score of 0.21 when compared with the Experts, Natives and Non-Natives groups.

	All F	**Experts F**	**Experts Native F**	**Experts non Native F**	**Naives F**	**Naives Native F**	**Naives non Native F**	**Natives F**	**Non Natives F**
All F	1.0	0.52	0.49	0.47	0.72	0.66	0.43	**0.80**	0.52
Experts F	0.45	1.0	**0.83**	0.47	0.52	0.47	**0.35**	0.46	0.51
Experts Native F	**0.64**	**0.67**	1.0	0.47	0.49	0.45	**0.33**	0.43	0.51
Experts non Native F	0.39	0.46	0.44	1.0	0.53	0.47	0.39	0.49	0.71
Naives F	0.54	0.43	0.60	0.41	1.0	0.78	0.42	0.69	0.57
Naives Native F	0.54	0.43	0.60	0.41	1.0	1.0	0.43	**0.80**	0.48
Naives non Native F	**0.13**	0.18	**0.15**	**0.15**	**0.15**	**0.15**	1.0	0.44	0.43
Natives F	0.63	0.48	0.63	0.43	0.57	0.57	0.18	1.0	0.50
Non Natives F	0.36	0.53	0.43	0.48	0.36	0.36	0.26	0.41	1.0

Table 5.3: Jaccard similarity coefficients for couples of HUM clusters, red for female speakers (top-right from the diagonal) and blue for male speakers (bottom-left). Clusters from female listeners.

When focusing only on the results for female listeners, in the female voices task we observe something similar to what happens in the CNN clustering groups with the minimum similarity rate being around 0.34. This is shown by the Non-Native Naives group that has a score of 0.33 when compared to Natives Experts and a score of 0.35 when compared with all Experts. The most similar clusterings are those obtained by Experts and Native

	All M	**(Experts) non Native M**	**Naives M**
All M	1.0	**0.20**	**0.65**
(Experts) non Native M	0.46	1.0	0.23
Naives M	**0.69**	0.44	1.0

Table 5.4: Jaccard similarity coefficients for couples of HUM clusters, red for female speakers (top-right from the diagonal) and blue for male speakers (bottom-left). Clusters from male listeners.

Experts with 0.83, considering the fact that only one female expert was Non-Native of French, we note that her answers cause an important decrease of the similarity between the two groups. The other most similar clusterings are those obtained by Natives and both all female listeners and Native Naives with a score of 0.80. In the male voices task, we observe an overall lower average similarity rate, 0.43 against 0.53 for the female speakers task. This is reiterated by the observed scores. Aside from the Experts and Native Experts similarity reaching a score of 0.67 the other most similar clusters are the ones from all female listeners and Native Experts with a score of 0.64. The Non-Native Naives group is the most dissimilar from all other groups with an average similarity rate of 0.15 and the lowest score of 0.13 compared to all listeners.

As already mentioned, male listeners were fewer in number than females and as such their results are scientifically less reliable, however, we still report them for completeness of description, hoping they might be assessed more firmly in future works. Male listeners clusterings of female voices present a lower average similarity, with a score of 0.36, than those from the male voices task, with a score of 0.53. Naives' clusterings show a greater similarity with the group of all listeners clustering, yielding a score of 0.65 in the female voices task and 0.69 for male voices task. We observe that similarity between the clusterings for the female voices task is lower with a score of 0.20 when comparing Non-Native Experts with all listeners and a score of 0.23 when comparing Non-Native Experts and Naive listeners. In the male voices task the same scores increase to 0.46 and 0.44, respectively.

These comparisons suggest how the listener in a voice characterisation task represents a complex variable that can be influenced by a large number of factors, e. g. expertise in voice studies or Native language in our case. The latter can add further complications in the analysis, by the means of how long the listener has been exposed to the spoken language of the listened voice or to what is their actual mother tongue. These and other considerations are part of our discussion in Chapter 8.

5.3.2 Phonetic clustering

The phonetic clusterings are obtained from distance matrices that are computed on the raw phonetics values, divided by component groups as shown in Table 5.1. Once the different groups of speakers have been formed using the HCA method, we compare the resulting clusterings in order to further understand the similarities that exist between

them. Similarity between the groups of clusters is, as mentioned, assessed by the means of a Jaccard coefficient. As a remember, a value of 1.0 corresponds to a perfect resemblance between the compared groups.

Table 5.5 summarises the distances between the different clusterings based on phonetic components, the values on the top right of the matrix's diagonal correspond to female speakers task while the left bottom values are those corresponding to male speakers task. The similarity scores show that the information is not modelled in the same way by the studied components for both sexes, with a few exceptions. Along the diagonal we observe some inversely proportional combination for the two sexes, e. g., Ltas-Hr is the lowest pair for female speakers and the Ltas-Qual pair shows a score of 0.5 while for male speakers in both cases the scores are consistently greater. The most similar clusterings are Form-Acoust and Glob-Ms with scores of 0.64 and 0.52 for female speakers and scores of 0.54 and 0.57 for males. This implies that the information carried by the Ms subset has a high weight on the global representation and in a similar way formants, that are present both in Form and Acoust, have a higher influence on the clustering. The least similar clusterings are MFCC-Rhythm with 0.16 as well as Ltas-Hadiff, Ltas-Hr and Ltas-Acoust with 0.17, showing that the information conveyed by Ltas is very different from all the other components for women while this is not the case for men's Ltas.

It has to be noted that the scores from the two global representations of the PHON groups, i. e. MFCC and Glob, behave in a completely opposite direction for female and male speakers. Their scores, 0.26 in the first case and 0.41 in the second one, confirm the trend of MFCC to be more similar to the considered phonetic components for male individuals. The average similarity score between clusterings for female speakers is 0.30 while it is 0.36 for male. This is similar to what we observe in the PCAs in Chapter 6 showing that the considered components convey more redundant information about male voices than for female.

	MFCC	**Glob**	**Form**	**f0**	**Acoust**	**Hadiff**	**Rhythm**	**Amp**	**Ms**	**Nrg**	**Hr**	**Ltas**	**Qual**
MFCC	1.0	0.26	0.25	0.27	0.24	0.29	**0.16**	0.20	0.25	0.25	0.29	0.34	0.24
Glob	0.41	1.0	0.43	0.27	0.36	0.29	0.28	0.34	**0.52**	0.35	0.30	0.21	0.24
Form	0.34	0.30	1.0	0.28	**0.64**	0.30	0.31	0.36	0.38	0.36	0.34	0.20	0.28
f0	0.27	**0.23**	**0.22**	1.0	0.38	0.33	0.26	0.30	0.34	0.27	0.25	0.30	0.45
Acoust	0.39	0.40	**0.54**	0.25	1.0	0.36	0.36	0.32	0.33	0.33	0.35	**0.17**	0.32
Hadiff	0.49	0.40	0.28	0.35	0.30	1.0	0.25	0.27	0.31	0.36	0.27	**0.16**	0.20
Rhythm	0.33	0.36	0.38	0.27	0.44	0.35	1.0	0.36	0.34	0.32	0.28	0.19	0.29
Amp	0.47	0.36	0.34	0.31	0.33	0.43	0.35	1.0	0.36	0.36	0.33	0.18	0.29
Ms	0.34	**0.57**	0.33	0.29	0.35	0.27	0.34	0.33	1.0	0.36	0.40	0.23	0.29
Nrg	0.38	0.45	0.34	0.25	0.41	0.33	0.40	0.43	0.40	1.0	0.32	0.35	0.37
Hr	0.26	0.32	0.36	0.25	0.38	**0.23**	0.41	0.28	0.33	0.30	1.0	**0.16**	0.25
Ltas	0.35	0.37	0.33	0.26	0.44	0.34	0.42	0.37	0.33	0.44	0.35	1.0	0.5
Qual	0.35	0.37	0.33	0.26	0.44	0.34	0.42	0.37	0.33	0.44	0.35	**1.0**	1.0

Table 5.5: Jaccard similarity coefficients for pairs of PHON clusterings, red for female speakers (top-right from the diagonal) and blue for male speakers (bottom-left).

5.3.3 Automatic clustering

In order to obtain what we call the automatic clustering, CNN from now on, we treat the confusion matrices from the CNN tasks, see Chapter 7, as distance matrices and apply the clustering methods described earlier. We obtain an average score of 0.48 for both

female and male speakers. It appears that more than a third of the information is shared by all clusterings since the minimal distance is above 0.30 for both sexes, but there are no pairs of clusterings that present similarity rates higher than 0.63.

Table 5.6 summarises the results. In detail, we observe that the least similar clusterings for female speakers are in combination with f0, in opposition to PHON clusterings, with f0-Form having a 0.34 score and f0-Ltas a score of 0.33. The most similar clusterings are Form-Hadiff with 0.58, f0-Amp with 0.59 and Nrg-Rhythm with 0.62. For male speakers the least similar clusterings are f0-Ltas with 0.37 and f0-Form with 0.38 while the most similar are Hadiff-Acoust with 0.59 and Nrg-Hadiff with 0.63. The similar clusterings reflect the similarities described by the other scores in Chapter 7, with the Rhythm and Nrg groups having similar important roles in the classification of female voices, while for males we observe that the Hadiff groups shares an important amount of information with the global representation, just as Hr does with all the other clusterings for both sexes.

	Spectros	**MFCC**	**Glob**	**Form**	**f0**	**Acoust**	**Hadiff**	**Rhythm**	**Amp**	**Ms**	**Nrg**	**Hr**	**Ltas**	**Qual**
Spectros	1.0	0.55	0.49	0.52	0.38	0.42	0.51	0.46	0.43	0.47	0.48	0.51	0.48	0.47
MFCC	0.45	1.0	0.50	0.51	0.40	0.52	0.46	0.45	0.52	0.46	0.47	0.57	0.48	0.55
Glob	0.47	0.49	1.0	0.56	0.38	0.51	0.52	0.48	0.44	0.47	0.46	0.48	0.44	0.49
Form	0.45	0.46	0.46	1.0	**0.34**	0.56	**0.58**	0.46	0.43	0.56	0.45	0.55	0.53	0.50
f0	0.39	0.38	0.50	**0.38**	1.0	0.40	0.40	0.45	0.59	0.38	0.50	0.46	**0.33**	0.40
Acoust	0.54	0.48	0.49	0.51	0.41	1.0	0.50	0.43	0.51	0.45	0.41	0.48	0.45	0.59
Hadiff	0.48	0.47	0.51	0.54	0.47	**0.59**	1.0	0.56	0.48	0.59	0.52	0.51	0.53	0.45
Rhythm	0.41	0.44	0.44	0.42	0.39	0.45	0.42	1.0	0.50	0.45	**0.62**	0.54	0.48	0.49
Amp	0.44	0.49	0.50	0.44	0.50	0.52	0.58	0.47	1.0	0.43	0.52	0.57	0.42	0.50
Ms	0.43	0.54	0.45	0.50	0.41	0.50	0.55	0.48	0.44	1.0	0.43	0.47	0.50	0.44
Nrg	0.45	0.45	0.46	0.46	0.49	0.49	**0.63**	0.48	0.51	0.52	1.0	0.52	0.53	0.48
Hr	0.50	0.52	0.58	0.48	0.46	0.51	0.54	0.42	0.48	0.53	0.46	1.0	0.45	0.53
Ltas	0.51	0.45	0.45	0.49	**0.37**	0.55	0.54	0.40	0.45	0.54	0.48	0.47	1.0	0.50
Qual	0.46	0.47	0.52	0.50	0.49	0.53	0.49	0.47	0.47	0.49	0.51	0.52	0.44	1.0

Table 5.6: Jaccard similarity coefficients for pairs of CNN clusterings, red for female speakers (top-right from the diagonal) and blue for male speakers (bottom-left).

5.3.4 PHON-HUM-CNN similarities

Listeners' responses are important in order to explore the variability of human perception in regards to control variables such as those we cited in the previous sections. Nevertheless, the comparison between human performance on voice characterisation and what has been obtained through different approaches can be used as a reference for a deeper understanding of voice characteristics' modelling.

We hereafter summarise the results of the similarity between the three clustering groups: PHON, CNN and HUM. The values reported are listed in the extensive tables in Appendix E. The thirteen PHON and the fourteen CNN clusterings, obtained from the different studied components and global representations, have been compared to the seven, nine and six clusterings obtained with the listeners, corresponding respectively to both sexes combined, female listeners only and male listeners only. This process leads us to formalise how the chosen control variables interact with the proposed phonetic modelling. However, as mentioned hereinabove, results from male listeners only represent hypothetical trends and cannot be considered as reliable as the female listeners' results due to the smaller sample we have been able to gather.

Before describing the three-point comparison of PHON-CNN-HUM clusterings, we report

the similarities from CNN-PHON comparison. To compare these groups we use the same similarity score which varies from 0 to 1, a score of 1 meaning that two clusterings are completely similar. We observe that, between these groups, the average similarity score for the clustering of both female and male voices is 0.34. The female speakers clustering has lower minimum scores around 0.19 for PHON-Ltas when compared to CNN-MFCC, Glob and Form. For male speakers clustering, PHON-f0 has the lowest score with 0.22 when compared to CNN-Form, Spectros, Acoust and Qual. The higher similarity scores we observe for female speakers clusterings are obtained by comparing PHON-Acoust to CNN-Ltas, 0.52, and Qual, 0.50. CNN-Acoust has a similarity of 0.47 with PHON-Rhythm's clustering. For the male speakers task the highest similarity score between CNN and PHON is registered by the comparison of Form clustering with a score of 0.55, the same score is obtained when comparing CNN-Form and PHON-Acoust, which, on average, is highly similar to all CNN answers. Comparing clustering from the same components' subsets but different modelling, CNN or PHON, we observe higher similarities for the male speakers task than for female speakers.

When we compare the clustering by listeners of both sexes to the PHON and CNN clusterings we observe, for female voices, that the most similar clustering corresponds to the clusters obtained with the Experts group. Indeed, the Jaccard coefficient results in scores of 0.46 with PHON-Form, 0.45 with PHON-Acoust and 0.55 with both CNN-Form and Ms. The other most similar clustering is the one obtained with the Naive group, in particular compared with the CNN-Hr clustering registering the second highest CNN score, with 0.54. The Non-Native Experts group reports the most dissimilarity when compared to CNN and HUM, followed by Non-Native listeners, with scores of 0.18 and 0.17 for MFCC, Ltas and Rhythm. The PHON-HUM clusterings comparison, when comparing Acoust to Non-Native and Non-Native Expert listeners, has the lowest scores, with 0.17. For the male voices task, we observe a different score distribution.

For the male voices, we observe a different score distribution. For example, we obtain scores lower than 0.30 when we compare clusterings obtained by Non-Native Experts with those from PHON-Hadiff, however, the lowest rate is obtained when compared with CNN-Ltas 0.16. The highest PHON-HUM similarities are observed with the comparison of Rhythm to Naives and Non-Native listeners, implying that rhythmic cues may have an important influence in voice characterisation for non Expert listeners. The CNN clusterings do not show very high similarity scores with the maximum being 0.48 with Expert listeners clustering compared to Hadiff or Non-Natives with Acoust and Qual.

When considering only female listeners, we obtain similar results, especially for the clusterings resulting from the male voices task, with one HUM clustering reporting the most dissimilar scores, i. e. Non-Native Naives, when compared to both PHON and CNN clusterings. In particular, with that group of listeners, the comparisons with CNN-Ltas and CNN-Rhythm get respectively a score of 0.11 and 0.12, and we obtain a score of 0.14 when compared with PHON-Qual and Rhythm. The comparison of the clusters obtained with PHON and Non-Native Experts are the most similar with Acoust, 0.40, Form, 0.38, as well as Nrg compared to Natives with 0.38. Concerning CNN clusterings, we observe that Native Experts' responses result in greater similarities with Hadiff and MFCC. The score distribution trends for the female voices are less reliable when compared with Ltas, showing low similarity scores with PHON clusterings but high ones for CNN. MFCC and Spectros have the highest similarity scores for CNN clusters with 0.54 and 0.56 respectively, while the highest PHON scores are registered by the Form group with an average

of 0.44.

When we compare male listeners' responses with the PHON and CNN clusterings, we obtain a slightly higher average similarity score, with 0.35 in opposition to 0.31 for the female listeners. The highest observed scores in this case are 0.49 with CNN-Hadiff and 0.46 with CNN-Acoust, PHON-Acoust also registers its highest score in this comparison with 0.41. The Naives listeners group is the second most similar to the PHON clusterings, with the highest score of 0.42 when compared to PHON-Rhythm. The lowest scores are obtained by Non-Native and Naive groups in comparison to Acoust and MFCC clusterings. The Native Naives group obtains the highest similarity scores when compared to CNN-Hadiff, 0.52, and PHON-Rhythm, 0.42, while the same group of listeners show the lowest scores against PHON-F0 clustering. Nevertheless, these results have to be taken with caution because of the small sample of male listeners.

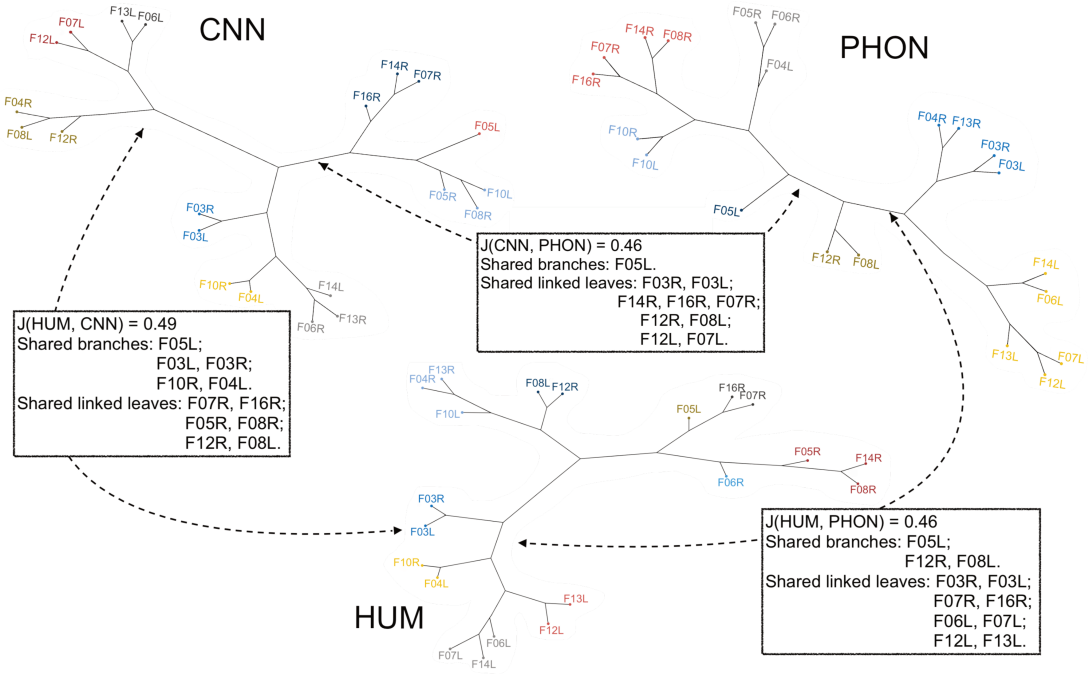


Figure 5.3: Similar clusters obtained from human Native Naives listeners responses in the female voices perception task. CNN and PHON correspond to clusterings based on formants. Colours represent the multiple branches.

Figure 5.3 gives an example of what the Jaccard similarity scores indicate. Three clusterings are represented, one from each major group, with colours indicating different branches, for HUM, the Native Naive group responses for the female voices task are represented. We have taken the clusterings resulting from formants for PHON and CNN since they have a similarity of 0.46 and, respectively, a score of 0.46 and 0.49 when compared to the selected HUM clustering. The different branches of the clusterings are depicted by different colours while every leaf represents a speaker, leaves present on a directly connected branch are considered more similar than others from the same leaf group. Inside the figure, we report the shared branches and leaves between each clustering.

We observe that one isolated speaker, F05L, is shared across the three groups. HUM and CNN also share two two-leaves branches, i. e. two pairs of speakers, while only one is shared between HUM and PHON. Therefore we have a higher similarity score for HUM-CNN comparison. Linkages between other speakers are shared even though they do not

result in completely equal branches. Three double linkages are shared between HUM and CNN clusterings, while we have four for HUM and PHON. The latter shares three double and one three-way linkages with the CNN clustering.

In Figure 5.4, the comparison between HUM clustering from Non-Native listeners for the male voices perception task, and the PHON and CNN clusterings obtained through the Qual group, is illustrated. The similarity scores are 0.40 between CNN and PHON, 0.42 between HUM and CNN and 0.48 between HUM and PHON. In the latter comparison we observe that there is one shared branch composed of 3 speakers and 12 shared leaves associated in two three-way and three double linkages. There is also a branch shared by HUM and CNN, however, in this case there are only seven leaves linked in a similar way. In the comparison of the two non-human based groups there are no full branches that are shared but 10 connected leaves, which explains why they obtain the lowest similarity rate.

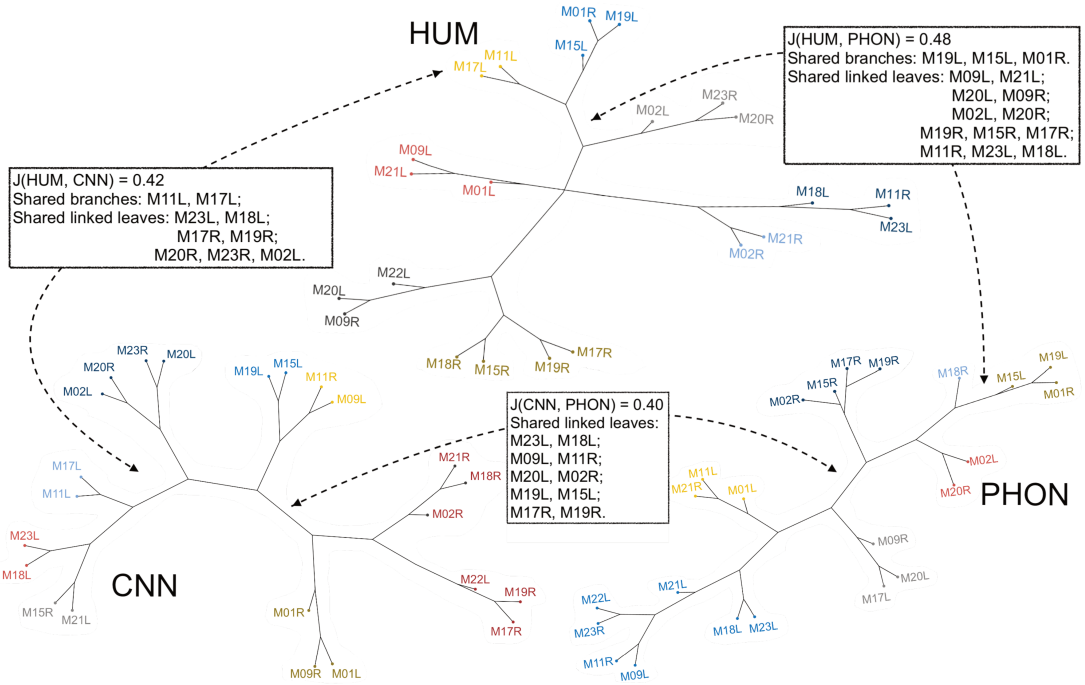


Figure 5.4: Similar clusters obtained from human Non-Native listeners responses in the male voices perception task. CNN and PHON correspond to clusterings based on Qual components group. Colours represent the multiple branches.

We obtain the completely opposite results, illustrated in Figure 5.5, with three highly dissimilar clusterings. In this case, Hadiff is used for the representation of both PHON and CNN, while responses from Non-Native Experts in the male voices task is used to represent for the HUM group. Beginning with the higher scores, in the PHON-CNN comparison we observe a similarity of 0.35, no branches are shared but 9 leaves in three two-way linkages and one triplet which, in the CNN clustering, represent a whole branch. 4 leaves are shared between the CNN and HUM groups, with a double linkage being also part of an entire branch in the CNN. This results in a total similarity score of 0.18. The other shown comparison has a similarity score of 0.16, we observe indeed that the PHON clustering presents a highly populated branch, resulting in a less coherent group. In this branch we find the majority of the leaves shared with HUM. 6 leaves out of the total 10 are present in the named branch, all in double linkages, corresponding to 3 different branches in the HUM cluster.

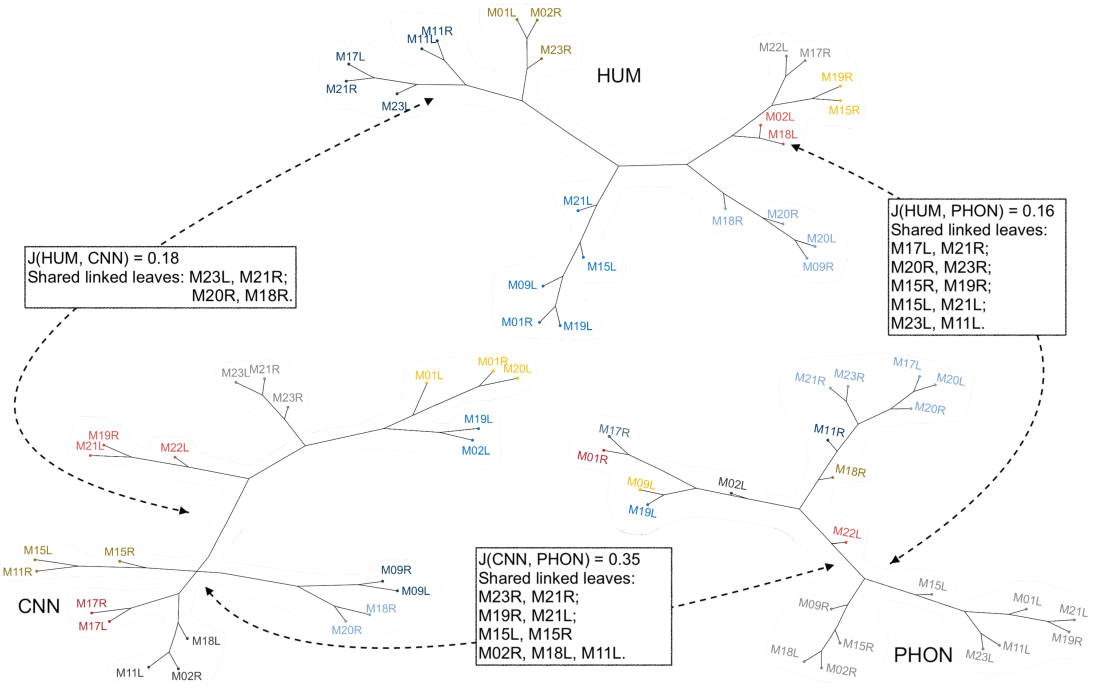


Figure 5.5: Dissimilar clusters obtained from human Non-Native Expert listeners responses in the male voices perception task. CNN and PHON correspond to clusterings based on Hadiff components group. Colours represent the multiple branches.

5.4 Chapter conclusions

The results of the listener variable in the characterisation of voices, presented above, aim to provide the path for the results and comparisons of the following chapters. The multiple influence factors that listeners carry, e. g., their age, sex or mother tongue, represent a complex set of variables to consider when working on a voice perception task. To improve decision making algorithms or clustering analysis methods and assess their reliability a human-machine comparison is fundamental. Sound modelling and characterisation is done by our brain at high performance levels but further investigations are needed in order to understand its mechanisms.

The use of clustering techniques is an important tool to better understand the distribution of speaker's characteristics. Phonetic- and CNN-driven clusterings show consistent differences in the distribution of information between the answers corresponding to their subsets. For the human answers we had access to a heterogeneous group of listeners in terms of expertise in speech domains. While the majority of subjects were French females we also had the possibility to compare Native answers with few Non-Natives. Further comparisons between the three main clustering groups show the potential for the analysis of influence factors in the perceptual characterisation of human voice. Computing metrics on clustering robustness and for distances between clusters provided a better understanding of which components may be used by human listeners and allowed us to assess the validity of CNN clusters based on phonetic knowledge. This is discussed in Chapter 8.

What follows is a summary, providing a report of the overall findings and discussion points of this chapter; more extensive discussions are present in Chapter 8.

5.4.1 Summary

- **PHON** clusters:
  - Acoust-Form and Ms-Glob are the most similar clusters combinations for both sexes confirming that Ms and Form have important roles in modelling the information shared with other components;
  - Ltas has a higher influence on male speakers' modelling. MFCC also share more information about male voices with global representations.
- **CNN** clusters:
  - Nrg and Hadiff sets show the most redundant information for both sexes;
  - Rhythm clustering has the most influence on the modelling of women's voices while it is Hr for men.
- **HUM** clusters:
  - For the female voices clusterings, all female listeners subsets share a third of the information, while for the male voices a higher variability is observed;
  - A same-sex effect is observed, since for the female voices task there is a higher average similarity between clusterings than those for male voices;
  - Expert listeners are less similar to Non-Native Experts than other groups;
  - Non-Native Naives have a different approach to the task in comparison to all other groups, suggesting that both sex and Native language have a major role in characterising voices even when the spoken language is well-known by Non-Native listeners.
- **PHON-CNN** comparison:
  - Clusterings for the female voices are less similar than for the male ones;
  - PHON-Acoust shares the most information with CNN clusterings for both sexes;
  - For male speakers PHON-f0 and -Amp clusterings share less information with other CNN clusterings;
- **PHON-CNN-HUM** comparison:
  - Form is the most similar clustering to the listeners responses for the female voices, while the most dissimilar is Ltas, this is the case when compared with both PHON and CNN;
  - Non-Natives share the least amount of information with PHON and CNN clusterings for both female and male voices task;
  - Experts' responses show a higher similarity to the clusterings obtained with harmonics and energy information for the characterisation of male speakers;
  - Naives' responses parallel the clustering obtained with Rhythm for the characterisation of female voices;
  - Non-Native female listeners clustering matches Qual, Hr and MFCC in the female voices task;

Chapter 6

Phonetics

The present chapter is dedicated to the results obtained through classical phonetic approaches, based on the state of the art described in Chapter 2, as well as part of Chapter 3. We start by introducing the extraction methods and grouping for the selected phonetic measurements. Following what has been discussed in the previous Part, four speech components are the main objects of our study, the next section presents which elements have been considered for each component.

In previous chapters, the main objective was to describe how individual differences have been studied throughout the decades in speech-related research. Hereafter, the main objective is to add our contribution to speaker characterisation domains starting with phonetic descriptions of some speech features, and by further exploring their interactions and understanding of their implications on individual differences. The following results are mainly based on PCA and other statistical modelling analysis carried on phonetic values.

Similar to Chapter 5, in this results chapter, we divided the extracted phonetic measurements into multiple groups aiming to represent a speech component or part of it. Table 6.4 summarises the groups used in the following sections. The extraction methods and data representation mainly remain the same as those presented in the previous chapter except when explicitly mentioned.

6.1 Reference values and linear models

In this section, before listing the results of our experiments, we provide a basic statistical description of the studied corpora. Our report firstly focuses on the comparison of our data with the reference data concerning the source and filter component, namely f0 and formants. The purpose of this description is to place our observations alongside those already present in the phonetic literature in order to assess the coherence of the studied data. In Appendix A an extensive version of both f0 and formants values tables are presented with speakers-specific values.

In Table 6.1 are presented mean and SD values that we have observed for f0. As mentioned before, NCCFr is a corpus of spontaneous speech while PTSVOX presents read speech.

From what has been discussed in Chapter 2, we expected PTSVOX to have higher mean values than NCCFr while the latter should have higher SDs. A clear difference between female and male speakers is observed in coherence with literature results with the female speakers having an average of 230 Hz and the male of 125 Hz. A higher SD, 31 Hz, is registered for female speakers while males have an average SD of 17. In the PTSVOX corpus the same tendency on mean values is reported with 218 Hz and 122 Hz respectively for women and men with microphone recordings, while with the telephone support we observe an increase to 228 Hz and 145 Hz.

Concerning SD we observe that in microphone recordings female speakers maintain higher values, 17 against 15 Hz, but the same is not true for the telephone condition where men have a SD of 23 against 18 for women. The latter phenomenon is certainly due to the fact that lower frequencies are more affected by signal degradation in telephone recordings. These observations confirm what has been discussed in the literature, Section 2.1.1.

The signal degradation is also explained by the results obtained with a linear mixed model that we conducted in order to understand the influence of recording conditions on f0 values. We used condition (microphone or telephone) and recording session as random effects with speakers as a fixed effect. A significant effect ($p < 10^{-}2$) is found on the condition predictor with a tendency to significance for the session (p close to $10^{-}2$). The first result is a direct consequence of what is observed in Table 6.1 with f0 values being very different depending on the recording condition for the same speakers and linguistic content. The second result, even though it shows a tendency and not significance, shows that the within-speaker variability is not to be neglected when studying speakers' characteristics since the complexity of the speaker identity cannot be fully understood without considering this factor.

	n	**Age**	**Mean**	**SD**
NCCFr				
f0 (female)	21	18-50 (avg 23)	230	31
f0 (male)	23	19-28 (avg 22)	125	17
$\text{PTSVOX}_{\text{mic}}$				
f0 (female)	12	19-21 (avg 19)	212	27
f0 (male)	12	18-22 (avg 19)	120	15
$\text{PTSVOX}_{\text{tel}}$				
f0 (female)	12	19-21 (avg 19)	222	37
f0 (male)	12	18-22 (avg 19)	133	52

Table 6.1: f0 mean and SD values for NCCFr and PTSVOX corpora, number of speakers and age's ranges are reported. For PTSVOX values from microphone and telephone recordings are separated.

All the formants' values have been obtained by averaging the measurements at 1/3, 1/2 and 2/3 in order to have more coherent results. Besides, formants show more unstable responses than f0. There is a clear difference between the frequencies belonging to spontaneous speech and read speech with the latter being more in line with literature observations. The other important difference observed concerns the transmission channel with values from microphone and telephone supports being very distinct even though they come from the same speakers attending the same task.

We compare these values with the references by [Georgeton et al., 2012], which used read

sentences by 40 female speakers, and by [Gendrot and Adda-Decker, 2005], which used 2 hours of speech mainly extracted from broadcast news from 15 female and 15 male speakers. In both works, the ten French vowels /i e ɛ a ɔ o u y øœ/ are used in order to provide reference values, i. e. mean, SD, maximum and minimum, for the first three formants of French language.. We mainly observe a coherence between the results, with some exceptions.

[Gendrot and Adda-Decker, 2005] provides a table of threshold values for filtering formant values in order to avoid the presence of acoustic inconsistencies. Only the first three formants have been studied since the computation of F4 shows intrinsic difficulties. We have used these values in order to filter our data. The filtering thresholds are reported in Table 6.2 and Table 6.3 alongside the mean and SD values of the two corpora. PTSVOX values for microphone and telephone data are reported separately.

For both female and male speakers, /i/, /y/ and /u/'s F1 are produced at much higher frequencies in spontaneous speech (NCCFr). This results in the front rounded vowel /y/ to be acoustically more open and to shift to the bottom of the vocalic triangle, especially for male speakers (Figure 6.1). The combination of higher F1 and F2 for /u/ results, for both sexes, in a more centralised production of this vowel. The same phenomenon of centralisation, higher F2 than in the reference values, is observed for /o/. F3 values show different tendencies for female and male speakers with /y/ and /o/ having higher realisations for female speakers. In men's case we observe higher F3 for both /y/ and /u/ while /i/, /e/ and /ɛ/ are produced in lower frequencies than expected. The latter phenomenon has been associated with rounding and lips protrusion effects on vowel production. Concerning F4, we observe that all female values are lower than the reference (only [Georgeton et al., 2012] report F4 values), in both spontaneous and read speech, but /o/, /ɔ/ and /u/ results are comparable with the cited studies.

These results enable us to address the main concern about spontaneous speech productions, which represent one of the main aspects of this work on speech sequences without taking into account a specific linguistic content. The high variability we observe in the frequency realisations from spontaneous speech makes, for example, the vowel /i/ more like a /u/ vowel, from a frequency point of view and the latter more like a central vowel, with vowels such as /e/ and /a/ also being procured in a more closed manner. However, thanks to the production context, the categorisation of these vowels remains possible. Looking at the different vocalic triangles from the 44 speakers in Figure 6.1, we can observe that multiple individual signatures in the vowel production strategies are reflected. The high variability of the produced vowel frequencies is just one of the elements that carry information on how speech is a continuously changing object especially when studying spontaneous realisations. As mentioned, individual production strategies are reflected through multiple speech components and focusing on categories that are highly inconsistent during spontaneous realisations can be a significant issue. In our studies, we try to overcome this by not focusing on the linguistic content of the analysed sequences, hypothetically resulting in an unbiased extraction of individual characteristics.

Similar to what has been done for f0, we used a linear mixed effect model in order to have a first basic understanding of how formants carry speaker information with the speaker as a fixed effect. No significant effect was found on the Vocalic Space Area, formants' range ratios and formants dispersion measurements for any of the three corpora. However, female speakers' F3 of /u/ and F2 of /i/ showed tendencies for speakers' characterisation.

We carried an additional experiment on the PTSVOX corpus using a LDA in order to study the discrimination power of formant values of /a/, the most present vowel. Following the methods described in [McDougall and Nolan, 2007], for each speaker's /a/ F1 and F2 measurements were taken at 9 moments of the vowel and combined in 3 series: all even moments (20 %, 40 %, 60 % and 80 %); all odd moments (10 %, 30 %, 50 %, 70 % and 90 %); the midpoint and the two even extremes (20 %, 50 %, 80 %). This is our first attempt to somehow represent the dynamics of the first two formants.

Our results do not achieve to replicate the scores from the cited study, average of 32 % for F1 and 42 % for F2 against an average of 12 % (14 % for female speakers and 9 % for male ones) for F1 and 15 % for F2 (female speakers at 16 %, while males ones at 15 %) in our case. It has to be noted that we used a smaller set of vowels with less controlled production contexts. However, the same tendencies are observed with the most discriminant values being the ones obtained from odd moments of the vowel.

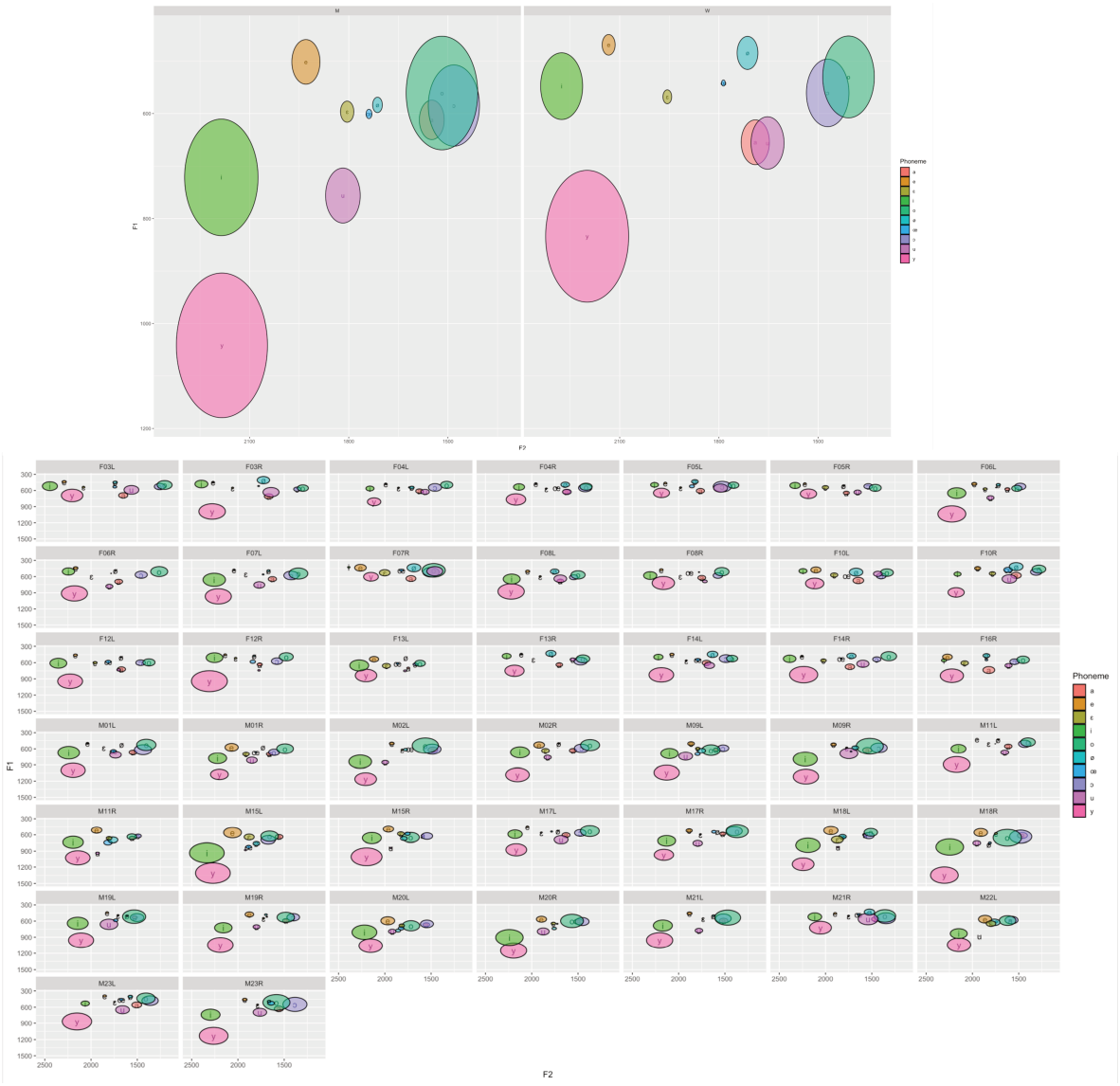


Figure 6.1: Vocalic triangles on F1-F2 space for the 44 NCCFr speakers, showing ellipses for each observed vowel. The two triangles on top are averaged from female and male speakers.

Vowel	Data set	Observations	**Mean F1 (SD)**	**Mean F2 (SD)**	**Mean F3 (SD)**	**Mean F4 (SD)**
			$F1 < 900$	$1600 < F2 < 3100$	$F3 > 2500$	
	NCCFr	11171	9551 | 473 (215)	10951 | 2268 (276)	11035 | 3158 (303)	4085 (308)
i	**PTSVOX**$_{\text{mic}}$	887	878 | 329 (89)	864 | 2336 (241)	881 | 3256 (258)	4235 (298)
	PTSVOX$_{\text{tel}}$	861	855 | 392 (89)	390 | 1825 (307)	559 | 2692 (222)	3269 (307)
			$F1 < 900$	$1400 < F2 < 3000$	$F3 > 2200$	
	NCCFr	12627	12194 | 455 (127)	12532 | 2127 (245)	12619 | 2934 (231)	3999 (296)
e	**PTSVOX**$_{\text{mic}}$	1397	1397 | 403 (59)	1386 | 2216 (229)	1396 | 3045 (207)	4177 (278)
	PTSVOX$_{\text{tel}}$	1284	1273 | 443 (73)	679 | 1663 (330)	1025 | 2507 (270)	3108 (301)
			$F1 < 1100$	$1400 < F2 < 2700$	$F3 > 2000$	
	NCCFr	17966	17291 | 554 (173)	17516 | 1954 (249)	17957 | 2872 (248)	3946 (316)
ɛ	**PTSVOX**$_{\text{mic}}$	1268	1268 | 477 (97)	1247 | 2028 (237)	1268 | 2963 (207)	4147 (321)
	PTSVOX$_{\text{tel}}$	1302	1298 | 509 (100)	567 | 1565 (240)	1026 | 2267 (275)	2956 (285)
			$F1 < 1100$	$900 < F2 < 2300$	$F3 > 1900$	
	NCCFr	24667	24088 | 647 (158)	24194 | 1691 (258)	24658 | 2817 (249)	3917 (315)
a	**PTSVOX**$_{\text{mic}}$	2038	2038 | 573 (118)	2022 | 1730 (262)	2034 | 2856 (222)	4105 (305)
	PTSVOX$_{\text{tel}}$	2022	2018 | 602 (108)	1887 | 1286 (285)	940 | 2046 (248)	2836 (271)
			$F1 < 900$	$1400 < F2 < 2800$	$F3 > 1800$	
	NCCFr	5607	3570 | 607 (257)	5327 | 2185 (315)	5607 | 3034 (355)	4021 (364)
y	**PTSVOX**$_{\text{mic}}$	633	607 | 354 (139)	515 | 1908 (361)	633 | 2742 (327)	3713 (343)
	PTSVOX$_{\text{tel}}$	615	565 | 470 (173)	447 | 1740 (308)	602 | 2458 (317)	3106 (306)
			$F1 < 1000$	$700 < F2 < 2300$	$F3 > 1800$	
	NCCFr	1260	1219 | 474 (148)	1232 | 1712 (254)	1260 | 2755 (233)	3840 (255)
ø	**PTSVOX**$_{\text{mic}}$	1343	1343 | 381 (59)	1331 | 1672 (288)	1342 | 2698 (257)	3835 (337)
	PTSVOX$_{\text{tel}}$	1270	1270 | 444 (57)	1229 | 1405 (332)	1025 | 2181 (347)	2969 (306)
			$F1 < 1100$	$800 < F2 < 2400$	$F3 > 2000$	
	NCCFr	5943	5537 | 515 (210)	5670 | 1776 (279)	5941 | 2826 (293)	3880 (324)
œ	**PTSVOX**$_{\text{mic}}$	226	226 | 516 (82)	226 | 1607 (193)	225 | 2737 (221)	3809 (298)
	PTSVOX$_{\text{tel}}$	223	223 | 550 (69)	221 | 1351 (269)	67 | 2128 (251)	2835 (251)
			$F1 < 1000$	$500 < F2 < 1500$	$F3 > 1800$	
	NCCFr	4198	3459 | 592 (252)	1796 | 1388 (207)	4198 | 2926 (345)	4006 (326)
u	**PTSVOX**$_{\text{mic}}$	381	376 | 344 (118)	320 | 1109 (264)	380 | 2726 (246)	3835 (238)
	PTSVOX$_{\text{tel}}$	377	362 | 482 (165)	324 | 1144 (247)	318 | 2287 (386)	3195 (330)
			$F1 < 1000$	$600 < F2 < 1600$	$F3 > 2100$	
	NCCFr	2821	2695 | 520 (167)	2102 | 1347 (262)	2819 | 2832 (267)	3924 (270)
o	**PTSVOX**$_{\text{mic}}$	623	623 | 463 (97)	534 | 1302 (198)	615 | 2724 (240)	3831 (280)
	PTSVOX$_{\text{tel}}$	537	537 | 529 (81)	518 | 1176 (210)	123 | 2192 (203)	2903 (294)
			$F1 < 1000$	$600 < F2 < 2000$	$F3 > 2000$	
	NCCFr	4153	4025 | 554 (151)	3925 | 1467 (275)	4152 | 2782 (262)	3870 (286)
ɔ	**PTSVOX**$_{\text{mic}}$	647	647 | 374 (56)	640 | 1106 (211)	645 | 2716 (206)	3876 (259)
	PTSVOX$_{\text{tel}}$	699	698 | 454 (68)	688 | 1115 (248)	409 | 2298 (338)	3172 (329)

Table 6.2: First four formants mean (average of measurements from 1/3, 1/2 and 2/3 of the vowel) and SD values for 10 French vowels from NCCFr and PTSVOX. Female speakers. For each vowel are reported the applied filters following [Gendrot and Adda-Decker, 2005] thresholds. The number of observations before application of the filters is reported in the *Observations* column, while the first number in each cell corresponds to the remaining observations after threshold filtering.

Vowel	Data set	Observations	**Mean F1 (SD)**	**Mean F2 (SD)**	**Mean F3 (SD)**	**Mean F4 (SD)**
			$F1 < 750$	$1500 < F2 < 2500$	$F3 > 2000$	
	NCCFr	13289	8845 | 493 (206)	10584 | 2133 (272)	13287 | 3070 (355)	3954 (389)
i	**PTSVOX**$_{\text{mic}}$	935	919 | 281 (82)	866 | 2064 (237)	934 | 2953 (276)	3774 (320)
	PTSVOX$_{\text{tel}}$	879	864 | 362 (90)	692 | 1849 (268)	833 | 2587 (311)	3229 (283)
			$F1 < 800$	$1100 < F2 < 2400$	$F3 > 2000$	
	NCCFr	14734	13002 | 444 (158)	13962 | 1916 (223)	14733 | 2752 (286)	3763 (299)
e	**PTSVOX**$_{\text{mic}}$	1480	1479 | 321 (46)	1463 | 1864 (233)	1480 | 2651 (206)	3684 (260)
	PTSVOX$_{\text{tel}}$	1366	1355 | 396 (70)	1273 | 1704 (282)	1160 | 2374 (257)	3048 (317)
			$F1 < 1000$	$1200 < F2 < 2300$	$F3 > 2000$	
	NCCFr	20325	17931 | 546 (208)	18506 | 1789 (258)	20309 | 2713 (332)	3756 (339)
ɛ	**PTSVOX**$_{\text{mic}}$	1293	1291 | 383 (72)	1240 | 1700 (234)	1284 | 2591 (208)	3657 (281)
	PTSVOX$_{\text{tel}}$	1274	1257 | 469 (103)	1114 | 1564 (224)	857 | 2271 (264)	2941 (324)
			$F1 < 1000$	$800 < F2 < 2300$	$F3 > 1800$	
	NCCFr	27088	25362 | 596 (161)	26210 | 1548 (273)	27082 | 2630 (288)	3705 (297)
a	**PTSVOX**$_{\text{mic}}$	2119	2118 | 429 (82)	2110 | 1422 (250)	2115 | 2530 (204)	3637 (265)
	PTSVOX$_{\text{tel}}$	2040	2037 | 534 (93)	2017 | 1361 (229)	1399 | 2109 (306)	2915 (344)
			$F1 < 800$	$1100 < F2 < 2400$	$F3 > 2000$	
	NCCFr	7464	3138 | 634 (217)	3858 | 2048 (208)	7464 | 3035 (370)	4006 (413)
y	**PTSVOX**$_{\text{mic}}$	677	618 | 340 (163)	532 | 1801 (282)	677 | 2596 (336)	3538 (379)
	PTSVOX$_{\text{tel}}$	650	571 | 448 (173)	590 | 1717 (229)	647 | 2326 (271)	3049 (319)
			$F1 < 900$	$700 < F2 < 2000$	$F3 > 1700$	
	NCCFr	1536	1278 | 518 (211)	1284 | 1666 (237)	1536 | 2710 (358)	3695 (362)
ø	**PTSVOX**$_{\text{mic}}$	1294	1293 | 326 (53)	1278 | 1463 (242)	1294 | 2513 (208)	3501 (261)
	PTSVOX$_{\text{tel}}$	1113	1108 | 399 (57)	1091 | 1450 (227)	991 | 2209 (326)	3013 (344)
			$F1 < 1000$	$800 < F2 < 2000$	$F3 > 2000$	
	NCCFr	7133	6044 | 543 (241)	5756 | 1680 (253)	7126 | 2730 (367)	3734 (367)
œ	**PTSVOX**$_{\text{mic}}$	241	241 | 416 (55)	241 | 1367 (161)	237 | 2445 (171)	3423 (244)
	PTSVOX$_{\text{tel}}$	230	230 | 485 (54)	229 | 1409 (160)	136 | 2191 (213)	2885 (327)
			$F1 < 900$	$400 < F2 < 1500$	$F3 > 1400$	
	NCCFr	4900	3395 | 650 (221)	1391 | 1426 (185)	4900 | 2996 (397)	4012 (370)
u	**PTSVOX**$_{\text{mic}}$	395	390 | 316 (98)	336 | 1052 (286)	395 | 2650 (261)	3622 (283)
	PTSVOX$_{\text{tel}}$	380	370 | 434 (137)	351 | 1086 (236)	377 | 2321 (322)	3217 (297)
			$F1 < 900$	$600 < F2 < 1600$	$F3 > 1500$	
	NCCFr	3512	3170 | 539 (183)	2222 | 1378 (267)	3512 | 2800 (384)	3805 (350)
o	**PTSVOX**$_{\text{mic}}$	645	644 | 396 (76)	626 | 1154 (198)	645 | 2492 (210)	3503 (216)
	PTSVOX$_{\text{tel}}$	555	554 | 484 (71)	551 | 1149 (175)	484 | 2081 (368)	2993 (344)
			$F1 < 900$	$600 < F2 < 1800$	$F3 > 1500$	
	NCCFr	4905	4463 | 563 (165)	4059 | 1434 (281)	4905 | 2672 (349)	3710 (336)
ɔ	**PTSVOX**$_{\text{mic}}$	702	702 | 331 (54)	693 | 990 (204)	702 | 2574 (206)	3558 (246)
	PTSVOX$_{\text{tel}}$	731	724 | 419 (83)	723 | 1047 (226)	697 | 2313 (377)	3182 (335)

Table 6.3: First four formants mean (average of measurements from 1/3, 1/2 and 2/3 of the vowel) and SD values for 10 French vowels from NCCFr and PTSVOX. Male speakers. For each vowel are reported the applied filters following [Gendrot and Adda-Decker, 2005] thresholds. The number of observations before application of the filters is reported in the *Observations* column, while the first number in each cell corresponds to the remaining observations after threshold filtering.

6.1.1 Temporal component

In order to study the influence of speakers' strategies of temporal organisation, we have studied multiple measurements relating to the temporal component on both our corpora. These include speech rate, measures of pauses, as well as the ones described in Section 2.2.1, i. e., consonant and vocalic PVI along other breath groups measurements. As already discussed in our literature review, speakers tend to personalise the patterns of speech temporal organisation, especially when planning spontaneous speech productions. This component's cues allow the temporal analysis even for degraded speech signals however the need for an accurate annotation is not negligible.

Our results confirm that when comparing the temporal measurements from speakers of similar ages and speaking the same language, the speech task, read or spontaneous, shows a significant effect ($p < 10^{-}2$) when used as random effects in a linear mixed model. In the reading task, no significant difference has been observed in the comparison of temporal measurements, the speakers seem to adopt very similar production strategies during this task. The 44 NCCFr speakers are a more suitable example for the study of temporal organisation strategies as individuals' characteristics. Indeed, we observe that the measurements playing more prominent roles in defining individual strategies are the lengths of pauses, as well as the first and last phonemes' duration. However, the speaking rate does not appear as consistent as the other considered cues across speakers.

6.2 Principal Components Analysis

What we described in the previous section are results obtained from specific linguistic content, e. g., comparing vowels by the formants measurements. As we proceed in the description of the other results we move away from the content influence on the measurements by considering speech sequences. Alongside the linguistic content another important factor of influence that has to be considered when studying individual characteristics is, as we already discussed, the within speaker variability. In the next sections, we present results from PCAs. This approach helps us understand the interactions between the different phonetic measurements we selected and the information carried by the values used to represent them. As mentioned above, we computed moving mean and CoV, as in [Lee et al., 2019], as well as the relative entropy for the analysed sequences. Following this method, mean, CoV and entropy values are computed as moving statistics every 30 ms after computation of measurements every ms, in order to take into account time-related variations. Mean values aim to represent the actual realisation of a measurement, CoV its stability, and entropy the disorder it carries, the latter however, is obtained, as mentioned above, on all observations from each sequence. These computations give us an average of 73416 observations for each NCCFr speaker and 40250 for each speaker of the PTSVOX corpus.

In order to statistically describe our data and understand the interactions between the selected characteristics and speaker variability, in the following sections we use different approaches. Statistical description has been mainly based on PCA and LDA, as mentioned before, they both represent common approaches in Phonetics, with the latter being also used as a classifier aiming to model the individual differences in our data. In addition, we used SVM as a more robust linear classifier. The segmented sequences are used to model

and study both the intra-speaker variation, by taking each sequence as a token, and the inter-speaker variability, having the whole variability matrices compared to each other.

Group names	**Measurements**	**Component**
Amp	Amplitudes of the harmonics near the first three formants (A1-3)	Source and filter
f0	f0 and its harmonics (H1, H2, H2k, H42k, H5k)	Source and filter
Form	First four formants (F1-4)	Source and filter
Hadiff	Differences between harmonics, and between harmonics and amplitudes (H1-A1, H1-A2, H1-A3, H1-H2, H2-H4, H2k-H5k)	Mode of vocal fold vibration
Hr	HNR at different pitch ranges (0-500 Hz HNR05, 0-1500 Hz HNR15, 0-2500 Hz HNR25, 0-3500 Hz HNR35), SHR	Mode of vocal fold vibration
MFCC	13 MFCC	?
Qual	Mean LTAS, CPP and energy measurements (RMS, soe, Praat-based energy)	Mode of vocal fold vibration
Rhythm	Intensity, ENV and TFS	Temporal
Temporal	Speech rate, consonant and vocalic PVI, pauses and segmental duration	Temporal
Dynamics	Open quotient-like measures on intensity, polynomial coefficients for f0 and formants	Source and filter

Table 6.4: Phonetic measurements groups used in Chapter 6 experiments and related components.

A different modelling approach is presented in Section 6.4.1, where we measured LLR between the read sequences using MVKD, as described by studies in Section 3.1. "The LLR is the numeric answer to the question: How much more likely are the observed differences between the known and questioned voice samples to occur under the hypothesis that the questioned sample has the same origin as the known sample than under the hypothesis that it has a different origin?" [Morrison, 2010]. We used only the read sequences because the linguistic material used to build the models through polynomial functions are more efficient than for spontaneous speech. The limits and advantages of this approach are part of the discussion.

An additional measurement described in this Chapter uses the concept of entropy. Other studies in speech science have used it in order to describe the variations of a phonetic phenomenon, see [Nilsson, 2006; Setiawan et al., 2009]. The term entropy relates to multiple scientific fields such as economics, sociology and information theory. Its basic idea is tied to the second law of thermodynamics, which states that the entropy of isolated systems left to spontaneous evolution cannot decrease with time, as they always arrive

at a state of thermodynamic equilibrium, where the entropy is highest. The formula that has been implemented in python is as follows:

$$H(X) = \sum p(x) * log(p(x))$$

Where X is a phonetic measurement with possible outcomes $x_1, ..., x_n\}$, which occur with probability $p(x_1), ..., p(x_n)$. We used base 2 logarithm as in the information theory variant of the formula by Shannon, in order to parallel the definition of entropy as the expected value of the self-information of a variable. Entropy is seen as a macroscopic entity that quantifies the average disorder in an isolated system at a microscopic level. It is not a property of the observed result but of the results we could have obtained instead. In other words, it qualifies the process by which the speech components (in our case) are generated. From our perspective, the disorder is represented by the values of phonetic measurements, which relate to multiple speech components, assumed when produced by a certain speaker rather than another.

6.2.1 Considerations on the within-speaker variability

We begin by presenting results that aim at describing the within-speaker variability with the different speech components we considered. In order to do this, the speech sequences have been analysed in a PCA for each speaker. We considered microphone and telephone sequences separately for the PTSVOX corpus so that the influence of the recording support could be observed. In the same way, the comparison between the two corpora allows us to understand how the speaking style can have an influence on the information carried by the speech components. Concerning the NCCFr analysis, we provide a comparison between phonetic measurements and MFCC as well, since a similar perspective is the object of discussion in the next Chapter.

Firstly, we report the number of Principal Components (PCs) with eigenvalues greater than 1 as they ensure factors accounting for an interpretable amount of variance in the data. However, we carry a detailed analysis only on the first 5 PCs as in general they represent most of the interpretable explained variance. Secondly, for the given PCs we extracted variables, with loadings at or exceeding 0.32 as it is usual practice, in order to understand both in a within- and between- speakers perspective which phonetic measurement contributes the most in characterising the studied voices.

We observe that on average for the PTSVOX corpus 7 PCs have an eigenvalue greater than 1 for both microphone and telephone sequences and both sexes. The average explained variance of these PCs is overall of 75 % with just the first 5 PCs accounting for an average of 71 %. The main difference concerns the first 2 PCs which for male speakers account for a lower average variance of 39 % in microphone sequences and 42 % in telephone against 43 % and 45 % for female speakers. The results from the NCCFr corpus show that the modelling for spontaneous speech in a PCA is very different. We observe that on average the number of PCs which have eigenvalues higher than 1 when considering only phonetic variables is 22 for both female and male speakers with a total explained variance of 68 %. Though, the first 5 PCs account for an average of 63 %. Considering MFCC an average of 11 PCs present considerable eigenvalues but the total explained variance decreases with an average of 58 %. Further results from coupling phonetic variables and MFCC show

an increasing number of interpretable PCs corresponding to 30 and an average variance of 60 %. In opposition to what has been observed for PTSVOX, the first 2 PCs for male speakers account for higher variances than for female ones.

We counted the number of times each phonetic measurement appeared with loadings higher than 0.32 in the first 5 PCs and reported them as their corresponding group, see Table 6.4. In this analysis we used a reduced number of speakers from the PTSVOX corpus (7 females and 8 males) in order to have a balanced set, and the same number of recording sessions, hence speech sequences. This impacted in particular the entropy computation which appears less reliable with a smaller number of observations. Therefore the results for this corpus exclude this value and present only moving mean and CoV values.

6.2.1.1 PTSVOX - female speakers

The distributions of the phonetic groups in the PCs differ consistently across speakers, Figure 6.2-Top shows the first PC corresponding to the microphone sequences of the 7 female speakers with relative explained variances, ranging from 21 % to 28 %. We observe that the mean values of the formants are largely shared by the majority of speakers, with CoV also playing an important role for 3 of them. The variables for voice quality are completely absent from more than half of these observations. When looking in detail at the measurements present for each group in this first PC, we see that the same variables are shared equally by the speakers. Hence, for the speakers that do not present a prominent characterisation of the first PC by formants only F1 plays a role. In opposition, for the fives speaker whose formants are prominent in characterising the PC, F1 is absent, while F2 and formant dispersion present a higher weight.

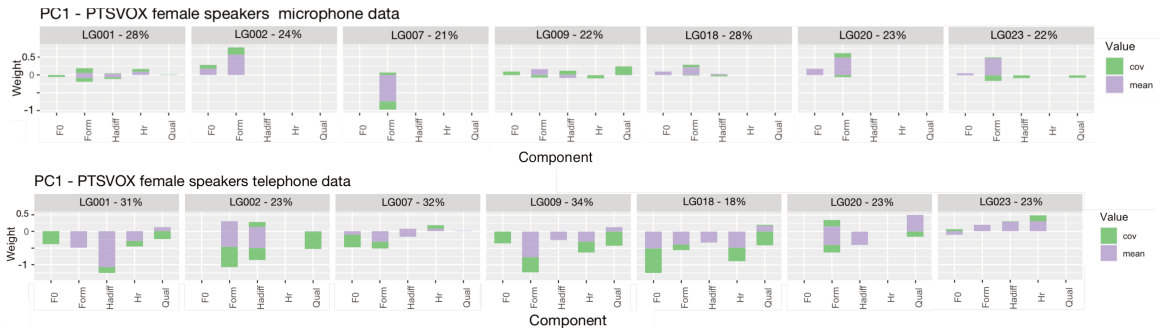


Figure 6.2: Weights distributions of phonetic measurements in the PC1 for the 7 PTSVOX female speakers. Top: microphone data. Bottom: telephone data.

Figure 6.2-Bottom shows the results for the same speakers but from the telephone recordings. We observe a total reduction of the weights of the formants values and an increased presence of voice quality characteristics. F1 continues to play its role for the speakers with less characteristic formants, such as LG023 and LG009, while for all the other female speakers it is the higher formants, F3 and F4, which play the most important role. A similar shift to higher frequencies in order to model the speakers 'variances is observed for f0 measurements. In microphone recordings only f0 was present, while for telephone recordings higher harmonics play a role in the characterisation.

The second PC, which takes into account the variances ranging from 14 to 20, shows complementary results to the one we have just seen. Speakers, such as LG001 and LG009,

for whom only F1 was present in PC1, are now highly characterised by F2 and F3 as well as by f0's high harmonics, i. e. H2k and H42k. As for the PC1 in Figure 6.2, Figure 6.3 compares the observations for the PC2 with similar results. Higher frequencies intervene to complete the information carried by the PC1. The combination of high frequency harmonics, H2k-H5k, appears in more than half of the speakers as a characteristic of this PC in both microphone and telephone observations.

Figure 6.3: Weights distributions of the speech components in the PC2 for the 7 PTSVOX female speakers. Top: microphone recordings. Bottom: telephone recordings.

The last 3 PCs account for 23 % to 30 % of the remaining variance with no particular trend concerning the speech components or the measurements involved except PC3, where the CPP plays a prominent role for 5 out of the 7 speakers. This suggests that while vocal tract and articulators movements play a major role in characterising female speakers' voices in a reading task, the second most important trait is represented by breathiness and voice quality.

6.2.1.2 PTSVOX - male speakers

Formants values appear in the PC1 of 6 out of the 8 male speakers when considering microphone recordings, in this case F2 and F3 are the most important variables for three speakers, the voice quality variables are those defining the PC1 while for one speaker the CoV values for all the considered speech components except f0 have the highest weight in the PC characterisation. Similarly to what has been observed for the female speakers, when comparing telephone and microphone results we see an understandable shift to higher frequencies for all the measurements present in this first PC. Apart from the lower presence of low frequency information we observe a greater similarity between the microphone and telephone results for male speakers. The information conveyed by the two recording media seems to be modelling more consistently. CoV plays a more reliable role for both media while the other values appear to vary more.

The variance explained by the second PC ranged from 15 % to 19 % in microphone results and 13 % to 18 % for the telephone results, while for the PC1 microphone sequences explained less variance with a range of 20 % to 26 % versus 21 % to 36 %. In this case we also observe similarities between the information carried by both recordings. Speakers who do not present f0 as an important characteristic in microphone recordings integrate only a small amount of information from high harmonics in the telephone modelling. F2 and F3 maintain their central role in characterising speakers who did not present formants values in the PC1. Breathiness characteristics such as the CPP are present for

the majority of speakers in the microphone results emphasising its role with respect to the results of female speakers.

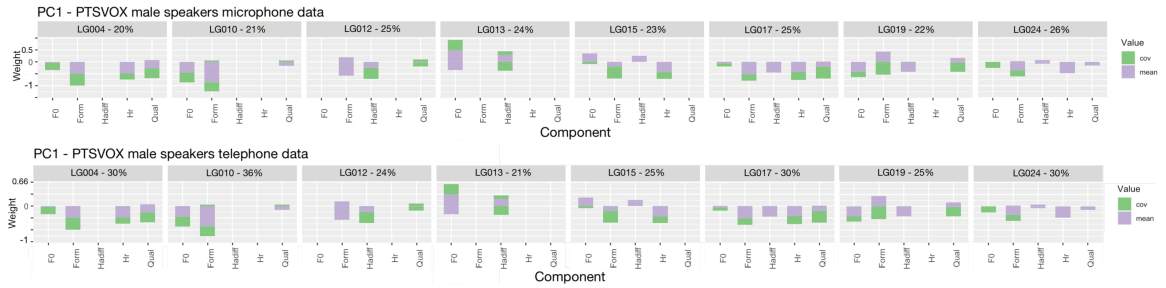


Figure 6.4: Weights distributions of phonetic measurements in the PC1 for the 8 PTSVOX male speakers. Top: microphone recordings. Bottom: telephone recordings.

Although the explained variance of the PC3 to PC5 is very similar between the two sexes, both having averages of 23 % and 27 % for microphone and telephone respectively, CoV and entropy values have a more important role for male speakers than for female speakers. The role played by these values suggests that, in order to characterise male voices, the indicators of their stability and randomness of their distribution can be considered as reliable as the values they produce.

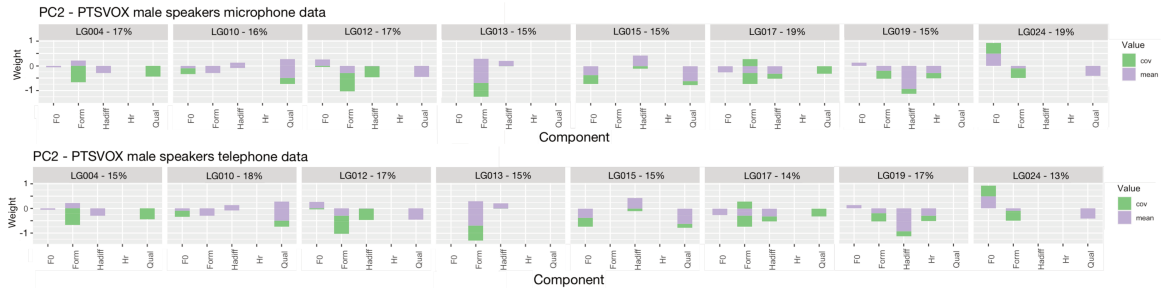


Figure 6.5: Weights distributions of phonetic measurements in the PC2 for the 8 PTSVOX male speakers. Top: microphone data. Bottom: telephone data.

6.2.1.3 NCCFr

The weight of phonetic measurements in the PCA for the NCCFr female speakers is less variable than what we observed for the PTSVOX speakers. As shown by Figure 6.6, the distributions appear to be more similar across speakers with minor exceptions, but when we focus on the detail of the measurements for each phonetic group we observe that not every variable plays the same role. In characterising speakers' voices from spontaneous speech, it also appears that the actual values the data assumes are more important than the distribution it may have.

In detail, we observe that the only group for which all relative measurements play a role in PC1 are Hr and Amp, however, CPP, F4, H2, H1A1 and H1A2 are also shared variables to a minor extent. Only half of the speakers present information on the values of the formants in PC1 corresponding to F4 CoV and entropy. For 17 out of the 21 speakers H2k values are shared as an important variable for this first PC while 2 additional speakers also show higher harmonics.

Regarding the 23 male speakers from the same corpus, we observe that some groups of measurements have a similar important role in describing the variance within-speakers: namely, the Hr group, related to voice hoarseness, and the CPP, index of breathiness, alongside the soe. Variables linked to the spectral shape of the source, notably H1H2 and H1A1, are shared by half of the male speakers in this PC1, as are H2k and H4. For two subsets of speakers, the lower harmonics, H1 and H2, as well as the amplitudes of the harmonics near F1 and F2 play an important role in statistical explanation of voice characteristics.

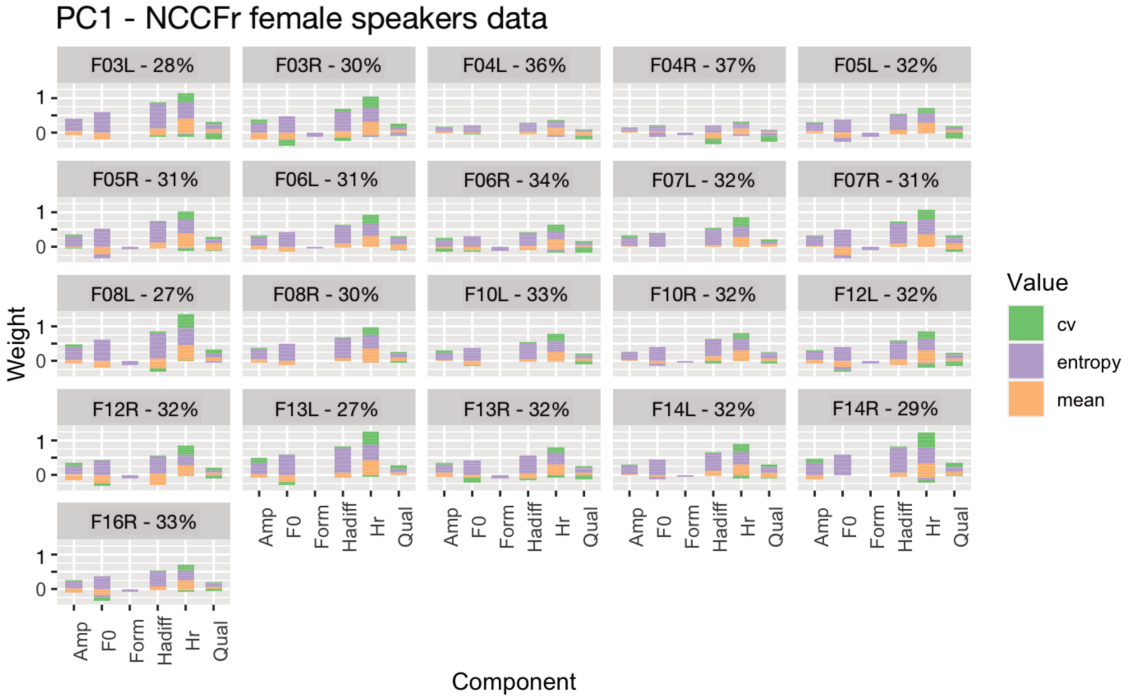


Figure 6.6: Weights distributions of phonetic measurements in the PC1 for the 21 NCCFr female speakers.

In general, we observe a higher similarity between the modelling of male and female speakers through spontaneous speech. Less than half of the total variance explained for both sexes is represented by PC1, ranging from 36 % to 45 % for female speakers and from 35 % to 44 % for male ones, Figure 6.6 and Figure 6.7. The variances for PC2 and PC3 are much lower, from 6 % to 12 % and 8 % to 13 %, sharing an opposite and complementary relation. Furthermore, the phonetic measurements that did not appear in the first PC play a role in defining the following two. For speakers of both sexes we see the emergence of information about rhythm and spectral shape, in particular the first two spectral moments, which are indices of energy distribution in the spectrum. F1 and F2 alongside LTAS measurements also appear for male speakers. The complementary relation between these two PCs is noted in particular by the fact that part of the speakers' variance is explained by the mean and CoV in PC2, while PC3, for the same speakers, uses entropy values for the same speech measurements and this is true the other way around for the other speakers.

As mentioned above, we performed an analysis on the MFCC in order to compare the phonetic results. However, MFCC do not show a high statistical power in describing the acoustic variance of the speakers. PC1 for both sexes have less than 20 % of average explained variance with the following PCs being even lower. No particular trend is observed

in the distribution of MFCC in the first PC, as all the 12 measurements are present. In PC2 and PC3 we observe an opposite distribution for the coefficients with odd rank coefficients present in PC3 and even rank coefficients in PC2. The combined analysis of MFCC and phonetic measurement show some interesting trends.

Even though the combination of MFCC and phonetic measurements does not increase the explained variance of the PCA, with an average of 32 % for PC1 in both sexes, we can make assumptions about the interpretation of the MFCC from the interaction they show with the observed classical phonetic measurements. In general, phonetic measurements have very similar distributions to the ones described in the analysis above. The lower rank MFCC appear in the PC1 for male speakers only, while for both sexes PC2 is defined by LTAS and the MFCC's entropy values with the addition of the Rhythm group measurements for male speakers. The complementary and opposite relation we observed for PC2 and PC3 in the pure phonetic analysis here is shifted to PC3 and PC4 with the emergence of f0 and intensity for female speakers, while for the PCs of male speakers both F1 and F2 have a significant weight. Higher formants appear for the description of variability only among female speakers in PC5.

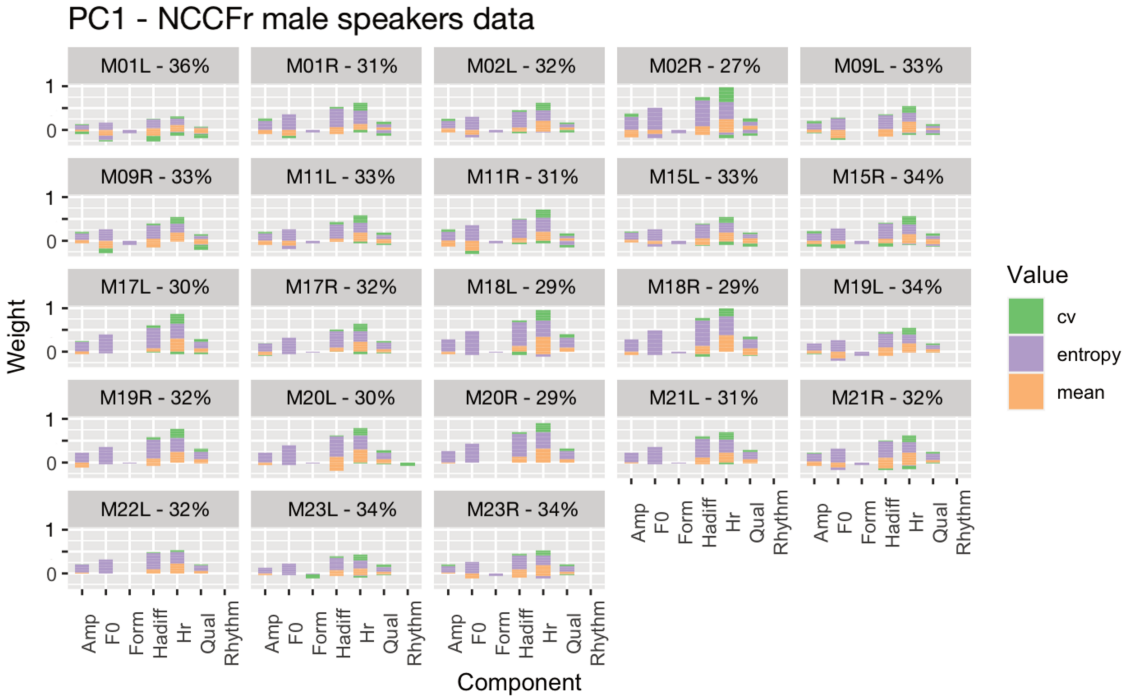


Figure 6.7: Weights distributions of phonetic measurements in the PC1 for the 23 NCCFr male speakers.

6.2.2 Between-speakers results

PCA makes it possible to reduce the dimensionality of the observed data and focus on what is really important to model the characteristics of the speakers. As already mentioned, understanding the within-speaker variability is important in order to understand how to differentiate the speakers from each other. The second part of our PCA description focuses on the between-speakers variability explained by these statistical models. As for the within-speakers analysis, an important part of the PTSVOX analysis is represented

by the microphone and telephone results comparison, in order to understand how different recording supports can influence the observed data.

The first main difference between spontaneous and read speech concerns the number of PCs that represent the explained variance and its percentage. They are lower in number for PTSVOX with only 5 for both sexes and both recording media but they have a cumulative explained variance above 90 % in all cases. For NCCFr the explained variance is around 70 %, and is higher when using phonetic measurements alone than when combining with MFCC. The opposite trend is observed for the number of PCs having an eigenvalue higher than 1, in the combined analysis they are 20 for both sexes but when only phonetic measurements are taken as variables for the PCA, they are 17 and 16 for female and male speakers respectively.

In this between-speakers perspective the results from microphone and telephone differ less than those we described above. Furthermore, phonetic measurements have similar weight distribution. The formants, F2 and F3, and high Harmonics are the most important variables defining the PC1 for female speakers of PTSVOX, while F1 emerges alongside lower source spectral information in the PC2. F4 and formants dispersion continue to behave in a similar way, showing a high influence on the PC3. There are no additional trends shown in the following PCs.

For male speakers we see the emergence of voice quality variables, i. e. CPP and the Hr group, in both microphone and telephone results. An important difference is observed in the role played by formants and f0 since in the microphone sequences PC1 is heavily defined by F3, F4 and FD whereas they only appear in PC2 for the telephone results. High harmonics, i. e. H42k and H4, are separated from f0 information accounting for PC2 and PC1 respectively in the telephone sequences while they remain both attached to PC2 in microphone results.

Similarly to the results from the within-speaker analysis, in the between-speaker description of the statistical variance for NCCFr, MFCC alone do not show a high statistical power in the results for both sexes. However, they are integrated into the phonetic analysis by interacting with part of the speech components. In both the phonetic and combined analysis, we observe, for female speakers, that the Hr group has the highest weight accounting for the PC1 variance alongside acoustic variables such as CPP, soe, high harmonics, i. e. H2k, H42k, H4, and high level source spectral information. PC2 is mostly defined by f0, Amp and Rhythm groups measurements to which low rank MFCC are added in the combined analysis. The third PC in the phonetic analysis regroups energy distribution and low level spectral shape information, while LTAS and F2, F3 emerge for the PC4. The PC3 and PC4 are repeated, shifted from one rank position since PC3 accounts for all MFCC, in the combined analysis with the integration of middle rank and high rank MFCC respectively.

For male speakers we observe a higher influence of rhythm measurements, i. e. env and intensity, in the PC1 alongside Hr, f0 and low harmonics. Complementarily, high harmonics, formants and TFS play a major role in defining the PC2. However in the combined analysis MFCC the second PC is shifted to the third rank in order to make place for spectral information by the means of MFCC, LTAS and spectral moments, which accounted for PC3 in the phonetic analysis. No particular trends are shown by PC4 and PC5 other than the emergence of F2 and FD alongside high rank MFCC in the combined analysis for male speakers.

In all our data, unlike the first two PCs, PC3-5 combined to account for just under half of the acoustic variance in the data that PCA was able to explain. In the NCCFr results, the variance was largely idiosyncratic with no particular acoustic category predominating apart from the Hr group variables. When the distributions of the groups of variables and their weights highly overlap, this suggests that the differences across both female and male voices must be found in the actual amount of variance for each measure rather than in their interactions. As the figures in the section above show, most of the variables are distributed approximately evenly across the PCs, with a few exceptions that appear only occasionally.

6.3 Classifications using phonetic measurements

The statistical description of speech components is only one aspect of understanding how speakers' voices are characterised and how information is conveyed in multiple ways during speech productions. An additional step in order to explore how speakers are characterised by speech components is to test the modelling data with classification algorithms. We used the above described values with LDA and SVM models, the results are summarised in Table 6.5. The classification rates are used as an indicator of the speakers characteristics' models robustness.

We used both LDA and SVM as their approaches to a classification problem are very different. LDA represents an analytic approach which assumes that all groups from the data are identically distributed and tries to estimate covariance matrices in order to linearly separate the data. SVM is based on an optimisation task where no assumptions about the data are made other than that all groups can be separated. The optimisation is performed on a subset of the data, the so-called support vectors, which lie on the separating margins, and are used to determine how the SVM discriminates the groups. LDA handles several classes well, as long as the assumptions are met, while SVM handles two-class classification problems better. Hence, in a multiclass classification several binary classifiers are created and tested against each other.

Therefore, the first test carried for the classification task was made using a LDA. The reported scores are F1-scores, the harmonic mean of Precision and Recall, obtained directly from the resulting confusion matrices. The classification through Linear modelling shows interesting results for the read speech corpus with a score of 0.75 for the female speakers using microphone data and 0.79 for the male speakers in the same recording conditions. The telephone data show a slight score degradation with female speakers losing 10 points, i. e. F1-score of 0.65 and with male speakers having a smaller deterioration with a score of 0.77. The same data shows a strong increase in classification rate when using a SVM algorithm with female speakers scoring 0.88 for both microphone and telephone data and male speakers scoring 0.92 and 0.86 respectively.

The NCCFr data does not replicate the same results, both algorithms performed very poorly. Indeed, for both female and male speakers the scores with LDA classification are under 0.10, while using a SVM they show only slightly higher scores with respectively 0.17 and 0.14. The same trend towards a lower score for spontaneous speech is found with the explained variance of the PCA, with an average decrease of 20 % compared to the read speech data.

These outcomes underline the idea that understanding how information is conveyed through the components of speech, particularly in spontaneous speech is an ongoing challenge. The statistical description of this highly variable object can only explain a small portion of the actual variability contained in its components. Read speech has proven to have a more stable structure, simpler to tackle in classification problems; it obtains significant results even under different recording conditions. However, as stated from the beginning of this thesis, characterising voices means taking into account multiple ways in which speech can be produced and spontaneous speech is the most common in real life situations. A complex input such as spontaneous speech productions appears to require a more complex and robust model.

data set	*Female speakers*			*Male speakers*		
	PCA	**LDA**	**SVM**	**PCA**	**LDA**	**SVM**
$\text{PTSVOX}_{\text{mic}}$	98 %	0.75	0.88	92 %	0.79	0.92
$\text{PTSVOX}_{\text{tel}}$	96 %	0.65	0.88	95 %	0.77	0.86
NCCFr	73 %	0.08	0.17	69 %	0.09	0.14

Table 6.5: Classification rates from SVM and LDA, expressed as F1-scores obtained from the confusion matrices on PTSVOX and NCCFr corpora using phonetic measurements values with relative cumulative explained variance by the PCA.

6.4 Modelling speech dynamics

The following step for our studies on voice characterisation concerns the representation of speech. Even though we computed moving means, CoV and entropy in order to have several indicators of the modulation of phonetic measurements in the analysed sequences, it appears that these values are not robust enough to represent the changes that occur during speech production. The statistical approach of a PCA has shown some limits as well. The explained variance is a good statistical description but does not fully correspond to what can be actually modelled in a classification problem.

As already discussed, whether we are studying read or spontaneous speech, we are studying a dynamic object, whose characteristics can be investigated through static parameters such as those described in the previous sections. However, as discussed in the literature review, the search for a dynamic representation of speech is important to further our understanding of how information is carried by the different phonetic measurements. The term speech dynamics refers to the changes of the studied speech phonetic measurements with respect to time and how these changes can be represented with minimal loss of information. We focused on two dynamics representations, the first one based on derivatives computation and open quotient-like measures, inspired by works such as [Kreiman and Shue, 2010; He and Dellwo, 2017; He et al., 2019], discussed in the present Section, while in the following section we present the results obtained with polynomial functions and Multivariate Kernel Density estimation (MVKD) approach.

In our first attempt to model speech dynamics we computed the derivatives of the intensity values for each of the 4 s sequences of the NCCFr corpus. As shown in Figure 6.8-Bottom, the derivatives give a representation of the speech in the form of peaks that relate to the velocity of the considered measurement. The same signal filter described by [He and

Dellwo, 2017], low-pass filtering at 10 Hz (Hann filter, roll-off 1/4 6 dB/octave), was applied to the original signal in order to maximise the syllable peaks information carried by the derivative curve. Further computations are made on the derivative curves in order to describe different aspects of the dynamic with relative mean and SD for each sequence: positive and negative peaks period (T and T_{neg}); value of the positive and negative derivative peaks (f and f_{neg}); interval between a positive peak and the following negative one (O_t), interval between a negative peak and the following positive one (O_q). We also tried to model the second derivatives, in order to obtain information on the acceleration of the produced speech but we did not obtain satisfactory results.

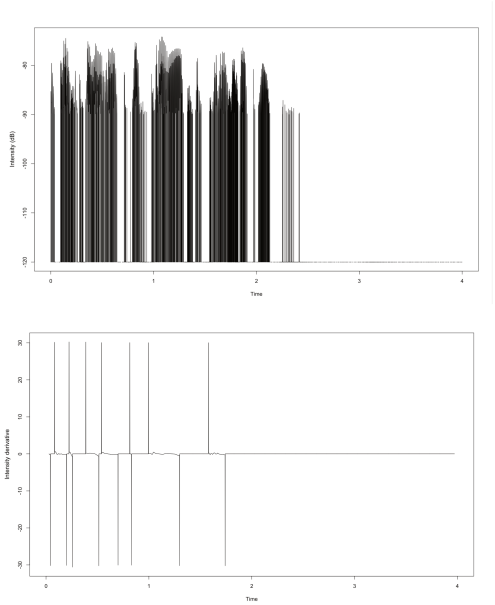


Figure 6.8: Top: intensity values from a 4 s sequence in the NCCFr corpus. Bottom: the filtered derivative wave of the same sequence that has been used to compute speech dynamics.

The choice for the Multinomial logistic regression approach is mainly related to the comparison with results obtained by [He and Dellwo, 2017; He et al., 2019] on similar dynamics representations. In a few words, Multinomial logistic regression predicts the probability of a dependent variable to belong to a group based on multiple independent variables using maximum likelihood estimation. It does necessitate a consequent sample size, hence our choice of working exclusively on the NCCFr corpus for this experiment. In our case, the main advantage of this statistical approach is that it does not assume normality, linearity, or homoscedasticity of the data. The results are expressed as the LLR between the submodels, Table 6.6 and Table 6.7 summarise what we have observed. As described in [He and Dellwo, 2017], the χ^2 value of the final model was calculated by taking the difference between the LLRs of the full model and the final model. In our case, the χ^2 value of each tested measure was calculated by taking the difference between the LLR of the final model and each reduced model. The explained variance was calculated by taking the percentage of the χ^2 value of each measure over the sum of all χ^2 values for all measures.

The total variance explained by these measurements is 64 % for female speakers and 77 % for males with a statistical significance for the full model. In detail, we notice that not every submodel is significant and not every measurement has the same weight on the analysis. The actual value of the derivative is significant for the male between-speakers

	LLR	χ^2	$p < 0.005$	Explained variance
(i) Model fitting information				
Null model	145.65			
Full model	65.728	79.922		64.25 %
(ii) Likelihood ratio test of each measure of intensity dynamics				
T	65.756	-0,028	NO	-0.02 %
f	66.007	-0,279	NO	-0.22 %
O_t	20.803	44,92	YES	36 %
O_q	35.513	30,214	YES	24.2 %
T_{neg}	62.309	3,418	YES	2.73 %
f_{neg}	67.575	-1,847	NO	-1.48 %
T (SD)	67.034	-1,306	NO	-1.04 %
f (SD)	65.741	-0,013	NO	-0.01 %
O_t (SD)	64.622	1,105	YES	0.88 %
O_q (SD)	64.437	1,290	YES	1.03 %
T_{neg} (SD)	63.146	2,581	YES	2.06 %
f_{neg} (SD)	65.629	0,098	YES	0.07 %
		$\sum \chi^2 = 80.156$		

Table 6.6: Results of multinomial logistic regression for female speakers of NCCFr.

	LLR	χ^2	$p < 0.005$	Explained variance
(i) Model fitting information				
Null model	166.504			
Full model	80,59	85.914		76.81 %
(ii) Likelihood ratio test of each measure of intensity dynamics				
T	80.986	-0.396	NO	-0.34 %
f	77.856	2.734	YES	2.39 %
O_t	3.388	77.201	YES	67.65 %
O_q	81.108	-0.518	NO	-0.45 %
T_{neg}	79.998	0.591	YES	0.51 %
f_{neg}	80.781	-0.191	NO	-0.16 %
T (SD)	81.097	-0.507	NO	-0.44 %
f (SD)	80.368	0.221	YES	0.19 %
O_t (SD)	79.216	1.373	YES	1.20 %
O_q (SD)	76.807	3.782	YES	3.31 %
T_{neg} (SD)	77.881	2.708	YES	2.37 %
f_{neg} (SD)	79.950	0.639	YES	0.56 %
		$\sum \chi^2 = 87.64$		

Table 6.7: Results of multinomial logistic regression for male speakers of NCCFr.

description but not for the female one. In a similar way we notice that the O_q has no impact for male speakers, while it is the most important variable in female between-speakers models. These differences emphasise the idea that sex is an important influence factor when considering speech production strategies. In particular the difference concerning O_t and O_q indicates that for female speakers the uprising of the speech production is more characteristic than for male speakers, whose closing trajectories have a higher discriminating power.

When faced with the prediction phase the regression model behaves in a similar way to the ones described above. Even though the statistical description has a high variance, the speakers' models are not robust enough to be able to classify them with consistent results. Similarly to the PCA's results, male speakers show a higher classification scores, i. e. 0.24, while female speakers do not exceed 0.20 of F1-score.

A relative increase in performance is shown when using dynamic computations in order to model speakers' characteristics in spontaneous speech. It has to be noted that the approach we applied is more demanding in terms of both computational power and pre-processing of the data. This leads to overall better results than the ones obtained through linear models and SVM classifications. It appears that using a more complex processing algorithm helps model an already complex input data such as the highly variable object which is spontaneous speech.

6.4.1 Voice comparison

For the second kind of experiments concerning the modelling of the speech dynamics we mainly based our study on the forensic phonetics approach, which has been discussed in Section 3.1. The phonetic measurements we focused on are f0 and the first four formants, since they are very commonly used in forensic voice comparison studies. We used polynomial functions to recreate the trajectories of the first chunk of each speaker of the PTSVOX corpus, for which duration was normalised as it is the standard for the polynomial approach. Some studies have also shown that better results are obtained when polynomials are calculated from normalised duration rather than raw duration [san Segundo and Yang, 2019; Rose and Wang, 2016].

Formants and f0 trajectories are extracted using Voicecause and model by permuting polynomials of degree corresponding to the number of syllables in the analysed chunk. Curve fitting procedures aim to perform data reduction by transforming a set of data points, constituting the formant trajectories, into a small set of coefficients. In our case 10 syllables were present to which we added two more orders since the polynomial function must include an offset or constant value and a slope coefficient. Afterwards, their coefficients are extracted for LLR processing using a MVKD analysis. We focused on PTSVOX since this particular approach demands exactly the same linguistic content to be compared. Even when using isolated words from NCCFr the acoustic trajectories appeared too unstable for the polynomials to perform a robust modelling.

LLRs were estimated using the formulas of MVKD described by [Aitken and Lucy, 2004] and implemented in R [R Core Team, 2021]. The Kernel estimation formula compares the *Mahalanobis* distance between the suspect and the offender mean vectors against measures derived from the same- and different-speaker (co)variances to determine the LLR. With this procedure any correlation between the variables is considered and the ratio of between- and within-speaker (co)variances acts as a key scaling factor. The output of the formula is scored by quantifying the ratio of the similarity of the difference between the suspect and offender samples and their typicality given a suitable reference sample. The latter is represented by a population made of all the corpus minus the compared speakers.

Each of the 15 retained speakers from PTSVOX completed two recording sessions, hence we had the possibility to compare the same speakers, 7 for female speakers and 8 for male

speakers, as well as comparisons of different speakers which lead respectively to 42 and 56 comparisons. Every comparison was done for both microphone and telephone sequences.

Scores for f0 and each formant were obtained separately and then combined in order to investigate their interactions. In Table 6.8 we report the number of comparisons done for this experiment and the results in terms of the C_{llr} metric, the lower the numbers are the better is the result. Following what has been observed in the statistical description of the same speakers, we notice that f0 has a higher influence for female speakers than for male speakers. Indeed, f0 for female speakers has the best score in our experiment with 0.14, while for male speakers the best score is achieved by F3 with 0.25. Higher formants have a better performance for male speakers when taken singularly in both recording supports. While lower frequencies show higher scores for female speakers in microphone results, F2 and F3 trajectories models perform best in telephone recordings.

	$PTSVOX_{mic}$		$PTSVOX_{tel}$	
	F	M	F	M
# of comparison	7; 42	8; 56	7; 42	8; 56
Single measuremnts				
f0	0.14	0.38	0.29	0.42
F1	0.40	0.40	0.44	0.40
F2	0.43	0.50	0.46	0.52
F3	0.43	0.25	0.48	0.32
F4	0.45	0.35	0.49	0.37
Double measuremnts				
f0-1	0.52	0.63	0.60	0.68
f0-2	0.50	0.62	0.69	0.66
f0-3	0.62	0.47	0.72	0.52
f0-4	0.43	0.41	0.50	0.48
F1-2	0.45	0.54	0.53	0.58
F1-3	0.55	0.53	0.59	0.56
F1-4	0.50	0.53	0.52	0.59
F2-3	0.69	0.37	0.80	0.40
F2-4	0.62	0.54	0.70	0.55
F3-4	0.48	0.56	0.59	0.56
More than 2 measuremnts combined				
f0-1-2	0.45	0.49	0.55	0.51
F1-2-3	0.45	0.45	0.55	0.47
F2-3-4	0.45	0.54	0.67	0.56
f0-1-2-3	0.50	0.45	0.52	0.46
F1-2-3-4	0.48	0.43	0.61	0.46
f0-1-2-3-4	0.52	0.50	0.54	0.53

Table 6.8: C_{llr} results and number of voice comparisons, same speaker and different speaker pairs, for the first read chunk from the PTSVOX corpus.

Overall, for both sexes, there is a deterioration in the score with telephone recordings. Furthermore, combining multiple trajectory coefficients in order to model more than one phonetic measurement does not increase the results. For female speakers, only the combination of non adjacent formants gets scores that are comparable to the ones obtained by single measurements. In a similar way, male speakers show a non redundant combination

for F2 and F3 as well as f0 and F4.

The reported poor interaction between these acoustic cues has already been observed in the previous analysis. In detail, we notice that formants and f0 do not behave in the same way for all speakers. The combination of f0 and both F1 and F2 perform better than average for a small group of female speakers, while the combination of F2 and F4 shows the same pattern for male speakers. These observations confirm that features influencing the acoustic outcome of speech production are not evenly distributed for all speakers. In addition, it should not be overlooked that polynomial coefficients may convey redundant information since they rely on a mathematical transformation rather than on a direct result of speech production.

Studies from the forensic phonetics domain usually focus on modelling isolated words or phonemes and diphthongs in order to assess similarity between speakers quantified by LLRs. In our study, we focused on larger speech sequences in order to model speakers' dynamics and to investigate how phonetic characteristics can be recovered by these models. Our results from f0 and formants trajectories confirm what we observed in the static investigations described above. Speakers' information is not evenly distributed on the frequency domain, clear differences are observed for female and male speakers. In addition, the recording support has an important influence on the degradation of results especially for female speakers.

The main downside of the presented approach is represented by the high demanding preprocessing and calibration of the analysis itself. However, even though these explorations of modelling speech dynamics do not lead to state of the art results, they provide us information which is further explored in the next chapter, especially with the idea of combining several analysis elements from different domains.

6.5 Chapter conclusions

This chapter's aims were to add a contribution to the phonetic research on speakers' characteristics. We have provided results from both spontaneous and read speech in French. PCA have been applied to carry the statistical description of multiple phonetic measurements but the variability that the described measures manage to provide is not balanced when faced with classification problems. More complex approaches have shown to increase the robustness of the speakers' characteristics' models but the search for representations capable of describing the high intrinsic variability of the speech remains a major concern in the continuation of our studies. These representations have to retain both discrimination power and speakers' information in order to be considered successful.

Overall, the experiments presented in this chapter use classical Phonetics' approaches. We have been able to confirm some findings from previous phonetic investigations. The influence of sex on the majority of phonetic measurements is a major point in all our experiments, however, this does not mean that speakers' characteristics are distributed in the same way for all subsets of speakers. We have seen that some characteristics are more prominent in some subsets but irrelevant for others, e. g. this was the case for spectral shape variability information. In a similar way, the influence of the recording medium and differences between speech types were assessed, providing evidence of the ability of

modelling the same information from less clear signals.

In order to conclude this section a summary of the entire chapter is presented, including the main findings and approaches that have been analysed throughout the presented experiments. In continuation of this phonetic based chapter, Chapter 8 provides additional discussions for both present results and those that are presented in the following chapters of the second part of this thesis.

6.5.1 Summary

- **Reference values**: overall, there is an important degradation of all phonetic measurements when comparing microphone and telephone recordings. However, male speaker information is more coherent between the two supports:
  - Overall coherent values for formant frequencies in line with the literature in French, vowel space areas do not appear to be significant between speakers;
  - High variability of /y/, /u/ and /o/ in spontaneous speech.
- **Temporal measurements**: while speech rate does not appear significant in both within- and between-speakers comparisons, duration of pauses, and of words' last and first phonemes show high within-speaker consistency;
- **PCA**: explained variance is higher for read than for spontaneous speech:
  - Confirmation that formants are more characteristic for female than for male speakers;
  - Female speakers use lower formants and f0 information, while voice quality, i. e. breathiness and hoarseness indicators have a much higher weight for male speakers characterisation;
  - MFCC have a lower statistical power than classical phonetic measurements;
  - Combining phonetic measurements and MFCC does not influence the distributions of the former in the PCs, neither increases the overall explained variance;
  - Interestingly, lower rank MFCC emerge alongside intensity and f0 for female speakers, while energy distribution and low level spectral shape information with middle rank MFCC for males;
  - The statistical descriptions do not result in models robust enough using the LDA. SVM modelling shows more promising results for the read speech sequences but the classification results remain very poor for spontaneous speech.
- **LLR approaches**: more complex analyses provide more consistent results:
  - For spontaneous speech interesting results are obtained using the open quotient-like approach. However, the problem of a high variance explained not balanced by scores in classification tasks still remains;
  - O_t is significant for female speakers and O_q for males, indicating that the uprising of the speech production is more characteristic for female speakers, while for males closing trajectories have a higher influence;
  - Using polynomial functions to recreate f0 and formants trajectories is highly demanding in terms of linguistic content coherence, preprocessing and actual processing of the analysis. However, promising results are obtained;
  - Clear differences between female and male speakers: f0 is the most performing measurement for the first ones and F3 for the latter;
  - Polynomial coefficients seem to convey redundant information, since no improvements are shown by the combination of multiple parameters.

Chapter 7

Natural Language Processing

The present chapter focuses on the results obtained following a NLP approach, in other words the transformation of phonetic measurement into representations suitable for an analysis through Deep Machine Learning. Three tasks, including speakers identification and verification, helped us understand how the CNN model speaker's information. As already mentioned, knowing where and how a speaker contributed to an audio recording plays a critical role in the resulting models and their robustness.

The fundamental idea of our study is that the majority of automatic systems use spectral measurements over small analysis windows, while merging phonetic and automatic approaches from a NLP perspective can help understand how different speech components contribute to the definition of a speaker model. The use of interpretable phonetic measurements representing different aspects of speech is not possible without phonetic knowledge. The automatic approach plays a role in applying precise representations and improving the robustness of their modelling. The two approaches constitute mutual helping aspects applied throughout this thesis and particularly underlying the results presented in the present chapter.

In the following sections, we first present the workflow of our speakers characterisation study and then we present all the obtained results. The extraction methods and regrouping of phonetic measurements correspond partially to what has been discussed in Chapter 6, to this we add the adaptation of those measurements to representations for our CNN. The ML metrics that are used to evaluate the results are explained along with the other methods and the differences between the three performed tasks. The presentation of the results is always divided by sex, like in the previous Chapter, and by task.

7.1 Convolutional Neural Networks methods

The workflow for the speaker characterisation studies through CNN approach is summarised in Figure 7.1, with each colour representing a different phase. In Phase 1 (yellow) we preprocess our obtained data for every speaker in the analysed corpus from the speech recordings, chunks of a duration of 4 s, mostly using NCCFr but in a preliminary study we used some modified data from PTSVox, see Section 7.3. The used chunks are the same as those analysed in the previous Chapter. Some of the experiments were performed on

specific phonemes, see Section 7.3, with no particular annotation except the one concerning the actual talking speaker and a minimum of 20 phonemes were considered in all of the others.

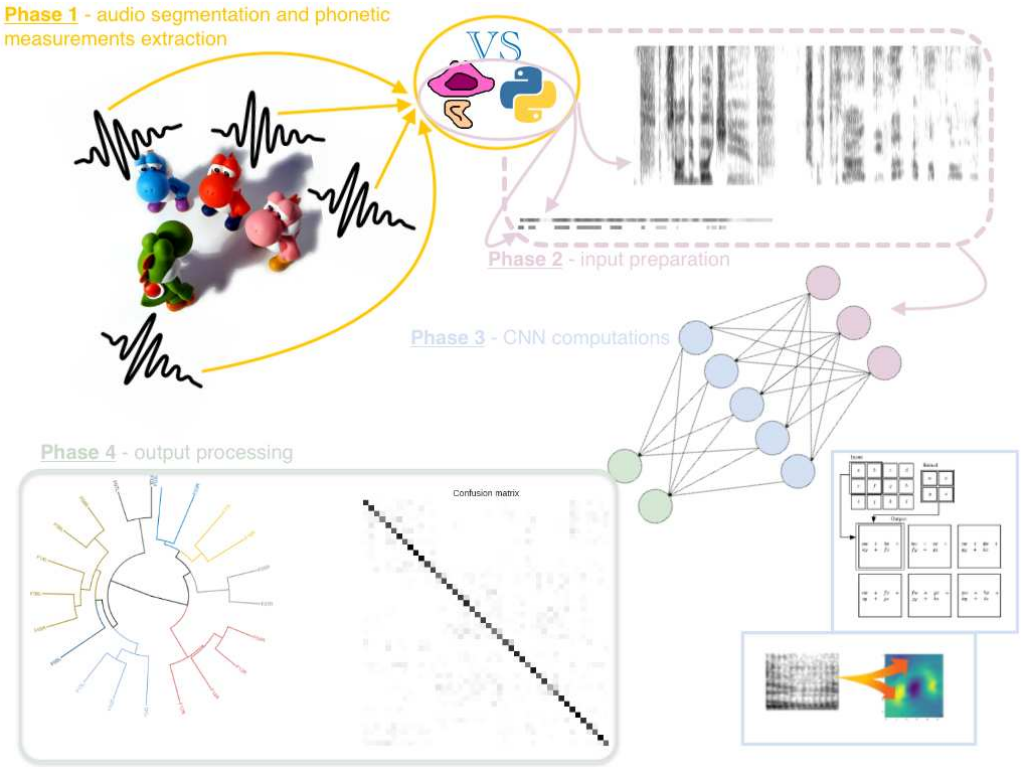


Figure 7.1: Workflow for speaker characterisation through Convolutional Neural Network.

We used Praat on the segmented chunks, through a custom script, and Voicesauce to compute every ms of the selected phonetic measurements, see Table 7.1, as well as 13 MFCC every 10 ms, with Python [van Rossum and Drake, 2009]. These extractions allow us to recreate the exact modulation of the different selected phonetic measurements. In Phase 2 (purple), we organise the inputs which are fed to the CNN, through a custom Python script converting the phonetic values in greyscale images as well as spectrograms from Praat. For spectrograms extraction, we chose to use 5.0625 ms frames and 0.5 ms hop size based on Praat default values, with a 16 kHz sampling rate. The speech sequences were element-wise multiplied by a Hamming window, from all the available window shapes in Praat (Gaussian, Square, Hamming, Bartlett, Welch and Hanning) it was the one giving the best results, no pre-emphasis was performed. The dynamic amplitude range was normalised to 70 dB to make sure that the dynamics did not bias discrimination.

As for Table 6.4 in Chapter 6, Table 7.1 below summarises the groups of measurements used in this chapter's experiments. For each group we report the measurements it includes as well as the related component, and the type of the representation: we define two types of global representations that translate to a wide representation of speech production, **global** for all production mechanisms e. g. in *Spectros*, and **component-global** for an entire component e. g. in *Qual*; whereas a **subset** representation means that only part of a component has been used in order to study the influence on speaker characterisation of specific measurements. The global representations play a major role as they are taken as baseline results. There is no explicit representation of temporal cues, however, in rendering the modulations of speech components we intrinsically incorporate elements

such as duration or pauses.

Group name	**Measurements**	**Component**	**Type**
Amp	Amplitudes of the harmonics near the first three formants (A1-3)	Source and filter	Subset
f0	f0 and its harmonics (H1, H2, H2k, H42k, H5k)	Source and filter	Subset
Form	First four formants (F1-4)	Source and filter	Subset
Acoust	Combination of *Amp*, *f0* and *Form*	Source and filter	Component-global
Int	f0 and intensity	Intonation	Subset
Env	ENV and TFS	Temporal	Subset
Pros	Combination of *Env* and *Int*	Prosody	Component-global
Nrg	RMS, soe, Praat-based energy	Source and filter | Mode of vocal fold vibration	Subset
Ms	Four spectral moments (center of gravity, standard deviation, kurtosis, skewness)	Source and filter | Mode of vocal fold vibration	Subset
Ha	Differences between harmonics (H1-H2, H2-H4, H2k-H5k)	Mode of vocal fold vibration	Subset
Hh	Differences between harmonics and amplitudes (H1-A1, H1-A2, H1-A3)	Mode of vocal fold vibration	Subset
Hadiff	Combination of *Ha* and *Hh*	Mode of vocal fold vibration	Component-global
Hr	HNR at different pitch ranges (0-500 Hz HNR05, 0-1500 Hz HNR15, 0-2500 Hz HNR25, 0-3500 Hz HNR35), SHR	Mode of vocal fold vibration	Component-global
Ltas	LTAS from four different frequency bandwidths between 1 and 5 kHz	Mode of vocal fold vibration	Component-global
Qual	Combination of *Ltas*, CPP and *Nrg*	Mode of vocal fold vibration	Component-global
MFCC	13 MFCC	?	Global
Glob	Combination of all phonetic measurements	All	Global
Spectros	Entire wide-band spectrograms	All	Global
MPS	Modulation Power Spectrum	Source and filter	Global

Table 7.1: Phonetic measurements groups used in Chapter 7 experiments, related components and type of representation.

The resulting input sizes of the spectrogram images correspond to 800×257 pixels (px), where 1 px corresponds to 5 ms in the time dimension and 31.13 Hz in terms of frequency. The resizing used for spectrogram images, bicubic interpolation, was essential for GPU memory handling. On the other hand, for the phonetic parameters images the horizontal axis was not resized in order to preserve all dynamic information. On the vertical axis pixel sizes correspond to the parameters number in a 1:1 ratio, e. g., images from the four formants were 4×4000 px, while for the MFCC the corresponding size was 12×400 px.

Every image was converted to 8-bit greyscale (0 to 255, from darker to lighter) allowing GPU memory to handle mini-batches of sufficient size. The greyscale levels were computed separately for every measurement using the absolute maximum as darker value (0) and absolute minimum as light grey (200). In order to differentiate NaN values from zeros they were taken out of the absolute minimum scale and assigned to the lighter value of

255. Figure 7.2 shows a visual example of the images representing the measurements' modulations that are obtained in Phase 2, and used as inputs for the CNN.

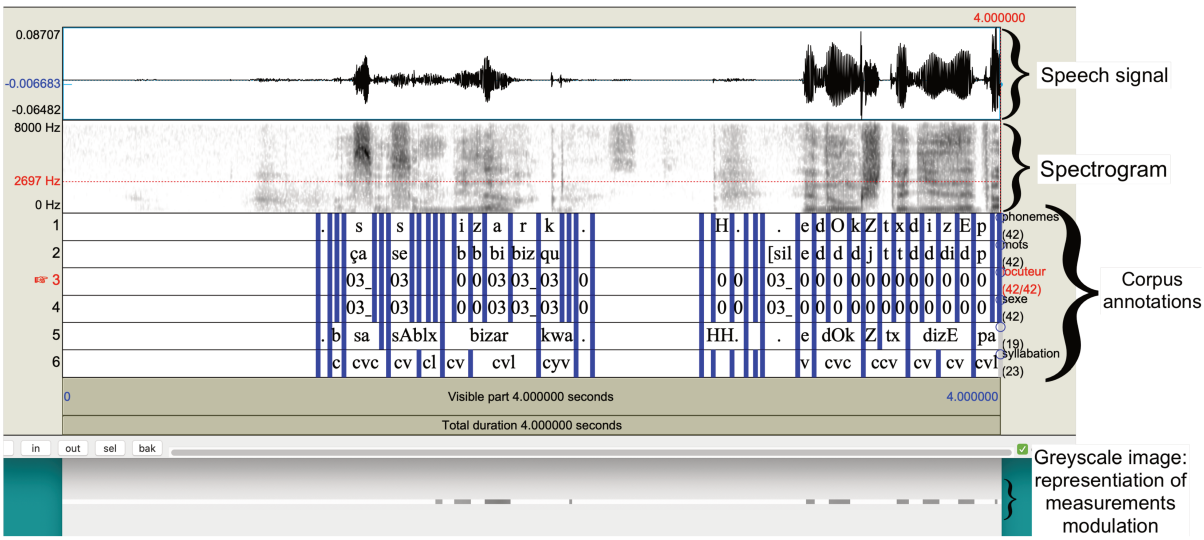


Figure 7.2: Examples of spectrogram and modulation representations for a 4s sequence from the NCCFr corpus.

Once all the input is processed, Phase 3 (blue) of the workflow consists of CNN computations. As mentioned in Chapter 3, CNN are the state of the art for image recognition, for our main experiments we used a slightly altered version of a ResNet-18 architecture [He et al., 2015a]. In preliminary studies, with sequences composed of only selected phonemes, we obtained consistent results using a basic LeNet architecture [LeCun and Bengio, 1995], see Section 7.3. By convolving a filter matrix on the input image, during the train phase, the CNN builds the model associated with the considered class, a speaker in our case, by extracting salient information. Depending on the task, the obtained models are compared in different ways. However, in every test and validation phase an overall success score is computed from the assigned probability of match between the tested token and the class. Further discussions on the differences between each task are presented in Section 7.1.1.

During all our experiments, speaker tokens were separated into training (70 %), validation (10 %) and test (20 %). All subsets of tokens were randomly extracted. The validation score was fundamental in order to prevent the CNN from overfitting, if the score reached a determined threshold, the training phase would come to a stop and the test phase would start. All the CNN models presented in the present Chapter have been trained and tested on a GTX 1080 Nvidia GPU using either MATLAB Deep Learning Toolbox [Mathworks, 2019] or custom scripts based on the Python library Keras [Chollet et al., 2015].

Processing the output of the test phase represents the final part, Phase 4 (green), of our workflow. We immediately compute a confusion matrix, using the assigned probabilities, from which multiple metrics are computed in order to evaluate the results, see section 7.1.2. Afterwards, the results of this NLP approach are compared to those obtained with the phonetic approach in order to understand the true reliability of the representation used to characterise the speakers' productions. Another use of the automatic modelling responses is the clustering of voice for a better investigation of (dis)similarity between the studied speakers. These last results are the main focus of discussion in the next Chapter.

7.1.1 Neural Networks tasks

As discussed in previous sections, speaker recognition can be studied through different experiments that involve speaker models comparison in different ways defined as recognition tasks. In a similar way, Neural Networks can be tuned to accomplish a multitude of tasks, for our main experiments we retained three protocols enabling us to study the interactions between speech components in regards to speaker characterisation.

The first task to be used is an **identification** task. During the training phase, the different speakers are modelled at the same time by the CNN, hence 23 models are created for the male speakers and 21 for the female speakers. During the test phase, an equal number of tokens for each speaker are compared with the learned models obtained through the training phase. In the last layer, the tested speaker is assigned with a probability of belonging to every possible class by the CNN algorithm. The classification rule, after applying a *softmax* function to normalise the sum of probabilities and generating probabilistic distributions, provides us with the highest scoring class, retained as the predicted speaker. If this predicted answer corresponds to the label of the tested class we consider that the classification is a success, however if the highest score is assigned to another speaker, the wrong answer is retained to feed the confusion matrix. This type of closed environment test allows us to understand how the studied traits, defined by the phonetic measurements, can be shared within a closed set of speakers. This task is the most basic and similar to baseline approaches in Phonetics. In order to understand the weight of the different cues, we reduced and altered the multiple subsets. All the different results are presented in the following sections giving an idea of how the considered characteristics interact.

In a different approach, we performed a **verification** task, for which each speaker is modelled along a reference population. The two resulting models are tested in a binary classification task. As in the identification task, the number of tested tokens per class is equal and randomly extracted. Every speaker is tested individually against a different population: the remaining speakers minus the target one. In this case, instead of learning multiple variation matrices, two of them are created, one related to a single speaker and the other one resulting from the addition of a greater variability matrix. In the binary classification that is the verification task, the algorithm applies a threshold to decide which class has to be retained as the predicted one. The main difference between the two tasks is the interpretation of their answers: the identification probabilities provide an insight on the similarity between the compared models emphasising the shared traits, since they are known by the model; the verification responses highlight the differences between the speakers, due to the comparison between a known smaller variability matrix and a large heterogeneous set of characteristics.

The third task we applied in our main experiments is the result from the verification principles and the answers from the identification task. We have decided to call this task **generalisation**. The underlying hypothesis is to test the robustness of the speaker model both against a larger population and against unknown speakers, which have formerly been considered similar to the target speaker, in order to generalise the speaker characteristics. Similarly to the verification task, in the training phase only two models are created: the target speaker and a population, this time consisting of all the remaining speakers minus the target one and the three speakers that, during the identification task, have been mostly confused with the target. The test phase performs a binary classification on three

sets of tokens, the target speaker, the reference population that have been modelled and the unknown speakers. All sets present an equal number of tokens.

The results from the generalisation task give us a complementary view on the study of speaker characterisation. The idea of comparing speaker models in an open environment, i. e., against other models, unknown by the classifier, is an important issue in all the domains that this thesis is inspired by. Finding that a particular speech characteristic is able to approximate a speaker's variation matrix represents an important issue for the forensic applications of voice comparisons, but also for the phonetic and NLP perspectives. However, we did not include every component nor measurement subsets for the generalisation task but only the most relevant, because of the consequent preprocessing needed to carry this task. We specify which representations have been selected in the results reported below.

7.1.2 Machine Learning evaluation metrics

In the previous chapter we have evaluated the performance of the different phonetic measurements in speaker characterisation using statistical modelling common in Phonetics. In the Machine Learning (ML) field, in order to evaluate the performance of an algorithm in resolving a *classification problem*, multiple indicators can be used. The principal aims are to evaluate and compare the performances of different models, as well as analyse multiple behaviour of a same model by tuning different parameters. Hence, we evaluate both the influence of the input and the ML technique itself. We briefly explain which are the ones we used in our experiments and what they tell us about the modelling performance, see [Powers, 2011; Grandini et al., 2020; Chicco et al., 2021] for more in-depth discussions.

The goal of a ML model is to obtain the best prediction $\hat{Y}$ of the outcome variable Y using the available modelled classes from the data $\mathbf{X}$. In multiclass classification, the response variable Y and the prediction y may be seen as two discrete random variables that assume values in $\{1, ..., K\}$, with K as the number of classes and each number representing a different one. A basic set of metrics can be obtained from the confusion matrix, since it encloses all the relevant information about the algorithm and classification rule performances. Accuracy is one of the major classification performance indicators. It returns an overall measure of the model's ability to correctly predict the classification of a single individual on the entire data set. *Null* or *Expected* accuracy, also called chance level, is considered when taking into account the accuracy metric. Other indicators of correctly identified cases are **Precision** (or **Positive Predictive Value**), measuring the correctly identified cases from all the predicted ones, and **Recall** (or **Sensitivity**), measuring the correctly identified cases from the actual class. However, considering these metrics alone does not provide us information about the underlying distribution of answers or the types of errors the classifier has made. Hence, we consider the **F1-score** as the basic evaluation metric, it is computed as the harmonic mean of Precision and Recall, and gives a better measure of the incorrectly classified cases. It ranges from 0 to 1, with a score of 1.0 corresponding to the highest performance from a ML model. The use of the harmonic mean penalises extreme values making the F1-score more suited to explain the influence of False Negatives and False Positives on the classifier, i. e., type I and type II errors. These are measured respectively by Recall and **Specificity**.

We also include the metrics for **Informedness** and **Markedness** when considering speak-

ers' results separately, their ranges are both $[-1; +1]$, a score closer to 1 is better. Informedness represents how the predictor, in this case our CNN, has been able to use the information of the considered class, the speaker in our case. It is computed as *Specificity + Sensitivity - 1*. Markedness, on the other hand, is *Positive Predictive Value + Negative Predictive Value - 1* and indicates how a speaker is marked by the predictor, in a sense how trustworthy the predictions are for the considered speaker. The worst value these metrics can take is 0 as it indicates the chance level, hence this score would indicate poor use of speaker information or poor characterisation of a speaker among the other modelled speakers.

The last indicators we report represent the measures of agreement between the distributions by pairs of predicted and actual values of the classifications. Firstly, The **Mattheus Correlation Coefficient (MCC)**, expresses the degree of correlation between the two named variables in a range $[-1; +1]$, much like the Pearson's Phi-Coefficient, pointing out different model behaviours during the training phase of the algorithm. It is considered, in binary classification problems, more informative than F1-score and accuracy, since it takes into account the balance ratios of the four confusion matrix categories (true positives, true negatives, false positives, false negatives). Secondly, the **Cohen's Kappa (K)**, which represents the dependence between the two distributions, allowing a reliable value to compare different models. It measures the degree of agreement between the true values and the predicted values, which can represent the classifier's performance. A score of 1 means a perfect match between true and predicted values while 0 is chance agreement. If the value is less than 0, something is wrong with the classifier as it is performing worse than chance agreement.

Finally it has to be noted that, of all the experiments that are presented in the following sections, in the preliminary experiments, as the name suggests, only the basic evaluation metrics have been computed as they were aimed to establish the validity of the protocol and its underlying directions.

7.2 Lexical distances

The content of the spoken message can, in some cases, be used to characterise speakers by their speech productions. In a speaker characterisation perspective, we could assume that speakers differ significantly in their productions at a lexical level as they do at the phonetic level. We all have theoretically the same articulators for the production of speech but the actual resulting sounds present different characteristics. In the same way, the combination of a finite set of elements can vary between-speakers by a large number of factors such as the use of foreign words and different dialects. In this section we discuss these considerations in regards to the corpora of our studies. Starting with PTSVOX, no lexical distances have been computed, since the reading task consisted in the same linguistic content for all the speakers, i. e. reading of three small texts (Appendix B). However, small differences in pronunciation have been observed between-speakers, they are irrelevant from the lexical standpoint in the same level as repeated words and hesitations, since in the latter case speakers were asked to repeat the sentence.

For the other corpus of our focus, NCCFr, some studies have already been done by the lexical standpoint in the original presentation article, [Torreira et al., 2010]. Hereafter,

we summarise these findings and add some new elements concerning speaker-specific differences. As said before, the data we analysed from NCCFr consists of 44 speakers during casual conversations. Some indicators, such as the considerable amount of overlapping speech (3 hours in total) confirm that the corpus contains highly interactive speech, a lexical analysis contributed to assess the actual casualness of the recorded speech. The presence of disfluencies, repeated words, the absence of double negation and other purely lexical elements such as swear words or the use of formal synonyms, e. g., "truc" instead of "chose" for "thing" have been assessed by the Authors as indicators of casualness. 5 speakers used more swear words than the rest, while 5 did not pronounce any. Casual word use was measured by adding the total number of tokens of casual and formal content words and then calculating the percentage of casual words over this total. The selected casual words were *cela/ça* (that), *chose(s)/truc(s)* (thing/stuff), *garçon(s)/gars* or *mec(s)* (boy/buddy), *oui/ouais* (yes/yep), *très/vachement* (very/freaking).

8 speakers did not pronounce any of the casual variant of these words. With respect to the function casual words, the word *ça* was used by all speakers, while few occurrences, 6, of the more formal variant *cela* were shared by 3 speakers. The word *ouais* showed more variability, with 32 speakers using it between 30 % to 95 %, with a mean of 69.8 %, and 14 speakers not using it at all. Interestingly, these 14 speakers used *oui* as often as the other participants. *Verlan*, a language game consisting in the inversion of segments and syllables in a word, was used by 60 % of the speakers. Other indicators also showed that most speakers used casual speech. Double negation was generally low across speakers, as expected from our previous analyses, with only a small number of significantly deviant speakers: 3 speakers displayed double negation rates between 15 % and 30 %, and two showed surprisingly high rates (38.9 % and 55.8 %). Furthermore, all speakers exhibited at least five repetition bigrams per thousand words and disfluency words were used by all speakers except two. No correlation was showed between the multiple indicators of casualness.

We used two additional methods in order to lexically compare the NCCFr speakers: the extraction of keywords, Table 7.2; and whole lexical similarities computed by the means of cosine metric, Figure 7.3. Concerning the first method, we reported 5 words per speaker in Table 7.2, with the relative keyness value via Term Frequency–Inverse Document Frequency (TF-IDF) statistics listed in Appendix C. The highest scoring words of a document are the most relevant to that document, and therefore they can be considered keywords for that document.

TF-IDF measures how important a word is to a document in a collection of documents, in our case the document corresponds to the spoken words during a speaker recording session, while the documents collection is the sum of spoken words of all speakers. This metric calculates the number of times a word appears in a text (term frequency) and compares it with the so called inverse document frequency, how rare or common that word is in the entire data set. Multiplying these two quantities provides the TF-IDF score of a word in a document. The higher the score is, the more relevant the word is to the document. In many cases, the words that appear more frequently in a group of documents are not necessarily the most relevant. A word pronounced by a single speaker but not by the remaining ones may be very important to characterise that speaker's lexicon.

All speakers attended the same task, however in different pairs, where they were asked to find a common answer to a set of questions on various themes, e. g. smoking ban in

public places or politics, we assume the lexical content differences are not so evident. However, the fact that they all already knew their interlocutors made them more likely to produce spontaneous conversations outside the demanded task. In this sense, we observed that female speakers pronounced the name of their interlocutor or of a third person more frequently than males, Table 7.2, indeed, 7 times female speakers present a name in their first 5 keywords against 3 from males. When analysing more than 5 keywords this trend was even more relevant, resulting with 4 more names for female speakers. Through keywords analysis we observe that the most common words are *ce* (this), *est* (is), *tu* (you), *je* (I), *oui* (yes). They are all function words not carrying a particular meaning by themselves. When comparing the actual keywords, the most common keyword appears to be *fille* (girl), followed by *France*, appearing respectively 12 and 10 times in the first 5 keywords for the 44 speakers.

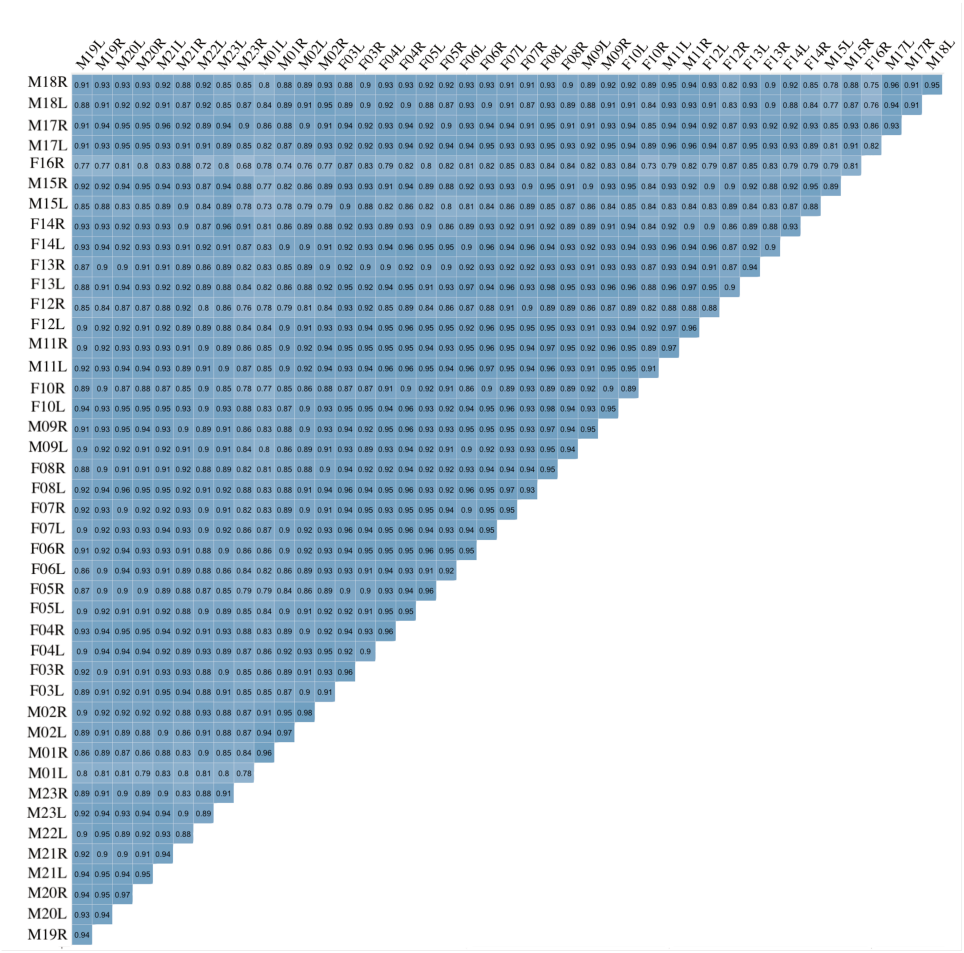


Figure 7.3: Cosine similarities between the 44 NCCFr speakers.

Other synonyms to define a female person in French are very common in the keywords list, even though some of them present a more specific semantics: *femme*, *gamine*, *copine*, *meuf* as well as *soeur* (sister), *maman* and *mère* (mother). As said, 42 out of the 44 speakers we studied are in pairs and we observed that these speakers share at least 1 keyword out of 5 within their pair. The presence of swear words is more prominent in male speakers, 3 of them presenting at least 1 swear word as a keyword, most likely used as an intercalary word.

Speaker	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5
M19L	théâtre	lettres	université	France	nombre
M19R	lettres	France	université	monde	Nobel
M20L	problème	BTS	école	Gatinot	Paris
M20R	Paris	envie	paysage	France	université
M21L	filles	monde	femme	copine	vie
M21R	filles	lu	lycée	films	besoin
M22L	salle	vie	France	espèce	art
M23L	politique	problème	UMP	gauche	travail
M23R	supporters	club	problème	gauche	Paris
M01L	Week-end	relou	matin	Paris	Marie
M01R	Sarkozy	France	euros	Virginie	anniversaire
M02L	France	Sarkozy	Jeanne	Ségolène	fille
M02R	Sarkozy	homme	France	différence	maison
F03L	Laurence	enfants	vie	théâtre	pauvre
F03R	femme	mère	Liliane	semaine	soeur
F04L	histoire	argent	prix	étoiles	chansons
F04R	moment	prix	envie	bizarre	vie
F05L	cafés	fille	genre	faim	Diams
F05R	soir	pied	Diams	frère	peur
F06L	mère	fille	Lolotte	gamine	parents
F06R	envie	mère	argent	Morgan	gamine
F07L	enfants	parents	problème	mercredi	filles
F07R	Marie	enfants	maman	garçon	interdiction
F08L	vie	Sandra	micro	garçons	pays
F08R	clope	soir	frères	vidéo	euros
M09L	intérêt	étudiants	France	Maine	garçons
M09R	France	dépénalisation	Toulouse	garçon	match
F10L	président	frère	compagnie	femmes	ordinateur
F10R	écran	personnes	lecture	vendredi	univers
M11L	putain	école	semaine	fin	euros
M11R	imagine	regarde	Sarkozy	soeur	métro
F12L	soeur	femmes	parité	mois	mètres
F12R	voiture	filles	fille	stage	Roubaix
F13L	Marion	filles	France	Paris	fille
F13R	filles	mignon	père	Marion	ans
F14L	film	limite	fin	journée	nuit
F14R	métro	Harry	mère	école	film
M15L	anthropologie	question	sociologie	étude	amie
M15R	Julie	voix	question	société	système
F16R	famille	copines	Laura	soir	feuille
M17L	femmes	balle	début	Ségolène	parents
M17R	euros	éducation	fou	soirée	femme
M18L	filles	euros	putain	niveau	regarde
M18R	putain	crème	euros	con	question

Table 7.2: Five keywords for each of the 44 NCCFr speakers, ranged by keyness score (TF-IDF), see Appendix C for relative values.

Figure 7.3 shows cosine similarities scores between the 44 speakers, these are obtained by vectorising the produced speech of each speaker and comparing them. The scores range from 0 to 1, a score of 1 corresponds to two speakers pronouncing exactly the same words the same amount of times. The average similarity is 0.9 with the highest similarity shown by the pair of speakers M02 with a rate of 0.98, while the lowest similarity is between-speaker F16R and M23R with 0.68. A high average score is expected due to the fact that the lexical core of all conversations remains highly constant between speakers and only a part of their lexicon varies due to some small changes in subject discussion. Considering only pairs of speakers the average similarity rate increases to 0.94 with a minimum of 0.88 registered by the F12 pair. The two isolated speakers (F16R and M22L) show different tendencies with the female one being on average less similar to all other speakers, 0.76, while M22L shows high similarity rates with all speakers, 0.88.

The words listed in Table 7.2 confirm the similarity rates shown by Figure 7.3. Aside from the French function words, which represent the common language spoken in every recording, each speaker presents little variations in their produced speech. However, cosine distance does not take into account the context in which the words are produced but only how many times they are present in a speaker conversation. Other distance computation methods such as Levenshtein could have served the purpose of weighting the context but the computational power needed to perform such a method for our corpus was too high. Even though pairs of speakers present 1 common keyword we still observe an important difference between the other keywords. Cosine similarity rate and keyword comparison can be considered satisfactory methods when characterising different speakers lexical contents since they do not require a high computational power and they are complementary in the information they convey about the lexical content.

7.3 Preliminary studies

This section rapidly presents the results from multiple experiments that have been carried on as preliminary studies in order to establish the main experiments protocol. Table 7.3, summarises some basic information such as the number of tokens for each speaker in the discussed studies, the corresponding input data, the total number of speakers and the analysed task. Only F1-scores are reported since, as mentioned above, it can be considered as a basic measure to evaluate results of a ML model. In the same table the last row concerns the main experiments that are presented in Section 7.4.

We first focused on phonemic level input data from the NCCFr corpus. Following the workflow presented in Figure 7.1, we extracted different spectrogram images of nasal vowels. The choice for this type of segment has been justified in Section 2.1.2. The spectrogram variations we worked on concerned both windowing techniques and subband analyses. The setup parameters described above have led to baseline results of 0.42 F1-score for both female and male speakers identification using /ɑ̃/ vowel. Adding other vowels material, i. e., /ɛ̃/ and /ɔ̃/, have shown higher scores of respectively 0.58 and 0.61 in classifying the 44 women and men from the considered corpus.

As shown throughout the previous chapter, in order to understand how speakers' information is conveyed by speech components we analysed sequences of speech production larger than isolated phonemes. Indeed, we used speech sequences from the reading task of the

PTSVOX corpus allowing us to have a more stable input data. Using spectrograms from 2 s speech sequences we obtained F1-scores of 0.61 and 0.66 for the 12 female and 12 male speakers. Given the limited amount of data from the analysed corpus, at the time of our first experiments we proceeded to add artificial noise to the original signal, e. g., keyboard tapping, rain and wind noises. The new data we analysed allowed us to investigate the influence of noise in the CNN modelling of speech characteristics. In a mismatch situation, where the noisy-augmented data is used only during the testing phase our models yield results of 0.41 and 0.60 respectively for female and male speakers. In addition, including the noisy data in the training phase as well led to a higher degradation of F1-scores with 0.36 for the 12 women and 0.49 for the men's identification. An alternative representation we used in our first exploratory approaches is the MPS of 2 s and 3 s speech sequences, in order to have a deeper in-sight of spectrum modulations. This representation has shown interesting results especially for the extraction of birdsong bioacoustics characteristics and speech intelligibility [Hsu et al., 2004; Elliott and Theunissen, 2009]. We classified 2 s samples of read speech data, obtaining scores of 0.38 for female speakers and 0.53 for male ones, with an increase of the F1-scores to respectively 0.58 and 0.56 with speech sequences of 3 s duration, hence more linguistic content to model. However similar results were not achieved when applied to the less stable spontaneous speech sequences.

One clear element emerging from our preliminary analysis concerns the higher scores obtained for male speakers in almost every comparison except with the CNN identification using /ã/ vowel spectrograms, where both sexes got the same scores. Adding noise to the original signal and using different speech representations also show higher performance degradations when modelling female characteristics

Corpus	**Speakers (F; M)**	**Tokens**	**Input data**	**Task**	**F1-score (F; M)**
NCCFr	44 (21; 23)	445	Nasal vowel /ã/ spectrograms	Identification	0.42; 0.42
NCCFr	44 (21; 23)	445	Nasal vowels /ã/, /ɛ̃/, /õ/ spectrograms	Identification	0.58; 0.61
PTSVOX	24 (12; 12)	180	2s spectrograms	Identification	0.61; 0.66
PTSVOX	24 (12; 12)	240	2s noise-added spectrograms (only testing phase)	Identification	0.41; 0.60
PTSVOX	24 (12; 12)	240	2s noise-added spectrograms (training and testing phase)	Identification	0.36; 0.49
PTSVOX	24 (12; 12)	200	2s MPS	Identification	0.38; 0.53
PTSVOX	24 (12; 12)	105	3s MPS	Identification	0.58; 0.56
NCCFr	44 (21; 23)	456	4s Spectrograms, MFCC and phonetic measurements images	Identification, verification, generalisation	See Section 7.4

Table 7.3: Summary of all the CNN studies for this thesis with relative information about the number of speakers and tokens per speaker, phonetic content of the named tokens and task carried on.

7.4 Main experiments

The following sections present the results of the main experiments of our study on speakers' characteristics. In more detail, we first compare in section 7.4.1 and section 7.4.2 the results from the three global representations: *Spectros*, *MFCC* and *Glob*. This allows us to explore the differences and similarities in how the speakers' information is conveyed by these different representations of the speech components extracted from the 4 s sequences. This section is the only one presenting results from the comparison of three different representations of speech production, the following sections report the results from the three main components cited in chapter 2 and their subsets of phonetic measurements: source and filter in sections 7.4.3 and 7.4.4; prosody in section 7.4.5 and 7.4.6; mode of vocal fold vibration in sections 7.4.7 and **??**. In each of these sections the identification and verification tasks results are presented and compared as well as the generalisation task, when performed.

Tables summarising the results are present in every subsection. The presented scores, Precision, Recall, F1-score, Specificity and K are associated with the speakers presenting the best Informedness value and the marked speakers. The last three metrics, as mentioned above, are especially important to understand how different models convey speaker information and what their comparison can tell us about specific components of speech.

7.4.1 Spectrograms as baseline - female speakers

Overall spectrograms show the best results in the identification task for both female and male speakers. In the first case, both Recall and Precision attain 0.96. Consequently, the average F1-score for female speakers has the same value. Specificity averaged 1.0 with only two speakers scoring below this values, indicating that the spectrograms generate a very low amount of False Positives. We notice that three female speakers achieve 1.0 for Informedness, and two others show a value of 0.99. Though, the lowest observed value is 0.9, which reflects the high performance of the spectrograms in an identification task using speakers' information extracted for the modelling. Looking at the marked speakers we notice that four of them achieve a score of 1.0, with one (F10L) being both among the informed and marked ones. However there are two other speakers showing a Markedness lower than 0.9, respectively F14R with 0.89 and F08R with 0.84. The same speakers are the only ones to generate False Positives, using the [Doddington et al., 1998] terminology, these are the *wolves* among a pack of *sheep* in the spectrogram identification.

The results from the Glob representation, all phonetic measurements, show lower rates with Precision and Recall respectively of 0.83 and 0.84, giving a F1-score of 0.83. We obtain the same score with the K statistic. The average Specificity decreases to 0.99 compared to the spectrograms results since only 8 out of 21 speakers attain a score of 1.0. When we look in detail at the Informedness and Markedness we observe some major differences. Only one speaker has a score of 0.99, comparable to the results obtained in spectrograms identification, and 5 out of 21 speakers have Informedness higher than 0.9. We report these speakers in Table 7.4, we can notice that only three are shared between the two identifications, which are F04L, F04R and F10L. The lowest score for Informedness in this identification task, 0.51, is obtained by F08R while the remaining speakers have average scores of 0.8. Concerning the marked speakers, we observe that

only F04R has a score of 1.0 while two other speakers have a score higher than 0.9. Among these three only one is marked by both Glob and Spectros. The least marked speakers are F13R and F07R with a score of 0.66 as well as F08R with 0.63.

The third global representation we look at in this first section is MFCC obtaining a F1-score of 0.85 resulting from identical scores of Precision and Recall. We obtain the same specificity score as Glob with, however, one less speaker having a score of 1.0 rate. The K statistic is higher, 0.86, corresponding to a more reliable model even though it still is less efficient than the Spectros. No speaker manages to obtain an Informedness score of 1.0 or 0.99 in this identification but the lowest rate, 0.64 for F08R, is however higher than the ones obtained by Glob. We find the same speaker, F08R, as the least marked with 0.66, standing out as the *wolf* for all three identification tests. However, comparing the speakers that have the best Informedness for each representation, we observe that three speakers are shared. MFCC's modelling appears to share more information with Glob than with Spectros, with more common speakers in both Informedness and Markedness results.

Input	**F1-score**	**Precision**	**Sensitivity**	**Specificity**	**K**	**Informed**	**Marked**
Spectros	0.96	0.96	0.96	1.0	0.97	F03L F04L F04R F05L F05R F10L	F07L F10L F10R F16R
Glob	0.83	0.84	0.83	0.99	0.83	F04R F04L F07L F10L F10R	F04R F10R F06R
MFCC	0.85	0.85	0.85	0.99	0.86	F04L F04R F07L F10L F10R F14R	F04L F04R F05R F10R

Table 7.4: Identification scores for the three global representations, female speakers.

Since the verification task involves a binary classification, we use MCC instead of F1-score in order to measure the different models' performances. We observe that Spectros general performance decreases, showing an average MCC of 0.79 for female speakers, while the Sensitivity and Specificity are at 0.91 and 0.88 respectively with a consequent increase of both False Negatives and False Positives. The marked speakers correspond to the ones obtaining the best Informedness but there is no correspondence with the ones appearing in the identification task except for F10R who is also among the most marked speakers in this task. Concerning the speakers with the lowest scores, we find again F08R, appearing with an Informedness of 0.69 alongside F13R and F14R, who both have scores of Informedness and Markedness lower than 0.7.

Glob is the representation showing the best performances in this task with a MCC of 0.85, even higher than the F1-score from the identification task. The Sensitivity is also increased, reaching 0.95, while Specificity is at 0.90. Similar to the Spectros results, the same speakers have the best Informedness and Markedness scores, but in an opposite trend these speakers are the same that have the higher Informedness in the identification task. F08R also appears to be the least recognised speaker with both scores of 0.7, alongside F13L. The latter and F12L are the speakers with the lowest Informedness and Markedness scores, 0.52, in the verification task using MFCCs. The best scores, as reported in Table 7.5, are obtained by F04R and F10R, who are the only speakers exceeding 0.9, this show that the MFCC modelling answers remain closer to the ones from Glob than from Spectros. MFCC obtain the lower average scores for this task with a MCC of 0.75,

a Sensitivity of 0.89 and a Specificity of 0.86.

The third task is the generalisation task, it is tightly linked to the verification task as it also performs a binary classification with a target speaker opposed to a population. However, speakers that were previously confused with the target speaker are excluded from the population and added during the test phase in order to test further the robustness of the representations.

For Spectros we observe a MCC score of 0.73, which is only a decrease of 0.06 points from the verification task, showing that spectrograms maintain an important robustness when faced to unknown speakers. The Sensitivity has a decrease of 0.11, with a score of 0.80 and a consequent augmentation of False Negatives, meaning that the target speakers are more frequently confused with the reference population. It appears that excluding from the population the speakers who were most often confused with the target speaker does not help the CNN to better discriminate between the two classes. Even though the similar speakers produce errors during the verification task, their contribution was more helpful than harmful in representing the characteristics of the target speaker. On the other hand Specificity shows a score of 0.90, hence an increase of 0.02, meaning a decrease of False Positive answers.

Concerning Informedness and Markedness of the target speakers we observe a similar trend to what we described for the verification task, the speakers having the highest scores are different from the speakers in the identification task except for one individual, F03L. F08R appears in this task as a marked speaker, this is an important example of how a speaker who performed poorly in the identification task gains a high benefit from the fact that its primary confusion targets are not part of the modelled population. It has to be noted that none of the scores exceed 0.8. The two speakers with the lowest scores for Informedness and Markedness, F05R and F10L, have both scores under 0.4. They appeared amongst the best scores of Informedness in the identification task, indicating that the presence of speakers having similar characteristics had an important weight on the decision.

A higher decrease of the overall performance between verification and generalisation task is shown with Glob with a MCC of 0.72 and a Sensitivity of 0.8, the Specificity remains the same as the previous score with 0.9. However, similar trends are shown by the three representations with 2 out of 3 speakers having the best information-related scores being shared between Spectros and Glob results. In a similar way one of the two speakers with the lowest score, F05R, is also shared by these two representations.

The MCC score registered for the generalisation task using MFCC, 0.55, is the lowest average score in the experiments concerning these three global representations, suggesting that MFCC are less suited to test with unknown speakers. Consequently, Specificity and Sensitivity also register the lowest scores, respectively 0.79 and 0.78. As in the identification task, concerning speakers with the best Informedness and Markedness, we see a combination of the speakers resulting from Spectros and Glob. Although, the highest score is 0.77 in this case. F05R also appears among the lowest scores, making her the female speakers' *wolf* from the generalisation task.

The core of the generalisation task is represented by the results for unknown speakers. The last column of Table 7.5 shows which speakers have the highest scores of Informedness and Markedness for each of the three global representations. We notice that MFCC's results

are in line with what is observed in other tasks, having a combination of the speakers from the other two tests. However, what is important for our analysis is the comparison between the previously marked speakers and those in these results, as this would mean that the CNN have performed extremely well and is able to identify the speaker's characteristics even without knowing them beforehand. No unknown speaker from the MFCC's results is present in those marked in the previous tasks, while for both Spectros and Glob at least two previously marked speakers are present. This implies that when the CNN are faced with tests that are not among its models, using both Glob and Spectros, it is capable of retrieving the information related to speakers whose characteristics stand out from the population. The same is not true for the MFCC's answers which appear to be more influenced by the reference modelling than by an absolute presence of speech characteristics.

Input	**Task**	**MCC**	**Sensitivity**	**Specificity**	**Informed + Marked**	**Informed + Marked (unknown)**
Spectros	Verification	0.79	0.91	0.88	F06R F10R	
	Generalisation	0.73	0.80	0.90	F03L F03R + F08R	F03L F04R F05L F05R F06R F08R F10R F12L F12R F14R
Glob	Verification	0.85	0.95	0.90	F04L F04R F07L F10R	
	Generalisation	0.72	0.80	0.90	F03L F03R	F03L F03R F05L F05L F06R F07L F10L F14L
MFCC	Verification	0.75	0.89	0.86	F04R F10R	
	Generalisation	0.55	0.79	0.78	F03L F08R F14R	F03R F05L F06L F07L F07R F08L F08R F10L F13L F13R F16R

Table 7.5: Verification and generalisation scores for the three global representations, female speakers. Informedness and Markedness for generalisation task report both best target speakers and unknown speakers.

On another note, we notice that having a target speaker which has similar characteristics as those unknown does not help the CNN to retrieve information about the speakers having very low scores of Informedness and Markedness. In this case we find both F08R and F05L for each of the tested representations. The presence of these speakers in both marked and unmarked results further emphasises the idea that the modelling reference has a strong influence on the studied characteristics. The fact that the unknown speakers show average results with a highly variable range of distributions emphasises the issue of modelling speech components and the variable nature of speech.

7.4.2 Spectrograms as baseline - male speakers

Male speakers identification results from the three global representations show that Spectros have a slightly lower overall score than those obtained by female speakers, while both Glob and MFCC have higher rates. 0.93 is the score that characterises the spectrograms' results, all the metrics obtain this score except Specificity which has a 1.0 score. A total of three out of the 23 male speakers obtain an Informedness score of 1.0, while two speakers stand out as the most marked. Concerning the lower scores of Informedness and Markedness we notice respectively two and three speakers average rates of 0.8. As we observed in the female speakers' identification, the remaining speakers have similar scores distributions, representing a population with similar recognisable speech traits, the *sheep*, in opposition to groups of speakers with evident characteristics and those with even more common traits that can be easily confused with other speakers, the *lambs*.

As for the female speakers, Table 7.6 summarises all the identification results. Glob representation scores rank second in this identification comparison with an F1-score of 0.88, consequence of a Precision of 0.89 and a Sensitivity of 0.88. The Specificity score drops 0.01 point with a 0.99 rate, as for MFCC, compared to Spectros, while the K results has a 0.89 score. The speakers who have higher Informedness and Markedness are completely different from the ones present in the Spectros results. Although, when comparing speakers with the lowest of these two scores we observe one *wolf*, M18R, who is common to every representation with the lowest Markedness score. In the same way, M20R registers the lowest Informedness rate for each of the three identifications.

MFCC show the lowest rates in this experiment, however the F1-score of 0.86 remains higher than the ones obtained by both MFCC and Glob for female speakers. The K statistic of 0.70 indicates a lower reliability for MFCC when used to model male speakers. Among the speakers with the higher Informedness scores there are two speakers, M21L and M23R, shared respectively with Glob and Spectros, confirming the already observed trend of MFCC to combine results from the other two representations with new elements. The same trend is not repeated for Markedness results, implying that the extracted information from the MFCC's modelling remains different from the other representations while the actual distribution of speaker's information is similar.

Input	**F1-score**	**Precision**	**Sensitivity**	**Specificity**	**K**	**Informed**	**Marked**
Spectros	0.93	0.93	0.93	1.0	0.93	M01R M19R M23R	M19L M22L
Glob	0.88	0.89	0.88	0.99	0.89	M15L M18L M21L	M01L M11R M15R M23L
MFCC	0.86	0.86	0.86	0.99	0.70	M21L M21R M23R	M11L M15R M23R

Table 7.6: Identification scores for the three global representations, male speakers.

We observe similar trends than with the female speakers in the verification task for male speakers. Glob obtains the best overall results, with a MCC of 0.84, followed by Spectros, 0.79, and MFCC, 0.75. Specificity scores maintain the same hierarchy while the Sensitivity ranks are inverted for MFCC and Spectros, hence a higher amount of False Negatives for the latter representation.

Comparing the speakers' specific results we notice that Spectros' marked speakers are

completely different from those marked with Glob and MFCC. These other two representations share the majority of their marked speakers. However, when we look at speakers with the lowest Informedness and Markedness scores there are more similarities between Glob and Spectros, 2 out of 3 speakers, than with MFCC which give a completely unseen set of speakers. This emphasises the fact that when classical phonetic measurements or representation are not able to successfully characterise speakers, MFCC are able to extract complementary information.

In the generalisation task, Glob maintains its overall higher scores with a MCC of 0.81 against 0.80 for Spectros. In addition, MFCC show the lowest scores with a MCC of 0.79 a Sensitivity of 0.85 and a Specificity of 0.92 while its counterparts show slightly higher scores. The Informedness and Markedness distributions are very similar between the three representations, with highest rates at 0.95 and the group of lowest rates at 0.60, with three or four intermediate groups. However the actual speakers' scores differ consequently. Indeed, among the speakers with the highest rates there are no shared speakers between the three sets of answers, which present similarities only by pairs. The same is true when looking at the speakers with the lowest rates except for the presence of a *wolf*, M09R, the least marked speakers are shared by the three representations.

Input	**Task**	**MCC**	**Sensitivity**	**Specificity**	**Informed + Marked**	**Informed + Marked (unknown)**
Spectros	Verification	0.79	0.82	0.87	M18L M19R M23R	
	Generalisation	0.80	0.86	0.93	M02L M19R M21L	M01L M02R M09L M09R M15L M15R M17L M18L M20L M21R
Glob	Verification	0.84	0.93	0.91	M01L M11R M23L	
	Generalisation	0.81	0.86	0.93	M01L M02L M15L + M23R	M02L M11L M19L M19R M21L M21R M22L M23L M23R
MFCC	Verification	0.75	0.89	0.86	M11R M22L M23L	
	Generalisation	0.79	0.85	0.92	M19L M19R M22L M23L M23R + M20R	M01L M01R M02L M02R M09L M17L M17R M20L

Table 7.7: Verification and generalisation scores for the three global representations, male speakers. Informedness and Markedness for generalisation task report both best target speakers and unknown speakers.

When we take a look at the results from the unknown speakers, we do not see the same trend as female speakers. The most marked speakers in this case are not the one coming directly from the least marked set. Indeed, most of the time they are the same as those already noticed during the previous tasks or completely new speakers that detach themselves from their *sheep* groups. A similar trend is observed for the unknown speakers

presenting the lowest Informedness and Markedness scores.

In conclusion, the representation of speech by spectrograms has been proven to perform best when applying an identification task and, for female speakers, when adding unknown speakers to the test phase. However, interesting results are obtained when comparing other global representations of speech sequences, namely MFCC and the use of phonetic-based measurements to represent the modulation of speech components throughout their production. In particular, the latter has better performances during the verification task, showing to be better suited for pair comparisons. This can be explained by the intrinsic noise of spectrograms compared to more focused representations of speakers' characteristics, which therefore, appear to have greater robustness when the task requires a more targeted answer.

On the other hand, MFCC show similar behaviour as Glob but with inferior performance in both verification and generalisation tasks, suggesting that the modelling made by these features remain less robust than the one obtained by phonetic measurements. In the following sections we analyse further how the information is conveyed by phonetic subsets in order to gain a clearer picture of the distributions of speaker characteristics.

7.4.3 Modulations of source and filter - female speakers

For the representation of the source and filter component, referring to the the phonetic measurements groups presented in Table **??** we have taken the first four formants (*Form*), the amplitudes of the harmonic near the first three formants (*Amp*), f0 and its harmonics (*f0*) and intensity. These multiple subsets and their differences are compared in the following subsections, we start by presenting the results of their combination for female speakers.

Table 7.8 summarises all the results of the group named *Acoust* which is the combination of all the phonetic measurements cited above. For female speakers we observe an F1-score of 0.61, which is the consequence of a Precision of 0.64 and a Recall of 0.62. The latter is related to a consequent amount of false negatives while, in contrast, we observe that the Specificity has an overall score of 0.98 which implies that predictions of false positives are less frequent in the identification task. Concerning the Informedness scores, two speakers stand out with the higher rates, F04R and F10R, and two others with the lower rates, F08L and F08R. The remaining 17 speakers are divided in three homogeneous groups following these metrics. Different results are observed through the Markedness scores where the number of speakers with low scores is higher, with four in total, and only two intermediate groups are present.

Speaker F04R has the highest scores for each metric in the verification task using source and filter component, with an overall average of 0.99. Although the overall performance of this representation corresponds to a MCC of 0.72, a Sensitivity of 0.88 and a Specificity of 0.84. Similar to what we observed for the identification task, considering the Informedness and Markedness scores, speakers arrange themselves in three homogeneous groups except for the F04R who stands out and a group of two speakers presenting very low scores: F12L and F13L.

7.4.3.1 Resonances

In order to represent the resonances we compare two representations, Amp and Form, both extracted via VoiceSauce as described in the previous chapter. In the identification task for the 21 female speakers Amp obtains a F1-score of 0.49, which corresponds to the lowest overall score from the subsets of this group, hence why we do not include it in the generalisation task. The same phenomenon is observed with the K statistic comparison where both Form and f0 scores outperform the 0.51 of Amp, even though Form does it on a smaller scale, with a score of 0.52 and 0.61 respectively.

Form scores are not consequently higher than Amp scores. In the identification task it shows middle range performances, the F1-score is 0.51, with a Precision of 0.54 and a Sensitivity of 0.52. A similar behaviour is obtained in the verification task with a MCC of 0.64, against 0.57 for Amp and 0.71 for f0. In the same task, Form shows the highest sensitivity, 0.83, of all the subsets related to Acoust, this indicates a low rate of false negatives for formants for female speakers when tested in binary classification problems. Form maintains the same good performance for both false negatives and positives in the generalisation task. In this case, it manages to outperform the other subsets of this group with a MCC of 0.61.

When we compare Informedness and Markedness scores for the two resonances representations we do not observe important similarities except for two speakers presenting the lower rates in the identification task, namely F08R and F12L. However, their values and score distributions differ. Indeed, for Amp, F08R appears as the less informative speaker while for Form it is the less marked with a consequent distance of 0.15 points from the subsequent group of unmarked speakers. Concerning the speaker F12L, she appears for Amp among the less marked speakers, while she is both less informative and unmarked for Form. However in the latter case she does not have the lowest score. We notice, in the verification task, that the number of speakers for whom Amp representation does not show high performances is consequent, 8 out of 21 have scores under 0.30. In contrast, for Form there is one speaker with a score of 0.37 and three homogeneous groups with scores ranging from 0.45 to 0.70.

Considering Form's generalisation results for the unknown speakers, a consequent amount of them benefit from the modelling approach applied. However, two speakers show the worst performance having respectively a negative score, for F05L, and a score near 0.01 for F04L. These rates indicate worse than chance and near chance behaviours.

7.4.3.2 Fundamental frequency

In the f0 subset, we observe similar trends to those observed in the PCA's results from the previous chapter when f0 and its harmonics have been tested as isolated representations. For the female speakers' data the H2 obtains the highest F1-score, with 0.73, in a simple identification task with f0 following with a score of 0.69 and both higher Sensitivity and Specificity.

Comparing the identification scores of the 21 female speakers, this subset shows a higher performance than the whole source and filter component in both F1-score and K statistic. There are very similar distributions of Informedness and Markedness as well, suggesting

the high weight f0 has over the other subsets. In the verification task f0 maintains its rank of best performing subset thanks to a F1-score of 0.71 and a Specificity of 0.82, however it shows a Sensitivity of 0.79, lower than both Form and Amp. As mentioned above, f0 is outperformed by Form in the generalisation task. Considering the Informedness and Markedness metrics in order to understand how well the speakers' information has been modelled we observe that, except for the four speakers with the highest scores, the remaining 17 have non homogeneous distributions.

Following what we observed in the previous section with considerations on the generalisation task from the global representations, we also notice, in this case, that an important number of speakers with poor results in the previous tasks appear as the most marked. This is the case for F06R, F08L, F12L and F14L. In addition, the marked unknown speakers from the generalisation task of f0 are half shared with those of the Form, this suggests only a partial redundancy of the modelled information between the two subsets. Overall, we observe one common trend for the majority of the scores with F08R showing low results, while F04R has among the highest ones.

Identification task

Input	**F1-score**	**Precision**	**Sensitivity**	**Specificity**	**K**	**Informed**	**Marked**
Acoust	0.61	0.64	0.62	0.98	0.58	F04R F10R	F04R F10R F16R
Amp	0.49	0.52	0.50	0.97	0.51	F04R	F04R
Form	0.51	0.54	0.52	0.98	0.52	F07L	F14R
f0	0.62	0.63	0.62	0.98	0.61	F04R	F04R F10R

Verification and generalisation tasks

Input	**Task**	**MCC**	**Sensitivity**	**Specificity**	**Informed + Marked**	**Informed + Marked (unknown)**
Acoust	Verification	0.72	0.88	0.84	F04R	
Amp	Verification	0.57	0.80	0.76	F04R	
Form	Verification	0.64	0.83	0.81	F05R	
	Generalisation	0.61	0.71	0.86	F03L	F04L F05R F06L F06R F08L F08R F10L F12L F14L
f0	Verification	0.71	0.79	0.82	F04R	
	Generalisation	0.57	0.69	0.85	F07L F12L F13R F16R	F03L F04L F06L F12L F12R F13L F13R F14L

Table 7.8: Identification, verification and generalisation results for the different representations of source and filter component, female speakers.

7.4.4 Modulations of source and filter - male speakers

What we present hereafter follows the same organisation as the previous section for female speakers' source and filter, we start with the results from the combination of all source and filter related measurements and then describe the multiple subsets in the subsequent sections. Table 7.9 serves as a summary for these component results for male speakers.

In the identification task for the 23 male speakers, Acoust shows a F1-score of 0.67, higher than the one obtained by female speakers. Small differences in score are observed when comparing the K statistic and the MCC from the verification task between the two sexes. In addition, another observed difference concerns the informative and marked speakers. In the results for female speakers, the two categories show important similarities, whereas in the male speakers tasks, these similarities are restrained to the Form identification results. The identification of male speakers using Acoust representation shows that uninformative

speakers correspond to the most marked speakers and vice versa. This suggests that, for male speakers, having characteristics related to the source and filter component does not imply that the created CNN model is able to efficiently use speaker information. This is emphasised by the distribution of Informedness and Markedness scores in the verification task where half of the speakers are ranged in a homogeneous group while the others have highly isolated rates.

7.4.4.1 Resonances

For the four formants, we have also performed additional tests in singular and coupled combinations. The results do not show specific trends, except for F3 which has slightly higher scores than the other formants and show speakers-specific scores more similar to those obtained in Form identification. F1-scores have an average of 0.30 for isolated formants for both female and male speakers, while 0.44 is the average when they are taken in pairs. Changing the position of a formant in the representation does not affect the CNN responses, i. e. for the F1-F2 pair, having F2 in the bottom pixel line and F1 on the top or vice versa produces the same scores.

The results from the male speakers identification are very similar for Amp and Form, the main difference concerns the K statistic which is less consequent for the latter representation, 0.48 against 0.60 for Amp. We observe that they share one of their most marked speakers as well as 3 out of 4 unmarked speakers. In the verification task, the scores remain similar in terms of values but the speakers-specific results show more differences. For instance, all speakers with higher Informedness and Markedness scores for one representation subset are modelled as sheep by the other representation, showing that the way CNN use the extracted information is different.

7.4.4.2 Fundamental frequency

f0 shows the highest K statistic, 0.76, of all the subsets from both female and male speakers. Similarly to what is observed above, its F1-score is higher when compared to Amp and Form in the identification task, as well as all the other metrics. In contrast to what is observed for female speakers, f0 has higher scores in both verification and generalisation tasks when compared to the other subsets. An additional difference concerns the measurements that show the higher influence on the f0 representation for male speakers. H1 is the one showing the highest influence, followed by H4, paralleling what has been observed in the phonetic analysis through PCA.

Verification and generalisation tasks show very similar distributions of Markedness scores for speakers using f0 representation. There is not a single large population, as we observed for the Form results and for all female results. Four homogeneous groups of four are formed, while the others have scores that are far apart. A very similar trend is observed using the clustering methods presented in the next chapter. Among the speakers with the lowest Informedness scores in the identification task we have found M19R who appears as a highly marked speaker in the generalisation task for both Form and f0. In a similar way, M11R, M21R and M17R show low Markedness with f0 in the identification task but high rates when speakers with similar source and filter characteristics are removed from the CNN model. As already observed for female speakers, nearly half of the unknown

speakers that appear to be marked in the generalisation task are shared by both Form and f0. However, the additional speakers differ consequently showing that the information these two representations carry is not the same while they are capable of complementing themselves as we observed in the Acoust results.

Identification task

Input	F1-score	Precision	Sensitivity	Specificity	K	Informed	Marked
Acoust	0.67	0.68	0.67	0.98	0.59	M01L M18L	M09R M19L M23L
Amp	0.52	0.56	0.53	0.98	0.60	M01L M18L	M09L M15R
Form	0.53	0.56	0.53	0.98	0.48	M11R M15L M15R	M23R M15R
f0	0.71	0.72	0.71	0.99	0.76	M18L	M01R

Verification and generalisation tasks

Input	Task	MCC	Sensitivity	Specificity	Informed + Marked	Informed + Marked (unknown)
Acoust	Verification	0.73	0.88	0.84	M18L	
Amp	Verification	0.61	0.83	0.77	M18L	
Form	Verification	0.62	0.81	0.80	M01L	
	Generalisation	0.63	0.73	0.87	M19L	M02L M02R M09L M11R M15L M17R M19R M20R M21L M21R M22L M23L
f0	Verification	0.74	0.88	0.86	M18L	
	Generalisation	0.71	0.80	0.90	M01R M02L M18R M19L M19R M23L M23R	M02L M09R M11L M11R M15R M17L M17R M18L M19L M20R M21R

Table 7.9: Identification, verification and generalisation results for the different representations of source and filter component, male speakers.

7.4.5 Prosody modulations - female speakers

The second component, prosody, aims at representing different prosodic aspects through two subsets: the first one involves Envelope and Temporal Fine Structure measurements in order to represent the respectively fast and slow changes of the temporal envelope during speech production, *Env*; the second one uses fundamental frequency and intensity in order to represent the modulations of the intonational contour, *Int*. As reported in Table 7.10, the combined representation for the present component is named *Pros*.

In the female speakers' identification task we observe similar scores for both subset, the F1-score of Env is 0.34 while it is 0.32 for Int, Precision is respectively 0.35 and 0.33. Int has a higher K statistic which makes its results more reliable than Env even though the latter presents overall higher scores. F04R is shared as the speaker with the highest Informedness rate, this is true for the verification task and Pros as well. The main difference concerning the subset comparison is the marked speakers which differ substantially between the two identification tasks. The score distributions are also consistently different, with Env presenting a higher number of isolated speakers, while Int speakers are combined in three similar groups outside the marked speakers and the most unmarked ones isolated. The unmarked speakers from Int identification results correspond to F06R, F13L and F03R which are respectively marked for Env and Pros in identification as well as in the generalisation task. This suggests, once again, that they all share the same modelling methods but the actual extracted information differs.

The change observed for F13L, unmarked in subsets of identifications and strongly marked in the component-global representation, is interesting in order to understand the extent

of complementarity that these measurements are capable of deploying when modelling speech components. In a similar way, among the unknown marked speakers from the generalisation results of Pros we find both some of the unmarked speakers from the identification task of the same representation and the marked speakers from the MFCC and Glob results.

Identification task

Input	**F1-score**	**Precision**	**Sensitivity**	**Specificity**	**K**	**Informed**	**Marked**
Pros	0.37	0.39	0.39	0.97	0.36	F04R	F04R F13L
Env	0.34	0.37	0.35	0.97	0.29	F04R F10L	F06R
Int	0.32	0.36	0.33	0.97	0.35	F04R	F04L F04R

Verification and generalisation tasks

Input	**Task**	**MCC**	**Sensitivity**	**Specificity**	**Informed + Marked**	**Informed + Marked (unknown)**
Pros	Verification	0.51	0.78	0.73	F04R	
	Generalisation	0.47	0.62	0.81	F03R F03L	F05R F06L F07L F08L F08R F10L F10R F12L F13R
Env	Verification	0.51	0.74	0.76	F04R	
Int	Verification	0.64	0.84	0.79	F04R	

Table 7.10: Identification, verification and generalisation results for the different representations of prosody component, female speakers.

7.4.6 Prosody modulations - male speakers

Male speakers' identification results for the prosody component show overall higher scores than for female speakers. Both Pros and Env F1-scores have an increase of 0.10 points, being respectively 0.47 and 0.45. In a similar way both Sensitivity and Specificity increase for these representations. However, only the K statistic of Env shows an increase compared to the female counterpart.

The comparison of the two prosody subsets of representation we use in our experiments show higher performances for Env in characterising male speakers sequences, as well as overall higher scores of Markedness. Indeed, the highest scoring speaker for Env is at 0.82, while Int highest score is 0.69. The number of marked speakers is also consequent when comparing between the two subsets. Most similarities can be found in the low scoring speakers with both M17L and M21R being the most unmarked speakers for both Int and Env and M20R being the less informative speaker in both identifications.

In the verification task the scores are overall higher for Int than for Env and we observe similar trends to those mentioned for the other task. The representation which performs the best has also the most marked speakers, which are partially shared with the less performing counterparts. This suggests that, in opposition to what we observed for female speakers, the amount of shared information in the prosody subsets is higher for male speakers. However, a more sparse score distribution is observed in comparison to the female speakers with a higher amount of isolated speakers.

Pros results from the generalisation task emphasise the role of the *wolf* speaker M21R, who relentlessly appears as the least marked one. In a similar way, M09R scores below and alongside M09L and M11L, they obtain high Informedness and Markedness rates with source and filter representations. The sudden changes observed for these speakers are an example of the variability of speech representation, with different components

representing the same data but giving different outcomes. The complementarity of these two components is demonstrated by the results of the global representations where all these speakers obtain similar high scores even if they are more accurately represented by different components.

The more prominent role of prosodic components in the characterisation of male speakers' voices has already been highlighted in the previous chapter's considerations. Here, Pros performances from both verification and generalisation tasks also show consequent increases in comparison to the female speakers results. The influence of sex on the role that specific speech components play in modelling speaker differences continues to be one of the influential factors that we are investigating in order to understand how speaker information is conveyed.

Identification task

Input	F1-score	Precision	Sensitivity	Specificity	K	Informed	Marked
Pros	0.47	0.51	0.48	0.98	0.36	M18L	M23L
Env	0.45	0.46	0.45	0.98	0.35	M18L	M09L M18L M23L M23R
Int	0.32	0.40	0.35	0.97	0.36	M18L	M21L

Verification and generalisation tasks

Input	Task	MCC	Sensitivity	Specificity	Informed + Marked	Informed + Marked (unknown)
Pros	Verification	0.54	0.78	0.76	M01L M18L	
	Generalisation	0.58	0.70	0.84	M19R M02L M18R	M01R M02L M11L M15L M17L M19R M21R M22L
Env	Verification	0.51	0.78	0.74	M18L M23L	
Int	Verification	0.61	0.83	0.78	M01L M18L M21L M23L	

Table 7.11: Identification, verification and generalisation results for the different representations of prosody component, male speakers.

7.4.7 Modulations of mode of vocal fold vibration - female speakers

The third set of phonetic measurements we explore in our CNN experiments corresponds to what has been discussed in previous chapters as the "mode of vocal fold vibration" as well as the idea of voice quality. In Table 7.12, as for each of the previously described components, we report the scores for the main subsets studied here. Unlike the two previous components, there is no representation equivalent to the global one for the voice quality component. However, *Qual* can be seen as partially fulfilling this role, even though some of the studied subsets are excluded from this group. For instance, we could have included Nrg or Ms as representations of the noise excitation part for the source and filter component, however energy measurements are commonly included as part of voice quality analyses as inharmonic source or spectral noise cues. As a consequence, we have included these two groups in the results for the voice quality component but considered that they may be moved as well to the source and filter results. Hereafter we compare six main representations and then analyse some of them in detail.

The female speakers' results from the identification task show that Hr is the representation with the highest F1-score with 0.76, as well as the highest K statistic with 0.74. The second higher score considering the value of K as the comparison metrics is obtained by Qual with 0.66, followed by Hadiff with 0.65. For both Precision and Sensitivity the

representations rank similarly except for Specificity where only Hr obtains 0.99 instead of 0.98. Concerning the speakers presenting the highest Informedness scores, we observe the major similarities between Qual and Ltas, which is not surprising due to the presence of some Ltas measurements in Qual. F10R and F04R are the most present speakers in the ones most informative or marked. Another interesting point is represented by the shared speaker, F16R, between Ms and Qual as well as the shared marked speaker, F07L between Ms and Hr. Looking at the lowest of these scores, we notice that F08R is the *lamb* of this component since she presents as the most unmarked and with the highest tendency to be confused with another. F13R and F05L are also present among the unmarked ones for Hr, Ltas and Nrg.

In the verification task, Qual obtains the best score showing a MCC of 0.74 and Ms ranking second with 0.72, while Nrg is the lowest performing representation for this task with 0.67. The three remaining representations have average scores of 0.70. Similar rankings are maintained when comparing the Sensitivity and Specificity scores as well. F10R is one of the shared marked speakers, between Qual, Ltas, Ms and Hadiff and Hr, while F04R is present for both Nrg and Hr, however, F06R is marked by Hadiff and Ms. This latter representation appears as the one sharing the most information with its counterparts. Similar to what has been observed in the identification task, F08R is the most unmarked speaker for Hr, Nrg and Hadiff results. F13L plays the same role for Ltas and Qual. Apart from the speakers with higher and lower Informedness and Markedness scores we mainly observe a large number of isolated speakers' scores in both tasks.

As mentioned before, only a part of these representations has been used in the generalisation task in order to have a reduced but still high variable set of representations. Qual, Hr and Ms are those that have been retained from this component. Qual and Ms both outperform Hr in this task and they appear to share a larger amount of information about the modelled speakers. Qual is the only representation to not have negative markedness values among the lowest scores of its speakers. This, as well as the fact that the lowest scores are 0.30, suggests that voice quality characteristics can be considered as more reliable even when the tested speakers are unknown.

7.4.7.1 Spectral and energy variations

We have carried further investigation on the different subsets from the mode of vocal fold vibration component. In particular, we aim to understand which measurements have the higher weight on the decisions and if there are differences when a particular feature is removed from the representation.

For instance, the Hadiff subset includes two types of measurements extracted by Voicesauce: the relative amplitudes between two harmonics, H1-H2, H2H4 and H2kH5k; the relative amplitudes of the first harmonic and the amplitudes of harmonics near the first three formants, H1-A1, H1-A2 and H1-A3. These two subsets, respectively *Hh* and *Ha*, of the Hadiff representation do not show consequent differences for female speakers. Their identification scores are very similar, F1-score of 0.55 for both and K statistics of 0.54 for Ha against 0.50 for Hh. Some differences appear in the comparison of marked speakers where Ha has results similar to the ones obtained with the Form representation from the source and filter component. This is not surprising considering the presence of amplitudes near the first three formants in the Ha representation. The same tendencies are shown by

the verification task where the MCC scores are 0.60 for Ha and 0.63 for Hh. The latter shows similar marked speakers to those shown by Pros representation, while Ha results follow those by Hadiff as well as those from the Form generalisation task. This suggests that the relative amplitudes measurements are capable of extracting characteristics related to resonances but which usually remain hidden through simple formant analysis.

The second subset we focus on, in regards to energy variation and spectral shape information, concerns different energy computations and their influences on the whole Qual representation. The three energy-related computations correspond to the default energy extracted in Praat, the one by VoiceSauce, and the soe measurement by VoiceSauce as well. Their identification results do not present any considerable difference in terms of score or marked speakers. In the verification task, while their speaker-specific results remain highly similar, their scores differ considerably with Qual_{soe} having higher performances and reliability with a MCC of 0.68 against 0.58 by the other two subset representations.

Spectral moments have been included as an additional representation in this component. Their inclusion in the study of an important component such as voice quality is justified by the presence of energy and other phonetic measurements of spectral shape information that have been found to correlate with voice qualities in the literature. This comparison aims to understand whether there is room for new measurements between those already established in the phonetic literature regarding the study of energy distribution in the spectrum. For female speakers we observe that the third spectral moment has a higher score when tested as a singular representation as well as a higher influence in the whole Ms set.

7.4.7.2 Long-Term Average Spectra

Similar to the information on the spectral shape provided by Hadiff, the second subset we investigate for this component is the *Ltas* representation. As mentioned before, the LTAS represents the average frequency distribution of the speech signal over a long portion of speech. The prominence of peaks in the LTAS between different bandwidths has been shown to provide general information about voice quality, namely resonances or sonority of the voice. We have compared results for different peaks from different bandwidths, between 1 and 5 kHz, including them alternatively in the Qual representation.

For the identification task of female speakers, we observe higher scores when the peak from medium-higher frequencies is used, the one between 3 and 4 kHz. The F1-score and K statistic of this representation is 0.77 against an average of 0.65 for all its counterparts. Similar marked and unmarked speakers are obtained by the multiple tested subsets, with the lower peak showing similar results to those obtained by Acoust and f0 in particular. However, in the verification task, it is the lowest peak that performs better and shows different results in terms of marked speakers. This suggests the complementarity of these measurements since the Ltas results show a complete combination of what we observed taking them separately.

7.4.7.3 Harmonics to noise ratios

The last representation that has been divided into subsets of measurements for further investigation is the *Hr*. As mentioned in the previous chapter, VoiceSauce provides different computations of the harmonics to noise ratio measurements, using different pitch ranges in order to compare the pitch component of the cepstrum with the energy of the harmonics at the noise floor. The possible pitch ranges are 0-500 Hz, 0-1500 Hz, 0-2500 Hz and 0-3500 Hz. In addition we included in this representation the Subharmonic-to-harmonic ratio (SHR) measurement, which quantifies the amplitude ratio between subharmonics and harmonics. In order to understand the weight these multiple measurements have on the whole Hr representation we carried tests excluding them alternatively and compared their results.

For female speakers the SHR measurement appears to provide the highest redundancy since its exclusion from the identification produced the highest F1-score and K with 0.71 and 0.72 respectively. The same scores are also obtained when the 0-15 kHz computation is excluded. In a similar way, the verification scores excluding SHR is the highest with 0.75. The HNR computation that appears to have the most influence is the one obtained using a pitch range of 0-500 Hz since, when excluded, the results decrease consistently for both identification and verification. As far as the marked speakers are concerned, no particular difference is observed.

Identification task

Input	**F1-score**	**Precision**	**Sensitivity**	**Specificity**	**K**	**Informed**	**Marked**
Qual	0.65	0.67	0.65	0.98	0.66	F05R F10L F10R F16R	F04R F10R F14R
Hadiff	0.63	0.63	0.64	0.98	0.65	F10R	F04R F10R
Nrg	0.66	0.63	0.98	0.62	0.58	F04R	F04L F04R
Ltas	0.65	0.63	0.98	0.60	0.56	F10L F10R	F12R F14L F16R
Ms	0.76	0.76	0.77	0.99	0.74	F04R F10R	F07L
Hr	0.66	0.67	0.66	0.98	0.62	F16R	F07L

Verification and generalisation tasks

Input	**Task**	**MCC**	**Sensitivity**	**Specificity**	**Informed + Marked**	**Informed + Marked (unknown)**
Qual	Verification	0.74	0.89	0.86	F10R	
	Generalisation	0.62	0.74	0.86	F03L F14R F16R	F03L F05L F05R F06L F06R F10L F10R F13L F14L
Hadiff	Verification	0.70	0.87	0.83	F06R F10R	
Nrg	Verification	0.67	0.86	0.81	F04R	
Ltas	Verification	0.69	0.87	0.81	F10R	
Ms	Verification	0.70	0.89	0.80	F04R	
	Generalisation	0.58	0.70	0.85	F03L F12R F14R	F03R F06R F07R F08R F10R F12R F13R F14L
Hr	Verification	0.72	0.88	0.83	F04L F06R F10R F14R	
	Generalisation	0.62	0.72	0.87	F03L F03R F14R F16R	F03L F03R F04L F04R F05L F06L F07R F08L F08R F12R F13L F13R

Table 7.12: Identification, verification and generalisation results for the different representations of voice qualities, spectral and energy variations, female speakers.

7.4.8 Modulations of mode of vocal fold vibration - male speakers

Overall the voice quality component registers higher F1-scores for male speakers identification in regards to female identification with Nrg being the only representation with

results under 0.70. Ms is the representation that performs the best with a score of 0.75, while the second one is Hadiff with 0.74. Taking the K statistic as a reliability index, their roles are inverted with Hadiff obtaining 0.81 and Ms 0.73, Nrg as well increases consequently with a score of 0.72. The three remaining representations rank under 0.70 with different scales.

The distributions of Informedness and Markedness scores appear less scattered than those for female speakers, in particular with a higher number of paired or tripled associated speakers showing the same confusions and rates. Ltas and Qual show less similarities than in the female speakers results, suggesting a lower influence of Ltas in the Qual modelling for male speakers. Indeed, they do not share any of the speakers either with the best or worst Informedness scores. M18L is the most informative speaker for the remaining representations, Nrg, Hadiff, Hr and Ms. The latter show the most similarities with Qual and in a lower scale with Nrg.

Hadiff maintains its prominent score in the verification task, emphasising the fact that glottal parameters and their correlated spectral slope variations are important elements for the characterisation of male speakers' voices. The MCC of 0.75 is not attained by any of the remaining voice quality representations, Ltas obtains 0.70 and Ms 0.71 while both Hr and Qual obtain 0.69. The same ranked performances are observed for both Sensitivity and Specificity results with Hadiff outperforming all the other representations. In terms of marked speakers, we do not have different tendencies than those mentioned in the identification task. Hadiff shows a small amount of similarities with its counterparts but overall the distribution is highly different. This implies that the way these different aspects of voice quality model the male speakers is effective in highlighting different characteristics. For unmarked speakers, we observe the same trend as with the identification results with important similarities between Qual, Nrg and Ms.

The three representations used for the generalisation task, Qual, Ms and Hr obtain similar scores, there is no prominent performance from any of them in particular. Indeed, the marked speakers differ from those from previous tasks. Ms shows similarities with both Hr and Qual regarding its most marked speakers, while Hr and Qual do not share any, at least when we do not consider the unknown speakers test results. When we look at the least marked speakers we notice that M09R is present in every generalisation result. The same speaker appeared as one of the most characterised by source and filter component, though, the fact that no low scores were registered using global representations, we can assume the complementarity between the opposite characteristics highlighted by these components. The results from the test using unknown speakers show, especially for Hr, a high number of speakers having consequent Informedness and Markedness rates. Among them we find the already marked speakers as well as those for whom CNN performed badly. This confirms the fact that multiple modelling excluding *wolves* from the reference populations enhances the CNN's performance making possible a larger appearance of speakers' characteristics.

7.4.8.1 Spectral and energy variations

Concerning the detail of the energy variations measurements for male speakers, Ms shows consistent differences between female and male speakers. Unlike what is mentioned above, the first two spectral moments have a higher influence on the representation results than

the other two. In addition, the fourth moment performs consequently lower performances when taken singularly.

The comparison of Hadiff's subsets shows that Hh has a higher reliability than Ha, with higher scores in both the identification, K of 0.79 against 0.67, and verification task, MCC of 0.69 against 0.65. The marked speakers and the scores distributions for the average population remain very similar between the two representations. However, the unmarked speakers are very different, namely with Ha showing more similarities with the f0 results while Hh appears to have a higher influence on the Hadiff results. The scores obtained by the isolated measurements also suggest that for male speakers H1-A1 and H1-H2 play a more prominent role than their counterparts, while for female speakers, each harmonic-based spectral shape measurement has the same weight.

In regards to the different energy computations, we observe that for the identification task the Qual_{vs} alternative is the one showing the lowest scores, 0.57 for both F1-score and K. Similar to what is observed for female speakers, Qual_{soe} shows the higher scores in both identification and verification tasks. The marked speakers confirm the idea that this representation is the one having the highest influence on the Qual results for male speakers as well. However, when comparing the unmarked speakers, we observe that $\text{Qual}_{\text{praat}}$ has higher similarities with the main results. This highlights the fact that when we study a highly variable object like speech, it is not as important to rely solely on successful features as it is to take a more global view of how different features can explain the variability.

7.4.8.2 Long-Term Average Spectra

The Ltas subsets investigation for male speakers show very different tendencies than the ones observed for female speakers. As expected, the peaks from the lowest bandwidths obtain better results, in particular the one from 2-3 kHz has a F1-score of 0.76 and a K of 0.81 and the use of the 1-2 kHz peak produced a F1-score of 0.74 and a K of 0.78. In opposition, the highest peaks obtain both a F1-score of 0.68 and K of 0.72. Concerning the distributions of Informedness and Markedness we see that the results are mirrored by non adjacent peaks, e. g. both marked and unmarked speakers distribution obtained using the 2-3 kHz peak are the same obtained with the 4-5 kHz. Hence, a redundancy of information is shown in the information distribution for male speakers, even though the actual scores are not repeated.

In the verification task, we observe the same trend, but this time it concerns the performance scores and not the speaker's distribution. The 1-2 kHz and 3-4 kHz pair shows a MCC of 0.69 while their counterpart obtains 0.72. The marked and unmarked speakers are shared between the 2-3 kHz and 3-4 kHz pair as well as the 1-2 kHz and 4-5 kHz.

7.4.8.3 Harmonics to noise ratios

The different HNR computations for the male speakers show some similarities with the results described for female speakers. The SHR measurement appears as the one with the least influence on both the identification and verification results. Its exclusion produces the highest results in both cases, with a F1-score of 0.76 and a K of 0.84 in the identifica-

tion task, and a MCC of 0.76 in the verification task. We observe the opposite for male speakers compared to female speakers concerning the pitch range-related computations. The HNR computation with a 0-500 Hz range has the second lowest influence on the results, since its exclusion produced a F1-score of 0.75 and a K of 0.80. However, the same is not true in the verification task, since excluding the HNR computed using a 0-500 Hz generates the lowest score with a MCC of 0.66.

For male speakers, the HNR computation relying on a pitch range of 0-2,5 kHz is the one showing the highest influence on all the results. In addition, the marked speakers distributions from the exclusion of this particular computation are the only to not match the Hr results.

Identification task

Input	**F1-score**	**Precision**	**Sensitivity**	**Specificity**	**K**	**Informed**	**Marked**
Qual	0.72	0.74	0.73	0.99	0.65	M17R	M15R M23R
Nrg	0.67	0.69	0.67	0.98	0.72	M18L	M01L M21L
Ltas	0.72	0.73	0.72	0.99	0.68	M01L	M23R
Hadiff	0.74	0.77	0.74	0.99	0.81	M18L	M15R M18L
Ms	0.71	0.74	0.71	0.99	0.62	M18L M23L	M09R M22L M23R
Hr	0.75	0.77	0.75	0.99	0.73	M18L M22L	M02L

Verification and generalisation tasks

Input	**Task**	**MCC**	**Sensitivity**	**Specificity**	**Informed + Marked**	**Informed + Marked (unknown)**
Qual	Verification	0.69	0.85	0.84	M11R M15R	
	Generalisation	0.72	0.81	0.89	M01L M02L M19R M21L	M01R M02L M09R M15L M17R M22L
Hadiff	Verification	0.75	0.89	0.86	M01R M09L M15R	
Nrg	Verification	0.69	0.87	0.83	M18L	
Ltas	Verification	0.70	0.86	0.84	M15L	
Ms	Verification	0.69	0.85	0.84	M01L M15R M22L M23L	
	Generalisation	0.72	0.80	0.90	M01R M22L M23L + M18R	M01L M01R M02R M11R M15L M15R M17L M18L M19L M19R M21L M21R M22L M23L
Hr	Verification	0.71	0.86	0.85	M01L M01R M23L	
	Generalisation	0.69	0.77	0.89	M01R M02L M21L	M11R M15R M17L M17R M20L M20R M21R M22L

Table 7.13: Identification, verification and generalisation results for the different representations of voice qualities, spectral and energy variations, male speakers.

7.5 Chapter Conclusions

The results presented in this chapter follow the idea of combining interpretable input data from the phonetic literature with a modern approach of modelling information from the said data. To do so, we extracted phonetic measurements and studied them in Chapter 6 in order to have a basic phonetic knowledge about the speakers we use in the characterisation investigations. The CNN approach carried in this chapter provides a more complex modelling of the data that the previous approach was lacking.

In this sense, the core of our CNN investigation is the fact that we provide the Machine Learning model with interpretable data of which we already understand the behaviour and processes. This enables a deeper understanding of the results, beyond the simple

performance scores which remain helpful for the reliability of the studied representation. Phonetic measurements are divided into three main components, i. e. source and filter, prosody, and mode of vocal fold vibration. Within each component we performed subsets tests in order to have an in-depth view of the interactions between the speakers' conveyed information.

Another aspect, for which the underlying phonetic knowledge plays an important role, is the interpretation of the different tasks results. As already mentioned, characterising a speaker's voice is not a simple process and understanding it involves mirroring multiple aspects of cognition. We have tried to recreate and combine some of these aspects in the different tasks performed by CNN.

The next section, provides a summary of the main findings from this chapter. The discussion about the results presented here is expanded in Part III of this thesis, the moment of *unity*, where all other chapters' elements are approached with a comparative take.

7.5.1 Summary

- **Lexical distances**: overall the lexical comparison does not appear to be a very robust measure in order to characterise speakers. However, keywords comparison covers the percentage of dissimilarity shown by cosine distances.
- **Preliminary studies**: the highest scores are obtained for the identification of male speakers in every tested representation, i. e. nasal vowels spectrograms for NCCFr, 2 s clear, noisy augmented and MPS for PTSVOX.
- **Global representations**:
  - For both sexes, spectrograms perform the best in the identification task with higher F1-scores and very low False Positives;
  - Information about speakers is not shared equally between Spectros and Glob, but MFCC combine results from both;
  - Very poor performance by MFCC in the generalisation task, with the best results obtained by Glob;
  - For the generalisation task of female speakers, the three representations show the same marked speakers. More important differences are observed for male speakers.
- **Identification, verification and generalisation tasks**:
  - Hr shows the best performance for female speakers with a F1-score of 0.76. SHR shows important redundancy with the other HNR measurements, while the HNR computation with 0-500 Hz range has the highest influence on the results;
  - Male speakers are most accurately identified by Hadiff at 0.81, with H1-A1 and H1-H2 showing the highest influence on the results;
  - Pros is the group showing the lowest overall scores and a high redundancy of extracted information for both female and male speakers;
  - One marked female speaker is shared in most of the results, but the other marked speakers vary consistently. Source and filter and voice quality components share a common unmarked speaker;
  - For male speakers there is a higher variability of score distributions than for females;
  - In the verification task, for female speakers Qual obtained the best results followed by Ms and Acoust. For male speakers Hadiff remains the best representation;
  - The generalisation task highlights higher variability in terms of characterisation of male speakers. Female marked speakers remain constant with verification and identification tasks results;
  - Male speakers also present a higher number of unknown speakers with high performance scores.

Part III

UNITY or Discussion and Conclusion

Chapter 8

What does characterise a speaker?

This chapter represents the final step of our investigation, the moment of *unity*. The results from the previous three chapters, moment of *instability*, are summarised and discussed in relation to what has been observed in the first part of this thesis, moment of *fixity*. Each of the three main areas of study that we have explored, i. e. Phonetics, NLP, Perception, represents the main subjects of discussion in the first three sections of the present chapter. However, as already mentioned throughout this thesis, speaker characterisation does not involve only one of these domains. This leads to a joint discussion in which the results from the previous chapters are summarised and discussed from a comparative perspective. The following joint discussion revisits the research questions stated in Section 1.3 to, hopefully, provide the necessary answers.

If we refer to the definition of "to characterise" given in the beginning of this thesis, we have "To mark something as a characteristic". In our case, it is a speech component or at least one of its aspects that we have to mark. In order to be marked it has to show important variation in its behaviour, which includes both the values the component takes in itself and the interactions it has with its counterparts. For this reason, the use of multiple phonetic measurements to represent different speech components is fundamental in the study of speakers' characteristics. Indeed, investigating how the speaker's information is conveyed by components, their redundancy and complementarity can be performed by different means.

How to efficiently represent the variability matrix associated with speakers' characteristics and its different levels has been one of the main questions in this thesis. Indeed, the variable nature of speech makes it likely to be transposed in multiple ways, as we have shown throughout Chapters 6 and 7. The use of mean values to represent f0 or the average of three points in formant analysis are some common examples from the phonetic point of view. A mean analysis on a fine description of time-related variations may be enough to incorporate the intrinsic variability, while complex representations of speech dynamics may result in poor renditions.

A visual tool to summarise components interactions and compare the characteristics between speakers is a radar chart. In Figure 8.1, four female speakers from the NCCFr corpus who showed characteristics related to different components are represented. The grey area in the chart represents the mean values of all the female speakers from the corpus while the red pattern corresponds to the considered speaker. In particular, we notice

that F04R and F10R, top left and bottom right speakers, have high performance in relation to prosody modulations in all CNN tasks. Their values of TFS, ENV and intensity differ consistently from the population average while less variation is observed for the other two represented speakers. They take the role of goats among the female speakers, at least concerning the prosody and energy-related components. The interactions of the single measurements representing a component can be simply understood through this representation. Indeed, F08R, bottom left chart, is a speaker that belongs to the class of sheep, those having poor performance because of characteristics being highly blended with the average population. The remaining example of F05R, on the top right, represents a speaker that is alternatively associated to a wolf and a sheep behaviour depending on the considered component. In fact, the formants and prosody characteristics show partially average and prominent values. We notice similarities with the formants pattern of the other speakers, however with higher observations which explain why this speaker is more prone to imitating others.

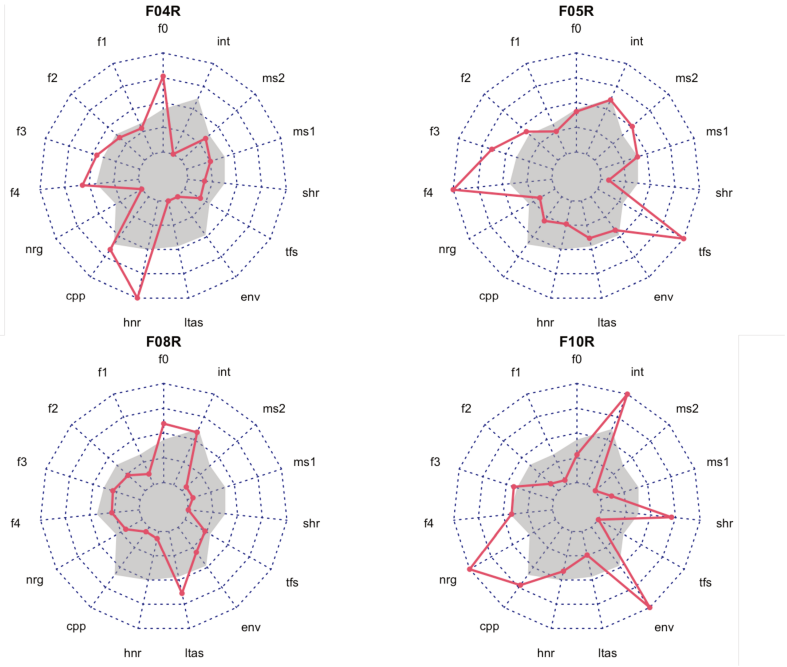


Figure 8.1: Radar charts for four female speakers of the NCCFr corpus that present characteristics related to different components. In grey are the mean values for the whole corpus while in red are the speaker-related values.

These examples show how speakers' voices are not a simple addition of variation around a prototype. The actual distributions of the multiple components seem more important than the values they assume. Speech productions are complex systems resulting from the interactions of speech components, which in turn are influenced by factors coming from multiple sources. The identity of the speaker is only one of the possible pieces of information we can retrieve, although, with varying success.

Our results confirm the idea that there is not only one component that easily characterises speakers. On the contrary, identifying the interactions of multiple aspects that are important for the characterisation of speakers is a fundamental step in understanding the distribution of speaker information in a complex variability matrix. A further understanding of how speaker variability is structured is an important step to describe the variations of phonetic characteristics inside a group of speakers [Tanner et al., 2020].

Modelling the interactions has to be part of the metric used in the analysis rather than being considered a confound. On the whole, speaker characteristics represent a unique combination of segmental, suprasegmental and paralinguistic features.

8.1 Building prior knowledge

The representation of speech components is one of the explored issues in Chapter 6. The different analyses presented, i. e. PCA, linear models, SVM, confirm the high versatility of speech and highlight the difficulty to have a comprehensive representation of its changes. The main results of the phonetic analysis are obtained through PCA. This analysis implies a dimensionality reduction, allowing the study of a large set of components like those we selected. However, it also means that the shape of the input plays an important role, e. g. the use of statistical values capable of rendering the time-related variations is fundamental to avoid information loss.

PCA between or within speakers can reveal the structure of the variation in the acoustic space of the stimuli, similar to the "telling voices apart" or "telling voices together" from [Lavan et al., 2018]. The implications of these results are important for prototype-based models of voice processing, as in [Lavner et al., 2000; Kreiman and Stidtis, 2011; Yovel and Belin, 2013], which lightly consider within-speaker variability. As the perceptual process must be adapted to the received acoustic input, the understanding of voice acoustic spaces' structures is fundamental to provide further insight to voice characterisation. Multidisciplinary takes to speaker characterisation have shown that to assess who is speaking both features and pattern analysis strategies are needed [Adank et al., 2009; Yovel and Belin, 2013; O'Brien et al., 2021]. Thus, the perception of unfamiliar voices, as it is the case in our study, requires both reference to a population of prototypes and evaluation of the manner in which the target voice deviates from that prototype. In our results, the individual prototypes are mainly influenced by the balance of higher-frequency harmonic versus inharmonic energy in the voice and formant dispersion. They are located in different groups of voice spaces with similar structures. However, these shared structures only describe a small portion of either between- or within-speaker variability. Furthermore, prototypes should not be considered as average tokens computed across complete acoustic signals, but rather as structures characterised by a variable number of attributes. These results further confirm that "deviations from the prototype" includes the two types of variability mentioned above, as in [Lavan et al., 2018; Lee et al., 2019]. The first represents differences within speakers with respect to their own prototype, and the second deviations of individual observations from a group prototype.

In this thesis, speech productions have been represented at different levels, i. e. spectrograms or MFCC incorporated a wide range of information about speech characteristics, while subsets of phonetic measurements have been used to represent specific components or part of them in the produced speech. The use of CNN in particular has made it possible to analyse modulations of speech components through precise representations created specifically for our investigations. Further application and comparison of this new way of representing speech characteristics is needed as part of the continuous challenge of understanding and rendering speech changes. The data studied here has been transformed through different representations in order to adapt to multiple levels of analysis. The creation of a consistent knowledge about the phonetic characteristics of the studied speakers

has been fundamental for all the investigations presented in this thesis. Given the strong influence of time on the observed changes in speech, the search for a representation of its dynamics capable of maintaining consistent results through an efficient computation has yielded contrasting results. An important part of the phonetic results concerns the assumptions about the actual reliability of the extracted data. For instance, this has been done through the comparison with reference values for formants and f0, and through the further comparison of already attested relationships between the selected components and our results.

Overall, the formants and f0 values for the PTSVOX are more in line with observations in the literature, with the reading task more suited for less unstable results. This is further confirmed by the low impact of the entropy computation on read speech data from the PCA. However, an important degradation of results is shown when the telephone is used as a recording medium, as also already observed in [Künzel et al., 1995; Eriksson, 2010; McDougall et al., 2015; Hughes et al., 2019]. The impact on f0 is on average 10Hz for both female and male speakers, while it is greater for the formants. The spontaneous speech observations of the formants show even higher differences from the attested reference values, [Gendrot and Adda-Decker, 2005; Georgeton et al., 2012] for formants and [Boë et al., 1975; Pépiot, 2013; Schmid et al., 2012; Gendrot et al., 2012]. The application of filters does not consistently reduce the amount of extracted data, which suggests that the observed values should be considered acceptable, even if an average increase of 200 Hz is observed compared to read speech. Indeed, a higher variability is observed especially for the closed vowels /y/ and /u/, which in spontaneous speech show F1 values similar to open vowels. This may be a direct consequence of the production context where speakers have a lower control on their articulation mechanisms, while from the auditory perspective the outcome is compensated. The age groups from the two corpora being comparable, no other factor than the production task seems to influence this variation.

The female and male speakers' results show differences that remain consistent between both corpora. The observations for female speakers seem to be more reliable than for males, for both f0 and formants, e. g. for the latter a higher number of formats fail to pass the filtering thresholds. The female speakers also show better results using linear models with formant-related measurements such as vowel space areas, even though they only translate as tendencies to significant results. In the PCA, however, we notice that male speakers' information remains more coherent between microphone and telephone in comparison to female speakers for whom the two recordings provide less similar results.

Confirming part of the findings by [Keating et al., 2017; Keating and Kreiman, 2016; Lee et al., 2019] on English language, we notice an overall important role of formants in the PCA from read data. In particular, the influence of FD as a characteristic for female speakers, and, even though to a lower extent, F3 for male speakers. It is not surprising that male speakers show higher characterisation using lower formants, F1 and F2 influence is observed in PC1 while F3 appears only in PC3. In contrast, female speakers are more characterised by F2 and the FD, being, as mentioned above, major influence factors for PC2. The relation between FD and the information about dominance in a conversation is discussed e. g. in [Matsumoto et al., 1973; Chhabra et al., 2012; Zuo and Mok, 2012]. In relation to our findings, this echoes the idea that for female speakers a one by one comparison is beneficial to identify individual characteristics.

Nevertheless, there are important differences between formants and f0 behaviour in the

PCA results. As noticed, the formants play important roles in the characterisation of speakers while f0 is a very marginal factor, only appearing in PCs higher than 3 for both females and males. Given the importance f0 has shown as a speaker' characteristics this is surprising. However, this may be explained by the shape of our data, taking moving mean values provides the PCA with time-related changes of the selected measurements, i. e. f0 observations could reproduce intonational contour patterns. Thus, the monotony of the reading task may influence negatively the importance of this factor, while the relation between formants and movements of the articulators are shown as more characteristic.

Information about f0 is present only as high harmonics, i. e. H42k and H4 in the telephone recordings, where lower formants have a lower weight for the male speakers. The same considerations given for f0 may be applied to voice quality characteristics as their role appears mostly marginal except for CPP, associated with breathiness information. The presence of the latter implies perturbations on f0, as shown in [Traill and Jackson, 1988; Culling and Darwin, 1993; Fraile and Godino-Llorente, 2014; Klug et al., 2019], which may explain the absence of f0 from most of the prominent characteristics.

Moreover, evidence about the link between variability across speech components and language-specific properties [Chodroff and Wilson, 2017; Tanner et al., 2020; Sóskuthy and Stuart-Smith, 2020], enforces our findings limited to French language, as most of the predictable variability across speakers is within a given group of phonetic cues. This could be further supporting evidence to explore how speakers are characterised by the alignment of multiple cues to produce speech contrasts simultaneously and to maximise the acoustic distinction between the categories. Thus, idiosynctratic patterns are maintained rather than emphasising one cue over another. The sample of speakers we analyse refers to a group with very little variation with respect to control variables such as ages and native language. This may be seen as a limit, however, even in this controlled population, the description of variability appears complex. Indeed, the absence of information about differences in variability across different homogeneous populations of speakers, and even speculation is lacking with regard to how many and what kinds of populations exist. This implies the convergence of variability within a population from variability across populations, similarly to what is discussed in [Keating et al., 2017; Keating and Kreiman, 2016; Lee et al., 2019]. The methods presented allow the study of variation in speaker characteristics, which is also fundamental for the improvement of voice perception models.

The multiple facets of the variability matrix

Source and filter components dominate the PTSVOX results from the PCA, whereas NCCFr speakers are characterised by a larger set of factors. Before discussing the differences from the PCA, the first difference observed is the significant results that the temporal cues obtain in spontaneous speech compared to the reading task. As mentioned above, the average age from the two corpora are comparable, hence the difference existing in the results from speech rate and other temporal cues is only imputable to the production task. Speech rates appear to be significantly different between PTSVOX and NCCFr. However, between speakers of the same corpus, no significant difference is observed, except for pauses, duration of first and last phonemes in NCCFr.

A main element influencing duration in the spontaneous task is the presence of hesitation markers, which have shown in the literature [Mary and Yegnanarayana, 2008; Künzel, 2013; Leemann and Kolly, 2015] a high idiosyncrasy with pauses lengths and pattern as well. Duration of last phonemes may be associated with another cue that emerges in the

feedback listeners have given about the perceptual experiment which is the changes in voice quality, in particular for male speakers, in word finals.

Three main aspects emerge from the PCA on NCCFr, differentiating the results from the read task. The first has already been mentioned and concerns the role of entropy in the spontaneous speech results. We have performed comparison analysis without the computation of entropy in both read and spontaneous speech. In the latter case, we have observed that entropy provides consistent additional information in the scale of 10 % increase of explained variance. The explanation may be found in the nature of the studied speech. The higher variability of results from the spontaneous productions seems to benefit from a value measuring its disorder [Shriberg, 2005]. Whereas in the case of a more stable object, such as read speech, this value only brings redundancy about the conveyed information. This is in line with what is observed in [Nilsson, 2006; Setiawan et al., 2009], where entropy is also associated with.

The second aspect, in both between and within speaker PCA, is linked to the emergence of voice quality characteristics which is noticeable, in particular with H1A1 and HNR shared by both female and male speakers. In the case of females we observe that H1A2 and the CPP play an important role in the definition of PC1, thus, indicating a higher presence of breathy voices during spontaneous speech. Other characteristics appear in the results of male speakers, e. g. H1A1, the soe and the LTAS that are linked to different voice quality cues. In addition, rhythm is another aspect that characterises male speakers, by the means of TFS and ENV, showing an overall wider range of characteristics used to describe males in comparison to female speakers.

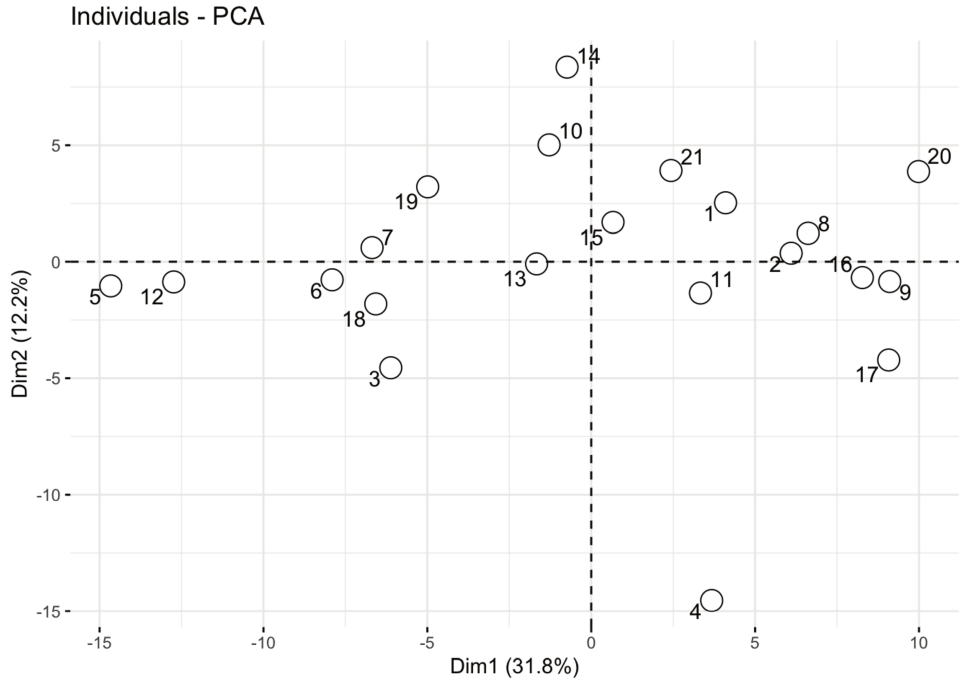


Figure 8.2: PC1-2 space for female speakers of the NCCFr corpus using both phonetic measurements and MFCC. Because of the large amount of data (73k observations per speaker) centroids are used to represent each speaker giving a visual idea of their distance.

The third aspect concerns the association of MFCC with phonetic measurements and their influence on the results. Female and male speakers are influenced in the same way

by these coefficients. In the latter case, it appears that MFCC have a high redundancy with the phonetic information. This is shown by Figure 8.3, like in Figure 8.2 we use centroids in order to illustrate the distances between speakers more visible.

Figure 8.3: PC1-2 space for male speakers of the NCCFr corpus using phonetic measurements and MFCC (Top) or only phonetic measurements (Bottom). Because of the large amount of data (73k observations per speaker) centroids are used to represent each speaker giving a visual idea of their distance.

Indeed, when MFCC are added to the analysis an important decrease of the explained variance is observed and some of the coefficients are integrated in the PCs along the phonetic measurements such as LTAS and spectral moments. Whereas for female speakers there is a small complementarity between MFCC and the other phonetic measurements,

since we observe that they are all placed in a specific PC with no additional phonetic information.

We have seen that some characteristics are prominent for some subsets of speakers and seem irrelevant for others, e. g. for spectral shape variability information, as in [Swerts et al., 2014; Dellwo et al., 2015; Hansen and Bořil, 2018; Eiswirth, 2020]. Furthermore, the results of diaphasic variation, i. e. differences related to speech styles, in particular in our investigation, highlight how multiple linguistic phenomena can be part of the same approach. Modelling variation above the phoneme level provides new perspectives on the location of the variation envelope, with the study of individual modelling serving as an abstraction for the analysis of variation at multiple levels [Michel and Jacqueline, 2000; Freydina, 2015; Brand et al., 2019]. In summary, we have observed that although significant between-speakers variability has been proven by statistical tests, it does not necessarily entail high recognition rates using the available classification algorithms. The high within-speaker variation highlights the importance of inferential statistics, in order to give a more in-depth understanding of the changes involved in speech productions. The use of a speech component modulation representation capable of providing consistent results and maintaining computational efficiency in terms remains a challenge to this day.

8.2 Evaluating the contribution of speech components

As shown in part of the results of the previous section, MFCC are not very successful in the statistical description of speakers' characteristics, yet they are the standard for ASR. In order to understand why they are so effective for ASR it is important to investigate what information they carry and how we can interpret them. In other words, what characteristics they can be linked to. The combination of MFCC and classical phonetic measurements has given some clues to answer these questions.

First of all, we have observed that their behaviour is not the same for female and male speakers. For females, they are linked to f0 and rhythmic cues, while for male speakers to energy distribution and spectral information. In addition, the fact that these characteristics have important roles for the descriptions of both sexes respectively confirms that MFCC, rather than representing a single cue, adapt to the relative characteristics of each speaker. Further investigations of the interactions between classical phonetic and automatic measurements are shown by the CNN results from Chapter 7.

Global representations and their implications for the rendering of speakers' variability matrices

Through the CNN approach we have used three major representations of the variability matrix associated with speakers' characteristics. These global representations have served the purpose of combining a large number of information about the changes in the considered speech sequences. The spectrograms have represented a standard in Phonetics for the visual study of speech for decades. Phoneticians rely on visual cues in order to describe multiple production mechanisms. Thus, it is not surprising that the Spectros representation has obtained the highest results in our experiments, outperforming MFCC and phonetic measurements in the SID task. Our results are aligned with other studies using the same data, i. e. [Gendrot et al., 2020], where sequences of 2 s duration are used

showing accuracy of 93.7 % against 93 % (females) and 96 % (males) for our 4 s sequences.

However, we have been faced with numerous difficulties in the tuning of spectrograms in order to acquire satisfactory results. Part of them are described in Section 7.3. For instance, the Hamming window and no pre-emphasis has shown the most consistent results in our examination, this is the case for ASR studies such as [Mary and Yegnanarayana, 2008; Sahidullah and Saha, 2012b]. Based on the compared results, we can also confirm that the duration of the tokens has a greater weight on the creation of the speaker model by the CNN than the amount of training samples. Furthermore, the promising results shown by other studies [Hsu et al., 2004; Elliott and Theunissen, 2009; Stilp and Kluender, 2010] and ours on the use of MPS to represent the modulation of spectral variations over time in a static representation, open the way for further research on how to more finely render speech changes.

The preliminary studies we have conducted to establish the parameters for our baseline representations also included spectrograms with artificially-added noise. In this case, we have been able to test both regular and mismatched conditions for the training and testing phases. The results show CNN's high reliability in identifying noise inside a spectrogram image when the noise only appears during the test phase. Noised spectrograms of 2 s duration achieve comparable results than nasal vowels spectrograms, the latter using nearly double the amount of training samples. This confirms the fact that higher training sets compensate for lower duration.

Once the spectrogram results were established as the baseline for our investigation, the representations for speakers' characteristics also required multiple tunings. The definitive choice has led us to a vector-like image allowing the use of different speech components while remaining efficient from the computational perspective. In the same image, one or multiple components can be represented with a minimal amount of noise that a wider spectrogram image might carry. This is confirmed by the verification task, where Glob achieves higher performance than Spectros and the generalisation task where the two show very similar results. However, as the SID results indicate, the information present in the spectrograms is still consequent, due to the way the spectrograms are able to render the interaction between the speech components.

The Glob vectors, and the MFCC, appear to lack some crucial aspects but show potential for the improvement of the representation of speakers' characteristics. The first because of the high interpretability they carry, the second because of the adaptation to the reference population they are based on. Even though MFCC show the lowest results for the three tasks we tested, they are able to combine trends observed in both Spectros and Glob results. This can be observed in Figure 8.4, where the confusion matrices for SID of the three global representations are shown. For instance, similar speakers achieved low scores for both MFCC and Glob, explaining their similar Informedness scores described in Chapter 7. However, the false positives and negatives patterns are mostly shared between Spectros and MFCC. This, in turn, explains why the Markedness results appeared as a combination of the other two representations.

Taking Doddington's terminology, Spectros results are characterised by an important presence of goats speakers. In both Glob and MFCC, we observe that some of these goats become wolves and lambs, since the lowest performances are observed for speakers who are more frequently confused and imitated with their counterparts.

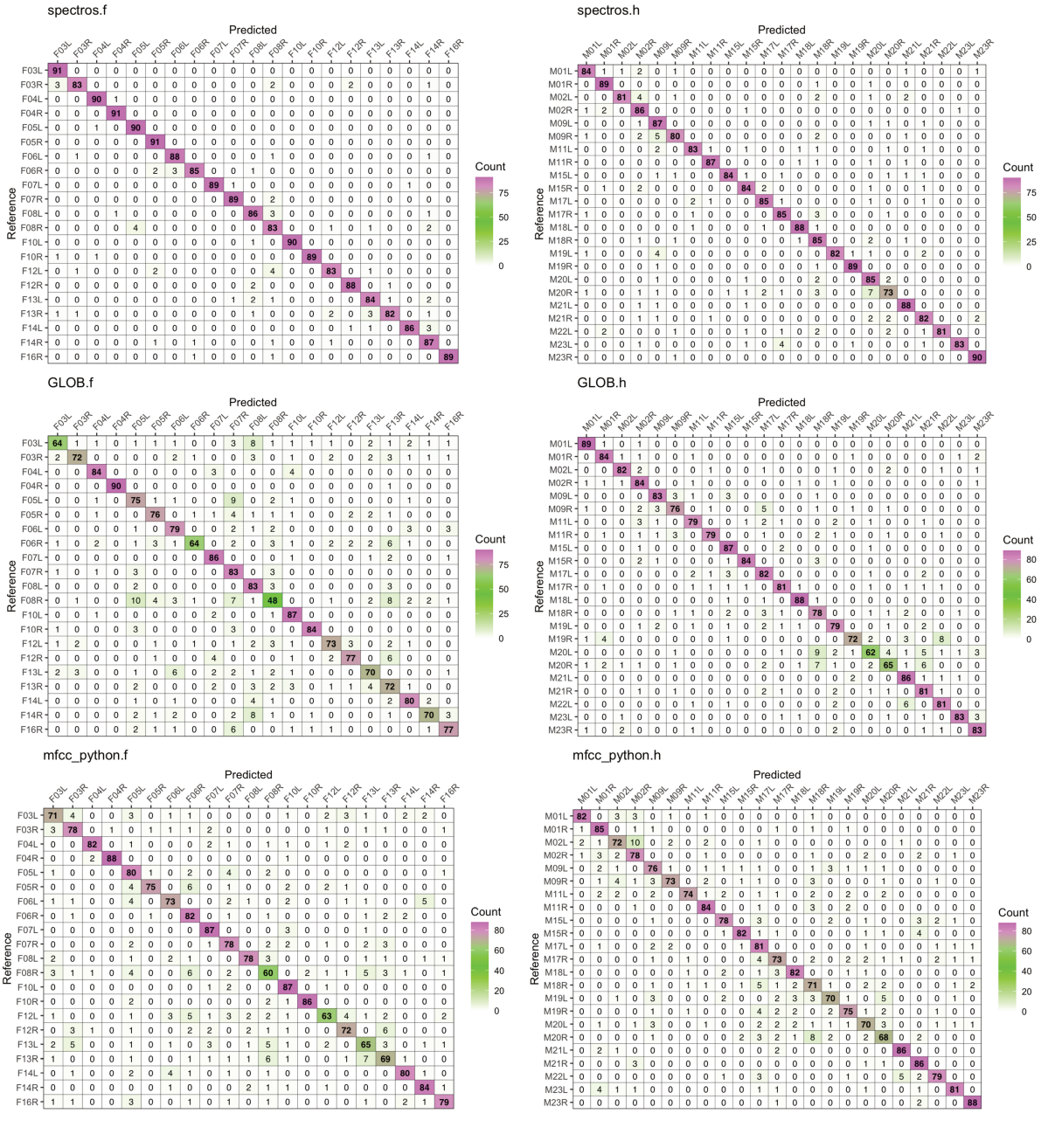


Figure 8.4: Confusion matrices from the SID tasks for the three global representations (Spectrograms, MFCC and phonetic measurements). On the left side female speakers, on the right male speakers.

Diagenic variation in the distribution of characteristics

In all our results, the tendency of the characterisation of male speakers achieving better performance is confirmed. The exceptions to this trend are shown by SID through spectrograms and in the verification task using formants and spectral moments. The latter result in particular confirms that spectral variations have a greater influence on modelling of female speakers. The greater extraction of formant frequencies for female speakers has been discussed above, as well as the higher weight of these cues in the PCA. The spectral moments carry basic information about the spectrum shape, i. e. mean, SD, skewness and kurtosis measures. As shown by our results, these elements are more consistent than other more classical phonetic measurements of spectral shape. In particular, the asymmetry coefficient showing a higher influence on the results indicates that the shift in energy to low or high frequencies is an important discriminant cue for female speakers.

The high influence of formants in the characterisation of female speakers, confirmed as well in the verification task, suggests that the way female speakers use their articulators is important in shaping their speech productions. Similarly, energy represents an important factor for the characterisation of female speakers. This is explained by both Ms and Qual outperforming the other components in SID and verification. In the Qual set we included the best performing LTAS measurements when tested individually, i. e. the peak between 3 and 4 kHz, as well as the best performing energy related cue of soe, and the CPP. The results are in line with what is observed within formants, which shows that higher frequencies include more consistent information about female speakers. The same trend is confirmed by the low performance of prosody measurement, taken at a very low level in the spectrum.

Furthermore, a consistent result from our experiments is the role played by f0 and its harmonics. Even though the scores do not outperform those of other measurements, such as HNR, which perform best for both female and male speakers characterisation, both the confusion patterns and marked speakers are highly related to what is observed in the Glob results. The CNN and PCA results both confirm the prominent weight of H2 and mid-high frequency harmonics, H2k, for female characterisation related to f0 cues. The opposite trend is clearly seen with H1 and H4 being in both CNN and PCA results marked as the harmonics with the highest weight for male speakers.

Overall, the fundamental difference that appears between female and male speakers, concerns the higher influence of the voice quality components on male characterisation. Indeed, the cues related to the harmonic source spectral shape outperform even the HNR measurements, which have shown consistent results for female speakers as well. The presence of the H1-H2 parameter has already emerged in the literature as a male speakers characteristic [Vaňková and Skarnitzl, 2014; Hughes et al., 2019], and it has been correlated with the open quotient [Kreiman and Shue, 2010]. H1-A1 appears as the second most influential indicator in the Hadiff set for male speakers, it is considered an indication of the bandwidth of F1, which in turn is an indicator of the degree of glottis opening during the closing phase. This is a further confirmation of what we observed using dynamics computation in Chapter 6. The analysis of peaks and troughs, using measurements derived from studies on open quotient, indicates that for female speakers the uprising of the speech production is more characteristic than for male for whom the closing trajectories have a higher differentiating power. Furthermore, the higher performance for male speakers indicates that voice quality is a more robust aspect for their characterisation than for

female speakers.

In regards to voice quality components which perform best in the characterisation of male speakers, the second best performing group of phonetic measurements is HNR, which also obtains consistent results for female speakers. In both cases the computation from the lower pitch range, i. e. 0-500 Hz, appears as the one providing the highest amount of information. However, for two oldest female speakers, it is the SHR measurement that has the most influence on the results. The two cited speakers are the only ones that belong to a completely different age range, 40 and 50 years old, than the rest of the NCCFr corpus, between 18 to 27 years old. Thus, the differences in voice quality, and in particular in the alternation of vocal cycles associated with SHR measurement, appear to be an indicator of age. This last assumption is made from a very small comparison sample and should be assessed by further investigations on more age-heterogeneous populations.

The interpretation of variation

The shared results between PCA and CNN approaches acknowledge two important facts: statistical models give a good representation of speech components' modulations, even though they use "static" values, i. e. mean, CoV and entropy; in order to take advantage of this knowledge the accuracy of representation needs to be tested in improved frameworks rather than classical linear models. The importance of entropy in statistical modelling, indicates how speaker characterisation does not rely on simple addition of well-performing measurements but rather on the interactions between the varying distributions of multiple components. The other models used to perform earlier classifications, e. g. LDA and SVM, on the data have not given consistent results, with slightly higher than chance level. However, the small trends that could be observed on the speakers' confusions parallel the more robust results we obtained later. Following the idea of analysing speech through a more coherent representation of its varying nature, the LLR approaches applied to source and filter characteristics have improved the quality of the representation but reduced the interpretability. The other downfall of the use of polynomial coefficients has been the redundancy of information and interactions regarding f0 and formants.

The additional aspect provided by the phonetic knowledge for the CNN approach is the direct interpretation of the different tasks results. Knowing what makes the input samples different from each other, thus how the phonetic information is differently distributed, implies that the interactions and influences of the outcomes can be finely interpreted. As already mentioned, characterising a speaker's voice is not a simple process and understanding it involves mirroring multiple aspects of cognition. The different approaches to the CNN tasks try to represent and to recreate some of these aspects. The combination with phonetic interpretation of the outcome makes the different tasks we perform a powerful tool to understand the characterisation process.

Furthermore, the evaluation of the different interactions between speakers shown in the three tasks is at the core of our CNN investigation. Beyond high performance scores, which are useful for the reliability of the studied representation but secondary for our aim, the aim being to deepen the understanding of the results through different metrics and their meanings. For instance, the use of Informedness and Markedness metrics makes it possible to understand how the information about the different components is extracted and used differently for each speaker or group of speakers [Powers, 2011; Stoll and Doddington, 2010]. The study of the groups that the model is able to create with similar speakers covered an important portion of our results, especially through the scope of the

generalisation task, which combines the outcomes of the other tasks. In the verification task, the CNN is tested on speaker identity in a binary framework. However, during the testing phase, samples of speakers previously confused with the target speaker are added.

The core of this task is represented by the results from the so-called unknown speakers. If these speakers, unknown to the CNN until the test phase, reach high Informedness or Markedness scores, we consider two hypothesis: if they already appeared among the informed/marked speakers in the previous tasks, it means that the CNN is able to characterise them even without prior knowledge; it could also mean that the knowledge of the target speaker helps the CNN to better/lesser use the information about speakers with similar characteristics.

Overall, the scores are slightly reduced with respect to the discrimination task. This indicates a certain robustness of speakers' model created by the CNN but also highlights the bias that the modelling can carry. For instance, we observe a clear difference between female and male speakers' results. Male marked and unmarked speakers remaining the same indicates that the model created is more "absolute". In contrast, female speakers modelling is highly influenced by the reference population. Thus, female speakers unknown to the CNN tend to have lower performance and continue to bias the results, while male speakers, even if they are unknown to the CNN, can achieve consistent results. An hypothesis to explain this behaviour may be that female speakers present a wider range of characteristics variation, while male speakers have more specific characteristics' distributions in their matrices. This is consistent with the results discussed in Chapter 3, showing that most ASR systems perform better with male speakers' stimuli. However, there is no agreement in the literature to what causes this result. Thus, the comparison of different ranges for female speakers is necessary to leverage their between-speakers differences. In contrast, male speakers rely on the prominence of single features to be characterised.

8.3 Connecting the dots

The perception results offer new insight into all the other findings, but they also bring a new variable to the problem of voice characterisation, the listener. As described in Chapter 4, despite a long history of active research, little is known about the cognitive and perceptual processes underlying voice discrimination and recognition. Most studies of voice perception have focused on stimulus characteristics, rather than on listener behaviour. Researchers have traditionally favoured designs in which single characteristics or stimuli conditions are varied and listeners' performance is measured as a function of these variations. Indeed, changes in speaker recognition performance emerge because the selected dimensions are considered important rather than focusing on the listeners' behaviour [Lavner et al., 2000; Eisner, 2015; Baumann and Belin, 2010; Gerlacha et al., 2020]. We performed a preliminary perceptual investigation, where simple cues such as segments duration and intonational contours were exchanged between two speakers. Even though the first results showed some trends also found in our final perceptual design, i. e. a same-sex effect, we have decided to focus on a wider perception analysis rather than to use specific cues.

However, the multiple influence factors that listeners carry by themselves, e. g., their age, sex or mother tongue, represent a complex set of variables to consider when working

on a voice perception task. A human-machine comparison is fundamental to improve decision making algorithms or clustering analysis methods and assess their reliability. The modelling and characterisation of sound is performed by our brain at high performance levels, but further research is needed in order to understand its mechanisms. In this regard, the comparison of the three clustering, CNN-PHON-HUM, further confirms some of the previous findings. For instance, we observe that the higher similarity for the female speakers' clustering is with formants based responses. This confirms the important role played by formants for the characterisation of females and not for male speakers.

Multidimensionality of clustering approach

As mentioned throughout this thesis, the interpretability of the studied measurements and their outcomes is another fundamental aspect that the perceptual approach addresses in particular. The question is to know whether what we observe in human perception can show us that some parameters are more interpretable or reliable than others. In this sense, the use of a clustering task allows us to observe how human listeners place similar and dissimilar voices in a variability space based on their self-created models without focusing on a specific characteristic. The use of a multidimensional space in the perception task could be even more relevant for the listeners to give a more accurate distance between the clusters. This could allow a further comparison with the hierarchy the HAC method provides.

Given that the CNN results are more robust for multiple speaker recognition tasks than those from statistical models presented in Chapter 6, it is not surprising that PHON clusters have lower clustering coefficients. However, these results remain important for our three-sided comparison, since they still provide some contributions to the outcome of our investigation. The phonetic investigation gives the actual interpretation and implications of speech components in speaker characterisation. Therefore, the diagenic differences between female and male speakers highlighted in automatic and phonetic results, are reiterated in perceptual answers. No main effects of phonetic convergence are observed as in [Pardo, 2006; Lee et al., 2021], even if the situational production might suggest it.

Overall, the clusters obtained by formants and MFCC show the most robust results for both females and males when the statistical modelling is used as the basis for the clustering analysis. The first result continues the trend of the important role played by formants in conveying information about female speakers. In addition, f0 appears to have a lower influence than F1-4 when combined for the cluster computation. Spectral moments show similar results for both female and male speakers, as well as an important similarity with Glob clustering. The influence of voice quality information for male speakers is further confirmed by the Ltas results having an important influence on their clustering. The PHON computation of clusters is obtained through distance matrices computed on the statistical description of the data. In contrast, the CNN clusters come from the combined matrices of the three tasks, used as distances between the different speakers. Differences exist between the two modelling, however some trends are shared. Indeed, we observe that male speakers clusters have overall higher similarity between PHON and CNN than female speakers. The least similar responses between the two groups of clusters were for f0 and Amp. This indicates that these components are more influenced by the used type of representation. Modulation of f0 and its harmonics, as in the CNN, appears to convey more information, e. g. pragmatic ones [Vaissière, 2004; Nebot, 2021], than the static description through statistical values. In the latter, we observe less impact

on the representation of speakers' characteristics and a lower similarity with the other components' clusters.

Moreover, the overall similarity of the global representations between PHON and CNN clusters show low scores, confirming the idea that a static description, even though through a high quality statistical model, still conveys the information very differently. These differences are important since they could provide complementary information that a single analysis cannot. The only way to fully understand the actual amount of complementarity between speech models and to get rid of noisy elements is validation through human perception.

Perceptual organisation

Nevertheless, a very important aspect of a perceptual analysis is that listeners may be unaware of what they are perceiving. For instance, feedback from the listeners of our experiments on the characteristics they used, or thought they used, to perform the task formed an important part of our hypothesis about the obtained clusters. The listeners were overall surprised to discover that there were actually 44 different speakers in the task. The feedback only partially reflects the similarity observed between the clustering results, and more consistently for female voices. More experienced listeners, such as some of the participants of our study, may have a higher level of consciousness about the selected traits used to characterise the speakers, as discussed in [Matsumoto et al., 1973; Kreiman et al., 1990; Latinus and Belin, 2011; Eisner, 2015]. The influence of mother tongue is more prominent for this group of listeners, as shown by the fact that the non native experts share the least amount of information with both other listeners and the non human clusters, similarly to findings by [van Lancker et al., 1985; Adank et al., 2009; Kelly et al., 2016b; Gerlacha et al., 2020]. This is true for results on female and male voices, as in [Baumann and Belin, 2010; Chhabra et al., 2012].

Some participants affirm to have relied on rhythmic cues such as intensity variations for the speaker to speaker comparison in order to feed the clusters. The presence of breathiness, hoarseness or other voice quality elements have emerged as a characteristic for the characterisation of male voices. In this sense, findings from [Kreiman et al., 1993; Lavner et al., 2000; Cutler et al., 2010; san Segundo and Mompean, 2017] show that voice quality cues have an important influence in characterising human voices. In contrast, for female speakers the listeners found the task more difficult, with voices being in general more similar, thus giving less possibilities to create multiple groups based on a first hearing. To our knowledge, there are no studies highlighting a similar tendency of female speakers having a single prototype that requires extensive processing to define its covariation. This can be seen in the time used to complete each task, since the task involving female speakers has a slightly longer average duration, 63 minutes against 55 minutes for male voices. However, this difference may also be explained by the fact that the male voices task was most commonly performed as second, hence making the listener more familiar with the operation.

Similarly to the other results obtained throughout our investigation, perceptual characterisation results confirm different behaviour in the characterisation of female and male speakers. In the case of expert listeners, results suggest that they use harmonics and energy information to characterise male speakers. In contrast, naive listeners' results are parallel to those of Rhythm components, suggesting that they represent more accessible cues for a less trained ear [Matsumoto et al., 1973; Lindh, 2009; Chhabra et al., 2012].

As mentioned above, our results show a same-sex effect in perceptual characterisation of voices. Indeed, all female listeners share at least a third of their clustering of female voices while less stable answers are observed for male voices. The same effect is observed with the responses provided by male listeners. Consistency of this results needs further questioning against a group-related effect that might take place when considering small testing populations. However, a consistent result shared by all listeners is the association of the two oldest speakers. As shown in the literature review, e. g. in [Moyse, 2014; Waller and Eriksson, 2016], the accuracy of age recognition by voice showed consistent results. In our case, having only two female speakers who do not belong to the same age group as the remaining 42 speakers may simplify their discrimination. However, the fact that they consistently appear together is a further confirmation of the influence of ageing on speech, and of the sensibility of humans to identify it.

However, we notice that non-native French listeners, even though they associated the said voices, have put them in larger clusters in contrast to native listeners that always isolate them in a single group. An even lower age effect is observed in PHON and CNN clusters. The latter only tend to isolate the two speakers in clusters obtained through spectrograms and MFCC, showing that the use of modulation of single components may imply loss of some speaker's information such as those associated with age. In this sense, we observe that this information is partially retained in the PHON modelling, however only the oldest of the two speakers is consistently isolated from the others.

An additional factor we have examined, in relation to the clustering analysis, is the potential influence of the speakers' origin. Ten speakers from the NCCFr corpus came from seven different French regions of Île de France. The presence of a regional accent can be identified as an important characteristic for some speakers. Our results do not show significant general trends for the male speakers. However, the two non-francilian female speakers are systematically associated by human listeners, even if in larger clusters rather than an isolated one. This trend is reiterated in the CNN cluster, but to a lesser extent, showing the two speakers associated only on the basis of intonation cues and formants. Further investigation may be needed to assess the actual identification of a different accent than the majority of the speakers.

One more thing... on combining multiple approaches

Speaker characterisation can be linked to different domains but its fundamentals cannot only be linked to one. To understand the nature of this type of study, like speech itself, one must examine its multiple facets. For instance, from a purely automatic perspective the analysis of about 60 speakers may seem insufficient to obtain consistent results. Similarly, having a wide range of components rather than examining a specific aspect in detail may seem too ambitious from a classical phonetic perspective. We consider that larger, heterogeneous groups of listeners are needed to develop convincing theories about the perceptual aspect of speaker characterisation.

In order to bridge an additional gap that exists between human and machine perception, we explore lexical distances. As mentioned earlier, when we perceive a large amount of information, we can retrieve a number of pieces of this information, including the content of the spoken message. In order to create a speaker model, humans also use segmental information, looking for characteristic ways to produce certain words or phonemes. The use of a particular expression or word chains may be another characteristic to distinguish speakers. In this perspective, we should consider the results of lexical distances of speak-

ers as an additional element to the variability matrix of speaker characteristics. Even though the overall lexical comparison does not seem very robust since the conversational content involves similar interactions, the keywords comparison covers the percentage of dissimilarity shown by the cosine distances.

Integrating knowledge from multiple domains should be the next step to explore more efficient models. For instance, we have a clear example of a representation of the modulation of speech components based on phonetic knowledge but too complex to be processed through classical modelling. Its behaviour could be described statistically, but a real fine representation of speech dynamics has resulted in extensive research requiring small contributions from multiple areas.

The underlying starting idea of this thesis is that the representation of speakers' characteristics is mostly based on spectral measurements, obtained from small analysis windows. Our work applies larger windows which can also include f0, intensity or spectral variations, in order to improve both interpretation and fidelity of the conveyed information. However, a solid basic knowledge is needed to allow this kind of approach. It is not just an analysis based on means from one or multiple measurements that allows the study of modulations over different windows. The next chapter serves as a conclusion to this third moment of our investigation. The final considerations on our work are stated, i. e. what has it done to expand the speaker's characterisation knowledge and what could be done in the future.

Chapter 9

Epilogue: do we have answers or more questions?

Variation is key to understanding and explaining how speech works. In Phonetics, it is common to refer to the simple fact that no two speech utterances are ever the same, even if they are produced by the same speaker. This fact is linked to the idea of within-speaker variation, which in turn, represents the first issue when studying speaker characterisation. Overcoming the within-speaker variation can represent a difficult task depending on what we focus on and which influencing factors are involved. For instance, this is the main reason that distinguishes forensic speech science from other forensic disciplines. For forensic evidence such as DNA, samples from the criminal and the suspect can be identical, and the criminal can therefore be identified. In speech comparison, the level of confidence in the results remains an important issue. Modern techniques to model speech productions are successful in meeting the challenge of describing the high variability matrix associated with speaker characteristics. However, further investigations are needed to extend the maximum range of variation that can be observed, using currently available models in linguistics.

To describe the multitude of sources of variation in the information that speech productions can convey about a speaker, considering the innate-learnt dichotomy seems reductive. Instead, the use of a physical-psychological-social trichotomy, such as the one shown in Table 1.1 of the Introduction, seems more appropriate. For instance, the different components of speech that can be associated with speaker' characteristics should not be considered as related to a single type of information, but should be studied from a multidimensional perspective. Throughout this thesis, we have highlighted the importance of understanding the interactions between the different components and influence factors in order to better anticipate the possible expressions of speakers' characteristics.

Given the view of a speaker identity as a complex system resembling a matrix of variability distributions, the addition of small contributions is fundamental to make progress in its description. The contribution may correspond to the consideration of a single or several parameters, and their influence on the speaker's model as well as their interactions with other factors. As a consequence, the computation of probability distributions for the variability matrices can gain in reliability.

Additional factors unrelated to the nature of speaker characteristics must be considered

when studying how speaker information is transmitted. Speech recordings can vary in quality due to, for example, a telephone transmission, distance between the speaker and the microphone, background noise. These and other influencing factors have been shown to weaken portions of the speech signal, causing unwanted changes. These are additional challenges that researchers need to overcome in order to fully resolve the issue of modelling speakers' characteristics.

Nevertheless, our investigation has provided some answers to questions regarding speaker characterisation, even though there is always margin for improvement and adjustment for future research. For instance, this work could have benefited from the use of articulatory measures, in order to provide a more in-depth understanding of the relation between speakers' anatomy and their produced speech. However, since this work was conducted on already existing data sets where such measurements are absent, their integration was impossible. Similarly, perception studies involving the isolation of single components could have provided additional information about the weight they have in perceptual characterisation. The number and variety of listeners is the other important issue concerning the perception task of this thesis. Male listeners represent a very small group compared to female listeners, six against 21. A larger number of non-native speakers might have shown more consistent trends in perception of speakers' characteristics from a multilingual perspective. This is why our perception results should be considered as showing trends that require further investigations for significant confirmations. During the exploratory phase for the creation of the perception experiment, we hypothesised to test a perception framework resembling the voice ratings described in Section 4.1.2. However, improving the classical voice rating by incorporating an abstract class may have helped the listeners classify similar voices based on their own symbolic perception. The choice of this additional class requires a fine analysis on the listeners' prior perceptual capacities in order to assess their reliability. Thus, the use of a clustering task without strict instructions on which feature to focus on appeared to be an acceptable middle ground.

Following this idea, the integration of clustering tasks directly on the CNN could have been explored. In particular, this could be done by improving the generalisation task we presented. CNN have already shown their potential application in open class problems [Shu et al., 2018]. The integration of LSTM architectures and more advanced attention mechanisms has been partially explored and could represent a real improvement for the modelling of speech component modulations. In this perspective, even more comparable results between human and machine could be obtained to find more efficient modelling techniques. Our CNN results used only coherent training and test data, except for some preliminary studies, and the generalisation task. Concerning the latter, it uses only a partial set of all our selected components, in order to reduce the redundancy of information that could cause all the iterations of the different subsets of components. In perspective, the idea of a representation made by the best performing components for the generalisation task may improve the efficiency of the characterisation. Indeed, the absence of more developed mismatch conditions, such as performing the training on a complete component and testing only a subset could represent a way to better understand the weight of each phonetic measurement.

Furthermore, as mentioned in the introduction of Chapter 7, the idea of different representations of speech variability has been another important focus of our work. Multiple representations have been used to capture the dynamic nature of speech productions. Consistent results have been obtained through the representation of speech by spectro-

grams and modulation vectors in order to focus on isolated components. However, a simple time shift in training and test data may lead to important challenges for such representations. Since temporal variations are an intrinsic characteristic of speech productions, their consideration must be done more efficiently by exploring more versatile representations. In this perspective, the potential shown by MPS results is the basis for new explorations in how the modulation of speech components is represented.

The research objectives addressed in the Introduction of this thesis included giving a wide overview of speech components' interactions and their roles in characterising speakers, in order to further understand their actual contribution to speaker characterisation. The comparison with human perception covered the third research aim to question the validity of the obtained results. Even though we have mainly fulfilled our aims, their complete achievement requires answers that may never be obtained, since each new explanation brings a new question. However, some results have emerged in this thesis that should be considered important for the speaker characterisation domains. (i) As shown in Chapter6 and discussed in Chapter 8, describing within-speaker variation is fundamental in order to understand between-speakers variability. The characteristics allowing the description of variation of speech components occurring in different stimuli by a single speaker appear consistent with the phonetic measurements that play an important role in the separation of stimuli by different speakers. This allows to create multiple groups of speakers inside the studied population that are characterised by similar distributions of speech components. In this sense, we observe that (ii) source and filter characteristics are more important in the description of female speakers' variation, while (iii) voice quality characteristics such as breathiness and hoarseness have a greater impact on male speakers. These results suggest that the characterisation of female speakers relies more on linguistic and articulation aspects, as well as the role of interlocutors in the conversation, i. e. the example of FD. Whereas it relies on paralinguistic aspects for the characterisation of male speakers, such as the level of confidentiality between speakers that changes in breathiness may convey. (iv) The perceptual responses confirm these tendencies, with the human-based clusterings showing consistency with CNN- and phonetic-based results. In particular, with the subsets related to the mentioned components that show greater similarities with the perceptual clustering corresponding to the female and male voices tasks. The clustering analysis further highlights (v) consistency of CNN results with the statistical analysis of speech components, supporting further application of these methods for phonetic studies. The last highlight of this thesis concerns the role of MFCC in comparison to classical phonetic measurements. (vi) They show a great adaptation to speakers' characteristics in the observed data, relating to different aspects for female and male speakers, and for the multiple groups of speakers present in our population. Rather than being representative of a specific trait, the MFCC are mainly linked to intensity and f0 for female speakers characterisation, while to the distributions of energy and low level spectral shape for male speakers.

In conclusion, this thesis has explored phonetic characterisation of speakers with the objective of combining phonetic knowledge on the interactions of speech components, with advanced modelling techniques through CNN, in order to create a mutually beneficial analysis. We hope that these results encourage discussion to solve issues associated with the description of speaker characteristics.

Appendix A

Speakers-specific values for F0 and first four formants

Speaker (F)	Age	Mean F0 (SD)	Speaker (M)	Age	Mean F0 (SD)
F03L	50	231 (35)	**M01L**	23	108 (18)
F03R	40	214 (33)	**M01R**	25	108 (14)
F04L	25	198 (22)	**M02L**	24	132 (21)
F04R	25	256 (31)	**M02R**	24	132 (20)
F05L	19	233 (19)	**M09L**	24	120 (20)
F05R	18	228 (20)	**M09R**	24	107 (16)
F06L	21	219 (27)	**M11L**	18	138 (19)
F06R	21	215 (31)	**M11R**	18	132 (20)
F07L	20	217 (34)	**M15L**	27	114 (15)
F07R	20	246 (38)	**M15R**	23	119 (13)
F08L	21	267 (37)	**M17L**	20	135 (19)
F08R	20	243 (36)	**M17R**	20	125 (13)
F10L	19	275 (42)	**M18L**	20	165 (21)
F10R	20	218 (34)	**M18R**	20	133 (24)
F12L	22	220 (34)	**M19L**	19	127 (15)
F12R	21	232 (30)	**M19R**	26	117 (17)
F13L	19	223 (36)	**M20L**	22	127 (19)
F13R	18	241 (28)	**M20R**	22	136 (20)
F14L	20	196 (20)	**M21L**	21	105 (15)
F14R	23	232 (27)	**M21R**	19	125 (12)
F16R	21	227 (32)	**M22L**	19	106 (12)
			M23L	20	124 (12)
			M23R	23	130 (15)
Mean all	23	230 (31)	Mean all	22	125 (17)

Table A.1: F0 values for NCCFr speakers.

Speaker (F)	Age	Mean F0 (SD)	Speaker (M)	Age	Mean F0 (SD)
LG001	19	234 (33)	**LG004**	22	120 (20)
LG002	21	230 (27)	**LG005**	18	98 (9)
LG003	20	226 (32)	**LG008**	23	117 (22)
LG006	24	144 (27)	**LG010**	18	110 (13)
LG007	18	196 (22)	**LG012**	18	154 (17)
LG009	19	217 (24)	**LG013**	21	133 (19)
LG011	18	191 (32)	**LG015**	19	117 (11)
LG014	19	198 (30)	**LG016**	19	144 (14)
LG018	19	220 (30)	**LG017**	20	108 (11)
LG020	19	230 (27)	**LG019**	19	111 (18)
LG022	19	215 (19)	**LG021**	20	119 (12)
LG023	19	242 (24)	**LG024**	19	105 (12)
Mean all	19	212 (27)	Mean all	19	120 (15)

Table A.2: F0 values for PTSVOX speakers, microphone recordings.

Speaker (F)	Age	Mean F0 (SD)	Speaker (M)	Age	Mean F0 (SD)
LG001	19	236 (39)	**LG004**	22	141 (72)
LG002	21	252 (41)	**LG005**	18	108 (54)
LG003	20	237 (44)	**LG008**	23	142 (80)
LG006	24	165 (62)	**LG010**	18	117 (44)
LG007	18	201 (28)	**LG012**	18	162 (42)
LG009	19	221 (27)	**LG013**	21	165 (79)
LG011	18	210 (36)	**LG015**	19	124 (36)
LG014	19	205 (40)	**LG016**	19	154 (38)
LG018	19	228 (35)	**LG017**	20	117 (46)
LG020	19	242 (36)	**LG019**	19	117 (40)
LG022	19	224 (23)	**LG021**	20	129 (30)
LG023	19	244 (28)	**LG024**	19	123 (58)
Mean all	19	222 (37)	Mean all	19	133 (52)

Table A.3: F0 values for PTSVOX speakers, telephone recordings.

Appendix B

Read passages from PTSVOX

Hereafter the three passages read by speakers from the PTSVOX corpus. Each row represents the actual chunk in which we segmented the whole recordings for the analysis discussed in Section 6.4.1.

Passage 1

Au nord du pays on trouve une espèce de chat dont la queue est très courte
Ils sont noirs avec deux taches blanches sur le dos
Leur poile est beau et doux
Juste à côté vit une colonie d'oiseaux
Dont les nids sont accrochés au bord de la falaise
Ils doivent faire attention à ne pas faire tomber leurs oeufs dans la mer
Ma soeur n'a qu'à traverser la rue pour rencontrer ces deux espèces
Vivant en harmonie au coeur d'un parc naturel
Régulièrement sur le coup de midi après avoir bu un bon thé
Nous sortons de chez elle pour aller observer ces animaux.

Passage 2

Ma soeur est venue chez moi hier pour prendre le the
Elle me parlait de ses vacances en mer du nord
Lorsque dans notre dos tomba un petit oiseau
Ses deux ailes etaient blessées et il avait reçu un coup violent sur la queue
Son coeur battait tres vite mais il etait en vie
Son plumage etait beau et doux
Je m'approchait du bord de la fenetre pour regarder dans la rue
Un chat s'eloignait d'un nid perché sur un arbre
Il avait du faire fuire l'oiseau apres l'avoir attaqué.

Passage 3

La bise et le soleil se disputait
Chacun assurant qu'il était le plus fort
Quand ils ont vu un voyageur qui s'avançait enveloppé dans son manteau
Ils sont tombé d'accord que celui qui arriverait le premier
À faire ôter son manteau au voyageur serait regardé comme le plus fort
Alors la bise s'est mise à souffler de toutes ses forces
Mais plus elle soufflait plus le voyageur serrait son manteau autour de lui
Et à la fin la bise a renoncé à le lui faire ôter
Alors le soleil a commencé à briller et au bout d'un moment le voyageur rechauffé a ôté son manteau
Ainsi la bise a dû reconnaitre que le soleil était le plus fort des deux.

Appendix C

TF-IDF values for 5 keywords from the 44 NCCFr speakers

Speaker	Keyword 1 Keyword 4	TF-IDF 1 TF-IDF 4	Keyword 2 Keyword 5	TF-IDF 2 TF-IDF 5	Keyword 3	TF-IDF 3
F03L	Laurence théâtre	8.92 2.73	enfants pauvre	3.96 2.56	vie	2.83
F03R	femme semaine	4.11 3.01	mère soeur	3.68 2.69	Liliane	3.67
F04L	histoire étoiles	2.64 2.18	argent chansons	2.59 2.11	prix	2.53
F04R	moment bizarre	5.12 3.19	prix vie	4.22 2.98	envie	3.69
F05L	cafés faim	4.25 2.44	fille Diams	3.55 2.18	genre	2.66
F05R	soir frère	2.66 2.39	pied peur	2.62 2.24	Diams	2.43
F06L	mère gamine	4.65 3.18	fille parents	3.78 3.00	Lolotte	3.60
F06R	envie Morgan	5.18 3.64	mère gamine	4.93 3.62	argent	4.44
F07L	enfants mercredi	6.22 3.10	parents filles	4.38 3.01	problème	3.64
F07R	Marie garçon	5.40 2.74	enfants interdiction	4.46 2.66	maman	2.95
F08L	vie garçons	4.46 3.21	Sandra pays	3.59 3.09	micro	3.35
F08R	clope vidéo	3.33 2.88	soir euros	3.10 2.58	frères	3.03
F10L	président femmes	3.40 2.86	frère ordinateur	2.93 2.26	compagnie	2.90
F10R	écran vendredi	2.24 1.73	personnes univers	1.95 1.67	lecture	1.79
F12L	soeur mois	4.05 3.25	femmes mètres	3.70 3.2¢	parité	3.26

F12R	voiture	5.01	filles	3.80	fille	2.93
	stage	2.39	Roubaix	2.37		
F13L	Marion	7.25	filles	5.63	France	4.07
	Paris	3.96	fille	3.96		
F13R	filles	3.00	mignon	2.77	père	2.47
	Marion	2.22	ans	2.08		
F14L	film	3.64	limite	2.74	fin	2.58
	journée	2.53	nuit	2.40		
F14R	métro	4.65	Harry	4.04	mère	3.68
	école 3.63	film	3.45			
F16R	famille	2.22	copines	2.03	Laura	1.97
	soir	1.83	feuille	1.24		
M01L	week-end	3.67	relou	2.97	matin	2.70
	Paris	2.63	Marie	2.25		
M01R	Sarkozy	3.90	France	3.22	euros	2.57
	Virginie	2.45	anniversaire	2.16		
M02L	France	5.89	Sarkozy	3.79	Jeanne	3.51
	Ségolène	3.50	fille	3.40		
M02R	Sarkozy	3.50	homme	3.33	France	2.95
	différence	2.93	maison	2.88		
M09L	intérêt	2.22	étudiants	2.18	France	2.11
	Maine	2.02	garçons	1.97		
M09R	France	4.43	dépénalisation	2.48	Toulouse	2.23
	garçon	2.22	match	2.20		
M11L	putain	5.92	école	4.06	semaine	4.04
	fin	3.33	euros	2.71		
M11R	imagine	4.41	regarde	4.30	Sarkozy	3.53
	soeur	3.31	métro	3.26		
M15L	anthropologie	5.27	question	3.31	sociologie	3.07
	étude	2.57	amie	2.01		
M15R	Julie	4.42	voix	3.90	question	3.50
	société	3.39	système	3.36		
M17L	femmes	3.3£	balle	3.07	début	2.84
	Ségolène	2.55	parents	2.46		
M17R	euros	4.46	éducation	3.74	fou	3.44
	soirée	3.07	femme	2.97		
M18L	filles	4.18	euros	4.01	putain	3.81
	niveau	3.49	regarde	3.13		
M18R	putain	6.34	crème	4.35	euros	4.13
	con	3.86	question	3.61		
M19L	théâtre	4.17	lettres	2.43	université	2.22
	France	2.03	nombre	1.98		
M19R	lettres	5.30	France	4.89	université	4.52
	monde	3.32	Nobel	2.4		
M20L	problème	3.99	BTS	2.77	école	2.60
	Gatinot	2.55	Paris	2.52		
M20R	Paris	5.99	envie	4.51	paysage	2.63

	France	2.63	université	2.49		
M21L	filles	4.62	monde	4.08	femme	3.52
	copine	2.87	vie	2.65		
M21R	filles	2.62	lu	2.03	lycée	2.03
	films	2.01	besoin	1.99		
M22L	salle	3.51	vie	2.59	France	2.25
	espèce	2.25	art	1.95		
M23L	politique	4.61	problème	3.85	UMP	3.63
	gauche	3.19	travail	3.16		
M23R	supporters	8.11	club	6.30	problème	5.77
	gauche	4.55	Paris	4.50		

Table C.1: Five keywords for each of the 44 NCCFr speakers with relative TF-IDF values.

Appendix D

Reference table for the perception task

ID	Sex	Age	Mother tongue(s)	Expert	Sessions duration		Sessions interval
01	W	26	French	0	00:26	00:27	> a week
02	W	26	French	0	00:37	00:26	> a week
03	W	32	Chinese	1	01:07	00:57	< a week
04	W	27	French	0	00:22	00:29	Same day
05	W	26	French	0	00:39	00:30	> a week
06	W	30	French	0	00:43	00:21	> a week
07	M	77	French	0	00:39	00:38	Same day
08	W	26	French	0	00:26	00:29	> a week
09	W	48	French	1	00:13	00:13	> a week
10	M	29	French	0	00:33	00:20	> a week
11	W	25	Arab	0	00:42	00:20	> a week
12	W	50	French	1	01:22	00:39	Same day
13	M	32	Chinese	1	00:39	00:28	> a week
14	W	26	French	1	00:55	00:50	< a week
15	W	21	French	1	01:48	01:20	> a week
16	M	22	French	0	00:23	00:17	> a week
18	W	24	French	1	00:16	00:32	< a week
20	W	27	French	0	00:16	00:13	Same day
21	W	28	French	1	00:19	00:22	> a week
22	M	32	French	0	00:35	00:40	> a week
25	W	35	French	0	00:22	00:29	< a week
27	W	27	French	1	00:58	00:49	Same day
28	W	23	French	1	00:28	00:22	> a week
29	W	30	French	0	00:36	01:59	> a week
30	W	23	French	1	00:34	00:20	> a week
31	W	27	French	1	00:27	00:24	Same day
32	M	28	Chinese	1	00:30	00:09	Same day

Table D.1: Reference table for the perception task anonymously reporting all information gathered from retained listeners. All used as control variables during the analysis described in Chapter 5.

Appendix E

Jaccard similarity rates for PHON-CNN-HUM clusters comparisons

Female speakers task

	Spectros	MFCC	Glob	Form	F0	Acoust	Hadiff	Rhythm	Amp	Ms	Nrg	Hr	Ltas	Qual
All	- ; 0.46	0.22; 0.45	0.37; 0.41	**0.48**; 0.44	0.35; 0.39	0.39; 0.41	0.32; 0.44	0.35; 0.41	0.34; 0.45	0.38; 0.48	0.36; 0.48	0.36; 0.48	0.23; 0.41	0.28; 0.43
Experts	- ; 0.46	0.21; 0.53	0.34; 0.42	**0.46**; 0.55	0.34; 0.35	0.45; 0.48	0.25; 0.52	0.38; 0.48	0.36; 0.47	0.35; **0.55**	0.34; 0.50	0.34; 0.53	0.20; 0.46	0.32; 0.49
Experts non native	- ; 0.21	0.24; **0.18**	0.28; 0.22	0.27; **0.18**	0.25; 0.31	**0.17**; 0.22	0.24; 0.20	**0.18**; 0.22	0.27; 0.24	0.27; 0.23	0.28; 0.23	0.23; 0.21	0.31; **0.18**	0.23; 0.20
Naives	- ; 0.46	0.20; 0.52	0.36; 0.42	0.44; 0.47	0.35; 0.37	0.42; 0.49	0.29; 0.43	0.43; 0.44	0.33; 0.48	0.34; 0.47	0.33; 0.42	0.31; **0.54**	0.19; 0.43	0.29; 0.49
Naives native	- ; **0.54**	0.20; 0.50	0.41; 0.43	**0.46**; 0.49	0.31; 0.34	0.45; 0.43	0.29; 0.45	0.38; 0.46	0.38; 0.42	0.34; 0.48	0.33; 0.45	0.34; 0.52	**0.17**; 0.45	0.29; 0.48
Natives	- ; 0.52	0.22; 0.49	0.37; 0.44	**0.46**; 0.47	0.33; 0.40	0.44; 0.44	0.33; 0.48	0.41; 0.51	0.35; 0.45	0.39; 0.46	0.39; 0.51	0.39; **0.54**	0.21; 0.44	0.30; 0.49
Non natives	- ; 0.20	0.23; **0.18**	0.27; 0.22	0.27; **0.19**	0.25; 0.31	**0.17**; 0.22	0.24; 0.20	**0.18**; 0.22	0.27; 0.24	0.27; 0.22	0.28; 0.23	0.23; 0.21	0.31; **0.18**	0.23; 0.20

Male speakers task

	Spectros	MFCC	Glob	Form	F0	Acoust	Hadiff	Rhythm	Amp	Ms	Nrg	Hr	Ltas	Qual
All	- ; 0.43	0.40; 0.40	0.33; 0.40	0.35; 0.43	0.23; 0.38	0.42; 0.44	0.31; **0.48**	0.42; 0.41	0.43; 0.41	0.31; 0.41	0.41; 0.43	0.37; 0.45	0.38; 0.42	0.38; **0.48**
Experts	- ; 0.42	0.38; 0.38	0.36; 0.41	0.33; 0.40	0.27; 0.41	0.40; 0.42	0.28; **0.48**	0.42; 0.41	0.38; 0.42	0.36; 0.42	0.41; 0.43	0.39; 0.43	0.38; 0.44	0.38; **0.46**
Experts non native	- ; **0.18**	**0.18**; **0.18**	0.21; 0.25	0.23; 0.20	0.30; 0.28	**0.18**; **0.18**	**0.16**; **0.18**	0.20; 0.21	0.20; 0.20	0.30; 0.20	0.22; 0.20	0.21; **0.18**	0.20; **0.17**	0.20; 0.20
Naives	- ; 0.40	0.32; 0.38	0.34; 0.43	0.38; 0.42	0.29; 0.43	0.39; 0.41	0.30; 0.45	**0.45**; 0.41	0.36; 0.41	0.37; 0.43	0.39; 0.44	0.35; 0.43	0.38; 0.41	0.38; 0.44
Naives native	- ; 0.39	0.33; 0.37	0.33; 0.41	0.38; 0.43	0.30; 0.41	0.39; 0.42	0.32; 0.44	0.44; 0.42	0.37; 0.40	0.36; 0.41	0.39; 0.44	0.34; 0.40	0.38; 0.40	0.38; 0.45
Natives	- ; 0.39	0.36; 0.38	0.34; 0.42	0.37; 0.45	0.28; 0.40	0.41; 0.42	0.33; 0.42	0.43; 0.43	0.30; 0.40	0.35; 0.41	0.36; 0.41	0.34; 0.41	0.35; 0.39	0.35; **0.48**
Non natives	- ; 0.42	0.32; 0.41	0.35; 0.41	0.36; 0.44	0.25; 0.44	0.40; **0.48**	0.30; 0.42	**0.47**; 0.41	0.31; 0.41	0.35; 0.45	0.39; 0.44	0.39; **0.46**	0.42; 0.42	0.42; **0.48**

Table E.1: Jaccard similarity between CNN, first score, PHON, second score in cell, and HUM clusters (both sexes listeners), Top: female speakers task. Bottom: male speakers task.

Female speakers task

	Spectros	MFCC	Glob	Form	F0	Acoust	Hadiff	Rhythm	Amp	Ms	Nrg	Hr	Ltas	Qual
All F	- ; 0.49	0.22; 0.48	**0.45**; 0.45	0.44; 0.46	0.32; 0.36	0.43; 0.44	0.31; 0.42	0.39; 0.46	0.33; 0.40	0.35; 0.45	0.35; 0.46	0.36; 0.50	**0.18**; 0.41	0.27; 0.48
Experts F	- ; 0.43	0.23; 0.44	0.35; 0.45	0.43; 0.45	0.36; 0.39	0.41; 0.43	0.32; 0.46	0.34; 0.47	0.34; 0.41	0.37; 0.46	0.35; 0.52	0.33; 0.46	0.21; 0.40	0.28; 0.42
Experts native F	- ; 0.40	0.23; 0.41	0.37; 0.41	**0.44**; 0.41	0.33; 0.39	0.36; 0.39	0.32; 0.37	0.31; 0.39	0.35; 0.39	0.35; 0.47	0.33; 0.43	0.31; 0.44	0.21; 0.36	0.25; 0.39
Experts non native F	- ; **0.56**	0.23; **0.54**	0.35; 0.44	**0.45**; 0.50	0.33; 0.36	**0.45**; 0.48	0.27; 0.46	0.39; 0.47	0.33; 0.43	0.42; 0.49	0.32; 0.43	0.36; 0.53	**0.18**; 0.45	0.28; 0.51
Naives F	- ; 0.46	0.20; 0.52	0.36; 0.42	**0.44**; 0.47	0.35; 0.37	0.42; 0.49	0.29; 0.43	0.43; 0.44	0.33; 0.48	0.34; 0.47	0.33; 0.42	0.31; **0.54**	**0.19**; 0.43	0.29; 0.49
Naives native F	- ; 0.49	0.21; 0.47	0.35; 0.41	**0.44**; 0.47	0.36; **0.34**	0.42; 0.43	0.30; 0.43	0.36; 0.43	0.43; 0.42	0.35; 0.46	0.33; 0.42	0.31; 0.49	**0.19**; 0.44	0.31; 0.43
Naives non native F	- ; 0.42	0.29; 0.43	0.42; 0.45	0.43; 0.41	0.30; 0.35	0.40; 0.40	0.34; 0.43	0.37; 0.47	0.37; 0.39	0.40; 0.41	0.39; 0.43	**0.44**; 0.41	**0.19**; 0.44	0.29; 0.49
Natives F	- ; 0.51	0.20; 0.50	0.41; 0.43	**0.45**; 0.50	0.32; **0.33**	0.44; 0.43	0.28; 0.47	0.37; 0.48	0.38 ; 0.42	0.35; 0.49	0.35; 0.44	0.36; 0.52	**0.19**; 0.45	0.31; 0.48
Non natives F	- ; 0.48	0.25; 0.52	0.37; 0.43	0.41; 0.47	0.36; 0.37	0.42; 0.51	0.29; 0.46	0.41; 0.41	0.33; 0.45	0.36; 0.48	0.34; 0.43	0.32; 0.52	0.21; 0.44	0.32; **0.53**

Male speakers task

	Spectros	MFCC	Glob	Form	F0	Acoust	Hadiff	Rhythm	Amp	Ms	Nrg	Hr	Ltas	Qual
All F	- ; 0.38	0.25; 0.43	0.30; 0.36	**0.38**; 0.43	0.36; 0.35	0.36; 0.40	0.21; 0.44	0.36; 0.40	0.28; 0.43	0.34; 0.45	0.32; 0.41	0.28; 0.38	0.26; 0.41	**0.38**; 0.43
Experts F	- ; 0.35	0.20; 0.33	0.35; 0.37	0.34; 0.33	0.29; 0.36	0.32; 0.35	0.25; 0.37	0.33; 0.39	0.34; 0.35	0.36; 0.38	0.34; 0.36	0.30; 0.34	0.23; 0.33	0.28; 0.33
Experts native F	- ; 0.39	0.23; 0.40	0.35; 0.39	0.36; 0.38	0.32; 0.37	0.37; 0.37	0.23; **0.49**	0.36; **0.47**	0.34; 0.39	0.37; 0.44	0.36; 0.44	0.32; 0.39	0.24; 0.44	0.33; 0.39
Experts non native F	- ; 0.37	0.24; 0.38	0.36; 0.41	**0.38**; 0.37	0.30; 0.36	**0.40**; 0.40	0.27; 0.37	0.36; 0.38	0.33; 0.39	0.35; 0.39	0.36; 0.41	0.32; 0.39	0.20; 0.38	0.28; 0.41
Naives F	- ; 0.38	0.22; 0.39	0.31; 0.37	0.35; 0.37	0.29; 0.35	0.37; 0.36	0.24; 0.45	0.40; 0.42	0.34; 0.36	0.35; 0.40	0.34; 0.39	0.37; 0.37	0.23; 0.43	0.33; 0.37
Naives native F	- ; 0.38	0.22; 0.39	0.31; 0.37	0.35; 0.37	0.29; 0.35	0.37; 0.36	0.24; 0.45	0.40; 0.42	0.34; 0.36	0.35; 0.40	0.34; 0.39	0.37; 0.37	0.23; 0.43	0.33; 0.37
Naives non native F	- ; 0.15	**0.15**; 0.13	**0.15**; 0.13	**0.15**; 0.13	0.31; 0.18	0.17; 0.14	0.31; 0.13	**0.14**; **0.12**	0.18; 0.14	0.16; 0.14	**0.15**; 0.14	0.13; 0.15	0.23; **0.11**	**0.14**; 0.13
Natives F	- ; 0.38	0.27; 0.36	0.34; 0.40	0.36; 0.36	0.31; 0.39	0.36; 0.36	0.26; 0.46	0.32; 0.47	0.33; 0.39	0.37; 0.39	**0.38**; 0.46	0.33; 0.35	0.28; 0.41	0.33; 0.37
Non natives F	- ; 0.29	0.21; 0.27	0.32; 0.31	0.31; 0.29	0.31; 0.37	0.26; 0.31	0.27; 0.31	0.29; 0.31	0.33; 0.31	0.34; 0.32	0.35; 0.32	0.26; 0.29	0.29; 0.29	0.27; 0.29

Table E.2: Jaccard similarity between CNN, first score, PHON, second score in cell, and HUM clusters (female listeners), Top: female speakers task. Bottom: male speakers task.

Female speakers task

	Spectros	MFCC	Glob	Form	F0	Acoust	Hadiff	Rhythm	Amp	Ms	Nrg	Hr	Ltas	Qual
All M	- ; 0.45	0.38; 0.40	0.36; 0.39	0.35; 0.41	0.21; 0.37	**0.41**; **0.46**	0.34; **0.49**	0.39; 0.37	0.35; 0.44	0.32; 0.40	0.40; 0.42	0.36; 0.44	0.33; 0.40	0.33; 0.42
(Experts) non native M	- ; 0.19	**0.17**; **0.18**	0.21; 0.27	0.23; 0.20	0.29; 0.26	0.18; **0.18**	0.16; 0.18	0.19; 0.21	0.21; 0.19	0.29; 0.20	0.22; 0.19	0.21; 0.19	0.19; **0.17**	0.19; 0.20
Naives M	- ; 0.44	0.33; 0.36	0.34; 0.38	0.36; 0.40	0.23; 0.38	0.39; 0.37	0.26; **0.46**	**0.42**; 0.35	0.35; 0.39	0.32; 0.37	0.38; 0.42	0.37; 0.40	0.33; 0.35	0.33; 0.41

Male speakers task

	Spectros	MFCC	Glob	Form	F0	Acoust	Hadiff	Rhythm	Amp	Ms	Nrg	Hr	Ltas	Qual
All M	- ; 0.44	0.36; 0.40	0.34; 0.40	0.36; 0.46	**0.20**; **0.36**	0.41; 0.45	0.35; **0.49**	0.40; 0.42	**0.42**; 0.41	0.30; 0.43	0.41; 0.43	0.32; 0.46	0.39; 0.43	0.39; 0.44
(Experts) non native M	- ; 0.40	0.31; 0.41	0.35; 0.42	0.40; 0.40	0.25; 0.42	0.41; 0.45	0.28; 0.41	0.41; 0.42	0.33; 0.43	0.36; 0.45	0.39; 0.42	0.41; 0.44	0.38; 0.42	0.38; 0.44
Naives M	- ; 0.44	0.41; 0.40	0.36; 0.44	0.35; 0.43	**0.23**; 0.41	0.42; 0.45	0.31; **0.52**	**0.42**; 0.44	0.36; 0.41	0.34; 0.42	0.40; 0.44	0.34; 0.46	0.38; 0.43	0.38; 0.44

Table E.3: Jaccard similarity between CNN, first score, PHON, second score in cell, and HUM clusters (male listeners), Top: female speakers task. Bottom: male speakers task.

Female speakers task

	MFCC	**Glob**	**Form**	**F0**	**Acoust**	**Hadiff**	**Rhythm**	**Amp**	**Ms**	**Nrg**	**Hr**	**Ltas**	**Qual**
Spectros	0.31	0.40	0.42	0.35	0.46	0.33	0.42	0.38	0.40	0.38	0.33	0.26	0.32
MFCC	0.25	0.37	0.45	0.32	0.49	0.29	0.45	0.36	0.36	0.33	0.35	**0.18**	0.31
Glob	0.27	0.40	0.43	0.29	0.44	0.34	0.35	0.34	0.36	0.38	0.37	**0.19**	0.27
Form	0.25	0.35	0.46	0.34	0.47	0.32	0.40	0.35	0.37	0.38	0.34	**0.20**	0.30
F0	0.22	0.34	0.34	0.28	0.30	0.28	0.41	0.41	0.38	0.40	0.30	0.29	0.36
Acoust	0.22	0.37	0.42	0.32	0.44	0.32	**0.47**	0.34	0.36	0.39	0.31	0.21	0.32
Hadiff	0.23	0.35	0.40	0.32	0.42	0.31	0.42	0.36	0.39	0.46	0.35	0.28	0.39
Rhythm	0.24	0.37	0.38	0.31	0.41	0.29	0.44	0.33	0.46	0.40	0.43	0.21	0.29
Amp	0.25	0.35	0.41	0.30	0.38	0.29	0.43	0.34	0.38	0.39	0.31	0.24	0.33
Ms	0.27	0.39	0.44	0.33	0.40	0.26	0.38	0.37	0.38	0.36	0.32	**0.20**	0.29
Nrg	0.28	0.39	0.44	0.31	0.46	0.32	0.38	0.37	0.38	0.41	0.39	0.22	0.32
Hr	0.23	0.34	0.41	0.31	0.42	0.31	0.45	0.35	0.37	0.39	0.35	**0.20**	0.31
Ltas	0.34	0.37	0.47	0.37	**0.52**	0.29	0.36	0.35	0.38	0.37	0.33	0.22	0.36
Qual	0.27	0.42	0.48	0.33	**0.50**	0.31	0.43	0.33	0.39	0.37	0.39	**0.20**	0.30

Male speakers task

	MFCC	**Glob**	**Form**	**F0**	**Acoust**	**Hadiff**	**Rhythm**	**Amp**	**Ms**	**Nrg**	**Hr**	**Ltas**	**Qual**
Spectros	0.32	0.36	0.34	**0.23**	0.42	0.32	0.39	0.33	0.34	0.45	0.42	0.41	0.41
MFCC	0.38	0.43	0.35	0.24	0.48	0.35	0.44	0.33	0.36	0.43	0.49	0.40	0.40
Glob	0.35	0.39	0.39	0.25	0.43	0.30	0.40	0.35	0.38	0.47	0.44	0.37	0.37
Form	0.32	0.40	**0.55**	**0.22**	**0.55**	0.32	0.43	0.31	0.35	0.42	0.37	0.42	0.42
F0	0.30	0.33	0.37	0.27	0.40	0.24	0.41	0.31	0.38	0.39	0.43	0.38	0.38
Acoust	0.36	0.40	0.34	**0.23**	0.46	0.36	0.43	0.34	0.35	0.46	0.39	0.42	0.42
Hadiff	0.39	0.38	0.36	0.24	0.44	0.35	0.47	0.41	0.34	0.44	0.38	0.40	0.40
Rhythm	0.37	0.40	0.39	0.37	0.44	0.37	0.44	0.31	0.34	0.38	0.39	0.36	0.36
Amp	0.41	0.40	0.41	0.24	0.54	0.33	0.45	0.35	0.35	0.41	0.40	0.37	0.37
Ms	0.35	0.40	0.35	0.28	0.46	0.36	0.47	0.36	0.37	0.42	0.38	0.40	0.41
Nrg	0.37	0.40	0.39	0.28	0.44	0.38	0.50	0.37	0.38	0.47	0.39	0.40	0.40
Hr	0.35	0.36	0.39	0.24	0.46	0.34	0.47	0.39	0.34	0.44	0.46	0.41	0.41
Ltas	0.33	0.37	0.36	0.27	**0.50**	0.35	0.42	0.38	0.35	0.48	0.38	0.44	0.44
Qual	0.35	0.39	0.38	**0.23**	0.46	0.34	0.46	0.36	0.35	0.43	0.42	0.39	0.39

Table E.4: Jaccard similarity coefficient between CNN and PHON clusters, Top: female speakers task. Bottom: male speakers task.

Bibliography

Adami, A. and Hermansky, H. (2003). Segmentation of speech for speaker and language recognition. *Proceedings of EUROSPEECH*.

Adami, A. G., Mihaescu, R., Reynolds, D. A., and Godfrey, J. J. (2003). Modeling prosodic dynamics for speaker recognition. *Proceedings of ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, pages 788–791.

Adank, P., Evans, B., Stuart-Smith, J., and Scotti, S. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 35.

Aitken, C. G. G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*.

Ajili, M., Bonastre, J.-F., Kheder, W. B., Rossato, S., and Kahn, J. (2016). Comparaison des voix dans le cadre judiciaire : influence du contenu phonétique. *JEP (Journées d'Etudes sur la Parole)*.

Andreeva, B., Barry, W., and Steiner, I. (2007). Producing phrasal prominence in german. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Ardoint, M. and Lorenzi, C. (2009). Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. *Hearing Research*.

Arvaniti, A. (2013). The role of rhythm class, speaking rate and f0 in language discrimination. *Laboratory Phonology*.

Asadi, H., Nourbakhsh, M., Sasani, F., and Dellwo, V. (2018). Examining long-term formant frequency as a forensic cue for speaker identification: An experiment on persian. *International Conference on Laboratory Phonetics and Phonology*.

Asu, E. and Nolan, F. (2005). Estonian rhythm and the pairwise variability index. *Proceedings of FONETIKr*, pages 29–32.

Aubanel, V., Cooke, M., Davis, C., and Kim, J. (2018). Temporal factors in cochlea-scaled entropy and intensity-based intelligibility predictions. *Journal of Acoustical Society of America*, 143:443–448.

Audibert, N., Fougeron, C., Gendrot, C., and Adda-Decker, M. (2015). Duration- vs. style-dependent vowel variation: a multiparametric investigation. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Ball, M., Esling, J., and Dickson, B. (2018). Revisions to the voqs system for the transcription of voice quality. *Journal of the International Phonetic Association*, 48:165–171.

Ball, M., Esling, J., and Dickson, C. (1995). The voqs system for the transcription of voice quality. *Journal of the International Phonetic Association*.

Barlow, M. G. and Wagner, M. (1988). Prosody as a basis for determining speaker characteristics. *Proceedings of the Second Australian International Conference on Speech Science and Technology*.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33:5–22.

Barry, W., Andreeva, B., Dimitrova, M. R. S., and Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 2693–2696.

Barry, W. and Russo, M. (2003). Measuring rhythm: Is it separable from speech rate? *AAI Workshop, Prosodic Interfaces*, pages 15–20.

Baumann, O. and Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, 74:110–120.

Bee, M. and Gerhardt, H. (2002). Individual voice recognition in a territorial frog (rana catesbeiana). *Proceedings of the Royal Society of London B: Biological Science*, 269.

Bele, I. V. (2006). The speaker's formant. *Journal of Voice*, 20.

Belin, P. and Grosbras, M. (2010). Before speech: cerebral voice processing in infants. *Neuron*, 65.

Belkin, K., Martin, R., Kemp, S., and Gilbert, A. (1997). Auditory pitch as a perceptual analogue to odor quality. *Psychological science*, 8.

Ben-David, S., Hrubes, P., Moran, S., Shpilka, A., and Yehudayoff, A. (2019). Learnability can be undecidable. *Nature Machine Intelligence*.

Bernardoni, N. H., d'Alessandro, C., Doval, B., and Castellengo, M. (2005). Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *Journal of the Acoustical Society of America*, 117:1417–1430.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska.Delacrétaz, D., and Reynolds, D. (2004). A tutorial on text-independent speaker verification. *Applied Signal Processing*, 4:430–451.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5:9/10:341–345.

Bolt, R., Cooper, F., David, E., Denes, P., Pickett, J., and Stevens, K. (1970). Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes. *Journal of the Acoustical Society of America*, 47:597–612.

Bonastre, J.-F., Bimbot, F., Boë, L.-J., Campbell, J., Reynolds, D., and Magrin-Chagnolleau, I. (2003). Person authentication by voice: A need for caution. *Proceedings of EUROSPEECH*.

Bonastre, J.-F., Kahn, J., Rossato, S., and Ajili, M. (2015a). Forensic speaker recognition: Mirages and reality. *Individual differences in speech production and perception*.

Bonastre, J.-F., Kahn, J., Rossato, S., and Ajili, M. (2015b). Forensic speaker recognition: Mirages and reality. *Individual differences in speech production and perception*.

Bousquet, P.-M., Larcher, A., Matrouf, D., Bonastre, J.-F., and Plchot, O. (2012). Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. *Proceedings of Odyssey*.

Boë, L.-J. (2000). Forensic voice identification in france. *Speech Communication*, 31:205–224.

Boë, L.-J., Bimbot, F., Bonastre, J.-F., and Dupont, P. (1999). Des évaluations des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. *Cahiers d'Etudes et de Recherches Francophones. Langues*, 2:270–288.

Boë, L.-J. and Bonastre, J.-F. (2012). L'identification du locuteur : 20 ans de témoignage dans les cours de justice le cas du lipsadon « laboratoire indépendant de police scientifique ». *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, pages 417–424.

Boë, L.-J., Contini, M., and Rakotofiringa, H. (1975). Etude statistique de la fréquence laryngienne. *Phonetica*, 32:1–23.

Brand, J., Hay, J., Clark, L., Watson, K., and Sóskuthy, M. (2019). Systematic covariation of monophthongs across speakers of new zealand english. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Braun, A. and Rosin, A. (2015). On the speaker-specificity of hesitation markers. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Bronkhorst, A. (2000). The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*.

Brummer, N. and de Villiers, E. (2013). The bosaris toolkit: Theory, algorithms and code for surviving the new dcf. *arXiv e-prints*.

Brunner, J., Perrier, P., and Fuchs, S. (2006). Influence de la forme du palais sur la variabilité articulatoire. *JEP (Journées d'Etudes sur la Parole)*.

Burke, E. and Murphy, C. (2007). How female barking tree frogs, hyla gratiosa, use multiple call characteristics to select a mate. *Animal Behaviour*, 74.

Butler, J., Vigário, M., and Frota, S. (2016). Infants' perception of the intonation of broad and narrow focus. *Language Learning and Development*, 12:1–13.

Byrne, D. and al. (1996). An international comparison of long-term average speech spectra. *Journal of Acoustical Society of America*, 96:2108–2120.

Calliope (1989). La parole et son traitement automatique. *Masson*.

Cambier-Langeveld, T., van Rossum, M., and Vermeulen, J. (2014). Who's voice is that? challenges in forensic phonetics. *Above and Beyond the Segments. Experimental Linguistics and Phonetics, Amsterdam: John Benjamins Publishing Company*, pages 14–27.

Cangemi, F. (2009). Phonetic detail in intonation contour dynamics. *AISV (Associazione Italiana di Scienze della Voce)*.

Cangemi, F., Krüger, M., and Grice, M. (2015). Listener-specific perception of speaker-specific productions in intonation. *Individual differences in speech production and perception*.

Chanclu, A., Amor, I. B., Gendrot, C., Ferragne, E., and Bonastre, J.-F. (2021). Automatic classification of phonation types in spontaneous speech: Towards a new workflow for the characterization of speakers' voice quality. *Proceedings of INTERSPEECH*.

Chanclu, A., Georgeton, L., Fredouille, C., and Bonastre, J.-F. (2020). Ptsvox : une base de données pour la comparaison de voix dans le cadre judiciaire (ptsvox : a speech database for forensic voice comparison). *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*.

Chardenon, E., Fougeron, C., Audibert, N., and Gendrot, C. (2020). Dis-moi comment tu varies ton débit, je te dirai qui tu es. *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*.

Chen, K. and Salman, A. (2011). Learning speaker-specific characteristics with a deep neural architecture. *IEEE Transactions on Neural Networks*, 22.

Chhabra, S., Badcock, J., Maybery, M., and Leung, D. (2012). Voice identity discrimination in schizophrenia. *Neuropsychologia*, 50:2730–2735.

Chicco, D., Tötsch, N., and Jurman, G. (2021). The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*.

Chignoli, G. (2018). *Reconnaissance Automatique du Locuteur, mécanisme humain et tache informatique : application de méthodes statistiques*. PhD thesis, Université Sorbonne Nouvelle.

Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant vot in american english. *Journal of Phonetics*, 61.

Chollet, F. et al. (2015). Keras. https://keras.io.

Christophe, A., Mehler, J., and Sebastián-Gallés, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy*, 2:385–394.

Clark, J., Adams, S., Dykstra, A., Moodie, S., and Jog, M. (2014). Loudness perception and speech intensity control in parkinson's disease. *Journal of Communication Disorders*, 51:1–12.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: a survey. *arXiv e-prints*.

Culling, J. and Darwin, C. (1993). The role of timbre in the segregation of simultaneous voices with intersecting f0 contours. *Perception Psychophysics*, 54:303–309.

Cutler, A., Eisner, F., McQueen, J., and Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory Phonology*, 10.

de Jong, G., McDougall, K., Hudson, T., and Nolan, F. (2007). The speaker discriminating power of sounds undergoing historical change: a formant-based study. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1813–1816.

Dediu, D., Janssen, R., and Moisik, S. (2017). Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language. *Language Communication*, 54:9–20.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., and Dumouchel, P. (2009). Support vector machine versus fast scoring in the low-dimensional total variability space for speaker verification. *Proceedings of INTERSPEECH*.

Dehak, N., Kenny, P., Dumouchel, P., Dehak, R., and Ouellet, P. (2011). Front-end factor analysis for speaker verification » in ieee transactions on audio. *Speech and Language Processing*.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for dc. *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium*, pages 231–241.

Dellwo, V. (2010). *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence*. PhD thesis, Rheinischen Friedrich-Wilhelms-Universität zu Bonn.

Dellwo, V., Kolly, M.-J., and Leemann, A. (2012). Speaker idiosyncratic rhythmic features in the speech signal. *Proceedings of INTERSPEECH*, pages 1584–1587.

Dellwo, V., Leemann, A., and Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *Journal of the Acoustical Society of America*, 137:1513–1528.

Deterding, D. (2001). The measurement of rhythm: A comparison of singapore and british english. *Journal of Phonetics*, 29:217–230.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. (1998). Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. *Proceedings of ICSLP (International Conference on Speech and Language Processing)*.

Dou, Z. and Zhang, Z. (2018). Hierarchical attention: what really counts in various nlp tasks. *arXiv eprints*.

Dumpala, S., Panda, A., and Kopparapu, S. (2017). Improved i-vector-based speaker recognition for utterances with speaker generated non-speech sounds. *arXiv eprints*.

Eichhorn, J., Kent, R., Austin, D., and Vorperian, H. (2018). Effects of aging on focal fundamental frequency and vowel formants in men and women. *Journal of Voice*, 32.

Eisner, F. (2015). Perceptual adjustments to speaker variation. *Individual differences in speech production and perception*.

Eiswirth, M. (2020). Increasing interactional accountability in the quantitative analysis of sociolinguistic variation. *Journal of Pragmatics*, 170.

el Gamal, E. (2015). *Speaker identification based on temporal parameters*. PhD thesis, Phonetics and Linguistics Department of the University of Alexandria.

Elie, J. and Theunissen, F. (2018). Zebra finches identify individuals using vocal signatures unique to each call type. *Nature Communications*, 9.

Elliott, T. and Theunissen, F. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*.

Eriksson, A. (2010). The disguised voice: Imitating accents or speech styles and impersonating individuals. *Language and identities, Chapter: 8, Edinburgh University Press*, pages 86–96.

Eriksson, A. and Wretling, P. (1997). How flexible is human voice? - a case study of mimicry. *Proceedings of EUROSPEECH*.

Esling, J. (1994). Voice quality. *The Encyclopedia of Language and Linguistics*.

Farrús, M., Hernando, J., and Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. *Proceedings of INTERSPEECH*.

Ferragne, E., Gendrot, C., and Pellegrini, T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Ferragne, E. and Pellegrino, F. (2004). A comparative account of the suprasegmental and rhythmic features of british english dialects. *Actes de Modelisations pour l'Identification des Langues*, pages 121–126.

Ferrer, L. and McLaren, M. (2018). A generalization of plda for joint modeling of speaker identity and multiple nuisance conditions. *Proceedings of INTERSPEECH*.

Ferrer, L., Shriberg, E., Kajarekar, S., and Sönmez, K. (2007). Parametrization of prosodic feature distributions for svm modeling in speaker recognition. *Proceedings of ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, 4.

Fontaine, B., Steinberg, L., and Peña, J. (2013). Sound envelope extraction in cochlear nucleus neurons: modulation filterbank and cellular mechanism. *Twenty Second Annual Computational Neuroscience Meeting*.

Fraile, R. and Godino-Llorente, J. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, 14:42–54.

Fredouille, C., Pouchoulin, G., Bonastre, J.-F., Azzarello, M., Giovanni, A., and Ghio, A. (2005). Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia). *Proceedings of INTERSPEECH*.

French, J., Foulkes, P., Harrison, P., Hughes, V., and Stevens, L. (2015a). The vocal tract as a biometric: output measures, interrelationships and efficacy. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

French, P., Foulkes, P., Harrison, P., Hughes, V., san Segundo, E., and Stevens, L. (2015b). The vocal tract as a biometric: output measures, interrelationships, and efficacy. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Freydina, E. (2015). Prosodic variation in academic public presentations. *Journal of Language and Education*, 1.

Fuchs, S. and Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? *Turbulent Sounds: An Interdisciplinary Guide*.

Fuchs, S., Winkler, R., and Perrier, P. (2008). Do speakers' vocal tract geometries shape their articulatory vowel space? *Proceedings of ISSP (International Seminar on Speech Production)*.

Furuyama, T., Kobayasi, K., and Riquimaroux, H. (2016). Role of vocal tract characteristics in individual discrimination by japanese macaques (macaca fuscata). *Nature Scientific reports*, 6.

Garcia-Romero, D., Snyder, D., Sell, G., McCree, A., Povey, D., and Khudanpur, S. (2019). X-vector dnn refinement with full-length recordings for speaker recognition. *Proceedings of INTERSPEECH*.

Gelfer, M. P. and Bennett, Q. E. (2013). Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender. *Journal of Voice*, 27:556–566.

Gendrot, C. and Adda-Decker, M. (2005). Impact of duration on f1/f2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in french and german. *Proceedings of INTERSPEECH*.

Gendrot, C., Adda-Decker, M., and Schmid, C. (2012). Comparaison de parole journalistique et de parole spontanée : analyses de séquences entre pauses. *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*.

Gendrot, C., Chignoli, G., Audibert, N., and Fougeron, C. (2018). Variabilité inter et intra locuteurs de mesures spectrales et prosodiques en parole lue. *JEP (Journées d'Etudes sur la Parole)*.

Gendrot, C., Ferragne, E., and Pellegrini, T. (2019). Deep learning and voice comparison: phonetically-motivated vs. automatically-learned features. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Gendrot, C., Ferragne, E., and Pellegrini, T. (2020). Information segmentale pour la caractérisation phonétique du locuteur: variabilité inter- et intra- locuteurs. *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le*

Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues).

Georgea, K., Kumara, C., Sivadasb, S., Ramachandrana, K., and Panda, A. (2018). Analysis of cosine distance features for speaker verification. *Pattern Recognition Letters*, 112.

Georgeton, L., Paillereau, N., Landron, S., and Kamiyama, T. (2012). Analyse form antique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants du fle. *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, pages 145–152.

Gerlacha, L., McDougall, K., Kelly, F., Alexander, A., and Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Communication*, 124.

Gick, B., Wilson, I., and Derrick, D. (2013). Articulatory phonetics. *Journal of the International Phonetic Association*.

Gold, E. (2014). *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. PhD thesis, University of York.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Grabe, E. and Low, E. (2002). Durational variability in speech and the rhythm class hypothesis. *Laboratory Phonology*, 7:515–546.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv e-prints*.

Greenberg, C., Martin, A., Branschain, L., Campbell, J., Cieri, C., Doddington, G., and Godfrey, J. (2010). Human assisted speaker recognition in nist sre10. *Proceedings of Odyssey*.

Gresse, A., Quillot, M., Dufour, R., Labatut, V., and Bonastre, J.-F. (2019). Similarity metric based on siamese neural networks for voice casting. *Proceedings of ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*.

Grossmann, T., Oberecker, R., Koch, S., and Friederici (2010). The developmental origins of voice processing in the human brain. *Neuron*, 65.

Han, J., Kamber, M., and Pei, J. (2012). 10 - cluster analysis: Basic concepts and methods. *Data Mining (Third Edition), The Morgan Kaufmann Series in Data Management Systems*.

Hansen, J. and Bořil, H. (2018). On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks. *Speech Communication*, 101:94–108.

Harmegnies, B. (1992). Les sources de variation du spectre à long terme de parole: revue de la littérature. *Canadian Acoustics*, 20:9–35.

Hautamäki, R., Sahidullah, M., Hautamäki, V., and Kinnunen, T. (2017). Acoustical and perceptual study of voice disguise by age modification in speaker verification. *Speech Communication*, 95:1–15.

He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Deep residual learning for image recognition. *arXiv e-prints*.

He, L. and Dellwo, V. (2016). A praat-based algorithm to extract the amplitude envelope and temporal fine structure using the hilbert transform. *Proceedings of INTERSPEECH*, pages 530–534.

He, L. and Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *Journal of Acoustical Society of America*, 141:488–494.

He, L., Glavitsch, U., and Dellwo, V. (2015b). Comparisons of speaker recognition strenghts using suprasegmental duration and intesity variability: an artificiale neural networks approach. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

He, L., Zhang, Y., and Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. *Journal of Acoustical Society of America*, 145:209–214.

Henrich, N., d'Alessandro, C., Doval, B., and Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of Acoustical Society of America*, 115:1321–1332.

Hewlett, N. and Cohen, W. (1993). Voicing distinctions in electrolarynx speech. *Third Congress of the International Clinical Phonetics and Linguistics Association*.

Hillenbrand, J., Cleveland, R., and Erickson, R. (1994). Acoustic correlates of breathy voice quality. *Journal of Speech, Language and Hearing Research*, 37.

Hirano, M. (1981). Clinical examination of voice. *Vienna/New York: Springer-Verlag*.

Hollien, H., Majewski, W., and Doherty, E. (1982). Perceptual identification of voices under normal, stress and disguise speaker conditions. *Journal of Phonetics*, 10.

Honikman, B. (1964). Articulatory settings. *David Abercrombie, D.B. Fry, P.A.D. MacCarthy, N. Scottand, J.L.M. Trim (eds.) In Honour of Daniel Jones*, pages 73–84.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:259–366.

Hsu, A., Woolley, S., Fremouw, T., and Theunissen, F. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *The Journal of Neuroscience*, 24:9201–9211.

Hudson, T., de Jong, G., McDougall, K., Harrison, P., and Nolan, F. (2007). F0 statistics for 100 young male speakers of standard southern british english. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1809–1812.

Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., and Gully, A. (2019). Forensic voice comparison using long-term acoustic measures of laryngeal voice quality. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Hughes, V., Harrison, P., French, P. F. J., and san Segundo, C. K. E. (2017). Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of INTERSPEECH*, pages 3892–3896.

Höhle, B. (2009). Bootstrapping mechanisms in first language acquisition. *Linguistics*, 47:359–382.

Iseli, M., Shue, Y.-L., and Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *Journal of Acoustical Society of America*.

Jessen, M., Meir, G., and Solewicz, Y. (2019). Evaluation of nuance forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic$_e val_0$1). *SpeechCommunication*, 110.

Jialin, P. and Rose, P. (2012). Likelihood ratio-based forensic voice comparison with the cantonese diphthong /ei/ f-pattern. *Proceedings of SST (Austrasian Speech Science and Technology)*.

Jiang, C., Wang, D., Huang, J., Marcus, P., and Nießner, M. (2019). Convolutional neural network on non-uniform geometrical signals using euclidean spectral transformation. *Proceedings of ICLR (International Conference on Learning Representations)*.

Johnson, C. and Hollien, H. (1984). Speaker indentification [sic!] utilizing selected temporal speech features. *Journal of Phonetics*, 12:319–326.

Johnson, E. and Seidl, A. (2008). Clause segmentation by 6-month-old infants: a crosslinguistic perspective. *Infancy*, 13:440–455.

Jotz, G., Cervantes, O., Abrahão, M., Settanni, F. P., and de Angelis, E. C. (2002). Noise-to-harmonics ratio as anacoustic measure of voice disorders in boys. *Journal of Voice*, 16:28–31.

Jusczyk, P., Cutler, A., and Redanz, N. (1993). Infants' preference for the predominant stress patterns of english words. *Child Development*, 64:675–687.

Kahn, J. (2011). *Parole de locuteur : performance et confiance en identification biométrique vocale*. PhD thesis, ED 536.

Kahn, J., Audibert, N., Bonastre, J.-F., and Rossato, S. (2011). Inter and intra-speaker variability in french: an analysis of oral vowels and its implication for automatic speaker verification. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Kajarekar, S., Ferrer, L., Venkataraman, A., Sonmez, K., Shriberg, E., Stolcke, A., Bratt, H., and Gadde, R. R. (2003). Speaker recognition using prosodic and lexical features. *Proceedings of ASRU (Automatic Speech Recognition and Understanding)*.

Keating, P., Esposito, C., Garellek, M., Khan, S., and Kuang, J. (2011). Phonation contrasts across languages. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Keating, P. and Kreiman, J. (2016). Acoustic similarity among female voices. *The Journal of Acoustical Society of America*, 140.

Keating, P., Kreiman, J., and Alwan, A. (2019). A new speech database for within- and between-speaker variability. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Keating, P., Kreiman, J., and Vasselinova, N. (2017). Acoustic similarities among voices. part 2: Male speakers. *The Journal of Acoustical Society of America*, 142.

Keating, P. and Kuo, G. (2012). Comparison of speaking fundamental frequency in english and mandarin. *Journal of Acoustical Society of America*, 132:1050–1060.

Kelly, F., Alexander, A., Forth, O., Kent, S., Lindh, J., and Åkesson, J. (2016a). Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features. *Proceedings of INTERSPEECH*, pages 1567–1568.

Kelly, F., Alexander, A., Forth, O., Kent, S., Lindh, J., and Åkesson, J. (2016b). Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features. *Proceedings of INTERSPEECH*.

Khosravani, A., Glackin, C., Dugan, N., Chollet, G., and Cannings, N. (2016). The intelligent voice 2016 speaker recognition system. *arXiv eprints*.

King, S. L., Connor, R. C., Krützen, M., and Allen, S. J. (2021). Cooperation-based concept formation in male bottlenose dolphins. *Nature communications*, 12.

Kinnunen, T., Kilpeläinen, T., and Fränti, P. (2000). Comparison of clustering algorithms in speaker identification. *Proceedings of IASTED Internatioanl Conference on Signal Processing and Communications*.

Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52.

Kinoshita, Y. and Osanai, T. (2006). Within speaker variation in diphthongal dynamics: what can we compare? *Proceedings of the 11th Australian International Conference on Speech Science and Technology*, pages 112–117.

Klug, K., Kirchhübel, C., Foulkes, P., and French, P. (2019). Analysing breathy voice in forensic speaker comparison. using acoustics to confirm perception. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 795–799.

Kockmann, M., Burget, L., Glembek, O., Ferrer, L., and Cernocky, J. (2010). Prosodic speaker verification using subspace multinomial models with intersession compensation. *Proceedings of INTERSPEECH*.

Kolly, M.-J., de Mareüil, P. B., Leemann, A., and Dellwo, V. (2017). Listeners use temporal information to identify french- and english-accented speech. *Speech Communication*, 86:121–134.

Kolly, M.-J., Leemann, A., de Mareüil, P. B., and Dellwo, V. (2015). Speaker-idiosyncrasy in pausing behavior: evidence from a cross-linguistic study. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Kreiman, J. (1997). Listening to voices: theory and practice in voice perception research. *Johnson K. Mullenix J. Talker Variability in Speech Research. Academic Press*.

Kreiman, J. and Gerratt, B. (2012). Perceptual interaction of the harmonic source and noise in voice. *Journal of Acoustical Society of America*, 131.

Kreiman, J., Gerratt, B., Kempster, G., Erman, A., and Berke, G. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *American Speech-Language-Hearing Association*, 36:21–40.

Kreiman, J., Gerratt, B., and Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research*, 33.

Kreiman, J., Gerratt, B., Precoda, K., and Berke, G. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research*, 35.

Kreiman, J. and Shue, Y.-L. (2010). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *Journal of Acoustical Society of America*, 132:2625–2632.

Kreiman, J. and Stidtis, D. (2011). Voices and listeners: toward a model of voice perception. *Acoustics Today*, 7:7–15.

Krizhevsky, A., Sutskever, I., , and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*.

Künzel, H. (2013). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech Language and the Law*, 4:48–83.

Künzel, H., Gonzalez-Rodriguez, J., and Ortega-García, J. (2004). Effect of voice disguise on the performance of a forensic automatic speaker recognition system. *Proceedings of Odyssey*.

Künzel, H., Masthoff, H., and Köster, J. (1995). The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition. *Science and Justice*, 35:291–295.

Ladd, D., Silverman, K., Tolkmitt, F., Bergmann, G., and Scherer, K. (1985). Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect. *Journal of the Acoustical Society of America*, 78:435–444.

Lancia, L., Krasovitskiy, G., and Stuntebeck, F. (2017). Coordinative patterns underlying speech rhythm. *PAPE (Phonetics and Phonology in Europe)*.

Latinus, M. and Belin, P. (2011). Human voice perception. *Current Biology*, 21.

Lavan, N., Burston, L., and Garrido, L. (2018). How many voices did you hear? natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, 110.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. PhD thesis, Cambridge University Press.

Lavner, Y., Gath, I., and Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, 30.

LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361.

Lee, K., Kinnunen, T., Colibro, D., Vair, C., Nautsch, A., Sun, H., He, L., Liang, T., Wang, Q., Rouvier, M., Bousquet, P.-M., Das, R. K., Bailo, I. V., Liu, M., Deldago, H., Liu, X., Sahidullah, M., Cumani, S., Zhang, B., Okabe, K., Yamamoto, H., Tao, R., Li, H., Giménez, A. O., Wang, L., and Buera, L. (2020). I4u system description for nist 2020 sre cts challenge. *NIST20 SRE CTS*.

Lee, Y., Goldstein, L., Parrell, B., and Byrd, D. (2021). Who converges? variation reveals individual speaker adaptability. *Speech Communication*, 131.

Lee, Y., Keating, P., and Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of Acoustical Society of America*, 146:1569–1579.

Leemann, A. and Kolly, M.-J. (2015). Speaker-invariant suprasegmental temporal features in normal and disguised speech. *Speech Communication*, 75:97–122.

Leemann, A., Kolly, M.-J., and Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, 238:59–67.

Lehiste, I., Olive, J., and Streeter, L. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of Acoustical Society of America*, 60:1199–1202.

Levandowsky, M. and Winter, D. (1971). Distance between sets. *Nature*, 234.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistics inquiry*, 8.

Lindh, J. (2006). Preliminary descriptive f0-statistics for young male speakers. *Working Papers*, 52.

Lindh, J. (2009). Perception of voice similarity and the results of a voice line-up. *Proceedings of FONETIK*.

Lindh, J. and Eriksson, A. (2007). Robustness of long time measures of fundamental frequency. *Proceedings of INTERSPEECH*, pages 2025–2028.

Lindh, J. and Eriksson, A. (2010). Voice similarity - a comparison between judgements by human listeners and automatic voice comparison. *Computer Science*.

Liu, L. and Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition*, 133:385–394.

Long, Y., Yan, Z.-J., Soong, F. K., Dai, L., and Guo, W. (2011). Speaker characterization using spectral subband energy ratio based on harmonic plus noise model. *Acoustics, Speech, and Signal Processing*, 1988:4520–4523.

Loukina, A., Kochanski, G., Rosner, B., Keane, E., and Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of Acoustical Society of America*, 129:3258–3270.

Lukic, Y., Vogt, C., Dürr, O., and Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. *IEEE International workshop on Machine Learning for signal processing*.

Mao, H., Shi, Y., Y. Liu, L. W., Li, Y., and Long, Y. (2020). Short-time speaker verification with different speaking style utterances. *PLoS ONE*.

Mary, L. and Yegnanarayana, B. (2008). Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, 50:782–796.

Mathevon, N., Koralek, A., Weldele, M., Glickman, S., and Theunissen, F. (2010). What the hyena's laugh tells: Sex, age, dominance and individual signature in the giggling call of crocuta crocuta. *BMC Ecology*, 10.

Mathworks (2019). *MATLAB Deep Learning Toolbox R2019a*. Mathworks, Natick, MA, USA.

Matsumoto, H., Hiki, S., Sone, T., and Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics*, 21:428–436.

Mattock, K. and Burnham, D. (2006). Chinese and english infants' tone perception: evidence for perceptual reorganization. *Infancy*, 10:241–265.

Mattock, K., Molnar, M., Polka, L., and Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, 106:1367–1381.

McDougall, K. (2003). Individual differences in the formant dynamics of vowels at different levels of stress. *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1611–1614.

McDougall, K. (2005). *The role of formant dynamics in determining speaker identity*. PhD thesis, Unpublished PhD thesis, University of Cambridge.

McDougall, K. (2006). Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies. *International Journal of Speech*, 13:89–126.

McDougall, K. and Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u/ in british english. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1825–1828.

McDougall, K., Nolan, F., and Hudson, T. (2015). Telephone transmission and earwitnesses: performance on voice parades controlled for voice similarity. *Phonetica*, 72:257–272.

McLaren, M., Lei, Y., Scheffer, N., and Ferrer, L. (2014). Application of convolutional neural networks to speaker recognition in noisy conditions. *Proceedings of INTERSPEECH*.

Mendoza, E., Valencia, N., Mufioz, J., and Trujillo, H. (1996). Differences in voice quality between men and women: Use of the long-term average spectrum (ltas). *Journal of Voice*, 10:59–66.

Michel, F. and Jacqueline, L. (2000). L'analyse de conversation, de l'ethnomethodologie à la linguistique interactionnelle. *Histoire Épistémologie Langage*, 22.

Mol, C., Chen, A., Kager, R., and ter Haar, S. (2017). Prosody in birdsong: a review and perspective. *Neuroscience and biobehavioral reviews*, 81:167–180.

Moon, J. and Hong, S. H. (2014). What is temporal fine structure and why is it important? *Korean journal of audiology*, 18:1–7.

Moore, B. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology*, 9.

Morgan, J. (1996). *Language and cognitive processes*, chapter Prosody and the Roots of Parsing. Psychology Press.

Morgan, J. and Wong, R. (2012). Development of cell types and synaptic connections in the retina. *Webvision The Organisation of the Retina and Visual System*.

Morrison, G. (2010). Forensic voice comparison. *Expert Evidence, Thomson Reuters, Sydney*.

Morrison, G. and Enzinger, E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic$_e val_0$1)˘$introduction. SpeechCommunication, 85 : 119 - -126$.

Morrison, G., Zhang, C., and Rose, P. (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International*, 208:59–65.

Morrison, G. S. and Thompson, W. C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science Technology Law Review*, 18:326–434.

Moulton, R. and Jiang, Y. (2018). Maximally consistent sampling and the jaccard index of probability distributions. *Proceedings of ICDMW*.

Mouterde, S., Theunissen, F., Elie, J., Vignal, C., and Mathevon, N. (2014). Acoustic communication and sound degradation: how do the individual signatures of male and female zebra finch calls transmit over distance? *PLoS-ONE*, 9.

Moyse, E. (2014). Age estimation from faces and voices: a review. *Psychologica Belgica*, 54.

Mugitani, R., Pons, F., Fais, L., Dietrich, C., Werker, J., and Amano, S. (2009). Perception of vowel length by japanese- and english-learning infants. *Developmental Psychology*, 45:236–247.

Najnin, S. and Banerjee, B. (2019). Speech recognition using cepstral articulatory features. *Speech Communication*, 107.

Nayana, P., Dominic, M., and Abraham, T. (2017). Comparison of text independent speaker identification systems using gmm and i-vector methods. *Procedia Computer Science*, 115:47–54.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998a). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24:756–766.

Nazzi, T., Floccia, C., and Bertoncini, J. (1998b). Discrimination of pitch contours by neonates. *Infant Behaviour and Development*, 21:779–784.

Nazzi, T., Jusczyk, P., and Johnson, E. (2000). Language discrimination by english-learning 5-month-olds: effects of rhythm and familiarity. *Journal of Memory and Language*, 43:1–19.

Nazzi, T. and Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, 41:233–243.

Nebot, A. (2021). Prosodic modulation as a mark to express pragmatic values: The case of mitigation in spanish. *Journal of Pragmatics*, 181.

Nespor, M. and Vogel, I. (1986). Prosodic phonology. *Dordrecht: Foris Publications*.

Niebuhr, O. and Skarnitzl, R. (2019). Measuring a speaker's acoustic correlates of pitch - but which? a constrastive analysis based on perceived speaker charisma. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1774–1778.

Nilsson, M. (2006). *Entropy and Speech*. PhD thesis, School of Electrical Engineering KTH (Royal Institute of Technology), Stockholm.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge University Press.

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. *Proceedings of the conference Law and Language: Prospect and Retrospect*.

Nolan, F., McDougall, K., de Jong, G., and Hudson, T. (2006). A forensic phonetic study of 'dynamic' sources of variability in speech: The dyvis project. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, pages 13–18.

Nolan, F., McDougall, K., and Hudson, T. (2011). Some acoustic correlates of perceived (dis)similarity between same-accent voices. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1506–1509.

Nooteboom, S. and Quené, H. (2021). On the mental representations of speech sounds. *Workshop in honor of John Ohala*.

Ochi, K., Mori, K., Sakai, N., and Aoki-Ogura, J. (2015). Accuracy of articulation rate control with visual feedback in persons who do and do not stutter. *Procedia - Social and Behavioral Sciences*, 193:217–222.

Ohala, J. (1983). Cross-language use of pitch: an ethological view. *Phonetica*, 40:1–18.

Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. *Sound symbolism*, pages 325–347.

Ouyang, I. and Kaiser, E. (2015). Individual differences in the prosodic encoding of informativity. *Individual differences in speech production and perception*.

O'Brien, B., Ghio, A., Fredouille, C., Bonastre, J.-F., and Meunier, C. (2020). Discriminating speakers using perceptual clustering interface. *XVII AISV Conference*.

O'Brien, B., Meunier, C., and Ghio, A. (2021). Presentation matters: Evaluating speaker identification tasks. *Proceedings of INTERSPEECH*.

Palaz, D., Magimai.-Doss, M., and Collobert, R. (2015). Analysis of cnn-based speech recognition system using raw speech as input. *Proceedings of INTERSPEECH*.

Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of Acoustical Society of America*, pages 2382–2393.

Peiffer-Smadja, N. and Cohen, L. (2019). The cerebral bases of the bouba-kiki effect. *NeuroImage*, 186.

Perrier, P. and Winkler, R. (2015). Biomechanics of the orofacial motor system: Influence of speaker-specific characteristics on speech production. *Individual differences in speech production and perception*.

Pettirossi, A., Audibert, N., and Crevier-Buchman, L. (2017). Le spectre moyen à long-terme corrigé en f0 (pitch-corrected ltas) : une mesure robuste de la voix dysphonique sur des énoncés de nature variable. *7èmes Journées de phonétique clinique (JPC7)*.

Poddar, A., Sahidullah, M., and Saha, G. (2019). Quality measures for speaker verification with short utterances. *Digital Signal Processing*.

Powers, D. (2011). Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation. *International Journal of Machine Learning Technology*.

Pépiot, E. (2013). *Voix de femmes, voix d'hommes: différences acoustiques, identification du genre par la voix et implications psycholinguistiques chez les locuteurs anglophones et francophones*. PhD thesis, Université Paris VIII Vincennes-Saint Denis.

Querleu, D., Renard, X., Versyp, F., Paris-Delrue, L., and Crèpin, G. (1988). Fetal hearing. *European Journal of Obstetrics Gynecology and Reproductive Biology*, 29:191–212.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramus, F. and Mehler, J. (1998). Language identification with suprasegmental cues: a study based on speech resynthesis. *Journal of Acoustical Society of America*, 105:512–521.

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73:265–292.

Ravanelli, M. and Bengio, Y. (2018a). Learning speaker representations with mutual information. *arXiv prints*.

Ravanelli, M. and Bengio, Y. (2018b). Speaker recognition from raw waveform with sincnet. *arXiv e-prints*.

Rendall, D. (2003). Acoustic correlates of caller identity and affect intensity in the vowel-like grunts vocalizations of baboons. *Journal of Acoustical Society of America*, 113:3390–3402.

Reynolds, D. (1995). Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17:91–108.

Reynolds, D., Campbell, J., Campbell, B., Dunn, B., Gleason, T., Jones, D., Quatieri, T., Quillen, C., Sturim, D., and Torres-carrasquillo, P. (2003). Beyond cepstra: Exploiting highlevel information in speaker recognition. *Proceedings of Workshop on Multimodal User Authentication*, pages 223–229.

Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10:19–41.

Richardson, F., Reynolds, D., and Dehak, N. (2015). A unified deep neural network for speaker and language recognition. *arXiv eprints*.

Rocco, I., Arandjelovic, R., and Sivic, J. (2019). Convolutional neural network architecture for geometric matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2553–2567.

Rose, P. (1999). Differences and distinguishability in the acoustic characteristics of hello in voices of similar sounding speakers: a forensic phonetic investigation. *Australian Review of Applied Linguistics*, 22.

Rose, P. (2002). *Forensic Speaker Identification*. Taylor Francis Forensic Science Series.

Rose, P. and Wang, X. (2016). Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using f-pattern and tonal f0 trajectories over a disyllabic hexaphone. *Proceedings of Odyssey*, pages 326–333.

Rose, P. and Winter, E. (2010). Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analyses. *Proceedings of SST (Austrasian Speech Science and Technology)*.

Rouvier, M., Bousquet, P.-M., Ajili, M., Kheder, W. B., Matrouf, D., and Bonastre, J.-F. (2016). Lia system description for nist sre 2016. *arXiv eprints*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., and Fei-Fei, L. (2015). magenet large scale visual recognition challenge. *JCV (International Journal of Computer Vision)*, 115.

Sadjadi, S., Pelecanos, J., and Ganapathy, S. (2016). The ibm speaker recognition system: recent advances and error analysis. *Proceedings of INTERSPEECH*.

Sahidullah, M. and Saha, G. (2012a). Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech Communication*, 54:543–565.

Sahidullah, M. and Saha, G. (2012b). A novel windowing technique for efficient computation of mfcc for speaker recognition. *arXiv eprints*.

san Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., and Kavanagh, C. (2019). The use of the vocal profile analysis for speaker characterization: Methodological proposals. *Journal of the International Phonetic Association*, 49:353–380.

san Segundo, E. and Mompean, J. (2017). A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity. *Journal of Voice*, 31.

san Segundo, E., Tsanas, A., and Gómez-Vilda, P. (2017). Euclidean distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. *Forensic Science International*, 270.

san Segundo, E. and Yang, J. (2019). Formant dynamics of spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation. *Journal of Phonetics*, 75.

Sato, Y., Sogabe, Y., and Mazuka, R. (2010). Discrimination of phonemic vowel length by japanese infants. *Developmental Psychology*, 46:106–1119.

Scheffer, N., Bonastre, J.-F., Ghio, A., and Teston, B. (2004). Gémellité et reconnaissance automatique du locuteur. *JEP (Journées d'Etudes sur la Parole)*.

Schmid, C., Gendrot, C., and Adda-Decker, M. (2012). Une comparaison de la déclinaison de f0 entre le français et l'allemand. *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*.

Schotz, S. (2007). Analysis and synthesis of speaker age. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Schreiner, C. and Urbas, J. (1986). Representation of amplitude modulation in the auditory cortex of the cat. i. the anterior auditory field (aaf). *Hearing Research*, 21:227–241.

Schreiner, C. and Urbas, J. (1988). Representation of amplitude modulation in the auditory cortex of the cat. ii. comparison between cortical fields. *Hearing Research*, 32:49–64.

Schrimpf, M., Kubilius, J., Lee, M., Murty, N. R., Ajemian, R., and DiCarlo, J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108.

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language*, 57:24–48.

Serrurier, A., Badin, P., Boë, L.-J., Lamalle, L., and Neuschaefer-Rube, C. (2017). Inter-speaker variability: speaker normalisation and quantitative estimation of articulatory invariants in speech production for french. *Proceedings of INTERSPEECH*.

Setiawan, P., Höge, H., and Fingscheidt, T. (2009). Entropy-based feature analysis for speech recognition. *Proceedings of INTERSPEECH*.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., and Kamitani, Y. (2019). End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13.

Shewell, C. (2009). Voice work: Art and science in changing voices. *Oxford: Wiley-Blackwell*.

Shon, S. and Ko, H. (2017). Ku-ispl speaker recognition systems under language mismatch condition for nist 2016 speaker recognition evaluation. *arXiv eprints*.

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31:153–169.

Shriberg, E. (2005). Spontaneous speech: how people really talk and why engineers should care. *Proceedings of INTERSPEECH*.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., and Stolcke, A. (2005). Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46.

Shriberg, E., Ferrer, L., Venkataraman, A., and Kajarekar, S. (2004). Svm modeling of "snerg-grams" for speaker recognition. *Proceedings of INTERSPEECH*.

Shu, L., Xu, H., and Liu, B. (2018). Unseen class discovery in open-world classification. *arXiv e-prints*.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (2011). Voicesauce: A program for voice analysis. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1846–1849.

Smith, H., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., and Stacey, P. (2020). Voice parade procedures: optimising witness performance. *Memory*, 28.

Snyder, D., Garcia-Romero, D., Sell, G., and Povey, D. (2018). X-vectors: Robust dnn embeddings for speaker recognition. *Proceedings of ICASSP (International Conference on Acoustics, Speech and Signal Processing)*.

Soh, M. (2016). Learning cnn-lstm architectures for image caption generation. *Stanford e-prints*.

Song, X., Osmanski, M., Guo, Y., and Wang, X. (2016). Complex pitch perception mechanisms are shared by humans and a new world monkey. *PNAS*, 113.

Sorin, C. (1981). Functions, roles and treatments of intensity in speech. *Journal of Phonetics*, 9:359–374.

Stilp, C. and Kluender, K. (2010). Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *PNAS*, 107:12387–12392.

Stoll, L. and Doddington, G. (2010). Hunting for wolves in speaker recognition. *Proceedings of Odyssey*, pages 159–164.

Subcommittee, I. (1969). Ieee subcommittee on subjective measurements ieee recommended practices for speech quality measurements. *IEEE Trans. Signal Process.*, 17:227–246.

Sundberg, J., Ternstrom, S., Perkins, W., and Gramming, P. (1988). Long-term average spectrum analysis of phonatory effects of noise and filtered auditory feedback. *Journal of Phonetics*, 16:203–219.

Swerts, M., Strangert, E., and Heldnert, M. (2014). F0 declination in read-aloud and spontaneous speech. *International Conference on Spoken Language Processing*.

Sóskuthy, M. and Stuart-Smith, J. (2020). Voice quality and coda /r/ in glasgow english in the early 20th century. *Language Variation and Change*, 32.

Søndergaard, P. L., Decorsière, R., and Dau, T. (2011). On the relationship between multi-channel envelope and temporal fine structures. *Proceeding of Speech Perception and Auditory Disorders*, pages 363–370.

Tanner, J., Sonderegger, M., and Stuart-Smith, J. (2020). Structured speaker variability in japanese stops: relationships within versus across cues to stop voicing. *Journal of the Acoustical Society of America*, 148.

Teixeira, J. P., Oliveira, C., and Lopes, C. (2013). Vocal acoustic analysis - jitter, shimmer and hnr parameters. *Proceedia Technology*, 9:1112–1122.

Thaitechawat, S. and Foulkes, P. (2011). Discrimination of speakers using tone and formant dynamics in thai. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Tirumala, S., Shahamiri, S., Garhwal, A., and Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems With Applications*, 90:250–271.

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The nijmegen corpus of casual french. *Speech Communication*, 52:201–212.

Traill, A. and Jackson, M. (1988). Speaker variation and phonation type in tsonga nasals. *Journal of Phonetics*, 16.

Traunmüller, H. and Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. *(unpublished manuscript)*.

Tweedy, R. and Culling, J. (2014). Does the signal-to-noise ratio of an interlocutor influence a speaker's vocal intensity? *Computer Speech and Language*, 28:572–579.

Vaissière, J. (2004). *Handbook of Speech Perception*, chapter Perception of intonation. Oxford, Blackwell.

van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: patterns and parameters. part i: recognition of backward voices. *Journal of Phonetics*, 13.

van Rossum, G. and Drake, F. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Vaňková, J. and Skarnitzl, R. (2014). Within- and between-speaker variability of parameters expressing short-term voice quality. *Speech Prosody*, pages 1081—-1085.

Vestman, V., Gowda, D., Sahidullah, M., and P. Alku, T. K. (2019). Speaker recognition from whispered speech: a tutorial survey and an application of time-varying linear prediction. *Speech Communication*, 99:62–79.

Wagner, P. and Dellwo, V. (2004). Introducing yard (yet another rhythm determination) and re-introducing isochrony to rhythm research. *Speech Prosody*, pages 227–230.

Waller, S. and Eriksson, M. (2016). Vocal age disguise: the role of fundamental frequency and speech rate and its perceived effects. *Frontiers Psychology*, 7.

Weingartová, L. and Volín, J. (2013). Spectral measurements of vowels for speaker identification in czech. *Studies in Applied Linguistics*, pages 21—-36.

Weirich, M. (2015). Organic sources of inter-speaker variability in articulation: Insights from twin studies and male and female speech. *Individual differences in speech production and perception*.

Weirich, M. and Fuchs, S. (2013). Vocal tract morphology can influence speaker-specific realizations of phonemic contrasts. *Journal of Speech, Language and Hearing Research*, 56.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Broker, F., Thiele, S., Wood, S., and Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, 59.

Wiget, L., White, Schuppler, B., Rauch, I. G. O., and Mattys, S. (2010). How stable are acoustic metrics of contrastive speech rhythm? *Journal of Acoustical Society of America*, 127:1559–1569.

Wightman, R., Touvron, H., and Jégou, H. (2021). Resnet strikes back: an improved training procedure in timm. *arXiv eprints*.

Winkler, R. (2007). Influences of pitch and speech rate on the perception of age from voice. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Winkler, R., Fuchs, S., Perrier, P., and Riede, M. (2011a). Biomechanical tongue models: An approach to studying inter-speaker variability. *Proceedings of INTERSPEECH*.

Winkler, R., Fuchs, S., Perrier, P., and Tiede, M. (2011b). Speaker-specific biomechanical models: From acoustic variability via articulatory variability to the variability of motor commands in selected tongue mucles. *Proceedings of ISSP (International Seminar on Speech Production)*.

Yamins, D. and DiCarlo, J. J. (2016). Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, 37.

Yang, B. (1992). An acoustical study of korean monophthongs produced by male and female speakers. *Journal of Acoustical Society of America*, 91.

Yeung, H., Chen, K., and Werker, J. (2013). When does native language input affect phonetic perception? the precocious case of lexical tone. *Journal of Memory and Language*, 68:123–139.

Yoo, J., Oh, H., Jeong, S., and Jin, I. (2019). Comparison of speech rate and long-term average speech spectrum between korean clear speech and conversational speech. *Journal of Audiology Otology*, 23.

Yoon, T. (2010). Capturing inter-speaker invariance using statistical measures of speech rhythm. *Proceedings of Speech Prosody*, 5:1–4.

Yovel, G. and Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Science*, 17.

Yuan, J. and Liberman, M. (2014). F0 declination in english and mandarin broadcast news speech. *Speech Communication*, 65:67–74.

Yumoto, E. and Gould, W. J. (1982). Harmonics-to-noise ratio as an index of hoarseness. *Journal of Acoustical Society of America*, 71:1544–1550.

Zhang, C. and Tan, T. (2008). Voice disguise and automatic speaker recognition. *Forensic Science International*, pages 118–122.

Zhang, M., Chen, Y., Li, L., and Wang, D. (2017). Speaker recognition with cough, laugh and "wei". *arXiv eprints*.

Zhang, X.-L. (2018). Linear regression for speaker verification. *arXiv eprints*.

Zhao, W., Gao, Y., and Singh, R. (2017). Speaker identification from the sound of the human breath. *arXiv eprints*.

Zheng, Y., Zhao, M., Ma, Y., Liu, M., Ma, X., Liang, T., Kong, T., He, L., and Xu, M. (2020). Thuee system description for nist 2020 sre cts challenge. *NIST20 SRE CTS*.

Zuo, D. and Mok, P. (2012). Formant dynamics of bilingual identical twins in non-contemporaneous speech. *Proceedings of SST (Austrasian Speech Science and Technology)*.

Coding references

The present document is based on a LaTeX template that has been altered in several ways that aim at improving usability.

All scripts that have been created and used for this thesis can be obtained by directly asking the Author. In case of a script inspired by other works this is mentioned in the document's header, however, if no other reference is present one can assume it is the result of the Author's writing.

Speech components in phonetic characterisation of speakers

a study on complementarity and redundancy of conveyed information

Abstract

The decomposition of the speech signal into phonetically meaningful units allows the analysis of between- and within- speaker variations. These are components associated with characteristics whose nature relates to the physical, psychological and social aspects of a speaker. In this thesis, we compare perceptual characterisation results with a phonetic analysis and advanced modelling techniques through Convolutional Neural Networks (CNN).

Clusterings' analysis shows that the perceptual results are coherent with those obtained by the CNN and phonetic approaches, which supports the application of these methods in Phonetics. Our results highlight that spectrograms are the most accurate speech representation for speaker identification (96 % correct answers on average). Higher formants and harmonics are more important in the characterisation of female voices. Whereas, voice quality characteristics, such as breathiness and hoarseness, play a major role in the characterisation of male speakers. The comparison between Mel Frequency Cepstral Coefficients (MFCC) and classical phonetic measurements is also examined. The MFCC are mainly linked to intensity and f0 in the characterisation of female speakers, while to the distributions of energy and low level spectral shape for male speakers.

Our findings confirm the importance of describing the within-speaker variation for a more complete understanding of between-speakers differences.

Keywords: speaker, characteristics, components, CNN, spontaneous, clustering, informedness, comparison

Les composantes de la parole dans la caractérisation phonétique du locuteur

étude sur la complémentarité et la redondance véhiculées des informations

Résumé

La décomposition du signal vocal en unités phonétiquement significatives permet d'analyser les variations inter- et intra- locuteur. Ces unités sont des composantes associées à des caractéristiques dont la nature est liée aux aspects physiques, psychologiques et sociaux d'un locuteur. Dans cette thèse, nous comparons une caractérisation perceptive, une analyse phonétique et des techniques de modélisation avancées par des réseaux de neurones à convolution (CNN).

L'analyse des clusterings montre que les résultats perceptifs sont cohérents avec ceux obtenus par les approches CNN et phonétique, ce qui soutient leurs applications en phonétique. Nos résultats mettent en évidence que les spectrogrammes sont la représentation de la parole la plus précise pour l'identification des locuteurs (96 % de bonnes réponses en moyenne). Les formants et des harmoniques plus élevés sont plus importants dans la caractérisation des voix féminines. En revanche, les caractéristiques de la qualité de la voix, telles que le souffle et la raucité, jouent un rôle majeur dans la caractérisation des voix masculines. Le lien entre les coefficients cepstraux à fréquence Mel (MFCC) et les mesures phonétiques classiques est également examiné. Les MFCC sont principalement liés à l'intensité et à f0 dans la caractérisation des voix féminines, tandis qu'aux distributions d'énergie et à la forme spectrale de bas niveau pour celle des voix masculines.

Nos résultats confirment l'importance de la description de la variation intra-locuteur pour une compréhension plus complète des différences entre locuteurs.

Mots-clés : locuteur, caractéristiques, composantes, CNN, spontané, clustering, informativité, comparaison

Université Sorbonne Nouvelle
Ecole Doctorale 622 - Sciences du langage
4, rue des Irlandais 75005 Paris