Universidade Estadual de Campinas

Instituto de Estudos da Linguagem

Julio Cesar Cavalcanti de Oliveira

MULTIPARAMETRIC ANALYSIS OF ACOUSTIC-PHONETIC MEASURES IN GENETICALLY AND NON-GENETICALLY RELATED SPEAKERS: IMPLICATIONS FOR FORENSIC SPEAKER COMPARISON

Análise fonético-acústica multiparamétrica em falantes geneticamente e não geneticamente relacionados: implicações para a comparação forense de locutor

CAMPINAS

2021

Julio Cesar Cavalcanti de Oliveira

MULTIPARAMETRIC ANALYSIS OF ACOUSTIC-PHONETIC MEASURES IN GENETICALLY AND NON-GENETICALLY RELATED SPEAKERS: IMPLICATIONS FOR FORENSIC SPEAKER COMPARISON

Análise fonético-acústica multiparamétrica em falantes geneticamente e não geneticamente relacionados: implicações para a comparação forense de locutor

Thesis presented to the Institute of Language Studies of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Linguistics.

Tese apresentada ao Instituto de Estudos da Linguagem da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Linguística.

Supervisor/Orientador: Plinio Almeida Barbosa

Co-supervisor/Coorientador: Anders Eriksson

Este exemplar corresponde à versão final da tese defendida pelo aluno Julio Cesar Cavalcanti de Oliveira, e orientada pelo Prof. Dr. Plinio Almeida Barbosa.

Campinas

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Estudos da Linguagem
Leandro dos Santos Nascimento - CRB 8/8343

C314m

Cavalcanti, Julio Cesar, 1992-
Multiparametric analysis of acoustic-phonetic measures in genetically and non-genetically related speakers : implications for forensic speaker comparison / Julio Cesar Cavalcanti de Oliveira. – Campinas, SP : [s.n.], 2021.

Orientador: Plinio Almeida Barbosa.
Coorientador: Anders Eriksson.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Estudos da Linguagem.

1. Fonética. 2. Fonética acústica. 3. Fonética Forense. 4. Prosódia. 5. Gêmeos idênticos. I. Barbosa, Plinio Almeida, 1966-. II. Eriksson, Anders, 1939-. III. Universidade Estadual de Campinas. Instituto de Estudos da Linguagem. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Análise fonético-acústica multiparamétrica em falantes geneticamente e não geneticamente relacionados : implicações para a comparação forense de locutor
Palavras-chave em inglês:
Phonetics
Acoustic phonetics
Forensic phonetics
Prosody
Identical twins
Área de concentração: Linguística
Titulação: Doutor em Linguística
Banca examinadora:
Plinio Almeida Barbosa [Orientador]
Livia Oushiro
Luciana Lucente
Sandra Madureira Fontes
Pablo Arantes
Data de defesa: 09-06-2021
Programa de Pós-Graduação: Linguística

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-4465-6242
- Currículo Lattes do autor: http://lattes.cnpq.br/7761155705294964

UNICAMP

BANCA EXAMINADORA

Plínio Almeida Barbosa

Livia Oushiro

Luciana Lucente

Sandra Madureira Fontes

Pablo Arantes

IEL/UNICAMP
2021

Ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós Graduação do IEL.

To my grandfather
Audalio Badega Cavalcanti
"In memoriam"

Acknowledgements

Throughout my doctorate studies, many notable people have crossed my path. Although it would be very difficult to name every single one of them, some have played an essential role during this particular period of my life. These are acknowledged in the following.

I want to thank my supervisor Prof. Dr. Plinio A. Barbosa. I have been the most fortunate student for being able to work and learn from him. I am grateful for the countless hours of work he has put into this thesis, all the discussions, guidance, patience, and for always being so kind and generous. I lack words to express how grateful I am for being one of his supervisees. Thank you for believing in me and in the research project I set out to conduct. I also think Rose Barbosa for all the affection and encouragement throughout the process.

Likewise, I thank my co-supervisor, Prof. Dr. Anders Eriksson. It was such an honor and privilege to work with someone like him. I am very thankful for all the valuable and singular contribution he brought to my research, all the discussions, his guidance, patience, humbleness, and for dedicating so much of his valuable time to co-supervising me. Thank you for inviting and having me at Stockholm University and for making me feel very welcomed there.

Moreover, I thank the Brazilian National Council for Scientific and Technological Development (CNPq) for partly funding this project (process: 140364/2017-0). Also, I thank the Coordination for the Improvement of Higher Education Personnel (CAPES) in cooperation with the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) for funding this research during the period I have stayed in Sweden (process: 88887.308270/2018-00). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

I acknowledge all the participants of the research (the twins pairs) who kindly agreed to take part in the study by offering their time, effort, and interest. Without their valuable collaboration, no research could have been done. For that, I am immensely and eternally grateful.

I thank my home university, the University of Campinas (UNICAMP) for the assistance I received throughout my doctorate. Furthermore, I am very thankful to all the brilliant professors

I had – special thanks to Prof. Dr. Livia Oushiro–for inspiring lectures.

I thank the Department of Linguistics at Stockholm University (SU) for providing me all the support I needed during the long period I have stayed in Sweden conducting the research, from staff to dear colleagues that I had the chance to learn from and share pleasant moments.

I thank Dr. Miguel Oliveira Jr. (Federal University of Alagoas - UFAL) for being a constant source of inspiration and encouragement. I have learnt some much from him. Also, thanks to the "Group of Studies in Phonetics and Phonology (fonUFAL)" for all the technical support and assistance I received during the data acquisition process.

Thanks to Dr. Cristina Felipeto (UFAL) for introducing me to the realm of linguistic studies while I was still an undergrad student, and Dr. Eduardo Calil (UFAL) for guiding me in my first years of research as a student, as well as Dr. Luzia Payão (UNCISAL) for fostering my interest in phonetics/phonology studies at the speech-language pathology and audiology faculty. Thanks to Dr. Ana Carolina Constantini for supervising me during my research qualification at UNICAMP.

Thanks to Dr. Justin Lo (University of York) for providing one of the scripts used to carry out the statistical analyses and kindly responding to my queries. Also, thanks to Prof. Dr. Marcin Wlodarczak (Stockholm University) for his valuable assistance concerning data visualization.

I want to express my gratitude to Dr. Sandra Madureira (PUC-SP), Dr. Livia Oushiro (UNICAMP), Prof. Dr. Luciana Lucente (UFMG), Dr. Pablo Arantes (UFSCar), Dr. Renata Passetti (PUC-SP) for reading the thesis dissertation and for presenting relevant comments/suggestions.

Thanks to Francisco Ribeiro, Paula Delfino, Rafael, Daniela, Aline, Emerson, Joana, Laís, Williane, Lucas, Tamara, Larissa, Priscilla, Cristiane (Kio), Carol, André, Ranúzia, and Nivaldo Badega for being part of the process.

My gratitude to Dr. Mikael Kalm for being a constant presence in my life and making my days so much happier. To Andreas Sallmunds for all the good moments we have shared and all the support. To Dr. Andrew Tait for the companionship and for always being there for me.

To my beloved parents, without the support, I could have never accomplished this, and also, to my brother. Thanks to Salezia Magna, Claudio da Costa, Bruna Oliveira for always being a source of inspiration and encouragement.

Finally, but not least, I want to thank José Alexandre D'Abbruzzo for all the support, patience, persistence, resilience, care, and companionship during all these years. It would not have been possible to accomplish this without him.

I want to express my gratitude to all people who, somehow, contributed to this work and that positively influenced my journey until now. To you all, my sincere appreciation and gratitude.

Resumo

A pesquisa desenvolvida na presente tese voltou-se para a análise multiparamétrica de medidas fonético-acústicas entre sujeitos geneticamente relacionados, i.e., gêmeos idênticos, e sujeitos não-geneticamente relacionados, i.e., comparações entre pares. De um modo geral, buscou-se responder a seguinte pergunta: "quais parâmetros fonético-acústicos e dimensões de análise são considerados notadamente discriminatórios em comparações realizadas entre indivíduos geneticamente relacionados e entre todos os falantes da pesquisa, e portanto, potencialmente relevantes para a aplicação forense?".

Parâmetros pertencentes a três diferentes dimensões fonético-acústicas foram analisados, a saber: frequências de formantes vocálicos (espectral), medidas temporais (temporal) e descritores da frequência fundamental (melódica), resultando em um total de 30 parâmetros fonético-acústicos analisados.

Os participantes da pesquisa foram 20 sujeitos, dez pares de gêmeos idênticos do sexo masculino, falantes do Português Brasileiro (PB) da mesma área dialetal, com idades entre 19 e 35 anos. O material de fala consistiu em conversas telefônicas espontâneas entre gêmeos, a partir de temas selecionados pelos próprios pares.

Em relação aos resultados, os achados sugeriram as frequências de formantes mais elevados, e.g., F3 e F4, como potencialmente mais discriminatórias em relação às frequências de formantes mais baixos, como indicado pela maior proporção de diferenças entre falantes e a análise do tamanho/extensão do efeito. Contuto, dentre todas as medidas, F3 apresentou as propriedades desejadas expressas pela combinação de menores valores de Cllr/EER, compatível com maior acurácia, e altos valores de AUC, compatível com um alto potencial discriminatório.

Em relação à qualidade da vogal, a vogal central baixa [a] e as vogais anteriores revelaram-se mais discriminatórias quando comparadas às vogais posteriores. Tais segmentos apresentaram também maiores distâncias euclidianas entre si, convidando a hipótese de uma possível relação entre a dispersão fonético-acústica das vogais e o nível de variação fonética admitido. Ademais, embora vogais tônicas tenham apresentado uma maior proporção de diferenças entre sujeitos, a combinação de vogais tônicas e átonas mostrou-se, em geral, mais discriminatória dos falantes.

Em se tratando de parâmetros temporais, as evidências sugerem a categoria de parâmetros macro-temporais, e.g., taxa de elocução e taxa de articulação, como mais discriminatórios e consistentes em condições de fala espontânea/não-controlada. Além disso, a análise de medidas temporais em gêmeos idênticos revelou um grau considerável de semelhanças intra-par, substancialmente mais elevado do que o observado para parâmetros não-temporais. Alguns fatores explicativos, incluindo a sobreposição dos fatores "entrelaçamento prosódico" e "padrões/representações temporais compartilhados" foram sugeridos para embasar tal convergência.

Quanto aos descritores da $f0$, a *baseline*, mediana, média, e valores extremos de $f0$ apresentaram as maiores proporções de diferenças intra-gêmeos e entre todos os sujeitos, acompanhadas de maiores tamanhos/extensões do efeito. Contrariamente, medidas de variação e modulação da $f0$ mostraram-se relativamente menos variáveis. Além disso, medidas da $f0$ analisadas na fala concatenada apresentaram, em geral, um melhor potencial discriminatório do que quando extraídas de vogais prolongadas. Os resultados reforçam a relevância da análise de descritores da $f0$ para fins forenses, especialmente da $f0$ *baseline*, medida com os menores valores de EER observados.

Palavras-chave:
Fonética, Fonética acústica, Fonética forense, Prosódia, Gêmeos idênticos

Abstract

The present thesis proposes a multiparametric analysis of acoustic-phonetic measures in comparisons performed with genetically related individuals, namely, identical twins, and across all subjects in the study (i.e., cross-pair comparisons). The central research question may be formulated as "which acoustic-phonetic parameters and analysis dimensions, namely the spectral, temporal, and melodic dimensions, are the most inter-speaker discriminatory in comparisons performed between genetically related individuals and across all speakers, and therefore potentially relevant for the forensic speaker comparison application?".

Parameters pertaining to three different acoustic-phonetic dimensions were analyzed: formant frequencies (spectral), speech timing (temporal), and fundamental frequency (melodic) estimates yielding a total of 30 acoustic-phonetic parameters.

The participants were 20 subjects, ten identical male twin pairs, Brazilian Portuguese (BP) speakers from the same dialectal area, aged between 19 and 35. The speech material consisted of spontaneous telephone conversations between twins, with dialogue topics decided by the pairs.

Concerning the main outcomes, evidence was found suggesting high-formant frequencies, namely F3 and F4, as potentially more speaker discriminatory than low-formant frequencies, as verified by the proportion of significant differences across speakers and the comparison of effect sizes. However, between these two formants, F3 has shown to possess the desired properties expressed by the combination of lower Cllr/EER, which is compatible with higher accuracy, and high AUC values, compatible with high discriminatory power.

Regarding vowel quality, the low central vowel [a] and front vowels appeared as the most speaker-discriminatory segments. These segments also seemed to display higher Euclidean distances from their neighbors, inviting the hypothesis of a probable relationship between vowel acoustic dispersion and the level of phonetic variation allowed by the phonological system. Furthermore, even though stressed vowels appeared more speaker-discriminatory than unstressed vowels, the combination of both vowel classes seemed to be more explanatory in terms of the observed inter-speaker differences.

As for speech timing parameters, evidence was found supporting the category of macro speech

timing parameters, mainly speech rate and articulation rate, as the most discriminatory and consistent parameters for speaker comparison applications under unscripted speech conditions. Moreover, the analysis of speech timing estimates in identical twin pairs revealed a remarkable level of intra-pair similarities, substantially higher than the observed for the same speakers' formant frequency patterns. Some explanatory factors, including the overlapping effects of "prosodic entrainment" and "shared speech timing patterns/representations", were suggested to account for such a high convergence.

As for $f0$ descriptors, $f0$ baseline, median, mean, and extreme values were found to display higher proportions of intra-twin pair and cross-pair differences while also presenting the largest effect sizes. Conversely, $f0$ variation and modulation estimates were found relatively more stable across different subjects. Moreover, $f0$ metrics assessed in connected speech tended to present a better discriminatory potential than lengthened vowels. The outcomes reinforce the relevance of analyzing long-term $f0$ metrics for forensic purposes, particularly $f0$ baseline, which displayed the lowest EER values among all tested $f0$ estimates.

Keywords:
Phonetics, Acoustic phonetics, Forensic phonetics, Prosody, Identical twins

List of figures

List of Tables

List of Abbreviations, Acronyms, and Symbols

Acoustic-phonetic parameters

F1 First speech formant frequency

F2 Second speech formant frequency

F3 Third speech formant frequency

F4 Fourth speech formant frequency

SRATE Speech rate

ARTRATE Articulation rate

SGDUR Stress group duration

V-V unit Vowel-to-vowel unit

VVDUR Vowel-to-vowel unit duration

VOWEL DUR Vowel duration

SILDUR Silent pause duration

IPI Inter-pause interval

$f0$ Fundamental frequency

f0mean $f0$ mean

f0med $f0$ median

f0base $f0$ baseline

f0min $f0$ minimum

f0max $f0$ maximum

f0sd $f0$ standard-deviation

f0SAQ $f0$ semi-amplitude between quartiles

f0M1 Standard-deviation of $f0$ maxima in semitones

f0M2 Standard-deviation of the F0 maxima positions in seconds

f0M3 standard-deviation of the F0 maxima positions in seconds

f0M4 1st-derivative f0 mean in Hertz/frame of the positive derivatives

f0M5 1st-derivative f0 mean in Hertz/frame of the negative derivatives

f0M6 1st-derivative F0 standard-deviation in Hertz/frame of the positive derivatives

f0M7 1st-derivative F0 standard-deviation in Hertz/frame of the negative derivatives

f0M8 Mean peakness of F0 max in semitones

Units of measurement

Hz Hertz

Bark Bark critical-band-rate

st semitones

s seconds

ms milliseconds

vv/s Vowel-to-vowel units per second

Statistical analysis

η^2 Eta-squared effect size

LR Likelihood ratio

MVKD Multivariate Kernel Density

Cllr Log-likelihood-ratio-cost

$\mathbf{Cllr_{raw}}$ Raw log-likelihood-ratio-cost

$\mathbf{Cllr_{cal}}$ Calibrated log-likelihood-ratio-cost

EER Equal error rate

ROC Receiver Operating Characteristics

AUC Area Under the Receiver Operating Characteristics curve

Euclid. dist Euclidean distance

Eff. size Effect size

Mag Magnitude

%diff Percentage of differences

sd Standard deviation

Other abbreviations

MZ Monozygotic

DZ Dyzygotic

FSC Forensic speaker comparison

JND Just-noticeable difference

Contents

Chapter 1

Introduction

The central theme of this thesis concerns the analysis of the speaker discriminatory potential of acoustic-phonetic parameters in comparisons performed between very similar speakers, i.e., identical twins, and across all individuals by following a multi-parametric perspective. The main motivation for carrying out this study lied on the necessity of advancing on the understanding of what may be regarded as relatively specific of an individual as far as acoustic-phonetic patterns are concerned. Such a knowledge bears some theoretical as well as practical implications, as will be mentioned further.

Considering the "*variable nature of speech*" in opposition to its "*less-varying facet*" as the main object of study, a particular emphasis is drawn on the comparison of genetically related speakers, such as identical twins. The remarkable high similarity observed between such individuals is taken here as a singular opportunity for assessing what may still be regarded as speaker contrasting when most sources of inter-subject variation are reduced, and, as a consequence, shedding light on the idiosyncratic nature of speech production.

In addition, given the potential of this study's outcomes for forensic speaker comparison ends, all analyses have been performed based on unscripted speech materials: spontaneous dialogues.

To the best of our knowledge, the present research represents the first experimental study in Brazilian Portuguese from a forensic phonetic analyses perspective carried out with a group consisting exclusively of male adult identical twins.

Estimates deriving from three different acoustic-phonetic dimensions were assessed (i.e., formant frequency, speech timing, and fundamental frequency dimensions), with the primary goal of providing a more comprehensive picture regarding individual productions, while allowing the identification and evaluation of potentially relevant parameters for the forensic speaker comparison application.

Overall, a set of 30 acoustic-phonetic parameters were explored in the present thesis, 4 speech formant frequencies assessed individually, as a function of vowel quality, and lexical stress; 11 speech timing parameters, compreehending macro, micro, and pause-related estimates; and 15 fundamental frequency acoustic descriptors, including range, centrality, baseline and modulation estimates.

It is noteworthy that acoustic parameters can either be assessed individually or in combination. However, given the fact that very little is known regarding the speech patterns of identical twins, mainly in Brazilian Portuguese, in the present experimental research an effort has been made to understand the individual contributions of different acoustic parameters, their level of inter-speaker variation, and the possible linguistic implications regarding their variation. Such a knowledge may help not only probe the individual linguistic significance of the parameters assessed, as they may also constitute the building blocks for future more complex models to be generated: one that better accounts for "*specificity*" in speech production.

This thesis is organized in six chapters which stand for different stages of the study. In the present chapter, *chapter 1*, the relevant literature is introduced, following the identification of their object of study, findings and main contributions to the field. It may be noted that, in all cases, the literature review departs from a broader to a narrower perspective, acknowledging essential aspects that are necessary for understanding the analyses conducted while offering the reader the possibility to become familiarized with the research theme. Research questions and hypotheses are also presented and motivated in view of previous studies and experimental tests.

In *chapter 2*, focus is drawn to the methodological approach and the adopted experimental design, including information about the participants, recording procedure, the segmentation and transcription process, the set of acoustic-phonetic parameters assessed, data extraction, analyses tools, and the statistical analyses conducted. The decision for adopting the term "identical" over "monozygotic" twins is motivated. Finally, system performance estimates, including those commonly applied within the forensic phonetics framework, are described.

Chapter 3 is dedicated to the analyses performed on the formant frequency dimension. The results are described in light of the speaker comparisons carried out with the first four vowel formant frequencies. The effects of the lexical stress component on the differences observed across speakers are also described. Furthermore, the results are discussed in consideration of the specialized literature concerning formant frequency patterns.

Chapter 4 concerns the analyses carried out on parameters related to the speech timing dimension. The results concerning the analyses performed with macro, micro, and pause-related speech timing parameters are presented. Some practical aspects regarding the sample size (n-size) influence on the statistical analyses are acknowledged and a compensation for such effect is applied.

Furthermore, the effects of "*prosodic entrainment*" are evoked to account for a remarkably higher congruence in speech timing patterns of identical twin pairs. The main findings are discussed in light of the available literature.

Chapter 5 regards the analyses conducted with acoustic descriptors of speaking fundamental frequency. The outcomes deriving from speaker comparisons based on fundamental frequency range, modulation, and variability parameters are described, and the implications of assessing fundamental frequency by employing two different types of *speech material* are exploited. A discussion is developed in view of forensic-related and general studies dealing with fundamental frequency patterns.

Chapter 6 is a closing chapter, where all the aforementioned analyses are put into perspective. General trends regarding the speaker-discriminatory potential of acoustic-phonetic estimates are presented and discussed. Some qualitative information regarding the participants, i.e., identical twin pairs, are retrieved, and an attempt to qualitatively correlate their individual answers to some of the acoustic-phonetic patterns observed is made. Finally, some new directions are pointed out concerning future developments of the present research.

1.1 Background

In the past few decades, attempts have been made towards a proper understanding as to whether human voice and speech features are endowed with absolute specificity. From a theoretical point of view, most of the experiments addressing individuality in voice and speech have tried at some level to shed light on the relevant and challenging question "does each person in the population have a measurably unique voice?" (NOLAN; OH, 1996). Some studies have presented answers to this question over the past few years, e.g., Loakes (2003, 2004, 2008), Loakes and McDougall (2010), Fernández (2012), San Segundo, Tsanas, and Gómez-Vilda (2017) and Weirich (2012), which corroborate this specificity at some level, with the magnitude of differences being dependant on the subjects under analysis.

Assessing aspects of individuality in speech and the limits of phonetic variation between very similar individuals require a considerably high experimental control level. As such, distances between individuals' due to biological/inherited (*nature*) and environmental/learned (*nurture*) factors have to be considerably shortened and controlled. In response to this challenge, identical twins have been the focus of scientific experiments in many different fields, including linguistic studies dealing with speech production.

The present study may be regarded as two-fold since it comprehends both practical as well as

scientific demands. The first demand is given by the necessity of identifying candidate acoustic measures for the speaker comparison application, adding to the body of knowledge of forensic phonetics research. The second regards the effort in understanding the variable nature of speech, identifying possible factors related to such variability. In this regard, understanding the underlying components related to inter-speaker variation is crucial when conducting forensic speaker comparison analyses. Furthermore, the identification of speech elements that display a great deal of variation among individuals also contributes to the understanding of what may be regarded as relatively stable, or in more suitable terms, "less varying", in the acoustic materialization of speech.

As to provide a proper background to the reader, general aspects regarding the research topic of the present thesis are introduced and commented on in the following.

1.1.1 Genetically related individuals

The search for elements responsible for shaping human characteristics represents a common goal for many scientists in many different research fields, such as medicine, psychology, anthropology, biology, sociology, and linguistics. Endeavors have been made towards the understanding and quantification of the contribution of genetic (*nature*) and environmental factors (*nurture*) in determining human traits, from a concrete (e.g., morphological aspects) to a more conceptual viewpoint (e.g., human behavior). As an answer for such a lack of knowledge, genetically identical and genetically related individuals have been submitted to systematic research over the past few years, what has resulted in a scientific method known as "*the twin method*".

As described by Vogel and Motulsky (1986), the twin method is founded on the biological fact that monozygotic twins (MZ), the so-called identical twins, originate from division of the same zygote. As a result, they are considered to be genetically identical. It follows that any phenotypic[1] differences between MZ twins must be explained by environmental influences, expressed by any factors that are not a priori fixed genetically. Conversely, dizygotic twins (DZ), also known as fraternal twins or non-identical twins, like any other siblings, share only half of their genes.

In terms of structural or anatomical aspects, identical twins are assumed to have very similar vocal tracts in size and shape. According to Beck (1997), studies of genetically identical speakers make it clear that a person's genetic makeup is a major factor in determining his or her overall size, shape, rate of growth and maturation. The characteristics of the human face, for instance, i.e.,

[1] As described in Korf (2004), "*phenotype*" is the term used in reference to the physical characteristics resulting from the genetic code or genotype. The interaction of several genes, with each other, and environmental factors is necessary to generate a certain *phenotype*, i.e., the expression of physical characteristics.

facial morphology, seem to be determined both by genetic and environmental factors, as remarked by Djordjevic et al. (2016). In a population-based Twin Study using 3D facial image landmarks of 1380 female twins, the authors found evidence suggesting that genetic factors seem to explain more than 70% of the *phenotypic* variation in facial size, nose (width, prominence and height), lips prominence and inter-ocular distance, where some traits have shown potential dominant genetic influence, such as the prominence and height of the nose, the lower lip prominence in relation to the chin and upper lip philtrum length. According to the authors, environmental contribution to facial variation seems to be the greatest for the mandibular ramus height and horizontal facial asymmetry. Such an observation is compatible with the suggested relationship between environmental factors on facial morphological characteristics, such as breathing patterns. While performing a systematic review and meta-analysis of lateral cephalometric data, Zheng et al. (2020) reported the tendency of mouth breathers to display a retrognathic maxilla and mandible, vertical growth pattern with high mandibular plane angle, along with other facial modifications.

The extension of genetic influences on anatomical aspects are not solely limited to peripheral structures. Consistent evidence of genetic influences on brain structure has also been systematically reported in the literature (THOMPSON et al., 2001; MAGGIONI et al., 2020). As remarked by Thompson et al. (2001), genetically identical twins are almost entirely correlated in their gray matter distribution, including areas related to language cortices, for instance. Through a magnetic resonance imaging (MRI)[2] experimental approach, the authors were able to identify a genetic continuum, in which brain structure appeared increasingly similar in subjects with increasing genetic affinity (e.g., unrelated subjects < non-identical twins < identical twins).

In the same direction, the review study conducted by Maggioni et al. (2020) on twin studies using MRI, suggested that the global brain morphology and network organization are highly heritable from early childhood to young adulthood. However, as pointed out by the authors, functional magnetic resonance imaging (fMRI) studies on MZ twins provided evidence of life experiences' influence on brain function through epigenetic changes, showing associations among genetic expression, hormone and serotonin levels and in turn brain emotion processing. Furthermore, it has been observed that genetic correlations among brain regions exhibit heterogeneous trajectories, and this heterogeneity reflects the progressive, experience-related increase in brain network complexity, pointing to the key role of environment in mediating brain network differentiation.

From a language development perspective, as pointed out by Rice (2020), twin studies have

[2] A non-invasive brain imaging technique that is able to provide multi-modal information, such as brain tissue morphology, structural composition, and spatio-temporal signals relating to its activity and connectivity. This inherent multi-modality makes MRI an interesting tool for the study of neurodevelopmental pathways (MAGGIONI et al., 2020).

allowed the identification of causal pathways for several language abilities, which have enriched our understanding of the sources of language abilities. The identification of a genetic component in different types of speech and language disorders, such as phonological disorder, learning disorders, and delayed speech production, has also been possible through studies following a twin design, where a higher agreement concerning the nature of the disorder has been found for monozygotic over dizygotic twins (LEWIS; THOMPSON, 1992). A "*twinning effect*" of risk for late language acquisition in twins has also been suggested in the literature, as observed by Rice (2020). Such an effect seems also greater for monozygotic than dizygotic twins. Overall, the studies' outcomes support a genetic component on language acquisition and the probable heritability of some specific language impairment. With regard to the present study, the assumption of twin pairs as displaying remarkable similarities in their outcomes during critical language acquisition and development periods is of unquestionable relevance, given the possible implications of such convergence in their linguistic performance in adulthood.

Although the genetic component (i.e., the genome structure) is able to account for a great deal of variation among individuals, it can not be regarded as absolute. As remarked by Maggioni et al. (2020), recent studies of genome structure have been integrated with the investigation of modifications in gene activity and expression that do not seem to affect the DNA sequence, namely, the epigenetic regulation.

As mentioned by Weinhold (2006), the term "*epigenetic*" can be literally translated to "in addition to changes in genetic sequence." According to the author, such term includes any process that changes gene activity without modifying the DNA sequence, leading to modifications that can be transmitted to other cells. The identification of such mechanism, which opened space for a brand new scientific branch in modern genetic science, has been challenging the scientists' understanding of the genetic regulation, as posed by the environmental effect on the gene function, i.e., the "*genetic expression*". In that regard, even with the identification of all human genes, we are still very far from understanding how their expressions are regulated, how gene products behave towards each other and towards the environment (KORF, 2004). However, as observed by Alberts et al. (2018), although research in this field is still in its early stages, the idea that environmental events can be permanently recorded by our cells is an exciting fact that challenges future generations of scientists.

However, it is important to note that similarities at the genetic level do not necessarily imply similarities in other complex domains, such as at the behavioral realm. As remarked by Vogel and Motulsky (1986), if we consider that the biological basis of human behavior is in the brain, intrinsic genetic variation may likely be observed as in any other human organ. In this sense, the overlapping of genetic and environmental factors in genetically identical subjects makes the interpretation of

behavioral information rather difficult. Such interpretation becomes even more challenging in light of recent outcomes suggesting different neural activity patterns in identical twins (MAGGIONI et al., 2020), which touches upon the complex and still not very well understood interplay between *structure* and *function*.

Assuming the linguistic component as intertwined with variables from social, cultural, and psychological orders, cf. Labov (2011) and Tomasello (2000), a considerably higher level of external/environmental influences may be expected concerning the level of variation observed for structural or organic components between identical twins. Notwithstanding, such individuals are assumed to display the highest possible level of convergence regarding the before-mentioned factors, which may influence their linguistic patterns, especially regarding their speaking manner. In the present study, the acoustic nature of their speech production is the object of interest.

As highlighted by Loakes (2008), research on the speech patterns of twins from a forensic phonetic perspective make it possible to understand the very limits of variation between speakers. The mere fact that most of these individuals present similar linguistic and environmental influences during their development and youthhood reflects a considerable reduction of common sources of inter-speaker variation. In that sense, the analysis of acoustic speech parameters from a forensic perspective with twin speakers represents a singular experimental condition to test their consistency and speaker-discriminatory potential.

1.2 On vowel formant frequency

In the following sections we comment on general aspects regarding the analysis of vowel formant frequencies, departing from a broader perspective, where common sources of variation in the acoustic production of formants are identified, to more specific conditions, as in the application of vowel formant analysis in the comparison of genetically-related speakers.

1.2.1 General aspects

In terms of speech production, vowel segments– as other speech sounds, are believed to convey information of three different dimensions: linguistic, social, and idiosyncratic dimensions, as mentioned by Ladefoged and Broadbent (1957) in a widely cited study. These dimensions are directly related to the variation in speech and responsible for shaping ones' "speech profile" at the acoustic, articulatory, and perceptual levels.

As pointed out by Ladefoged and Broadbent (1957), linguistic information refers to what is being said or "*the significance of the utterance*" conveyed through the shared linguistic system.

Together with the linguistic content, there is information related to the general background of a speaker, such as geographical origin, social class, social groups, level of education. These aspects are related to the social or socio-linguistic dimension. A third kind of information can also be identified, namely the idiosyncratic features of a person's speech, expressed by learned speech patterns acquired throughout life and by anatomical and physiological aspects, dependent on the shape of the vocal tract, its dimensions, and proportions. In this sense, both the linguistic and the socio-linguistic information conveyed by vowels might depend mainly on the relative positions of the formants or, in the authors' terms, on "*the relative formant structure of the vowels*". In contrast, personal information seems to depend partly on the absolute values of the formant frequencies, given that the frequency ranges in which someone speaks cannot be modified by natural means as they are related to anatomic and physiological properties. In this sense, physiological and anatomical factors are acknowledged sources of inter-speaker variation concerning vowel formant frequencies.

Other sources of inter-speaker variation are commonly reported in the literature, including differences as a function of sex, dialect, age, speaking style, and speech rate. For instance, acoustic space sizes based on F1 x F2 measures tend to be larger for women than for men, and formant values are considerably higher for females when compared to males (ESCUDERO et al., 2009; BEHLAU et al., 1988; KENT; VORPERIAN, 2018), which is also due to anatomical reasons, such as the descent of the larynx in males during puberty (TRAUNMÜLLER, 1984).

Moreover, differences in F1 and F2 formant frequencies have also been found during the lifespan, in which both formant frequencies and bandwidths tend to decrease during typical speech development. Additionally, decreases in voice fundamental frequency and F1 due to aging are reported to be more likely in females than males (KENT; VORPERIAN, 2018). There is also evidence pointing to a higher inter-speaker variability of acoustic vowel spaces between slow and fast speakers, with slow speakers displaying a larger average variability (TSAO; WEISMER; IQBAL, 2006), and differences related to speaking style, with a tendency to more centralized formant values in spontaneous speech when compared to word reading, for instance, cf. HARMEGNIES and POCH-OLIVÉ (1994).

Other linguistic factors have also been found to influence F1 and F2 values, and consequently, the dispersion of vowels in the vocalic space, such as the effect of lexical stress. In this regard, Silber-Varod et al. (2019) observed that lexical stress has different effects on different vowels in Hebrew, with the vowels /a/ and /e/ being the most clearly affected, with a tendency of centralization in the unstressed condition. The same centralization tendency was observed by Santiago and Mairano (2018) for the Spanish vowels /a/ and /o/, and by Barbosa (2012) for the Brazilian Portuguese vowel /a/, implying a higher dispersion of the matching stressed vowels in the vocalic space in

such languages.

Furthermore, a combined effect between speech tempo and stress on the overall size of vowel spaces has been verified by Fourakis (1991) in American English, with a tendency for larger vowel spaces for the slow stressed condition and smaller for the fast unstressed condition.

It is noteworthy the fact that formant frequencies are among the most frequently assessed parameters in the forensic speaker comparison (FSC) practice. In general, the FSC task is commonly carried out considering F1, F2, and F3, due to the inaccessibility imposed by the telephone bandwidth to higher frequencies (GOLD; FRENCH, 2011). Notwithstanding, recent technological advances in telephone communication (e.g., WhatsApp and Telegram) have widened the possibilities of using higher frequencies, such as F4 and F5, in evidence materials (CAO; DELLWO, 2019).

In terms of vowel acoustic specification, the quality of a vocalic segment is primarily correlated to the frequency of the first and second speech formants (ROSE, 2002), namely F1 and F2, produced by the proper manipulation of mouth opening and constriction location, respectively. Variations in these dimensions are related to the degree of articulatory precision required for producing a given vowel (STEVENS; HOUSE, 1955). Furthermore, higher formants such as F3 (i.e., somewhat related to vowel configuration, as in the case of front rounded vowels) and F4 are commonly referred to as being considerably speaker-specific, conveying more speaker discriminant information (CAO; DELLWO, 2019). These formants are also seen as related to voice quality aspects in spoken and singing voice (SUNDBERG, 2015).

According to Traunmüller (1984), the position of the higher formants in the spectrum, such as F3, and F4, is largely determined by the vocal tract length. Moreover, in the experiment conducted by Stevens and House (1955), the authors observed that while F2 tended to increase in frequency as the point of constriction moved forward from the glottis, there was only a small increase in F3 as the mouth opening increased in size and became less rounded during the referred movement. In general, the rate of the increase depended mostly on the size of the constriction.

Regarding the resonance production in the vocal tract, both one-dimensional and three-dimensional acoustic experiments confirm that the fourth formant frequency (F4) is generated in the proper larynx. The experiments also suggest that the laryngeal cavity (LC) is acoustically independent of other parts of the vocal tract (TAKEMOTO; ADACHI, et al., 2006; TAKEMOTO; MOKHTARI; KITAMURA, 2010). As reported by Takemoto, Mokhtari, and Kitamura (2010), experiments with acoustic models showed that the elimination of the LC also resulted in the suppression of F4 while retaining other formants. The same study also indicated that F4 is considerably sensitive to LC shape changes, as in the case of constrictions in the ventricular area, which increase F4 while other formants remain nearly stable. In the same direction, the experiment carried out by Cao

and Dellwo (2019) with sustained monophthongs also has shown that when F4 increased across different speakers, only F5 seemed to follow in most situations, revealing a positive correlation between these formants.

1.2.2 Vowel formant frequency analysis in genetically-related speakers

Relatively few studies have investigated the speech or voice characteristics of genetically-related speakers using acoustic analysis, as in MZ twins. Most of the research has been carried out with English-speaking subjects, limiting the possibilities of cross-language comparisons. Other languages, however, have been addressed in recent years, such as German (WEIRICH, 2010, 2012; WEIRICH; LANCIA; BRUNNER, 2013; WEIRICH; SIMPSON, 2014), European Spanish (FERNÁNDEZ, 2012; SAN SEGUNDO, 2014; SAN SEGUNDO; TSANAS; GÓMEZ-VILDA, 2017; SAN SEGUNDO; YANG, 2019), Shanghainese and Mandarin (ZUO; MOK, 2015). It is also relevant to mention that small-sized studies are persistent in research involving identical twin speakers, which does not invalidate the observations but suggests some caution regarding generalizations.

In a pioneer study, Nolan and Oh (1996) looked at vowel formant frequencies in a group composed of identical twins. In their analysis carried out with three pairs of identical twins, aged between 20 and 23, who had grown up together, it was possible to find discriminable differences in controlled speech material. The experiment was performed with readings of a word list containing /l/ and /r/ before vowels in words such as "*lip*", "*let*", "*lot*", "*lug*", "*rip*", "*rap*", and "*rock*" of Southern British English. In the experiment, variations for F1, F2, F3, and F4 were observed for specific twin pairs, which according to the authors, suggest that twins are not necessarily phonetically identical and can appropriate themselves of the same articulatory freedom as other speakers to opt for alternative phonetic realizations.

To the best of our knowledge, the only research in Brazilian Portuguese (BP), which approached identical twins within the forensic phonetic perspective, was conducted by Figueiredo (1994). In this study, the vowel formant frequencies (F1-F4) of only one identical twin pair were assessed through reading. The main results suggested that the mean formant frequency of the twins was considerably similar, especially for F2 and F4. Conversely, the identical twin pair was found to behave differently for F1 and F3, as evidenced by the statistical analysis. The results also indicated a significant interaction in both F1 and F3 regarding the variables "*speaker*" and "*vowel quality*", with only the nasal vowel [ɐ̃] displaying significant differences. According to Figueiredo (1994), it is likely that the speakers revealed different strategies for nasal quality production, involving, for instance, different degrees of velum opening.

In an acoustic and perceptual experiment with Australian-English vowels involving three pairs

of similar-sounding MZ twins during a conversational speech task, Loakes (2003) observed significant acoustic differences for the productions of /æ/ and /ʌ/ in terms of F1. No significant differences were found regarding F2 and F3 for any of the eleven vowel segments analyzed. Concerning F4, significant differences were found for all vowels except for the back vowels /u/ and /ʊ/. As argued by the author, the results cannot disprove the assertion that each person in the population has a measurably unique voice.

In a follow-up study carried out by Loakes (2004), with the goal of analyzing the degree of speaker-specificity of Australian-English monophthongs in the speech patterns of four male twin pairs (three MZ and one DZ) based on F2 and F3 estimates, it was observed that the highest discriminant parameters for this group were F2 and F3 of the close-front vowel /ɪ/, followed by F3 of the front vowel /æ/, F3 and F2 of the close front vowel /i/ and F3 of the front vowel /ɛ/. The results of this investigation have presented some clear evidence that front vowels in Australian-English were more speaker-specific than other vowels. Overall, the six most speaker-specific parameters in this study were from front vowels, and four of the five most speaker-specific parameters were from close-front vowels.

Similarly, in Loakes (2008), a comprehensive acoustic study with static formant analysis was carried out with similar-sounding twin pairs, in which a forensically realistic material consisting of spontaneous conversational speech, composed by direct (four twin pairs) and telephone recordings (five twin pairs), as well as non-contemporaneous data were assessed. As in the previous one, an important aspect of this study regards the methodological approach applied, carried out with the analysis of same-segment vowel tokens, rather than strictly controlling to phonetic context. The formant analysis (F1-F4) of lexically stressed Australian English vowels also included a variability of phonetic contexts. In the referred experiment, differences in vowel realization were found where some speakers had consistently more fronted vowels than their twins, which has also been confirmed through acoustic analysis. A re-analysis of the data using a likelihood ratio (LR) approach confirmed that twins' speech was much closer in F-patterns than pairs of unrelated speakers in the corpus. The results also revealed that the inter-speaker variation was greater than the intra-speaker variation concerning the parameters assessed.

Furthermore, an acoustic study was conducted by Whiteside and Rixon (2003) with a pair of Southern Irish male MZ twins (T1 and T2) aged 21 years old and their age and sex-matched sibling, who participated in the experiment two years later. It was observed by the examination of F2 vowel onsets and targets that the MZ twin pair displayed F2 values and coarticulation patterns that were more similar than those of their age- and sex-matched sibling. According to the authors, the higher correspondence between MZ twins might be explained by greater physical similarities between the vocal tracts of T1 and T2 when compared to their sibling.

Later on, Weirich (2010) explored the articulatory and acoustic inter-speaker variability in the production of German vowels in stressed and unstressed conditions. The vowels /i/, /u/, and /a/, which were embedded in verbs included in carrier sentences, were produced by two female and one male MZ twin pairs and two female DZ twin pairs, aged between 20 and 34. The formants F1-F4 of each vowel were measured and then compared within the pairs. The results demonstrated that the inter-speaker variability was equally distributed for two out of the three MZ pairs, regarding low (F1-F2) and high formant frequencies (F3-F4). In contrast, within the DZ twins, F1 and F2 accounted for approximately 35% of the differences. According to the author, since the size and form of the vocal tract have a strong influence on the higher formants of a speaker, MZ twins are expected to show less inter-speaker variability in F3 and F4 than for F1 and F2, depicting higher formants as being more dependent on physiology and less influenced by alternative articulatory strategies. However, a considerable number of higher formant differences were still found between MZ twin pairs.

Concerning the analysis of the lexical stress in Weirich (2010), DZ twin pairs revealed more inter-speaker variability in unstressed than in stressed syllables. Conversely, two out of three MZ twin pairs showed no differences in the unstressed condition, only in the stressed context. According to the researcher, physiology seems to have a stronger influence on the production of a vowel when it is produced in an unstressed syllable (i.e., less acoustically salient).

From a dynamic perspective, there is also evidence of vowel formant transitions as being remarkably speaker discriminatory, with identical twins displaying consistent variations in the production of vocalic sequences, such as diphthongs (SAN SEGUNDO, 2014; SAN SEGUNDO; YANG, 2019; ZUO; MOK, 2015).

Within a forensic-phonetic scope, San Segundo and Yang (2019) investigated the dynamic acoustic properties of 19 vocalic sequences of Standard Peninsular Spanish, assessing their potential for forensic speaker comparison while using curve-fitting estimated coefficients. The study was carried out with male MZ and DZ twins, brothers, and unrelated speakers, aged between 18 and 52. The experiment was designed in a way to elicit specific vocalic sequences during a collaborative task, namely, finding out the missing information in a fax copy. The main outcomes of this study were that the fusion of 19 vowel sequences outperformed the analysis carried out with individual transitions, and the geometric-mean combination method outperformed the logistic regression analysis. In the experiment, MZ twins were found to deteriorate the system's performance for all vocalic sequences. Moreover, the fact that higher or lower similarities were found depending on the specific twin pair would indicate that the parameters assessed are not completely and uniquely genetically influenced, as claimed by the authors.

Despite the acknowledged high speaker-discriminatory potential of dynamic parameters, such

as the analysis of formant transitions, some practical implications must be considered. From a forensic-phonetic perspective, it is relevant to mention that vocalic transitions may require more data to account for the variation observed among individuals. In contrast, monophthongs may be regarded as substantially more recurrent, as evidenced by the higher frequency of monopthongs in relation to diphthongs during the analysis of the present study. Also, for this reason, most of the studies carried out with dynamic parameters are undertaken with controlled speech material, as to prompt the very same vowel sequences across individuals, as in Zuo and Mok (2015); or require an ad hoc design to induce the intended sequences in spontaneous speech (SAN SEGUNDO; YANG, 2019). This factor may solely reduce the practical application of dynamic features in some forensic speaker comparison contexts.

It may also be relevant to mention that static and dynamic features can both help explain how different speakers are from each other, providing complementary and useful data. In Weirich (2012), for instance, while MZ twins were found similar and DZ twins different for dynamic formant patterns (F2 and F3 transitions in sibilant-schwa sequences), when static parameters were included in the analysis, namely the spectral center of gravity, mean spectral peak, and mean formant measures, both MZ and DZ twins were found substantially different. According to the authors, physiological factors might present more influence on dynamic parameters, which was furthermore corroborated in Weirich, Lancia, and Brunner (2013), where higher similarities for looping trajectories in tongue movements were observed for MZ twins when compared to DZ twins and unrelated speakers for VCV sequences (vowel-consonant-vowel), as evidenced through electromagnetic articulography (EMA).

With regard to phonetic studies in Brazilian Portuguese, there is a lack of evidence concerning how robust and consistent vowel formant frequencies are when considerably similar speakers are compared. Additionally, very little is known about the limits of phonetic variation for related individuals, as genetically-related subjects. The present study aims to advance this understanding.

In the next section, general aspects regarding the analysis of speech timing patterns, which is also a matter of research of the present thesis, will be addressed. Essential aspects concerning the production and perception of tempo in speech will be acknowledged to provide the background of this study's topic and the basis on which future analysis will be drawn.

1.3 On the speech timing domain

The concept of time is inherent to the description of any dynamic system, which also extends to the realm of speech production and perception. Notably, duration patterns can be identified in

many linguistic organization levels and are systematically exploited by languages when implementing contrast (SCHWARTZ et al., 1997a), from the segment, the syllable up to higher linguistic domains. In that regard, "*timing*", as described by the organization of duration throughout utterances, can be assessed at different linguistic levels depending on the researcher's interest.

Apart from widely acknowledged general linguistic temporal patterns, how do individuals vary in speech timing measures when speaking in the same language and dialect? Can such a variation, within limits imposed by the production system, be regarded as speaker-discriminatory? Moreover, what are the effects of reducing common sources of inter-speaker variation on speech tempo patterns? The present study represents an attempt at addressing such questions with special consideration to spontaneous speech.

1.3.1 General aspects

Measuring speech tempo from a signal-based approach requires some methodological criteria to be considered and some experimental decisions to be made. As pointed out by Jessen (2007), the primary decision concerns the linguistic unit based on which the parameter will be estimated, namely *the unit of measurement*. This could be segments, syllables, words. According to the author, although there are arguments for each of these choices, the most commonly used unit is the syllable, especially in the domain of forensic speaker analysis. Secondly, it must be decided which linguistic dimension the estimations will be based on, either based on the "canonical" (*abstract*) or "realized" (*concrete*) units. It is worth mentioning that, depending on the dimension chosen, different estimations may be obtained; as syllable reductions are reasonably common in spoken corpora, an analysis on the basis of produced syllables, for instance, often tend to result in lower speaking rates.

A third methodological aspect mentioned in Jessen (2007) regards the size and kind of speech unit used for the analysis. Concerning this aspect, estimations can be made over the entire duration of a recording, yielding a global measure expected to portray a speaker's habitual temporal speech behavior, or over smaller portions throughout the recording, expected to capture local temporal variations that may be relevant for the analysis. Notably, the *size* and kind of *speech unit* adopted also have practical consequences, as verified in the present study. Smaller units or units that are considerably more frequent tend to yield higher amounts of data, which, in practice, may enhance the statistical power of the analysis being performed, given that statistical models are unquestionably sensitive to the variable's number of observations (BERBEN; SEREIKA; ENGBERG, 2012). In that regard, an extraction based on syllables or vowel segments naturally tend to result in more data points than measures extracted from longer speech intervals/units, such as

words or intonation phrases.

In the present thesis, a syllable-sized duration unit named V-V unit (i.e., vowel-to-vowel phonetic unit) was adopted, on which basis speech rate, articulation rate, and syllable duration were assessed. The relevance of this unit in psychoacoustic terms is broadly discussed in Pompino-Marschall (1989), and its explanatory potential of the speech rhythm production explored in Barbosa (2007). Such a phonetic unit comprises all the segments uttered between two consecutive vowel onsets, with the onset of the following vowel defining the beginning of a new V-V unit. It has been studied and employed among others by Barbosa (2007), Constantini (2014), Gonçalves (2017) and Arantes, Eriksson, and Lima (2018), with its application tracing back at least to Manuel and Krakow (1984) and Öhman (1966).

Finally, adding to the before-mentioned criteria, another crucial aspect regards the treatment given to pauses. Notably, the inclusion or exclusion of pauses (i.e., silent and filled pauses) in the speaking rate estimations may yield different outcomes. Electing one parameter over the other should be motivated by what is being analyzed and the research goal. Moreover, this fundamental difference contrasts two of the most commonly used speech tempo parameters, namely speech rate, when pauses are kept in the intervals, and articulation rate, when pause duration is not included when calculating the total sample duration (KÜNZEL, 1997; TSAO; WEISMER, 1997). Regardless of whether silent pauses should be included or excluded in the analysis, their minimum length must be defined and controlled to prevent the inclusion or exclusion of silent intervals that are not related to pausing behavior (e.g., silent closure periods in the acoustic signal). As remarked by Künzel (1997), based on the findings in the literature concerning automatic measurements, a threshold value of 100 ms appears adequate in order to prevent counting occlusion phases of plosives as silent pauses.

1.3.2 Tempo in speech: aspects of perception

Even though the focus of the present thesis is on the produced speech timing differences, general aspects concerning how humans perceive differences in the temporal dimension of speech must be acknowledged, as the evaluation of speech samples from an auditory-perceptual perspective is a common and recommended practice in forensic phonetics; adding valuable information to the acoustic analysis, such as the impressionistic magnitude of the acoustic differences observed.

As far as the perception of tempo in speech is concerned, different factors are known to influence the way in which listeners perceive speech tempo differences between two or more stimuli, such as the relationship between intended and produced units, the magnitude of the difference, the direction of the change, and who is judging it, if the speaker himself/herself or other individuals.

In an experiment performed by Koreman (2006) with German speakers using spontaneous speech stimuli, the author aimed at assessing whether the perceived speaking rate is determined by the number of realized phones (*realized*) or is dependent on the intended speech production (*intended*), defined as the potentially realizable phones according to a canonical form of the words, possibly defined in the mental lexicon of speakers, the lexicon form. This research question is justified because realized forms, the speech material, can differ substantially from the underlying, abstract forms.

In this experiment, the intended and realized articulation rates were derived from intonation phrases' transcription by counting the number of intended and produced phones divided by its duration. Only phrases with a duration within 1 and 1.5 s were selected to control the effect of phrase duration. A perceptual scaling test was performed afterwards, in which listeners represented by 12 selected subjects had to judge each stimulus phrase, including their speech samples, in a continuum scale from -3 (too slow) to 3 (too fast), also including intermediate values (very slow, quite slow, normal, quite fast, very fast). This experiment's overall findings suggest that listener judgments of their own speech rate reflect both the realized and intended rates. Speakers clearly perceived differences between stimuli in both conditions. Even if only the intended (canonical) rate of two identical phrases differs from one another, listeners did tend to notice a difference in speech rate. In this specific situation, the phrase with the faster-intended rate, represented by a higher number of canonical forms, was perceived as faster. According to Koreman (2006), this experiment shows that the perceived speech rate is also determined by the listener's knowledge of what articulations are implied by a particular utterance.

In the experiment conducted by Koreman (2006), the author also investigated the hypothesis that speech rate perception is influenced by the listeners' own speaking habits, based on a hypothetical projection of their own speech behavior. In this regard, classifying individuals as fast or slow talkers would be subjected to one's own notion of what is to speak at a regular pace. The study's outcomes contradict this hypothesis. Subjectively fast speakers did not judge stimuli differently from subjectively slow speakers, which suggests that listeners' own speaking habits do not appear to affect their perception of speech rate. Conversely, the results of the experiment carried out later on by Bosker (2017) points to another direction. While also testing whether one's own speech also influences his or her perception of other talkers, the author found evidence suggesting that, in fact, variation in speech production may induce variation in speech perception, which tends to be attenuated when exposed to someone else's speech, possibly due to a speaking-induced suppression that acts by reducing one's sensitivity to their own speech rate. According to the author, since temporal characteristics of our own speech may affect our perception of others, dialogic communication may be facilitated when talkers converge toward their interlocutor's speech rate.

The realization that speakers vary in their speech tempo behavior is commonplace. Variation in the temporal dimension of speech can be found intra-individually, as for communicative and contextual demands, as well as in comparison inter-subjects, as the outcome of different speaking habits. Nonetheless, within the domain of speech science, observed and perceived differences are utterly different matters. In light of that, as an attempt to quantify the just-noticeable difference (JND) for speech tempo, Quené (2007) conducted a series of perceptual experiments in Dutch involving tempo-manipulated and non-manipulated stimuli. The stimuli consisted of 20 speech fragments, excerpted from longer text passages that resembled short news items, without major pauses. The speech fragments were then accelerated to the relative duration from 0.80 to 0.95 relative to the original duration and decelerated to the relative duration from 1.05 to 1.20, with uniform temporal compression or expansion throughout the fragments. Different subjects were recruited for each of the three experimental sessions. In the first session, the listeners' task was to indicate whether the two versions were the same or different (*same-different design: 2IAX*); in the second experiment, the listeners' task was to indicate whether the first or the second fragment was the faster one (*2IFC design*); finally, in the third experiment, their task was to rate the perceived naturalness of each stimulus on a 7-point scale (*scaling*). This session was a control experiment to confirm if speakers were basing their responses on acoustic artifacts generated by the manipulations. As argued by the authors, if the manipulated stimuli contained phonetic artifacts, then the manipulations would be easy to identify and result in lower naturalness ratings. This hypothesis was disproved. The results obtained suggested an estimated JND of about 5% of the base tempo of speech, which points that tempo variations exceeding this threshold are likely to be noticeable, and as a consequence, relevant in speech communication. Moreover, direction-of-change, if accelerated or decelerated, and amount-of-change also displayed a significant effect, suggesting that listeners behave differently depending on these factors. As for direction-of-change, estimated JNDs in speech tempo of about –4% for accelerations and +6% for decelerations were observed.

It is worth drawing attention to the acknowledged discrepancy between produced and perceived speech rate, as signaled by some studies (LANE; GROSJEAN, 1973; CARTWRIGHT; LASS, 1975). In an experiment involving a psychological scaling procedure for estimating the correspondence between the measured and perceived rate of continuous speech, the estimations observed by Cartwright and Lass (1975) indicated that although a linear trend was present, there was no one-to-one relationship between measured and perceived speech rate. In other terms, changes in perceptual magnitude were not equivalent to changes in physical magnitude, with changes in sensation growing substantially faster.

Also, Lane and Grosjean (1973) carried out a study to test the generalized hypothesis of

different growth in terms of speech rate perception depending on who is judging it, the speaker (*autophonic judgment*), or an external listener (*extraphonic judgment*). This investigation was done by scaling the autophonic and extraphonic reading rate, using methods of magnitude production and estimation. The obtained exponents of the two power functions show that when a speaker doubles his reading rate, he perceives a sixfold increase, whereas a listener perceives less than a threefold increase. Thus, the results corroborate the assumption that speech rate perception grows more rapidly for the speaker than for his listener. According to the authors, this disparity– also acknowledged for the voice loudness perception– indicates that the speaker's judgments of rate are not based solely on the acoustic signal. This may, in part, be explained by the fact that, when a speaker judges the characteristics of his speech, the cues available include tactile and proprioceptive feedback, as well as bone-and-air-conducted sidetone. Notwithstanding, when an individual is asked to judge someone else's speech, he is deprived of all these cues (except for the last one), and, as a listener, his judgment is conducted differently. Moreover, as mentioned by the authors, the autophonic rate adds to the list of other continua on which psychological magnitude is a power function of the stimulus, and also that the sensory mechanisms mediating the speaker's perception of his speech amplify constant stimulus ratios into much larger constant subjective ratios.

1.3.3 Tempo in speech: aspects of production

Many factors, from different orders, are known to account for the variation in produced speech rate, as in the case of linguistic and extra-linguistic factors. Some of them are speaking style, dialect, phrase length, age, sex, and speakers' emotional state. Other factors within the human communication health domain have also been suggested, such as cognitive decline and speech-language disorders (e.g., stuttering disorder).

In the study conducted by Bona (2014), speaking style was found to significantly affect the produced speech rate, articulation rate, frequency of pauses, pause ratio, mean pause duration, and the standard deviation of pause duration within speakers. Speech and articulation rates were lower in retelling (text recall) when compared to other speaking styles, such as reading and spontaneous conversation. Furthermore, pauses tended to be more frequent and also longer for the retelling condition. The highest speech and articulation rates and the shortest pauses were observed for the reading style.

In an experiment conducted by Jacewicz, Fox, and Wei (2010), the researchers also found significant differences when comparing articulation rate in spontaneous unconstrained talks and sentence reading. The results showed that speakers who had a faster speaking rate also had a faster

reading rate. When the reading rate increased by one syllable per second, speaking rate increased by 0.69 syllables per second. According to the authors, the outcomes reveal a relationship between the articulation rate in speaking and in reading, which suggest the existence of the same underlying motor control mechanism for speaker-specific rate.

Inter-speaker variations in temporal characteristics of speech have also been reported as a consequence of aging. In the experiment conducted by Bona (2014), while assessing different parameters, such as speech rate, articulation rate, frequency of pauses (per 100 words), pause ratio (%) in the total speech time, and mean silent and filled pauses duration, speaker's age had a significant effect on speech rate, articulation rate, and in the frequency of pauses, corroborating the study's hypothesis that the speech passe tend to decrease during the lifespan. No significant differences between age groups, however, were observed concerning filled pauses ratio and mean filled pauses duration.

As pointed out by Bona (2014), pauses can be used as an strategy for resolving difficulties in speech planning and articulation. The results obtained by the researcher suggests that although elderly speakers present more difficulties in tasks such as lexical access, when compared to younger speakers, they seem to employ speaking strategies to compensate for these limitations, which appear to reflect on parameters other than pauses.

In the same direction, while assessing average articulation rate (phonetic syllable-based) in American English under the influence of age, in which 192 speakers from two dialectal areas ranging between 8 and 91 years were recorded in spontaneous speech and sentence reading, Jacewicz, Fox, and Wei (2010) observed that elderly speakers in fact tend to produce lower articulation rates. In the experiment, differences in reading rate were significant only for young adults who read faster than speakers from all other age group and for the oldest adults aged over 66 years who read significantly slower than the remaining adults. According to the authors, the effects of age in reading rate might be potentially linked to reading skill variations among individuals. Since most of the young adults in the study were college students, the fact that they were engaged extensively in reading on a daily basis, could influence their performance. As for spontaneous speech, there was a tendency for articulation rate to increase as speaker age increased, achieving its peak value around the mid 40s, (i.e., 45 years-old). Nonetheless, the degree of articulation rate increase for each additional age year was not constant. The size of the increase reduced around the age peak value. After the peak, speaking rate was found to decrease unevenly for each year increment of age.

Similarly, Jacewicz, Fox, and Wei (2010) observed a significant tendency for older speakers to produce shorter phrases than younger subjects. The log of phrase length decreased by 0.0011 for each 1-year increment of age in their study. Quené (2008) also found evidence for a variation in

phrase length as a function of age, with phrase length decreasing with age for young adult speakers, and then gradually increasing for older adult speakers. The amount of within-speaker variation in phrase length also varied as a function of the speaker's age. Note, however, that there is a considerable age range difference between these two studies, with the first being conducted with speakers up till 90 years-old, whereas in the second the oldest person was 60 years-old. The fact that only teachers participated in the latter study (i.e., subjects who have experience with public speaking), may also have some implication on the interpretation of the results.

Neuromuscular and sociolinguistic components have also been suggested to determine an individual's habitual speaking rate, as suggested by Tsao and Weismer (1997). According to the authors, while the former component may explain why individuals with relatively slow and relatively fast habitual speaking rates would display different upper limits for maximum speaking rates, as imposed by certain neuromuscular constraints; the latter would consider the influence of both social learning and linguistic constraints on the habitual speaking rate. As remarked by the authors based on experimentation, despite the evidence suggesting that neuromuscular constraints seem to play a role in determining an individual's habitual speaking rate, none of these components alone (i.e., neuromuscular and sociolinguistic) can account for the control of individuals' speaking rates. Furthermore, additional evidence deriving from a follow-up study, cf. Tsao, Weismer, and Iqbal (2006), seemed to corroborate the referred biological basis for inter-talker rate differences, as evidenced by the analysis of a new set of speech materials.

In terms of the variables sex and dialect, Jacewicz, Fox, and Wei (2010) noted that males tended to speak significantly faster than females. However, this difference was only significant in spontaneous speech, not being verified for reading rate. Furthermore, although the temporal difference between males and females was present, it was smaller than that observed between dialects, namely American English spoken in Wisconsin and North Carolina. In the dialectal comparison, Wisconsin speakers had a significantly faster speaking rate than North Carolina speakers. Even though the referred study has not been carried out within a forensic-phonetic frame, this finding may point the relevance of considering the dialect impact when comparing individuals from different populations, or subjects who have potentially migrated to different dialectal areas.

Processing speed has also been suggested to impact on temporal aspects of speech tempo, as evidence from pathological speech clearly supports the intricate relation between cognition and speech tempo. While analyzing the association between articulation rate and processing speed in a group of 122 patients clinically diagnosed with multiple sclerosis, Friedova et al. (2019) found a significant association between these two components. As confirmed through linear regression analysis, slowed articulation rate was strongly associated with processing speed decline in the group of tested patients.

From a speech pathology perspective, as in the stuttering disorder, Arcuri et al. (2009) observed that, for Brazilian Portuguese speakers, the rate of fluent speech during a repetition task, measured in terms of speech rate considering the number of vowel-to-vowel segments (V-V units) divided by the total sum of the segments' duration, varied among individuals as a function of the stuttering severity. The group with mild/moderate stuttering displayed higher and similar speech rates compared to the group with severe stuttering. Individuals with diagnosed severe stuttering displayed lower speech rates compared to the other groups even during their fluent turns. This outcome suggests an inverse relation between stuttering severity and speech rate, in which the higher the severity of the disorder, the lower the speech rate. According to the authors, this difference seems to be related to difficulties in motor programming, affecting mainly the rhythm and the timing of speech.

Variation in speaking rate has also been found to have implications on speech intelligibility. It has been reported that words spoken at a slower speaking rate are significantly more intelligible than those spoken, as reported by a study carried out with hearing-impaired and non-impaired listeners in Japanese (HOSOI et al., 1992). Notwithstanding, it has also been suggested that "clear speech" can also be achieved in faster speaking rates, as mentioned by Krause and Braida (2002). According to the authors, intelligibility tests suggest that clear speech has some inherent acoustic properties that contribute to its high "clarity" without altering rate, and that the identification of these acoustic properties could lead to improved signal-processing schemes, as in the case of hearing aids.

Speakers' emotional state has also been reported to influence the speaking rate. As observed by Siegman (1978) when rehearsing about some experiments in which speakers were submitted to interviews, the authors observed that a subjective increase in anxiety on the individuals triggered by the intimacy level of the topic was associated with differences in average pause duration, pause duration ratio, and pause frequency ratio, reflecting differences in speech rate estimations. In that study, interview questions were classified as displaying topics/themes with different levels of intimacy, ranging from low to high.

Apart from the before-mentioned factors, variation in speech timing characteristics can also be explained as a function of linguistic factors. From this perspective, considerable attention has been paid on the effects of phrase length on speech tempo.

In the study conducted by Quené (2008) with Dutch speakers (teachers) in spontaneous speech, speech tempo, namely articulation rate, was analyzed by means of multilevel/mixed-effects modeling including the speaker's sex, age, country of origin, dialectal region and phrase length. The study's outcomes revealed that speech tempo was partly determined by phrase length, due to a mechanism known as anticipatory shortening, which seems to account for longer phrases, contain-

ing more syllables, being spoken at a faster rate and shorter average syllable duration.

Notwithstanding, there is no complete agreement in the literature whether an anticipatory shortening mechanism regarding speech tempo exists. Conversely to what was observed by Quené (2008) for Dutch speakers, an opposite trend was found in the experiment by Jacewicz, Fox, and Wei (2010) with American English speakers, in that longer phrases containing more syllables were spoken at a faster rate in Dutch but not in American English, where shorter phrases (containing less syllables) were found to be spoken faster. According to Jacewicz, Fox, and Wei (2010), one possible explanation for this cross-study divergence relates to the fact that while participants in the Dutch corpus were school teachers, who possibly have by practice a better command of spoken language in terms of articulatory planning, verbal monitoring and effective use of pauses, American English speakers varied in their professional and educational background, and consequently in their experience with spoken language usage.

Regarding the present study, it is worth mentioning that those variables acknowledged in the literature as bearing influence on speech timing measures, such as age, sex, speaking style, and dialect, are relatively well controlled, considering that only adult young male individuals from the same dialectal region were recruited. Moreover, all individuals were recorded in the same speaking style: a spontaneous dialogue over the mobile phone. Also, the possible effects of phrase length on the temporal measures are minimized since very representative data was used, containing many possible realizations from each speaker, extracted from different dialogue parts in the recordings, as better exploited in *Chapter 2* of the present thesis.

1.3.4 Forensic-phonetics studies on speech timing

Despite the relevance of temporal parameters for forensic speaker comparison, given their relative resistance to the limitations imposed by the telephone transmission system - one of the main drawbacks of spectral-based analysis, cf. Künzel (2001), very few studies have been conducted with temporal measures within the forensic phonetic scope. This observation motivates and justifies the relevance of the present study.

In order to provide an appropriate context, some of the relevant studies available on speech timing analysis within the forensic perspective are reviewed in the present section and their main outcomes are highlighted.

A comprehensive experimental study on articulation rate, within the forensic phonetic scope, was carried out by Jessen (2007) with a group of 100 German-speaking male subjects, ranging between 21 and 63 years of age– an average of 39 years. Articulation rate was assessed in three different conditions: face-to-face spontaneous speech, spontaneous speech over the telephone, and

reading. The speech context in the first two conditions was a descriptive task, namely the description of a set of pictures to a conversation partner. The analysis of articulation rate was carried out globally and locally considering only phonetic syllables. The speech unit chosen for the extraction of the measures was the "*memory stretch*", characterized by portions of fluent speech containing a number of syllables that can easily be retained in short-term memory. Global articulation rate was computed by the mean articulation rate across memory stretches, including standard deviation values. The number of syllables in a memory stretch ranged from 4 to 20 syllables, excluding filled, unfilled pauses, and syllable lengthening. The number of memory stretches were on average 25, 23, and 20 in task one, two, and three, respectively.

Regarding articulation rate mean values, it was found that both direct and telephone-transmitted speech deviated significantly from reading, in which the parameter was found higher for the latter condition. Considering the analysis of the measure's standard deviation, direct speech and telephone-transmitted speech were again found distinct from reading, in which a lower standard deviation was observed. Moreover, no significant differences were observed between direct and telephone-transmitted spontaneous speech. Finally, according to Jessen (2007), the observation that intra-speaker variability across reading and spontaneous speaking was greater for articulation rate (sd) than for its mean, suggests the analysis of mean articulation rate as more viable from a forensic viewpoint. Furthermore, as an attempt to compensate for a speaking style mismatch, this knowledge on the differences observed between reading and spontaneous speech could be used when guidelines for forensic speaker comparisons are proposed.

As for Brazilian Portuguese (BP), a phonetic experiment was conducted by Gonçalves (2017) on speech tempo parameters with a realistic forensic data-set. The speech material comprised of spontaneous samples from intercepted telephone conversations, and direct recordings carried out with the same speakers within a time gap ranging from 10 to 38 months later. Seven speakers were recruited, five males and two females, aging from 14 to 31 years (mean: 24 years) in the first speech sample, and from 15 to 33 years (mean: 26 years) in the second sample. Global and local speech and articulation rates were assessed in the study, in comparisons inter- and intra-subjects. Following a general and expected trend, higher values were obtained for articulation rate in comparison to speech rate, as consequence of pauses intervals in the latter measure. Although non-statistically significant, a higher variability was globally and locally observed for speech rate. With regard to the measurement procedures, namely global and local measurement method, differences were observed between global and local measures in speech rate, as evidenced through a paired t-test. As for articulation rate, both global and local measurement methods yielded similar outcomes. When assessing their variance through F-tests, it was noted that even though local measurements of speech and articulation rates tended to be less than in global measures, no statistical significance

was found. In terms of intra- and inter-speaker variability, the assessment of speaker-specificity through an intra-class correlation coefficient analysis (ICC) suggested that only articulation rate, (global and local) fulfilled the requirement of higher inter-subject variability in relation to the intra-subject variability.

The study conducted by Constantini (2014) within the forensic-phonetic domain with 35 male BP speakers from seven different regions in Brazil, aimed at assessing the speaker- and dialect-discriminatory power of eight acoustic parameters, including speech rate analysis, in different harmonic-to-noise ratios. The addition of different Gaussian noise levels (0.01 and 0.02 dB) to the recordings intended for observing how robust the parameters are to this kind of distortion, very often present in spoken forensic materials. In the referred experiment, the best performing parameters in differentiating the studied varieties were the spectral emphasis and the median speech fundamental frequency. No dialectal differences were observed in response to speech rate. As for the analysis considering the Gaussian noise addition, the first two parameters have shown to be systematically affected. The results revealed a rather more abrupt change in the spectral emphasis than for F0 median, reaching an increase of 55% (Gaussian 0.01) and 154% (Gaussian 0.02) in relation to the original recording. As for $f0$ median, the greatest change observed was of 3 Hz, which despite being statistically significant, possibly does not interfere in the discrimination of subject, according to the author. The outcomes of this study suggest that the analysis of speech rhythm-related parameters, as in the case of speech rate and vv units duration, is the most consistent approach when dealing with noisy audio samples.

The results obtained by Jessen (2007) and Constantini (2014) have important implications for the forensic speaker comparison practice, once it signals the relative resistance of temporal measures to variables commonly present in forensic casework, serving as an adequate alternative to situations where other parameters, as in the case of vowel formant analysis in telephone recordings, are not reliable, which is mainly due to limitations imposed by the so-called "*telephone effect*", represented by virtual changes in the acoustic outcomes as a consequence of the lower cut-off slope of the telephone band-pass, with direct consequences on the analysis being performed (KÜNZEL, 2001; BYRNE; FOULKES, 2007; PASSETTI, 2015).

Despite the considerable relevance of assessing temporal parameters in the forensic speaker comparison practice, given their relative resistance to the degradation imposed by the transmission system, few studies have addressed such measures from a forensic-phonetic perspective. This is especially true when considering spontaneous speech-oriented studies or experiments carried out with very similar speakers. The present study represents an attempt to fill this research gap.

As previously mentioned, the motivation for analyzing identical twins speakers' patterns resides on the fact that this group represents the extrapolation of what may be regarded as the highest

possible level of similarity across individuals. In that regard, knowing what is still particular of a person in such extreme conditions is a matter of interest for forensic phonetic science. Finally, it is noteworthy that very few twin studies have been carried out on the domain of speech timing. Therefore, the present thesis also intends to fill in such a gap.

1.4 On the fundamental frequency domain

Differences in voice fundamental frequency (hereafter, $f0$) average values and variability have been reported as a function of several components, ranging from physiological, psychological, linguistic, stylistic to socio-cultural factors (WILLIAMS; STEVENS, 1972; BRAUN, 1995; TRAUNMÜLLER; ERIKSSON, 1995; AWAN; MUELLER, 1996; HIGUCHI; HIRAI; SAGISAKA, 1997; ESCUDERO et al., 2009; ARANTES; ERIKSSON, 2019; PROBST; BRAUN, 2019; RILLIARD et al., 2013; SIGNORELLO et al., 2020). The different levels of contribution of these components on the voice/speech fundamental frequency reveal the multi-modal nature of such physical parameters, bearing repercussions on how individuals are perceived by listeners (KLOFSTAD; ANDERSON; NOWICKI, 2015).

1.4.1 General aspects

From a physiological viewpoint, the components underlying the determination of the voice fundamental frequency (i.e., the acoustic correlate of the vibration frequency of the vocal folds) have been decomposed and explored in detail by the classical *myoelastic-aerodynamic theory of voice production* proposed by Van den Berg (1958). As described by the author, this frequency depends on a number of five interdependent factors related to the vibrating part of the vocal folds, namely the effective mass; the effective tension; the effective area of the glottis during the cycle; the subglottal pressure; and the damping of the vocal folds. As pointed out by the author, when these factors are known, the frequency can be estimated.

Regarding the acoustic dimension, divergences in the voice fundamental frequency range as a function of age and gender is perhaps one of the most commonly acknowledged. As pointed out by Zhang (2016), children tend to have higher mean fundamental frequencies when compared to male and female adults, whereas females tend to have higher mean fundamental frequencies than males. The reason for such difference resides in the vocal fold geometry, including length, depth, and thickness of the vocal folds, with the tendency for larger vocal folds to display a lower frequency of vibration and vice-versa. Consequently, it can be said that an individual's fundamental frequency has an important genetic/organic motivation.

As for other factors bearing influence on $f0$ estimates, the influences of speaking style and speaking condition are of high importance, particularly for forensic studies. In this regard, while analyzing three distinct speech contexts, i.e., interview, telephone conversation with direct recording and telephone conversation with telephone-transmitted recording, Jong et al. (2011) observed that both telephone conditions resulted in slightly higher average values of the parameter in relation to the interview context, a difference of just 1 Hz for direct recording and of 5 Hz for telephone-transmitted recordings. According to the authors, two possible factors might be related to these outcomes in daily based situations; the first concerns the nature of telephone communication itself, represented by a limited bandwidth of the transmission channel and the absence of visual cues; and the fact that phone calls are often made in noisy environments, inducing "loud talking", which can generate an increase not only in the intensity measure, but also in the voice fundamental frequency, according to the authors. The effect of the telephone speaking style has also been investigated and these findings corroborated by Passetti (2018) for Brazilian Portuguese speakers.

The emotional state of an individual has also been proved to affect acoustic-phonetic parameters, among others, $f0$. In the study conducted by Higuchi, Hirai, and Sagisaka (1997), while analyzing $f0$ contours of Japanese sentences on the basis of $f0$ min, $f0$ amplitude at the phrase domain, and $f0$ amplitude at the lexical domain, the authors observed clear differences when comparing four speaking styles, e.g., unmarked, hurried, angry, and gentle. Amongst the analyzed speaking styles, more expressive differences were observed for the angry speech, which was characterized by a high $f0$ min and a minimum change in both phrase and lexical amplitudes, yielding reasonable flat $f0$ contours. The possible application of the outcomes in the domain of speech synthesis was also considered and suggested by the authors.

1.4.2 Forensic applications of $f0$

In terms of forensic-phonetic applications, the analysis of $f0$ may be regarded as one of the most frequently assessed parameters in speaker comparisons worldwide, applied by the majority of forensic experts (GOLD; FRENCH, 2011). One of the main justifications for such a high applicability regards the easiness with which the parameter can be assessed, being readily available even in small stretches of speech (LOAKES, 2006). Another crucial aspect regards the relative resistance of $f0$ measures to external variables, such as the microphone, the recording device, audio compression, and external noise levels, cf. Carson, Ingrisano, and Eggleston (2003), Fuchs and Maxwell (2016), Maryn et al. (2017), Jannetts et al. (2019) and Cavalcanti, Englert, et al. (2021).

As pointed out by Jessen (2009), in the forensic phonetic domain, $f0$ is mostly assessed by its "global" aspects, in which the average $f0$ and its standard deviation are among the most commonly explored measures. As the author mentions, while the average $f0$ in speech is regarded as displaying an important anatomical motivation, its variability is less dependent on structural factors and possibly better explained by other elements, such as individuals' manner of speaking.

In the past few years, a particular $f0$ estimate, the "base value of $f0$" or "$f0$ baseline" has been given attention in a number of forensic phonetic studies (ARANTES; ERIKSSON, 2014; SILVA et al., 2016; ARANTES; ERIKSSON; LIMA, 2018; PASSETTI, 2018), due to its higher resistance to many different sources of variability.

Initially proposed by Traunmüller and Eriksson (1995), and revisited in Lindh and Eriksson (2007), the base value[3] of $f0$ is grounded on a well-known observation in various types of motor activity, namely, the point of departure, a resting position (i.e., baseline). In that instance, as described by Lindh and Eriksson (2007), the $f0$ baseline can be regarded as a neutral mode and frequency of vibration to which the vocal folds return after prosodic or other types of frequency excursions. For that reason, such a measure is regarded as relatively more robust since it better represents the neutral articulation of a given individual. The speaker-discriminatory potential of such measure has also been corroborated by previous research (SILVA et al., 2016; ARANTES; ERIKSSON; LIMA, 2018).

More recently, in the experiment conducted by Arantes and Eriksson (2014) with the purpose of assessing the stabilization time of $f0$ long-term mean, median, and base value employing a change point analysis performed in recordings of the same text spoken in 26 languages, average stabilization points of an order of 5 s for the base value and 10 s for mean and median estimates were observed. Furthermore, the variance after the stabilization point was considerably reduced, with a reduction in variability of approximately 40 times for mean and median and more than 100 times for the base value. As remarked by the authors, these results show that stabilization points in long-term measures of $f0$ tend to occur earlier than what has been previously suggested.

It is noteworthy that most studies on $f0$ have been carried out with semi-spontaneous or read speech which justifies the need for a deeper understanding concerning the robustness levels of the parameters under unscripted speech conditions. Furthermore, the pertinence of assessing $f0$ extreme values and modulation estimates should also be examined, along with the implications of "*speech material*" on the discriminatory potential of $f0$ long-term estimates (e.g., directly com-

[3] In Traunmüller and Eriksson (1995) where the concept was first derived, the authors referred to it as the *base value* which is still the preferred concept name. The reader should be aware, however, that the same concept is also referred to as the baseline, as in Lindh and Eriksson (2007). In the present thesis, we use both terms to signify the same concept.

paring the discriminatory patterns of $f0$ long-term parameters extracted from lengthened vowels vs. connected speech).

1.4.3 Twin studies on $f0$

Regarding twin studies, within the speech and voice analysis domain, the voice fundamental frequency is probably one of the most commonly explored dimensions, turning it into a very fruitful field. Several studies have systematically reported a high correlation between monozygotic (i.e., identical) twins concerning the physical parameter of voice fundamental frequency. The majority of the studies have focused mainly on analyzing average and standard deviation values of the parameter, disregarding the possible speaker-contrasting potential of other fundamental frequency-related estimates, such as measures of $f0$ modulation. Some of these studies are commented on in the following.

With the purpose of assessing the application of $f0$ as a potential *phenotype*, Przybyla, Horii, and Crawford (1992) conducted a large-scale study with 122 twin pairs in American English, 50 pairs of female twins and 12 pairs of male twins aged between 15 to 75, from which only nine pairs were dizygotic (i.e., non-identical) twins. As is the case for most twin studies, there was also a higher number of female than male speakers, i.e., 50 pairs of females with a mean age of 41 years and 12 pairs of male twins with a mean age of 40. The analyzed speech samples consisted of readings, and the $f0$ (in Hz) was assessed following an automatic approach. Overall, when comparing monozygotic (MZ) and dizygotic (DZ) twins solely based on their $f0$ mean and standard deviation values, no significant inter-group differences were found. Furthermore, when calculating a matrix of correlations for $f0$ and age, weight, and stature, statistically significant associations of $f0$ with age and weight were found. When an adjustment for age and weight was applied, coefficients of correlations were separated according to zygosity, revealing a larger discrepancy for $f0$ measures in DZ twins. According to the authors, such a finding suggests a genetic component underlying the variation of the parameter. The results also appear to suggest the influence of organic covariates on $f0$, such as age and weight.

Furthermore, Debruyne et al. (2002) conducted an analysis on $f0$ and its intra-speaker variability during a reading task. A group of 30 female MZ twins and 30 DZ twins Dutch speakers, aged 15–29 years, and a control group consisting of 30 non-genetically related individuals of equal age were assessed. In the referred study, $f0$ was found considerably more similar in MZ twins when compared to DZ twin pairs, while no significant correlation was observed for unrelated peers, which according to the authors, is compatible with a genetic basis on the determination of the parameter. However, when intra-speaker variability of $f0$ was assessed, highly similar results were observed

between MZ and DZ twin pairs, suggesting that an individual's voice is determined by much more than genetic constitution alone.

The preliminary study carried out by Loakes (2006) with Australian English speakers also focused on assessing long-term fundamental frequency in a group composed of eight pairs of twins aged between 18 and 20. Three identical twin pairs and one non-identical twin pair were analyzed in reading and spontaneous speech across non-contemporaneous samples. The extraction of $f0$ mean, median, mode, and standard deviation values was performed in the midpoint of all labeled vowel tokens produced by the speakers. For that reason, the number of tokens sampled in spontaneous speech varied considerably across individuals. The outcomes revealed that speakers tend to fall within a specific $f0$ range and that twins tend to have a more similar mean long-term $f0$ than unrelated pairs. However, when individual $f0$ values were compared in a sequence/list, it was noted that twin pairs do not necessarily have the closest mean $f0$ values, with two pairs being closer to other speakers than to their twin brothers regarding their mean values. Such a difference turned out statistically significant. Evidence was also found suggesting that long-term $f0$ is relatively stable within-speakers in most cases, even when non-contemporaneous data sets were compared.

A comprehensive study on the voice quality characteristics of a group of 45 monozygotic twin pairs (MZ), consisting of 19 male and 26 female pairs, was carried out by Van Lierde et al. (2005) considering the effects of sex and age (i.e., male vs. female, and under vs. above 17 years old). One remarkable advantage of the study was the extensive age range. Speakers ranged from 8 to 61 years-old. The analysis consisted of both auditory perceptual evaluation using the *GRBAS* scale, cf. Wuyts et al. (2000), widely applied within the clinical voice setting, and acoustic measurements of a set of voice quality parameters, including $f0$ measurements in the middle points of sustained /a/ vowels. Overall, there was no significant influence of sex and age on the levels of vocal similarities in MZ, and high correlation scores were observed concerning their $f0$ estimates, which was also corroborated in the auditory perceptual evaluations. However, it is worth noting that, as it appears, the subjects were compared as a group (male vs. female, younger vs. older than 17) and not individually; hence, intra-pairs specificities were not the focus of the study.

A clinical case study on identical twins' voice characteristics was performed by Cielo, Agustini, and Finger (2012) with a small sample consisting of one male and one female MZ pairs of Brazilian Portuguese aged 20 and 28 years, respectively. The voice quality analysis was carried out through auditory-perceptual and acoustic analysis, including automatic measurements of $f0$ mean, median, standard deviation, maximum and minimum. A descriptive intra-twin pair comparison was made based on the twins' $f0$ values and their respective average differences. Moreover, a statistical analysis was performed to verify the extent to which the values deviated from standard normality scores, possibly helping the understanding of genetic and environmental influences on voice

patterns. Among the voice quality features assessed, the maximum phonation time was found to deviate from the normality for the female twin pair and one male twin, along with other dysphonic marks. According to the authors, environmental factors may be on the basis of the male twin pair difference, such as the practice of physical activity and the professional use of the voice by the twin with the highest maximum phonation time.

Another case study was conducted by Whiteside and Rixon (2013) on the $f0$ and temporal patterns of male Southern Irish monozygotic twins (T1 and T2) and an age- and sex-matched sibling (S) recorded two years later. The twin pair was aged 21 and their non-twin sibling 20 by the time of the recordings. Mean and standard deviation $f0$ values were automatically extracted from sentences in reading speech, and comparisons among the subjects were made. The results indicated significant differences regarding $f0$ mean values for all three genetically-related speakers. Conversely, no significant between-sibling differences were found for the $f0$ standard deviation regarding the comparison of the identical twin pair and between one of the twins (T2) and the non-twin sibling (S). Furthermore, when assessing the magnitude of the dissimilarities between sibling pairs through Euclidean distance measures, the smallest distances were observed between the identical twins (T1 and T2), mainly for $f0$ mean.

It is noteworthy that the commonly reported intra-identical twin pair similarities are not solely restricted to the $f0$ domain, as other voice quality parameters are also regarded as relatively similar for such individuals. In this instance, the experiment conducted by San Segundo and Gómez Vilda (2013) adopted a comprehensive approach on the analysis of twins' and non-twins' voice quality. Estimates of glottal source biomechanical parameters derived from vowel fillers, including $f0$, were used to assess phonation characteristics' similarities and differences in twins' voices. The participants consisted of 40 male native speakers of European Spanish, seven MZ pairs, five DZ pairs, four pairs of non-twin siblings, and four pairs of unrelated speakers (i.e., a control group) recorded during spontaneous conversation carried out in non-contemporaneous sessions. The main outcomes suggested a great influence of both genetic and environmental factors on the parameters assessed, as indicated by the relatively similar scores obtained for MZ and DZ twins, whereas non-relative subjects showed scores well around the background baseline. Furthermore, one MZ twin pair out of seven displayed low matching scores in an intra-twin comparison. According to the authors, such findings may suggest that phonation characteristics may be due to learned styles as much as to imprinted biological patterns. Later on, in San Segundo (2014), the contribution of both *"nature"* and *"nurture"* in the determination of one's voice characteristics was corroborated by the researchers when analyzing a considerably larger number of twin pairs while following the same approach.

1.5 Research questions and hypothesis

Considering that the present thesis comprises three large experiments, the research questions, and the hypotheses are presented independently for each experiment. The research questions comprehend both general and specific aspects, as follows:

1.5.1 General research question

In summary, the present thesis aims to provide an answer to the question: which acoustic-phonetic parameters and analysis dimensions, namely the spectral, temporal, and fundamental frequency dimensions, are the most inter-speaker discriminatory in comparisons performed between genetically related individuals and across all speakers (i.e., cross-pair comparisons), and therefore potentially relevant for the forensic speaker comparison application?

From a general viewpoint, the present thesis assumes that those parameters displaying the highest speaker discriminatory power in comparisons of genetically-related individuals (i.e., identical twins) could be regarded as substantially robust for the forensic speaker comparison application in Brazilian Portuguese.

1.5.2 Specific research questions

Experiment I: formant frequencies

- Which vowel formant frequencies (e.g., F1, F2, F3, and F4) are the most inter-speaker discriminatory in comparisons between genetically-related individuals and in cross-pair comparisons?
- Which vowel segments are the most speaker contrasting in Brazilian Portuguese?
- What are the effects of lexical stress on the discriminatory power of the formant frequencies assessed?
- Is it possible to differentiate identical twins on account of their vowel formant frequencies while taking part in spontaneous dialogue?

A few hypotheses have been suggested regarding the research questions of experiment I:

i. It is expected that higher formant frequencies will be relatively more speaker discriminatory than lower formant frequencies. ii. It is predicted that the cardinal vowels displaying higher distances from their neighbors in terms of F1 and F2 might be the most speaker discriminatory

ones. In that sense, the acoustic distance between vowels would play an essential role in determining the level of phonetic variation. iii. Regarding the stress component, based on previous literature findings, cf. Weirich (2012), it is expected that stressed vowels may display a higher speaker-discriminatory power in intra-twin pair comparisons, whereas unstressed vowels may be more discriminatory when considering comparison involving all speakers. iv. Although identical twins are expected to be remarkably similar regarding their formant frequency patterns, differences will still be observed between such individuals.

Experiment II: speech timing

- Which set of speech timing parameters (e.g., speech rate, articulation rate, V-V unit duration, silent pauses) are considerably more speaker-discriminatory in comparisons performed between genetically related individuals and across all speakers (i.e., cross-pair comparisons)?
- Which speech timing domain/category, namely, macro, micro, and pause-related, provides the most speaker contrasting and accurate metrics?
- Is it possible to differentiate identical twins on account of their speech timing measures while taking part in spontaneous dialogue?

A few hypotheses have been suggested regarding the research questions of experiment II:

i. It is anticipated that speech timing parameters pertaining to the micro temporal category may be more speaker discriminatory compared to the other parameters, perhaps signaling the implications of n-size (i.e., number of data points) in the discriminatory pattern outcomes. ii. As in experiment I, although identical twins are expected to be remarkably similar regarding their speech timing patterns, differences will still be observed between such individuals.

Experiment III: speaking fundamental frequency

- Which set of $f0$ parameters (e.g., $f0$ mean, median, baseline, standard deviation) are considerably more speaker-discriminatory in comparisons performed between genetically related individuals and across all speakers (i.e., cross-pair comparisons)?
- Which parameters category, namely, centrality, modulation, and variation, provides the most speaker contrasting and accurate estimates?
- Is it possible to differentiate identical twins on account of their $f0$ estimates while taking part in spontaneous dialogue?

- Is there any influence of "*speech material*" (i.e., $f0$ assessed in concatenate speech vs. lengthened vowels) on the contrasting power of $f0$ long-term measures?

A few hypotheses have been suggested regarding the research questions of experiment III:

i. Based on previous literature reports, the baseline value of $f0$ (i.e., $f0$ baseline) is expected to display the best speaker-discriminatory potential while displaying the best system performance estimates. ii. Given the relatively well-established effects of speaking style on $f0$ modulation and variation patterns, it is expected that $f0$ centrality measures may outperform the remaining parameters concerning its speaker discriminatory potential. iii. In view of the substantial physiological motivation regarding $f0$ centrality estimates, and the possible high influence of "learning" on $f0$ modulation patterns, identical twins are expected to be relatively more similar to each other regarding their $f0$ dimension than for any other dimension assessed here.

Chapter 2

Material and methods

The present chapter regards the methodological approach adopted in the present study, covering several instances, such as general information on the participants, the experimental design adopted, tools, data segmentation and transcription processes, data extraction, the set of acoustic-phonetic parameters assessed, the statistical analyses performed, and the criteria for establishing the discriminatory potential of the parameters.

2.1 Ethical statement

This study, registered under the protocol 95127418.7.0000.8142, was evaluated and approved by the ethical committee at The University of Campinas (UNICAMP). All participants voluntarily agreed to take part in the research verbally and by signing a participant consent form. All personal information regarding the participants is kept private.

2.2 Participants

The participants recruited for the present research are 20 subjects, ten male[1] identical twin pairs, all Brazilian Portuguese (BP) speakers, speaking the same dialect (BP spoken in Alagoas). The age of the participants ranged from 19 and 35 years old, with a mean of 26.4 years. All identical twin pairs were codified with letters and numbers, following the pattern: A1, A2, B1, B2, C1, C2, D1, D2, and so on. The same letters were used to identify identical twin pairs. More detailed information regarding the participants' age and degree of education is presented in Table 2.1.

[1]The analysis of speech/voice characteristics of groups composed mostly by male speakers is a common practice in forensic speaker comparison research. This is mainly due to the higher prevalence of law infractions by male

Table 2.1: **Age and education degree of twin pairs**

Twin Pair	Age when recorded	Education degree
A1 - A2	19 years old	In the university / In the university
B1 - B2	19 years old	Secondary school / Secondary school
C1 - C2	21 years old	Secondary school / Secondary school
D1 - D2	21 years old	Elementary school / Secondary school
E1 - E2	26 years old	In the university / Secondary school (incomplete)
F1 - F2	29 years old	University degree / University degree
G1 - G2	29 years old	University degree / University degree
H1 - H2	30 years old	Secondary school / In the university
I1 - I2	35 years old	University degree / University degree
J1 - J2	35 years old	Elementary school / Elementary school

The decision to adopt the term "identical twins" over "monozygotic twins" resides on a practical reason: the latter terminology implies assessing the twins' genetic material. As no laboratory genetic assessment was carried out, the first term will be preferred. However, it is worth noting that the expression "identical" does not imply that speakers are identical to each other, standing solely to the relative high physical similarities displayed by the twin pairs.

Speakers were recruited through a recruitment method known as chain sampling or "snowball", in which subjects are contacted among their acquaintances or by recommendation of other participants in the study. Each twin within the pair lived and resided in the same city/town. The pairs were recruited in five different cities in the state of Alagoas, which stands for the second smallest state in Brazil in terms of geographic area.

The inclusion criteria were: i. Identical twins; ii. male speakers; iii. same dialect; iv. aged between 18-45 years; v. with at least elementary school completed. The exclusion criteria were: i. Reported hearing loss or speech disorder, ii. identical twins raised apart; iii. identical twins that lived apart from each other for more than five years.

Moreover, all twin pairs in the present study reported that they had grown up together, were frequently in contact with each other, displayed a high-affinity level, and shared common social groups in most cases. Furthermore, the researchers made sure the participants were not going through a temporary cold or allergic reaction prior to the recordings, which would have caused the recording session to be postponed.

Regarding the participants' degree of education, all participants have at least finished elementary school (i.e. "ensino fundamental"). 17 participants (85%) finished secondary school (i.e. "ensino médio"), and 2 (10%) participants started but did not finish secondary school. 8 (40%)

individuals in comparison to female. As a consequence, they are more likely to be involved in a forensic investigation.

participants have a university degree, and 7 (35%) were still attending university.

Twins were, in most cases (i.e., 70% of cases), matched regarding their degree of education, not displaying a substantial gap between them (e.g., elementary school vs. university degree). Moreover, although most of the speakers (i.e., 75% of them) had some university background, they were not recruited in an academic context. They were primarily contacted for being identical twins, having their degree of education checked only afterward. This aspect may also add to the ecological nature of the present study's data.

2.2.1 Qualitative information about the twin pairs

One of the main advantages regarding the present study's design– as in other twin studies– regards the control of the variable "*familiarity*" between speakers, a very often neglected factor. Here, such factor has been qualitatively assessed[2] by applying a survey containing questions about the social dynamic of twin pairs and other pertinent information, including life-style habits that could imply changes in voice quality descriptors, such as the habit of smoking. All questions were presented in the speaker's native language. i.e., Brazilian Portuguese. Some of the questions and respective answers are depicted in Figures 2.1–2.12, with their corresponding English translation.

The speaker who answered affirmatively to the question in Figure 2.12, which concerns the habit of smoking, was: E1. A further investigation was conducted regarding the frequency of such a habit. It was reported that both twins occasionally smoked; however, E1 was reported to do it more frequently, mostly during weekends, for at least five years. Such an observation is expected to present implications on the results, particularly regarding experiment III. Notwithstanding, if differences are found, they may represent a piece of interesting evidence towards the impact of external factors, e.g., lifestyle or health-related habits, on voice acoustic patterns.

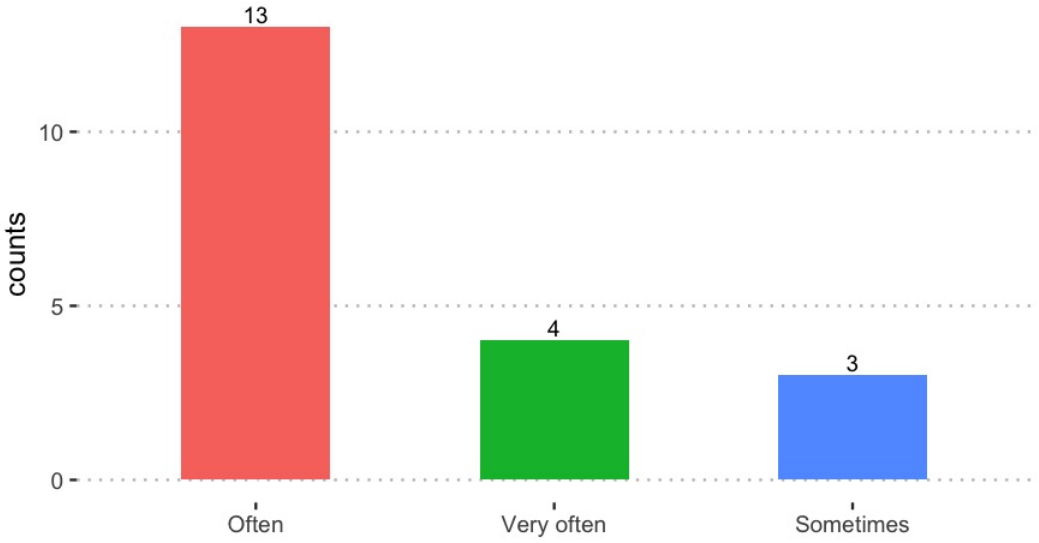
Other pertinent questions were also contained in the questionnaire, some of them targeting the same type of response, but posed in a different way as to allow consistency verification.

It should be remarked that the researcher only verified individual answers for qualitative questions after performing all analyses (e.g., the ones in Figures 2.1–2.10). This aimed to prevent any prior knowledge on the part of the experimenter that could interfere in the data analysis process.

As needed, possible intra-twin pair agreements/disagreements regarding the before-mentioned questions will be explored in the final chapter (*chapter 6*), where an overall appreciation of their acoustic-phonetic patterns will be drawn.

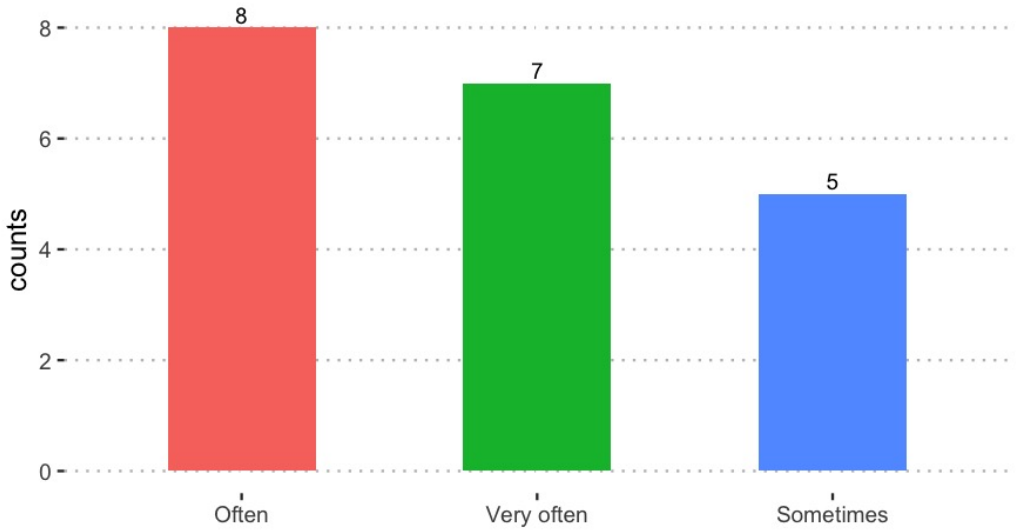
[2]No statistical analysis was performed using such information, as that they were not treated as predictor variables.

Figure 2.1: **Question I**



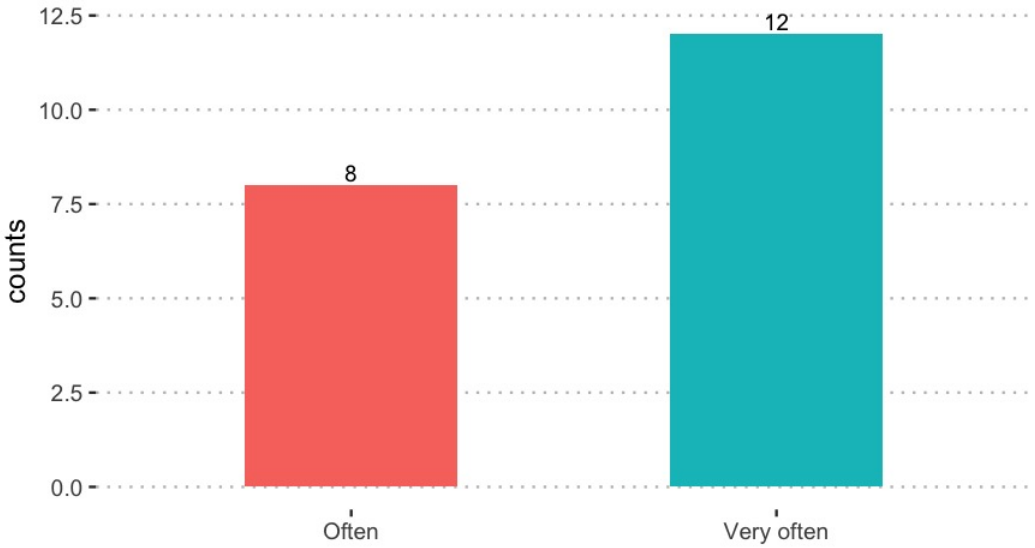
Question: How often are you mistaken for your brother because of your looks?
Answers: 65% often, 20% very often, 15% sometimes.

Figure 2.2: **Question II**



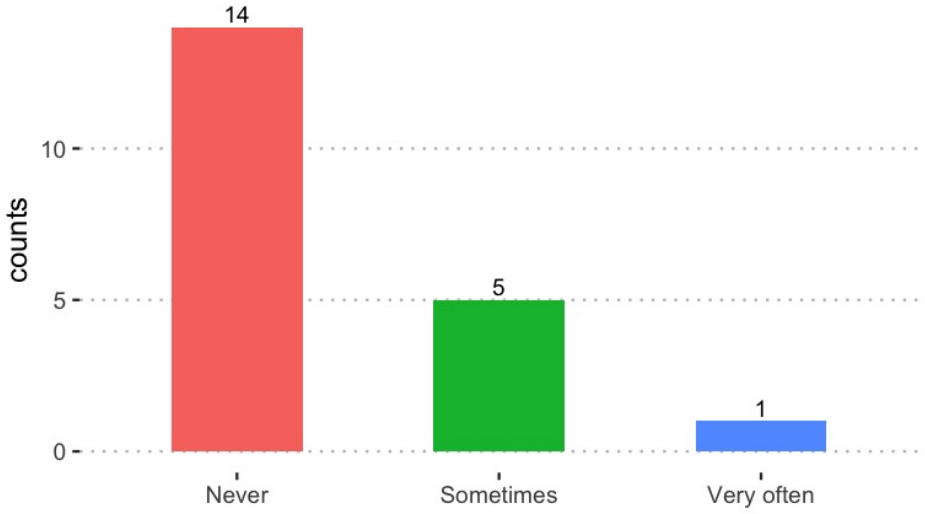
Question: How often are you mistaken for your brother because of your voice or the way you speak?
Answers: 40% often, 35% very often, 25% sometimes.

Figure 2.3: **Question III**



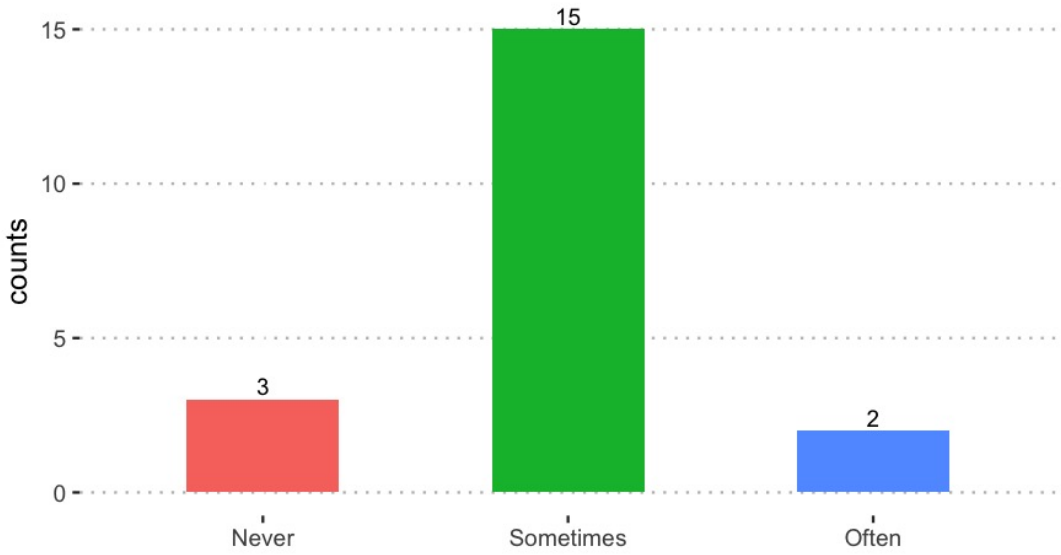
Question: How often were you mistaken for your brother because of your looks when you were teenagers?
Answers: 40% often, 60% very often.

Figure 2.4: **Question IV**



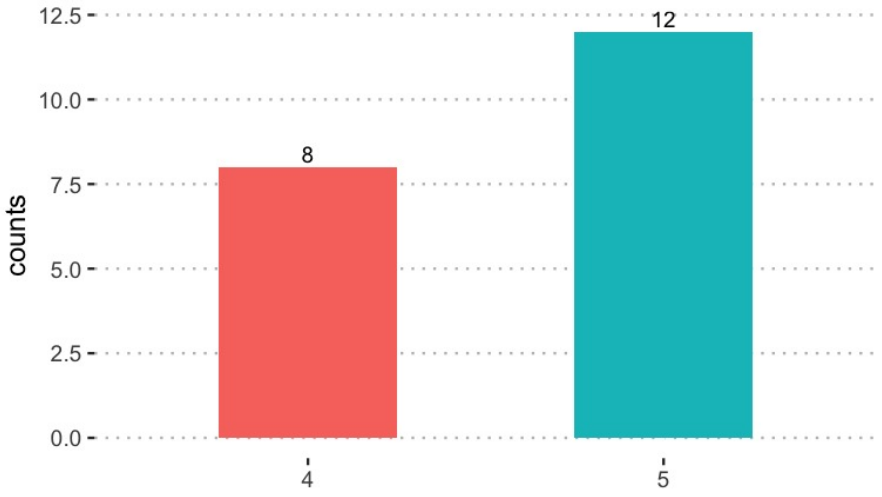
Question: Do you try or have you ever tried to sound different from your twin brother?
Answers: 70% never, 25% sometimes, 5% very often.

Figure 2.5: **Question V**



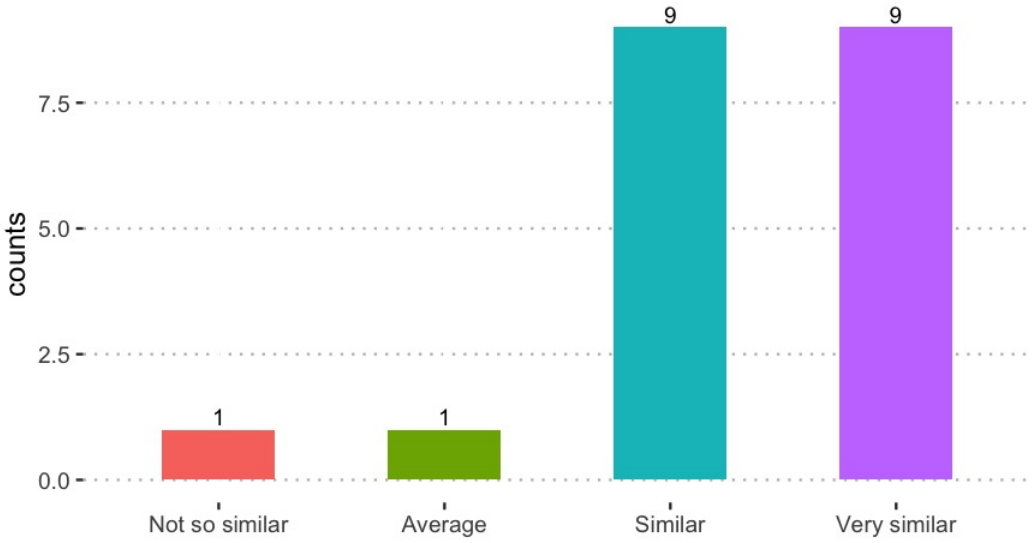
Question: Do you try or have you ever tried to look different from your twin brother?
Answers: 15% never, 75% sometimes, 10% often.

Figure 2.6: **Question VI**



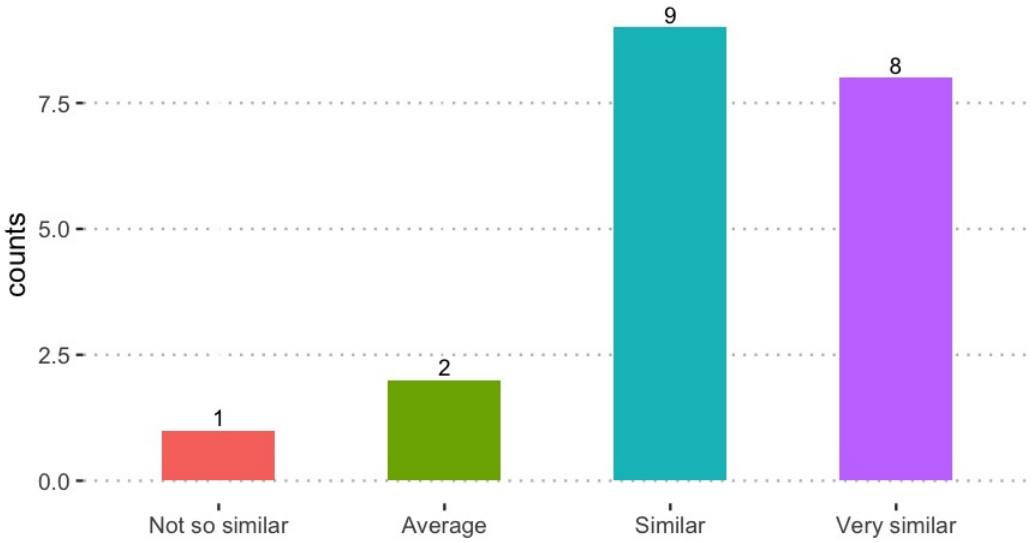
Question: What is your level of identification with your brother on a scale from 1 to 5, being 1 very low and 5 very high?
Answers: 40% chose “4”, 60% chose “5”.

Figure 2.7: **Question VII**



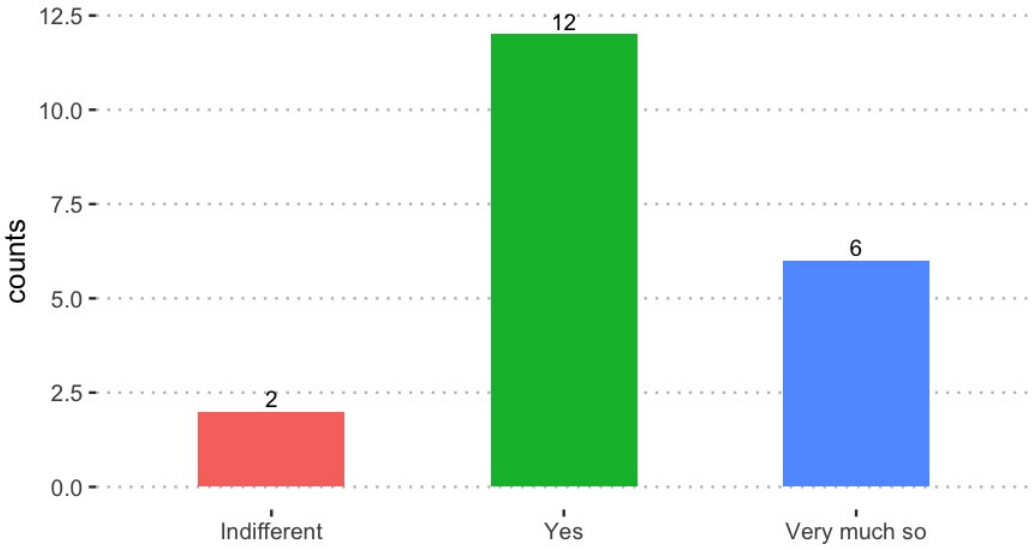
Question: How similar do you consider your voice compared to the voice of your twin brother?
Answers: 5% “not so similar”, 5% “average”, 45% “similar”, 45% “very similar”.

Figure 2.8: **Question VIII**



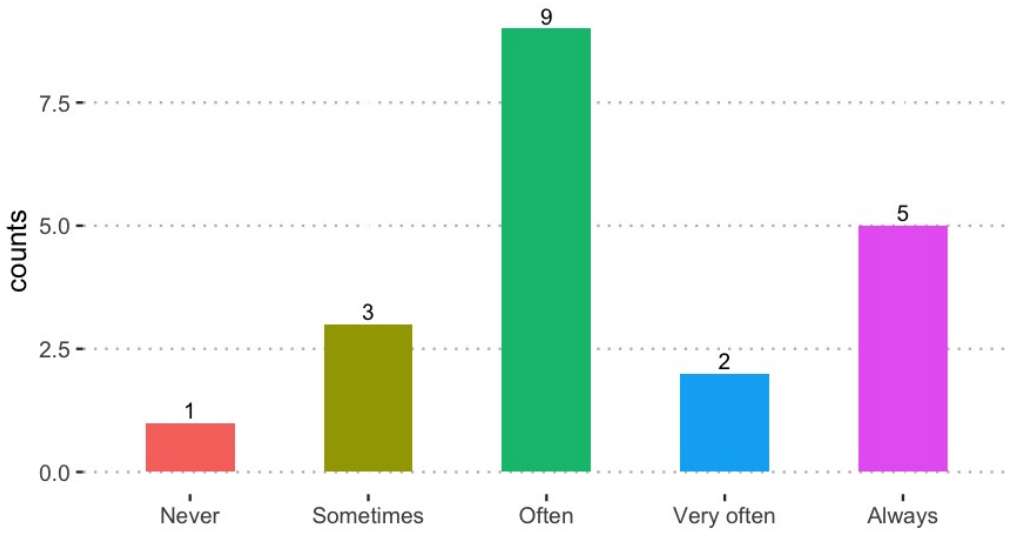
Question: How similar do you consider yourself compared to your twin brother?
Answers: 5% “not so similar”, 10% “average”, 45% “similar”, 40% “very similar”.

Figure 2.9: **Question IX**



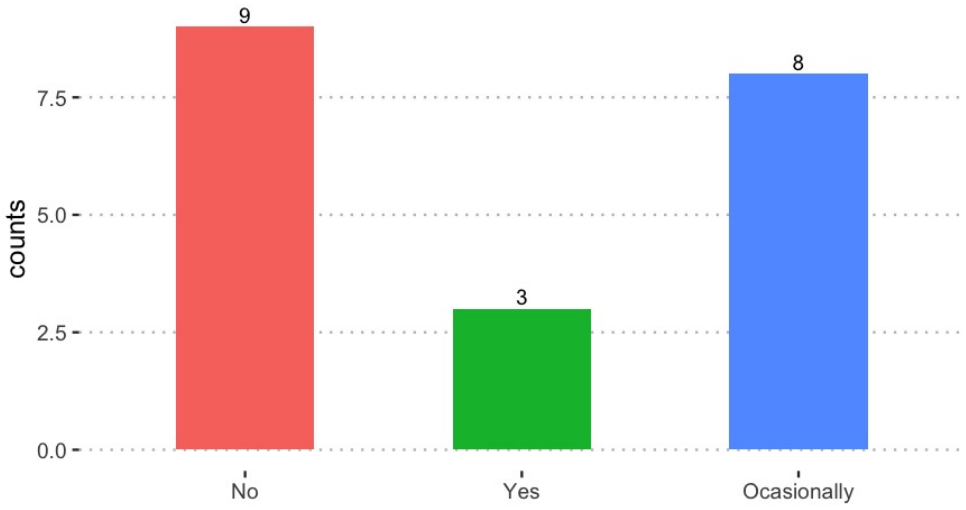
Question: Do you like the fact that you both are twins?
Answers: 10% “indifferent”, 60% “yes”, 30% “very much so”.

Figure 2.10: **Question X**



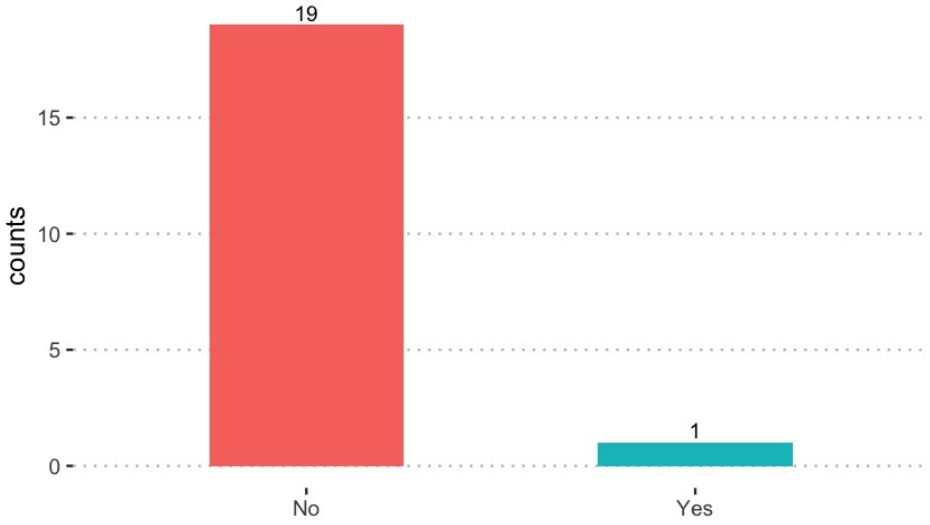
Question: Do you share the same social groups?
Answers: 5% “never”, 15% “sometimes”, 45% “often”, 10% “very often”, 25% “always”.

Figure 2.11: **Question XI**



Question: Do you drink alcohol?
Answers: 45% “no”, 15% “yes”, 40% “ocasionally”.

Figure 2.12: **Question XII**



Question: Do you smoke?
Answers: 95% “no”, 5% “yes”.

2.3 Experimental design

The recordings were carried out in silent rooms located in the cities where the twin pairs resided. The speech material employed in the present research consists of spontaneous telephone conversations between twins, with dialog topics being decided by the pairs, aiming at eliciting a more ecologically-valid data. During the recording sessions, twin pairs were placed in different rooms, not directly to see, hear, or interact with each other. The speakers were encouraged to start a conversation using a mobile phone while being simultaneously recorded by high-quality microphones. The audio signals were then processed and registered in two separate channels.

The mentioned recording approach aimed at eliciting a telephone speaking style and represents an attempt to approximate the experimental conditions to more realistic forensic circumstances, as conducted in San Segundo (2014). An illustration of the recording session is presented in Figure 2.13.

Figure 2.13: **Recording session: spontaneous dialogue**

As previously mentioned, a relevant aspect concerning the experimental design adopted in the present study regards the relative control of the variable "*familiarity*" between speakers, which is expected to approximate the research context to those involved in real forensic contexts. In a forensic situation, speakers usually are familiar with each other, which may bear influences on how they speak, and the level of speech monitoring they apply during the interactions. Since all pairs recruited in the present study were siblings who displayed a close relation, the impact of interlocutor-oriented self-monitoring may be regarded as considerably reduced, reflecting a more

ecological interaction context.

Notably, a potential limitation regarding the experimental design adopted here, in which twins are placed as a speaking duo, concerns the possible effects of "*prosodic entrainment*" to which such speakers may be submitted, potentially enhancing their level of phonetic similarity. However, if differences between twin pairs are observed even in such circumstances, these may be regarded of high relevance, given their possible less susceptible or resistant nature to external influences. The possible effects of *prosodic entrainment* are acknowledged in different parts of the thesis, mainly in *Chapter 4*, which regards the analysis of speech timing parameters.

All recordings were carried out with a sample rate of 44.1 kHz and 16-bit amplitude resolution, using an external audio card (Focusrite Scarlett 2i2[3]) and two headset condenser microphones with the same specifications (DPA 4066-B[4]), especially designed for speech capture. Microphones were positioned (approximately) at a 2 cm distance from the speakers' mouth. Speakers were instructed to hold the mobile phone on the opposite side of the microphone to avoid possible misadjustments and interference.

The unedited recordings had an average duration of about 14 minutes, and the transcribed material had an average duration of about 2.30 min per subject.

Although some topics have been suggested (e.g., remembering events from childhood, other events/moments shared by the twin pairs, ongoing issues they wanted to talk about), the conversation topics were, in all cases, decided by the twins before the recording sessions. They were also instructed to change topics whenever they wanted. The decision not to establish fixed dialogue topics aimed to guarantee that individuals spoke about a subject they wanted to speak about and not a subject they were compelled to talk about. Again, such a decision also aimed to preserve the "naturalness" of the interactions.

2.4 Data segmentation and transcription

The data annotation, as presented in Figure 2.14, comprised 11 distinct layers in the Praat textgrid (BOERSMA; WEENINK, 2018), as follows:

1. **Dialogue part:** different portions/parts of the dialogues throughout the recordings, e.g., beginning, middle, and final parts;

2. **Speech chunks:** speech intervals on average 3 s long, in most cases corresponding to inter-pause intervals (i.e., stretches of speech between long silent pauses);

[3]Frequency Response: 20Hz - 20kHz ± 0.1dB; Dynamic Range 111dB (A-weighted); Impedance: 3kΩ.
[4]Frequency response: 20 Hz - 20 kHz; Sensitivity: -44 dB re. 1 V/Pa; Rated output impedance 30 - 40 Ω.

3. **All vocalic segments:** all vocalic segments produced within speech chunks, including monophthongs, diphthongs, and nasalized vowels;
4. **Oral monophthongs:** oral monophthongs only;
5. **Oral diphthongs:** oral diphthongs only;
6. **Lengthened vowels:** vowel prolongations with a minimum duration threshold of 100 ms;
7. **Silent pauses:** silent pauses with a minimum duration threshold of 100 ms;
8. **All pauses:** combination of silent and filled pauses;
9. **Vowel-to-vowel intervals:** syllable-sized units defined as all the segments uttered between two consecutive vowel onsets;
10. **Smoothed z-scores peak values:** smoothed z-scores peak values at the end of the stress group;
11. **Stress group intervals:** intervals corresponding to two consecutive salient V-V intervals.

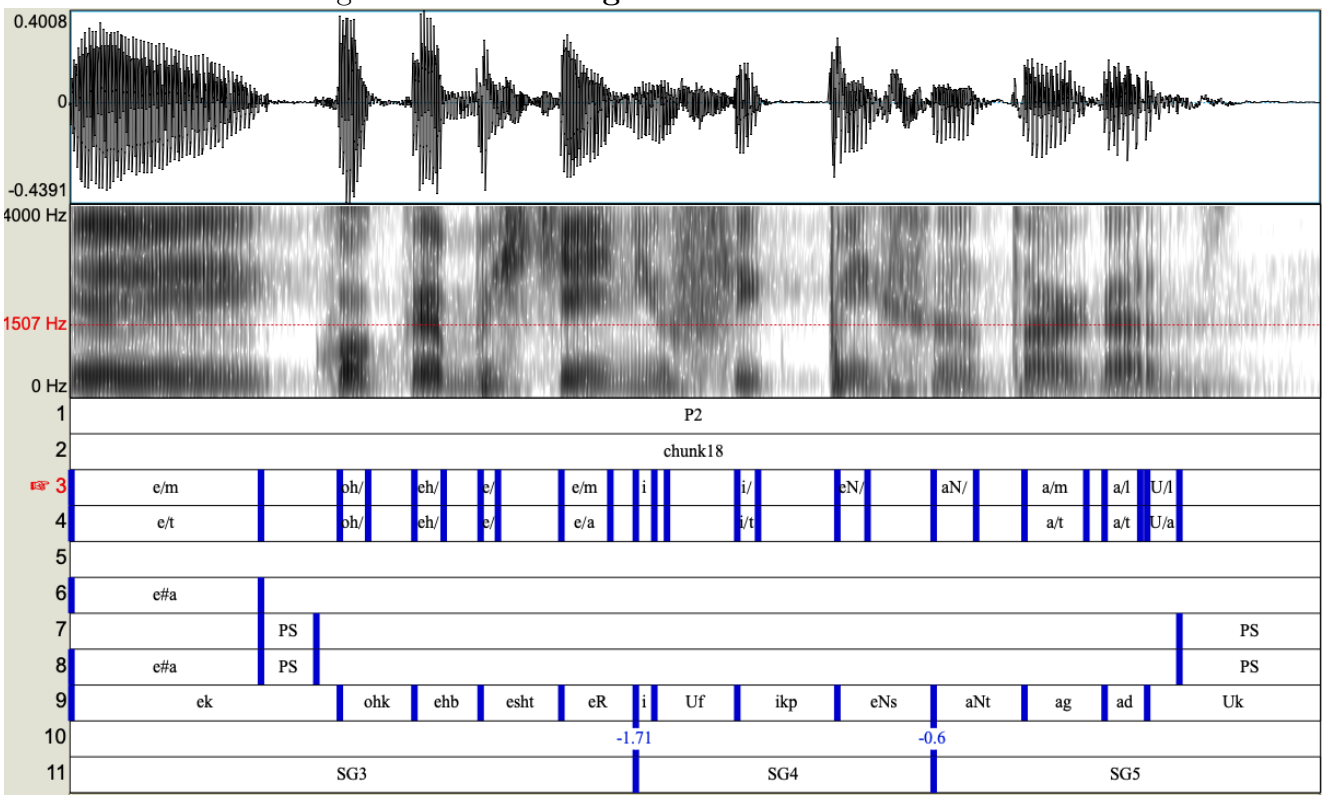
2.4.1 Experiment I: vowel formant analysis

All vowels were segmented and transcribed manually in the Praat software following auditory and acoustic criteria, namely the energy appearance/disappearance in the broad-band spectrogram. After the segmentation of all vowels in layer 3 of the Praat textgrid, oral monophthongs were segmented in a separate textgrid layer (see layer 4 in Figure 2.14), from which nasalized vowels, diphthongs, and triphthongs were disregarded. Oral monophthongs were then manually classified as stressed or unstressed as well as modal or laryngealized. The analysis of laryngealized events remains a topic for future investigation.

Since the recordings are of very high quality, the vowel labeling process was possible in most cases. This was also due to the fact that all speakers were recorded simultaneously in different channels, eliminating speech overlaps that could possibly disturb the labeling process. It is also worth mentioning that the author was present at the moment of the recordings and was responsible for all transcriptions, being able to base his judgment on the analysis of the speech context. In some rare cases, vowel segments were excluded when their identity was too difficult to define.

The F1, F2, F3, and F4 values were automatically extracted from the middle points in the labeled vowels through a Linear Predictive Coding (LPC) technique. The parameters extraction

Figure 2.14: **Data segmentation and annotation**



was done using the Praat script "*ProsodyTime*", cf. Barbosa (2015). The script generates a .txt file containing speaker identity, vowel names, vowel durations in seconds, mean formant frequencies in Hertz, and other parameters. Given that extractions were carried out automatically and that only male voices were used, the hypothetical influence of extraction errors can be regarded as reduced.

For the comparison of identical twin speakers in the Bark scale, cf. Zwicker (1961), formant frequencies in Hertz were transformed to Bark according to the following formula 2.4.1 (TRAUNMÜLLER, 1990):

$$z = [26.81/(1 + 1960/f)] - 0.53 \tag{2.4.1}$$

Comparisons of formants in the Bark scale between identical twins were carried out in order to assess whether the observed differences could be perceptually and linguistically relevant.

Initially proposed by Zwicker (1961), Bark is a critical-band rate based on psychoacoustic principles, which are essential for understanding some characteristics of the human hearing system (ZWICKER; FASTL, 2013). It is worth mentioning that not necessarily all differences observed in one scale are also significant in the other, once they are based on different acoustic principles.

Also, it must be recognized that not all variations in speech production's physical dimensions are relevant from a linguistic and perceptual viewpoint.

As to observe to what extent the acoustic distances of cardinal vowels in the vocalic space could be related to formant variation, the Euclidean distances between vowels were also assessed.

The Euclidean distances were measured by computing each segment's F1 and F2 coordinates in the two-dimensional vocalic space. Subsequently, the mean acoustic distances between neighboring vowels and between the extreme front vowels ([i] - [ɛ]) and extreme back vowels ([u] - [ɔ]) were also quantified. The formula 2.4.2 was applied, in which (x, y) stands for the coordinates of vowel 1 in the Euclidean plane, whereas (a, b) stands for the coordinates of vowel 2 in the Euclidean plane.

$$\text{dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2} \tag{2.4.2}$$

2.4.2 Experiment II: speech timing measures

In total, a set of 11 temporal speech parameters were analyzed, including macro, micro, and pause-related temporal parameters, as described below. Such a classification, based on the average duration of phonetic syllables (i.e., V-V units), aimed to help the reporting of the outcomes, and the development of further discussions, as better described in *Chapter 4*. All parameters were extracted automatically using the Praat script *ProsodyDescriptorExtractor*, cf. Barbosa (2020). Parameters' codes are presented within parenthesis:

Macro speech timing parameters:

- **Speech rate (SRATE):** defined as the number of V-V units in each speech chunk divided by its total duration (V-V units/second), including silent and filled pauses.
- **Articulation rate I (ARTRATE I):** defined as the number of V-V units in each speech chunk divided by its total duration (V-V units/second), excluding only silent pauses.
- **Articulation II (ARTRATE II):** defined as the number of V-V units contained in each speech chunk divided by its total duration (V-V units/second), excluding silent pauses and vowel lengthening.
- **Stress group duration (SGDUR):** defined as the interval corresponding to two consecutive salient V-V intervals (in second), i.e., those units for which a duration increase has been automatically detected. Each stress group ends with a salient V-V interval . This parameter was generated automatically.

Micro speech timing parameters:

- **V-V units duration (VVDUR I):** syllable-sized duration units defined as all the segments uttered between two consecutive vowel onsets (in millisecond). Both salient and non-salient V-V units are included in this parameter.
- **V-V units duration (VVDUR II):** the aforementioned phonetic unit corresponding solely to non-salient V-V units (in millisecond), represented by those units for which a duration increase has not been automatically detected.
- **Vowel duration (VOWEL DUR):** defined as the duration of produced oral monophthongs (in millisecond).

Pause-related parameters:

- **Silent pause duration (SILPAUSES):** silent pauses equal or superior to 100 ms, a threshold commonly applied in automatic measurements, as to prevent occlusion phases of plosives from being counted (KÜNZEL, 1997).
- **Filled pause duration (FILPAUSES):** defined as vowel prolongations equal or superior to 100 ms, perceived as hesitations/filled pauses (in millisecond).
- **All pauses (ALLPAUSES):** combination of silent and filled pauses (in millisecond), i.e., lengthened vowels before silent pauses or in hesitations.
- **Inter-silent pauses intervals (IPI):** defined as the interval comprising the speech production between two consecutive silent pauses (in second).

The durations of the selected speech chunks (see textgrid layer 2 in Figure 2.14) were around 3 s as an attempt to match "speech turn time" variation among speakers. There was an already expected tendency for some subjects to hold their turn for a longer period of time compared to others. Furthermore, inter-pause intervals were tracked throughout the transcriptions and employed as a more objective criterion for segmenting chunks (i.e., intervals between longer pauses, never containing less than three V-V units, to prevent from selecting intervals containing solely vowels under the effects of phrase-final lengthening). As for longer intervals, without perceived silent or filled pauses, they were preserved in their total duration, or in some cases, divided into smaller parts, maintaining the structure of intonational phrases. Moreover, considerable agreement regarding speech chuck boundaries and stress group boundaries was observed.

By the end of the transcription process, the individual chunks inside the selections were, on average, 3.12 s long. Note that this value corresponds to the right limit of 95% confidence intervals for stress group duration in Brazilian Portuguese, cf. Barbosa (2006). The average v-v unit number in each speech chunk was 9.9 units, with a minimum of 3 and a maximum of 32 units, with 14 units being the most recurrent number. Such a high variability regarding the number of syllables within the intervals is due to the fact that, in spontaneous speech, individuals tend to vary substantially both in type and extent of sentences they produce. A total of 851[5] speech chunks were analyzed, an average of 42 chunks per speaker with a standard deviation of 5.

The main acoustic criterion for segmenting silent and filled pauses (see layer 8 of the Praat textgrid in Figure 2.14) was their relative duration. Only silent pauses and vowel prolongations at least 100 ms long were included in the analysis. Moreover, breathing sounds, as in the case of inhalation noises, were included in the non-speech part. The referred 100 ms threshold was established based on the observation that most pauses produced by the speakers exceeded this limit, and also in light of previous studies (KÜNZEL, 1997). As for filled pauses (vowel lengthening), the referred 100 ms threshold was above the observed average duration for oral monophthong, i.e., 84 ms.

2.4.3 Experiment III: fundamental frequency descriptors

Following the manual segmentation and transcription of all speech material, as illustrated in Figure 2.14, the extraction of $f0$ parameters was carried out automatically using a modified version of the same Praat script *"ProsodyDescriptorExtractor"*, cf. Barbosa (2020). For the extraction, the $f0$ floor and ceiling were defined as 60 Hz–300 Hz. For F0 smoothing, a cut-off frequency filter of 2 Hz was used to compute $f0$ linguistically-relevant peaks throughout the utterances. An overall description of the parameters extracted are presented in the following.

$f0$ acoustic descriptors

A set of 15 f0 measures were considered for assessment in connected speech (i.e., at the domain of phrases), including descriptors of $f0$ dispersion, centrality, and modulation (f0M), as presented below. As for lengthened vowels, given the more stationary $f0$ patterns observed, only the first seven parameters were considered for the analysis, i.e., centrality and dispersion estimates. Parameters' descriptions start with their respective codes/labels:

- **f0mean:** $f0$ mean in semitones ref 1 Hz/ and in Hertz

[5] 853, after the exclusion of two possible outliers; 851.

- **f0med:** $f0$ median in semitones ref 1 Hz/ and in Hertz
- **f0min:** $f0$ minimum in semitones ref 1 Hz/ and in Hertz
- **f0max:** $f0$ maximum in semitones ref 1 Hz/ and in Hertz
- **f0sd:** $f0$ standard-deviation in semitones ref 1 Hz/ and in Hertz
- **f0base:** base value of $f0$ in semitones ref 1 Hz/ and in Hertz (i.e., equivalent to the 7.4th quantile of the $f0$ sample)
- **f0SAQ:** $f0$ semi-amplitude between quartiles in semitones ref 1 Hz/ and in Hertz (i.e., a non-parametric measure of $f0$ dispersion)
- **f0M1:** smoothed $f0$ peak rate in peaks per second (i.e., $f0$ peak rate/s)
- **f0M2:** standard-deviation of $f0$ maxima in semitones ref 1 Hz/ and in Hertz (i.e., when there is more than one peak in the interval)
- **f0M3:** standard-deviation of the $f0$ maxima positions in seconds (i.e., standard-deviation of peaks' duration)
- **f0M4:** 1st-derivative $f0$ mean in Hertz/frame of the positive derivatives (i.e., $f0$ rising rate in the peaks)
- **f0M5:** 1st-derivative $f0$ mean in Hertz/frame of the negative derivatives (i.e., $f0$ falling rate in the peaks)
- **f0M6:** 1st-derivative $f0$ standard-deviation in Hertz/frame of the positive derivatives (i.e., standard deviation of $f0$ rising rate)
- **f0M7:** 1st-derivative $f0$ standard-deviation in Hertz/frame of the negative derivatives (i.e., standard deviation of $f0$ falling rate)
- **f0M8:** Mean peakness of $f0$ max in semitones relatively to $f0$ range multiplied by 1000 (i.e., corresponding to the width of $f0$ peaks)

$f0$ in connected speech

Intervals of continuous speech with an average duration of 3 s were selected for the extraction of $f0$ descriptors in connected speech. Note that such intervals correspond to the same textgrid layer from which speech timing estimates were extracted: "speech chunk" (see textgrid layer 2 in Figure 2.14).

A total of 853 speech chunks were analyzed, an average of 42 chunks per speaker with a standard deviation of 5, from which $f0$ estimates in connected speech were computed. As for those intervals for which some estimates could not be computed, e.g., $f0$ peak rate descriptors, these were disregarded during the statistical testing.

$f0$ in lengthened vowels

All lengthened vowel segments produced by the speakers in different portions of the dialogues were also segmented and transcribed manually (see textgrid layer 6 in Figure 2.14). Furthermore, the vowel segments most frequently lengthened in the corpus were identified, after performing their duration extraction. These were: [a:], [ɛ:], and [i:], in a decreasing frequency of occurrence, being oftentimes perceived as filled pauses. Because these vowels were found more often prolonged, they were elected for the extraction of $f0$ descriptors, given their potential forensic applicability (e.g., the extraction of glottal source parameters). It is noteworthy that, besides being the most commonly lengthened vowel in the corpus, the central prolonged vowel [a:] is also one of the most commonly used for the assessment of voice quality aspects within the clinical setting, e.g., Cielo, Agustini, and Finger (2012), Moon et al. (2012), Cavalcanti, Englert, et al. (2021) and Sotome et al. (2021). A minimum duration threshold of 160 ms was established for the selection of vowel segments, as in San Segundo (2014).

A total of 399[6] lengthened [a:], [ɛ:], and [i:] vowels were analyzed, a mean of 20 vowels per speaker and a standard deviation of 7 vowels. The lengthened vowels displayed a mean duration of around 250 ms (median of 212 ms), with a standard deviation of 31.7 ms. All vowels produced with a creaky phonation were excluded from the analysis. As for longer vowels produced with both modal and creaky portions, only the modal portion was considered.

[6]Note that this number is lower than the number of filled pauses transcribed "560", which is mainly due to the fact that only the most frequently lengthened vowels were selected for such an analysis, i.e., [a:], [ɛ:], and [i:].

2.5 Statistical analysis

All statistical analyses was carried out in the R software (R CORE TEAM, 2020). As most of the data do not to fit the normal distribution, as verified through the Shapiro-Wilk normality test ($\alpha < 0.05$), the statistical testing was performed by means of non-parametric methods. The Kruskal-Wallis rank-sum test was applied to verify possible differences in each tested parameter, followed by the post hoc analysis with the Dunn's Multiple Comparison Test (two-tailed). The Bonferroni correction was automatically performed to adjust the alpha threshold due to multiple comparisons, based on the 190 comparisons among all individuals. Therefore, it can be said that all differences reported here are based on and reflect the same statistical criteria.

Following the comparison of all subjects, intra-identical twin pair differences were identified and systematically reported. It is worth mentioning that such comparison is already expected to yield a great deal of inter-speaker similarity, given that, by taking part in the same dialogue, twin pairs may naturally be under some level of *prosodic entrainment*, which may, in part, account for their possibly high congruence.

Furthermore, differences observed for comparisons carried out across all speakers were also considered and reported (hereafter, cross-pair comparisons). For instance, A1 - B1; A1 - B2; A2 - B1, A2 - B2, and so on. Overall, 190 comparisons among speakers were carried out for each tested parameter. The main justification for including such a comparison is that they may be regarded as more realistic from a forensic phonetic perspective, in which individuals may be similar regarding several aspects– such as sex, age, dialect, education degree– but not as similar as identical twins. In all cases, twins were compared to other twins while interacting with someone they were accustomed to (i.e., their own siblings). As such, the variable inter-speaker "*familiarity*" may be regarded as equally controlled as far as cross-pair comparisons are concerned.

Effect size estimates were computed for all tested parameters. Such a metric adds to the understanding of how much of the variation observed can be attributed to the variable "*speaker*". For the estimation of the Kruskal-Wallis Effect Size, the following Formula 2.5.1 was applied, where H is the value obtained in the Kruskal-Wallis test; k is the number of groups; n is the total number of observations:

$$\eta^2 = (H - k + 1)/(n - k) \tag{2.5.1}$$

The magnitude of the effects were attributed automatically by the R package "*rstatix*" (KASSAMBARA, 2020) in the R software, in view of the values commonly reported in the literature for the eta-squared (η^2) : 0.01$\leq$ 0.06 (small effect), 0.06 $\leq$ 0.14 (moderate effect), and $\geq$ 0.14 (large

effect). As such, the effect size index assumes values ranging from 0 to 1, which when multiplied by 100% indicates the percentage of variance in the dependent variable explained by the independent variable, cf. Tomczak and Tomczak (2014) and Fritz, Morris, and Richler (2012).

Finally, to assess whether stressed and unstressed vowels, as well as whether front and back vowels present different variances, the Fligner-Killeen test of homogeneity of variances was performed in experiment I, which tests the null hypothesis that the variances in each of the groups are equal. Moreover, Wilcoxon Rank Sum Tests followed by Bonferroni corrections were carried out to probe possible differences for vowel height (F1) between front and back vowels occupying the same horizontal plane, and between the entire front and back vertical vocalic space dimensions.

2.5.1 The speaker discriminatory performance of acoustic-phonetic parameters

Regarding the assessment of the suitability of the acoustic-phonetic parameters for forensic speaker comparisons, three estimates were examined as a function of the comparisons among all speakers, i.e., cross-pair comparisons. The first estimate is the *Log-likelihood-ratio-cost function* (Cllr), an empirical estimate of the precision of likelihood ratios proposed by Brümmer and Du Preez (2006), and applied, among others, by Morrison (2009). It is a measure of *accuracy*, initially developed for use in automatic speaker recognition and subsequently incorporated into the forensic framework. It is given by the Formula 2.5.2:

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}}\log_2\left(1+\frac{1}{LR_{ss_i}}\right)+\frac{1}{N_{ds}}\sum_{j=1}^{N_{ds}}\log_2\left(1+LR_{ds_j}\right)\right) \tag{2.5.2}$$

In the Formula 2.5.2, *Nss* and *Nds* are the number of same-speaker and different-speaker comparisons, and *LRss* and *LRds* are the likelihood ratios derived from same speaker and different speaker comparisons. A same-origin penalty value is $log_2(1+1/LRs)$, and a different-origin penalty value is $log_2(1+LRd)$. As such, Cllr is a continuous function which is small for correct likelihood ratios, and large for incorrect likelihood ratios. Therefore, the lower the Cllr, the better the performance of the system (MORRISON; ZHANG; ROSE, 2011; MORRISON; ZHANG; ENZINGER, 2019).

According to Morrison, Zhang, and Rose (2011), the Cllr estimate has the desired properties of being based on likelihood ratios, being continuous, and more heavily penalizing worse results[7]. For computing such estimate, likelihood ratios were calculated through Multivariate Kernel Density

[7]Providing less support for the consistent-with-fact hypothesis or more support for the contrary-to-fact hypothesis, cf. Morrison, Zhang, and Rose (2011).

analysis - MVKD (AITKEN; LUCY, 2004), i.e., a non-parametric approach, implemented in the R package "*fvclrr*" (LO, 2020). Multiple pairwise comparisons were performed across individuals in which the background sample consisted of data from all speakers, except those being directly compared (i.e., a cross-validation procedure).

Likelihood ratios were calculated using the Formula 2.5.3, where, as described by Morrison, Zhang, and Enzinger (2019), LR is the likelihood ratio; E is the evidence, i.e., the measured properties of the voice on the questioned-speaker recording; p(E|H) is *the probability of E given H*; respectively Hs is the same-speaker hypothesis, and Hd is the different-speaker hypothesis (i.e., same-origin and different-origin hypotheses). In the referred formula, the numerator of the likelihood ratio stands for a *similarity* term, whereas the denominator stands for a *typicality* term.

$$LR = \frac{p\left(E \mid H_S\right)}{p\left(E \mid H_d\right)} \tag{2.5.3}$$

To be able to perform the analyses of formant frequencies in experiment I, without taking into account "vowel quality", *fusion* and *calibration* procedures were employed. Such a process results in the combination of likelihood ratio (LR) scores from (multiple) test systems to provide a single set of fused LR scores, based on a logistic-regression model trained with the same set of data (i.e., self-calibration). Such a procedure is well exploited in Morrison, Zhang, and Rose (2011), and is regarded as an adequate solution when combining multiple estimates of likelihood ratios on the same data, such as different vowels, cf. Morrison, Zhang, and Enzinger (2019). Such a procedure is also implemented in the aforementioned R package "*fvclrr*".

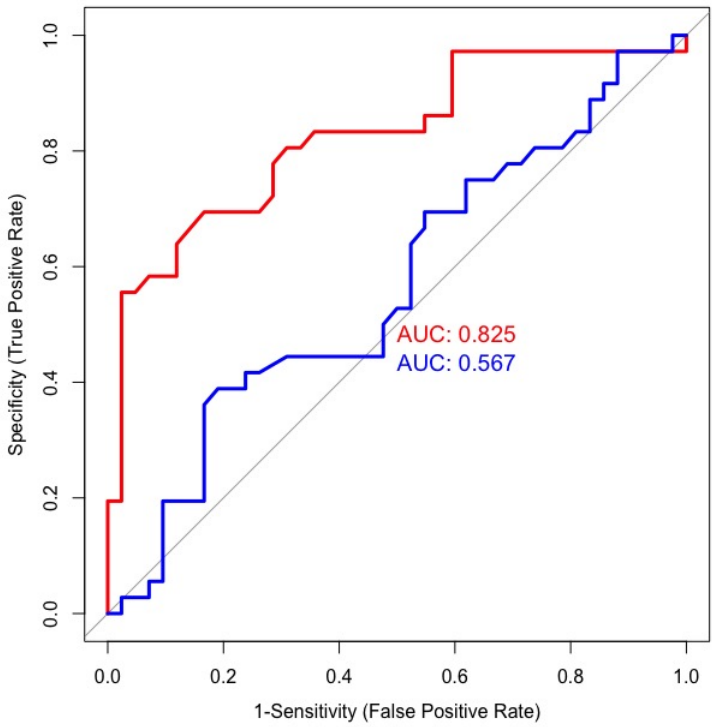
The second estimate is the *Equal Error Rate* (EER), which represents the point where the false reject rate (type I error) and false accept rate (type II error) are equal, being used to describe the overall accuracy of biometric systems (CONRAD; MISENAR; FELDMAN, 2012). This estimate was generated along with the Cllr. Lower EER values are compatible with better accuracy, whereas higher EER values suggest worse discriminatory performance. Both Cllr and EER values are reported as average values after performing several tests.

Finally, in order to observe the performance of acoustic parameters in terms of their binary classification (i.e., predictive power), *Receiver Operating Characteristics* (ROC) graphs were plotted, using the R "*pROC*" package (ROBIN et al., 2011). ROC plots are two-dimensional graphs commonly used in signal detection theory to depict the relative trade-offs between benefits (*true positives*) and costs (*false positives*), providing an estimate that allows the comparison across models/parameters, the *Area Under the ROC curve* (AUC) estimate, cf. Fawcett (2006). Moreover, the multi-class ROC function, as formulated by Hand and Till (2001), was applied to compute the multi-class AUC, which provides the averaging of several AUC estimates. Because of the

mathematical solution applied, no ROC curve can be visualized for multi-class AUCs.

An example of a ROC graph is given in Figure 2.15. In such a figure, two curves can be visualized, each of them representing an estimate/classifier: a red and blue curve, along with their corresponding AUC values. In such graph, the lower left-hand point (0,0) represents the strategy of never registering a true or false positive classification (a "conservative" approach); the upper right-hand point (1,1) represents the opposite strategy, of unconditionally registering a true or false positive classifications (a "liberal" approach); finally, the upper left-hand point (0,1) represents "perfect classification", whereas the diagonal line (y=x) represents a random performance (equivalent to "guessing"), cf. Fawcett (2006).

Figure 2.15: **Receiver Operating Characteristics (ROC) graph**



In Figure 2.15, the red curve is the one closer to the upper left-hand point (0,1), whereas the blue curve is the one closer to the diagonal line. Also, when comparing their AUC values (i.e., the portion of the Area of the unit square), it is clear that the classifier represented by the red curve has a greater area, and therefore a better average classification performance than the one represented by the blue curve. Note that the possible maximum value for AUCs is 1.

For the sake of interpretation, it should be observed that an ideal parameter for the forensic application should depict relatively low Cllr/EER values while displaying relatively high AUC

values in relation to the other parameters under comparison. Notably, since different criteria are considered for computing the aforementioned estimates, some degree of variation is expected regarding their classification performance. Assuming that a metric that satisfies most conditions is reasonably better than one that is approach-dependent, the overall performance of all metrics was taken into account.

Chapter 3

The discriminatory patterns of formant frequencies

3.1 Introduction

The present chapter[1] regards the analysis of the speaker discriminatory potential of vowel formant frequencies, namely the first four speech formants (F1, F2, F3, and F4), in comparisons of genetically-related speakers (i.e., identical twins) and across-pairs. Both vowel quality and lexical stress variables were considered in the experiment. Speaker comparisons were performed based on formant frequencies derived from the seven phonetic vowels of Brazilian Portuguese (BP) insofar as their classification as stressed or unstressed were also taken into account. The main outcomes are presented and discussed in the following.

3.2 Results

A total of 9,446 vowels were analyzed in the present study, of which 5,487 (58%) were classified as stressed and 3,959 (42%) as unstressed. The most frequent vowels in the corpus were the central vowel [a] with 3497 occurrences, [i] with 1677, [u] with 1116, [ɛ] with 1015 and [e] with 988 occurrences. The less frequent vowels were the back vowels [ɔ] with 531 occurrences and [o] with 622 occurrences. The number of vowel tokens by speaker varied from 402 to 588 tokens, a mean of 472 and a standard deviation of 61.8.

[1]The outcomes deriving from this section of the thesis have been (partly) published in the peer-reviewed open access journal Plos ONE in the form of a research article, cf. Cavalcanti, Eriksson, and Barbosa (2021a). The full length research article can be retrieved from https://doi.org/10.1371/journal.pone.0246645.

It may be noted that the three most frequent vowels were the most extreme and contrasting ones in the BP vocalic system, as revealed by the following order of occurrence: [a] > [i] > [u].

Considering that the vowels [ɛ] and [ɔ] are generally stressed in BP and produced as unstressed only in a few contexts for the analyzed dialect, a relatively small number of unstressed tokens for the corresponding segments was already predicted. Even so, all vowels were included in the analysis.

The reporting of vowel formant differences among speakers will consider the subdivision between vowel quality-related measures in BP— F1 and F2, and high formant frequencies — F3 and F4. Moreover, the articulatory phonetic distinction between front and back vowels will be acknowledged when reporting the differences. The main results are described in the following.

3.2.1 The inter-speaker discriminatory potential of F1-F4 frequencies

Tables 3.1, 3.2, 3.3, and 3.4 depict statistically significant differences for mean formant frequencies observed between speakers. The differences are reported per individual vowels as well as per individual parameters, regardless of the vowel identity. The overall percentage of differences observed for individual formants are presented in the column "%diff", which considers the number of statistically significant differences for all seven vowels presented in "total", in relation to the total number of comparisons carried out, in this case, 70 (10 x 7) for identical twin pairs (Tables 3.1 and 3.3), and 1.330 (190 x 7) for all speakers, i.e., cross-pair comparisons (Tables 3.2 and 3.4). By calculating the totality of the differences observed for each formant, it is intended to verify the contrasting potential of individual formants independently from vowel specificity. The same approach was applied to the analysis of the vowels' discriminatory potential.

As can be noted in Table 3.1, the statistical analysis revealed a consistent pattern in comparing primary vowel quality-related and high-frequency formants in identical twin pairs, with high formant frequencies displaying a higher proportion of inter-speaker differences in comparison to low formant frequencies.

Furthermore, by inspecting Table 3.1, it can be seen that the overall discriminatory potential of the horizontal and vertical articulatory dimensions, assessed through F1 and F2, yielded a total of 11.5% and 15.5% of differences in twin pairs, respectively. Therefore, such speakers appeared to be relatively more similar regarding their F1 patterns than for F2.

When considering vowel specificity, five identical twin pairs have shown to contrast significantly through the analysis of both F1 and F2 in specific vowels, namely, B1-B2; C1-C2; D1-D2; E1-E2; G1-G2.

As for cross-pair differences presented in Table 3.2, speaker comparisons as a function of F1 and

Table 3.1: **Significant (p<0.05/2) and marginally significant differences (p<0.10/2) within identical twin pairs for comparisons considering both stressed and unstressed vowels. Two-tailed test with Bonferroni correction.**

	Front vowels				Back vowels			Differences	
	i	**e**	ɛ	**a**	ɔ	**o**	**u**	**Total**	**%Diff**
F1 (Hz)	B1-B2 G1-G2		C1-C2	B1-B2 C1-C2 G1-G2	*G1-G2* *(p=0.026)*	C1-C2		8	11%
F2 (Hz)	B1-B2 D1-D2 E1-E2 G1-G2	*D1-D2* *(p= 0.04)* *E1-E2*	B1-B2 E1-E2	B1-B2 E1-E2			*C1-C2* *(p=0.04)*	11	15%
F3 (Hz)	D1-D2 E1-E2 G1-G2	C1-C2 G1-G2 (p=0.03)	H1-H2	C1-C2 G1-G2 H1-H2	G1-G2 H1-H2		G1-G2 H1-H2	13	18%
F4 (Hz)	D1-D2 E1-E2 G1-G2 J1-J2	E1-E2 G1-G2 (p=0.03)	C1-C2 E1-E2 G1-G2 H1-H2	C1-C2 D1-D2 E1-E2 F1-F2 G1-G2 H1-H2	E1-E2	E1-E2	E1-E2	19	27%
Total	13	6	8	14	4	2	4		
		Mean= 9				Mean= 3		mean=	18%

F2 yielded a percentage difference of 37% and 17%, respectively. As noted, in such a comparison, speakers appeared to be more similar for F2 than for F1.

With regard to intra-twin pair comparisons as a function of high formant frequencies and vowel quality, both F3 and F4 parameters seemed to display a relatively higher proportion of inter-speaker differences when compared with lower formant frequencies. High formant frequencies, i.e., F3 and F4, were able to differentiate seven twin pairs for some specific vowels (C1-C2; D1-D2; E1-E2; F1-F2; G1-G2; H1-H2; J1-J2), as can be seen in Table 3.1.

Moreover, when formant frequencies were assessed independently of vowel quality, it was observed that twin pairs appeared considerably more distinct for F4 when compared to F3, with a percentage of 27% differences observed for the former and 18% for the latter parameter. However, such a discrepancy in the proportion of intra-twin pair differences regarding F3 and F4 appeared to be mostly explained on account of one specific pair, E1-E2. Note that this pair, in particular, was found more dissimilar for F4 than for F3.

As for the comparison between F3 and F4 as a function of cross-pair comparisons (see Table 3.2),

similar proportions of differences have been found for the two formants: 39% and 38%, respectively. Notably, higher proportions than the observed for intra-twin pair comparisons; a result that could have been more or less anticipated, given the nature of the material assessed.

Table 3.2: **Significant (p<0.05/2) and marginally significant differences (p<0.10/2) among all speakers (cross-pairs) for comparisons considering both stressed and unstressed vowels. Two-tailed test with Bonferroni correction.**

	Front vowels				**Back vowels**			Differences	
	i	**e**	**ɛ**	**a**	**ɔ**	**o**	**u**	**Total**	**%diff**
F1 (Hz)	93	81	74	84	49	55	63	499	37%
F2 (Hz)	59	35	42	67	10	1	14	228	17%
F3 (Hz)	51	49	71	128	78	62	82	521	39%
F4 (Hz)	105	65	63	123	52	48	59	515	38%
Total	308	230	250	402	189	166	218		
	Mean= 262				Mean= 191				Mean= 32%

Table 3.3: **Significant (p<0.05/2) and marginally significant differences (p<0.10/2) within identical twin pairs for F1 x F2 between identical twins for stressed (top), and unstressed vowels (bottom). Two-tailed test with Bonferroni correction.**

				Stressed					Differences
	i	**e**	**ɛ**	**a**	**ɔ**	**o**	**u**	**total**	**%diff**
F1 (Hz)	B1-B2		C1-C2	B1-B2 C1-C2 G1-G2		C1-C2		6	8 %
F2 (Hz)	B1-B2 E1-E2	E1-E2	B1-B2 E1-E2	B1-B2 E1-E2			*J1-J2* *(p=0.03)*	8	11%
Total	3	1	3	5	0	1	1		Mean= 10 %
				Unstressed					Differences
	i	**e**	**ɛ**	**a**	**ɔ**	**o**	**u**	**total**	**%diff**
F1 (Hz)				C1-C2 G1-G2				2	3 %
F2 (Hz)	D1-D2		E1-E2	*B1-B2* *(p= 0.026)* *E1-E2*				4	5 %
Total	1	0	1	4	0	0	0		Mean= 4 %

Effect size estimates– which serves as a statistical metric allowing the assessment of the direction and strength regarding the relationship between variables (BERBEN; SEREIKA; ENGBERG, 2012)– are presented as a function of individual vowels and formant frequencies in Table 3.5. By inspecting such a Table, it can be verified that, for comparisons performed while combining stressed

Table 3.4: **Number of significant differences for F1-F2 among all speakers (cross-pairs) for stressed, and unstressed vowels (p<0.05/2). Two-tailed test with Bonferroni correction.**

	Stressed							Differences	
	i	**e**	ɛ	**a**	ɔ	**o**	**u**	**total**	**%diff**
F1 (Hz)	72	65	48	73	28	35	30	351	26 %
F2 (Hz)	45	29	30	52	4	-	4	164	12 %
Total	117	94	78	125	32	35	34		19 %
	Unstressed							Differences	
	i	**e**	ɛ	**a**	ɔ	**o**	**u**	**total**	**%diff**
F1 (Hz)	42	19	23	55	7	12	25	183	13 %
F2 (Hz)	18	2	6	33	**-**	**-**	**-**	59	4 %
Total	60	21	29	88	7	12	25		9 %

and unstressed vowel, F4 was the only parameter displaying large effect sizes for all vowels assessed, followed by F3 and F1, which varied from mostly large to moderate effect sizes. In contrast, F2 displayed the smallest effect sizes, varying from mostly moderate to large effect sizes in front vowels to moderate and small effect sizes in back vowels.

In addition, when comparing global effect sizes as a function of formant frequencies, regardless of vowel specificity, it can be observed that, among all formants, only F3 and F4 displayed large effect sizes for inter-speaker comparisons, whereas both F1 and F2 displayed small global effect sizes (see column "all vowels" in Table 3.5).

Overall, the analysis of F1-F4 mean formant frequencies was able to contrast eight pairs of identical twins out of ten. However, two identical twin pairs were still considerably similar regarding their outcomes, not being effectively contrasted through their formant frequency patterns; namely: A1-A2 and I1-I2. In addition, the outcomes seem to suggest that, amongst all measures, the high-frequency formants F3 and F4 were potentially the most inter-speaker discriminatory in intra-twin pair comparisons, whereas F1, F3, and F4 appeared to be the most discriminatory in cross-pair comparisons.

Note, however, that the overall performance of such parameters when more standardized assessment criteria are adopted, e.g., EER, Cllr, and AUC metrics, has yet to be explored, as will be done further.

3.2.2 Differences in the Bark critical-band scale

Regarding intra-twin pair comparisons by employing the Bark critical-band scale, some dissimilarities have been observed compared to the results obtained while using the Hertz scale. All

Table 3.5: **Effect size (η^2) estimates for vowel segments and formant frequencies (large effect sizes in blue).**

	i	e	ɛ	a	ɔ	o	u	All vowels
Stressed and unstressed vowels								
F1	**0.237**	**0.312**	**0.253**	0.106	**0.294**	**0.292**	**0.167**	0.0363
	large	**large**	**large**	mod	**large**	**large**	**large**	small
F2	0.128	0.133	**0.154**	0.0965	0.105	0.0499	0.0560	0.0261
	mod	mod	**large**	mod	mod	small	small	small
F3	0.0972	**0.174**	**0.284**	**0.346**	**0.566**	**0.365**	**0.294**	**0.191**
	mod	**large**	**large**	**large**	**large**	**large**	**large**	**large**
F4	**0.326**	**0.294**	**0.265**	**0.211**	**0.311**	**0.275**	**0.172**	**0.202**
	large	**large**	**large**	**large**	**large**	**large**	**large**	**large**
Stressed vowels								
F1	**0.263**	**0.295**	**0.224**	**0.188**	**0.263**	**0.262**	**0.207**	0.0312
	large	**large**	**large**	**large**	**large**	**large**	**large**	small
F2	**0.158**	0.130	**0.174**	**0.159**	0.0808	0.0615	0.0867	0.0267
	large	mod	**large**	**large**	mod	mod	mod	small
Unstressed vowels								
F1	**0.209**	**0.397**	**0.330**	0.102	**0.444**	**0.414**	0.137	0.0496
	large	**large**	**large**	mod	**large**	**large**	mod	small
F2	0.108	**0.173**	0.121	0.0606	**0.203**	0.0594	0.0385	0.0273
	mod	**large**	mod	mod	**large**	small	mod	small

dissimilarities corresponded to the front vowel [i], occurring mostly for F2, F3, and F4.

Four comparisons that were significant in Hertz have shown to be only marginally significant in Bark ($p \leq 0.10/2$): D1-D2 (F2); E1-E2 (F3); G1G2 (F4); J1-J2 (F4). Also, three significant comparisons in Hertz were found to be non-significant in Bark for the same vowel [i]: G1-G2 (for F2); D1-D2 (for F3); D1-D2 (for F4).

Such dissimilarities regarding the non-significant differences in Bark stand for only 6.5% of the total number of significant differences observed in Hertz. Apart from these observations, there seemed to be a noteworthy agreement regarding the comparison of statistical results across the two scales. Overall, 93.5% of differences observed in Hertz were also significant or marginally significant in Bark.

3.2.3 The discriminatory patterns of phonetic vowels in BP

Significant differences within identical twin pairs and in cross-pair comporisons as a function of vowel quality are reported in Tables 3.1 and 3.2, respectively, as well as the total number of

differences for each vowel. Numbers of significant differences are also displayed according to an articulatory phonetic criterion, namely the distinction amongst front, central, and back vowels.

As can be seen in Tables 3.1 and 3.2, amongst the phonetic vowels of BP, the central vowel [a] was the one displaying the highest number of significant differences in intra-twin pair and cross-pair comparisons, followed by the front vowels [i], [ɛ] and [e], respectively. Furthermore, from the back vowels class, the vowel [u] seemed to display the highest proportions of inter-speaker differences across all subjects, whereas for identical twins, both [ɔ] and [u] were found to portray fewer inter-speaker differences.

As observed in Table 3.1 for both F1/F2 and F3/F4 formant groups, front vowels appeared to be, in general, considerably more discriminatory than back vowels. In terms of (primary) vowel quality-related formants (i.e., F1-F2), only two identical twin pairs were able to be differentiated by the analysis of the back articulatory dimension, whereas five twin pairs were effectively contrasted through the comparison of front vowels. When considering all formant frequencies (F1-F4), an average of nine significant intra-twin combinations for the former and three for the latter vowel category were observed. Notably, F2 estimates from back vowels appeared to be the parameter displaying the lowest inter-speaker discriminatory potential concerning identical twin pairs.

Although in a relatively smaller proportion, a discrepancy in inter-speaker differences for front and back vowels was also found for cross-pair comparisons, as shown in Table 3.2. Again, F2 from back vowels was the parameter displaying the lowest proportion of inter-speaker differences.

In Table 3.5, effect size estimations can be visualized as a function of vowel quality in comparisons carried out among all speakers. Regarding the combination of stressed and unstressed vowels, the front vowel [ɛ] was the only segment displaying only large effect sizes, suggesting higher differences in mean values among individuals for this segment, as well as a greater explanatory potential of the variable "*speaker*" on the observed differences. In summary, all reported differences concerning vowel quality ranged from moderate to large effect sizes, except for the back vowels [o] and [u], for which small magnitudes regarding F2 were found. This outcome suggests less substantial differences in terms of F2 average values among speakers for the before-mentioned vowel segments.

Representations of the two-dimensional vocalic spaces from twin pairs are depicted in Figure 3.1. By visually inspecting the over-plots, it is possible to observe how closely related such individuals are in terms of their vocalic acoustic patterns, mainly in terms of their F1 and F2 mean values. Concerning the phonetic variability in the front/back articulatory dimensions, diagrams representing the areas corresponding to a confidence level interval of 95% are displayed for each vowel and each twin pair in Figures 3.2 and 3.3. As such, the area inside the ellipses corresponds to 95% of observed data points, in which average values are expressed by the vowel letters (please,

note that: eh = ɛ, and oh = ɔ).

Figure 3.1: **Two-dimensional vocalic space comparisons (intra-twin pairs)**

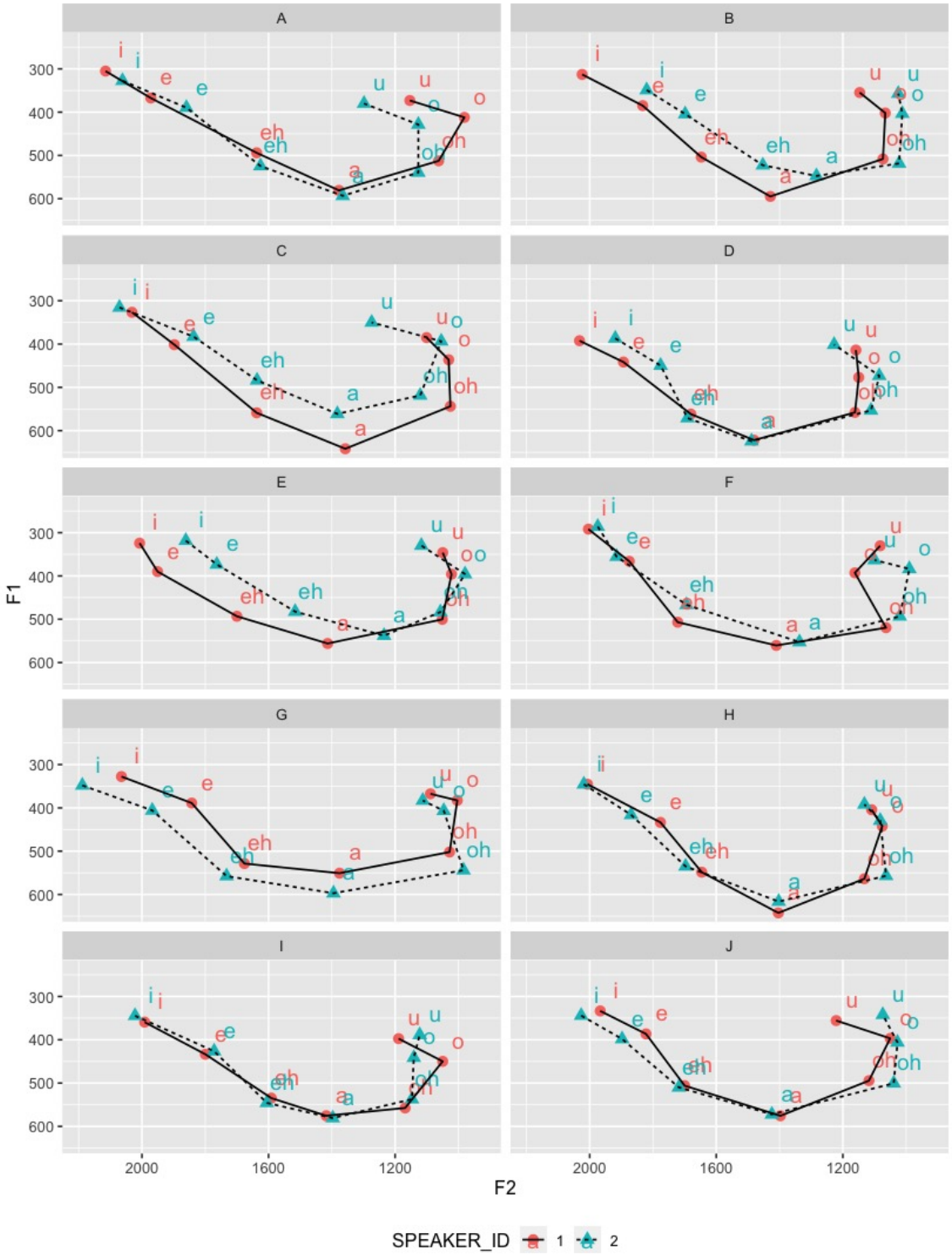


Figure 3.2: **Front vowels vocalic spaces' means and confidence interval for intra-twin pair comparisons.**

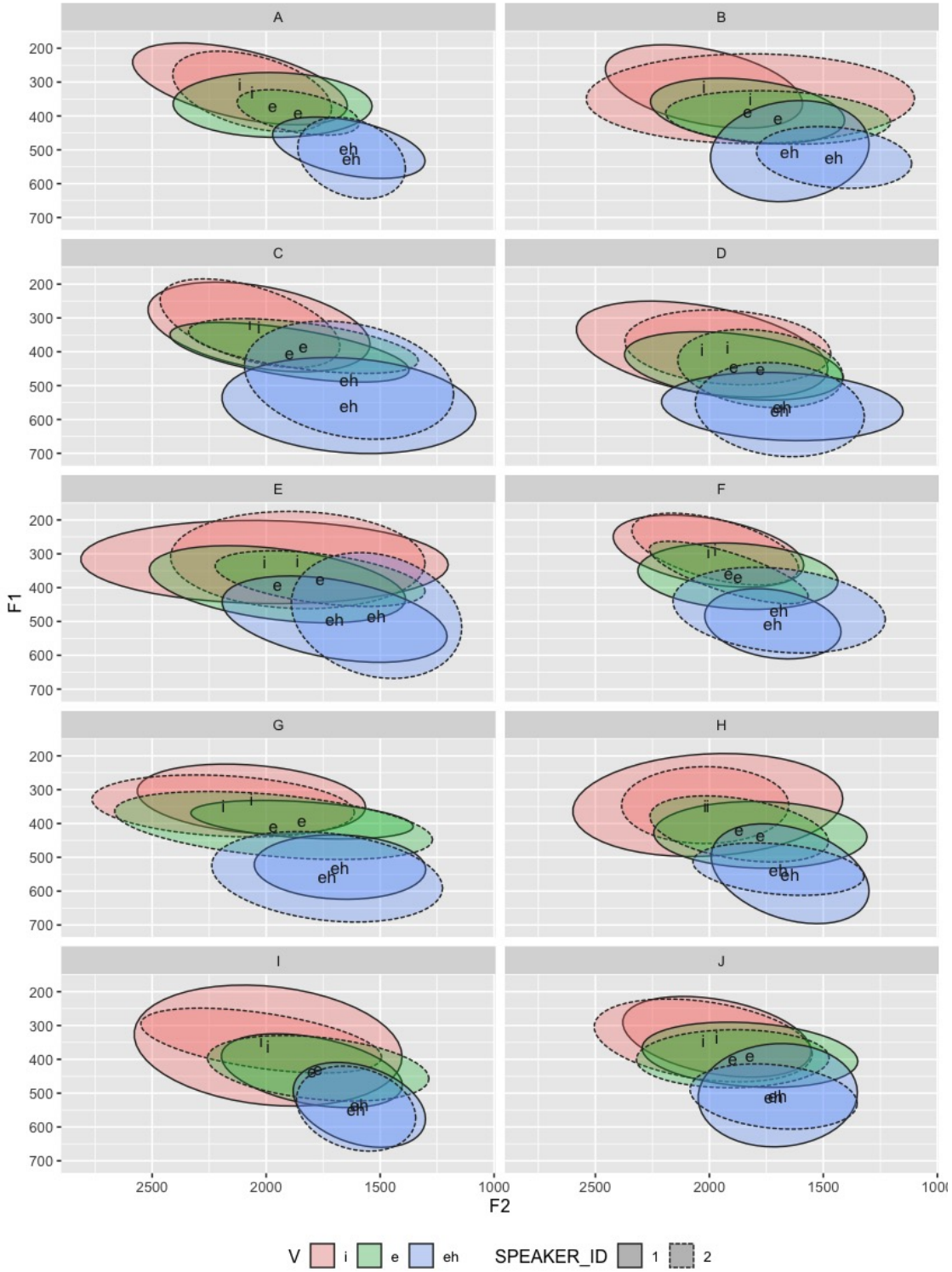
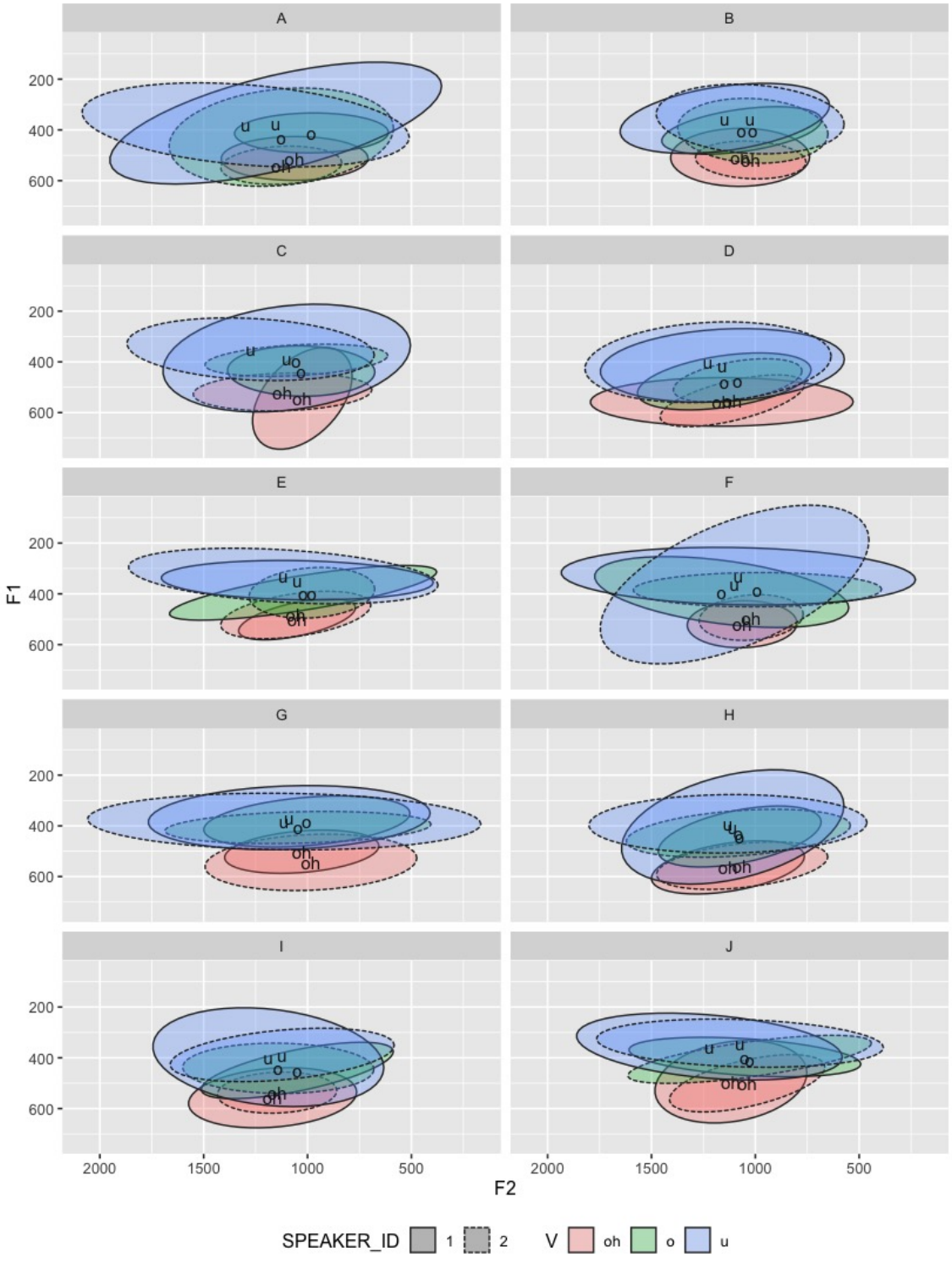


Figure 3.3: **Back vowels vocalic spaces' means and confidence interval for intra-twin pair comparisons.**



In Figure 3.3, a higher degree of overlap between the ellipses for back vowels suggests closer acoustic-phonetic proximity for these vowels compared to the front ones. Furthermore, such acoustic-phonetic proximity was confirmed by the assessment of the Euclidean distances between neighboring vowels in the cardinal vowel space.

As shown in Table 3.6, which presents the intervocalic Euclidean distances and mean F1-F2 values between phonetic vowels, there was a tendency for front vowels to be considerably more dispersed when compared to their back counterparts. The Wilcoxon test revealed statistically significant differences for vowel height (F1) in the comparison between [u] and [i], [o] and [e] ($p < 0.001$) with the exception of [ɔ] and [ɛ] ($p = 0.7$). In this case, the back vowels [u] and [o] were considerably lower than their front vowel counterparts, resulting in a vocalic asymmetry.

Furthermore, the entire front and back vertical articulatory dimensions also appeared to exhibit a height discrepancy, in which the F1 mean difference in distance between the extreme front vowels [i]-[ɛ] and the extreme back ones [u]-[ɔ] was found significantly discrepant ($p < 0.001$). The comparison of the Euclidean distances between the cardinal vowels displayed in Table 3.6 also seemed to corroborate this discrepancy related to articulatory working space, particularly when considering the extreme vowels in the front articulatory dimension [i]-[ɛ] and back [u]-[ɔ].

With regard to vowel formant variability, front vowels were found more variable for both F1 and F2, with mean standard-deviations of 94.8 Hz and 250.9 Hz, respectively, compared to back vowels, for which mean standard-deviations of 84.2 Hz and 227.1 Hz were observed, respectively. The reported variance differences were found statistically significant when performing the Fligner-Killeen test of homogeneity of variances, suggesting different variances for the front and back vowels (F1: $\chi^2 = 61.731$, $p < 0.001$; and F2: $\chi^2 = 75.129$, $p < 0.001$).

Table 3.6: **Intervocalic Euclidean distances and mean F1-F2 differences between phonetic vowels.**

Class	Vowels	Euclid. dist (Hz)	Euclid. dist (Bark)	F1 diff (Hz)	F2 diff (Hz)
	i – e	175	0.84	65	163
	e – ɛ	235	1.33	123	200
Front	i – ɛ	409	2.18	188	363
	u – o	111	0.59	44	82
	o – ɔ	93	0.98	109	24
Back	u – ɔ	164	1.42	153	58
	a – ɛ	273	1.27	61	266
Central	a – ɔ	317	1.67	59	312

3.2.4 The lexical stress effect

Contrary to what was initially hypothesized regarding the stress effect on intra-twins and cross-pair formant frequency comporisons, the numbers reported in Tables 3.3 and 3.4 suggest the same trend in terms of acoustic-phonetic differences for both groups. In general, stressed vowels displayed a higher proportion of inter-speaker differences than unstressed vowels. Furthermore, all identical twin pairs that were contrasted through unstressed vowels were also discriminated through stressed vowels comparisons, except for one pair: D1-D2.

From a more comprehensive approach, by considering the combination of stressed and unstressed vowels, as presented in Tables 3.1 and 3.2, an even higher number of significant differences were found for both types of comparisons (i.e., intra-twin and cross-pair comparisons), which may suggest the combination of stressed and unstressed vowels as being more explanatory in terms of the differences observed amongst individuals. However, note that such a condition also involves a higher number of data points.

Through the inspection of Figure 3.4, which displays the data points corresponding to stressed and unstressed segments produced by each speaker, it can be observed how their vowels are dispersed in the vocalic space. A higher concentration of stressed vowels in peripheral regions, as well as a more centralized dispersion of unstressed vowels, could be visualized. This tendency was confirmed when mean F1 and F2 frequencies are plotted in the cardinal space. As can be visualized in Figure 3.5, unstressed vowels' formant frequencies displayed a tendency to be more centralized than stressed vowels, which seemed to be particularly the case for the central vowel [a], resulting in a vertical reduction of the vowel space. Such a tendency seems to apply to all speakers. By visually comparing the global distribution of stressed and non-stressed vowel data points (Figure 3.6), their association (Figure 3.7A), and global average values (Figure 3.7B), the difference between these two vowel classes is suggested. The patterns in Figure 3.7B also suggest the combination of stressed and non-stressed vowels behaving as an intermediate state, as represented by the in-between lines, particularly for the vowel [a].

When analyzing effect sizes as a function of lexical stress in Table 3.5, it is possible to observe magnitudes ranging from moderate to large in stressed vowels and small to large in unstressed vowels. Moreover, F1 of stressed vowels was the only parameter displaying only large effect sizes. In general, in both stressed and unstressed conditions, effect sizes were smaller for F2 than for F1. Finally, regarding individual vowel segments, the stressed vowels [a], [i], [ɛ], and the unstressed vowels [e] and [ɔ], were the segments displaying only large effect sizes. That is three front vowels, one central, and one back. Once again, inter-speakers effect sizes were numeric larger than in intra-speaker comparisons.

Figure 3.4: **Stressed and unstressed vowel data points**

Figure 3.5: **The effects of stress in the two-dimensional vocalic space intra- and inter-speakers.**

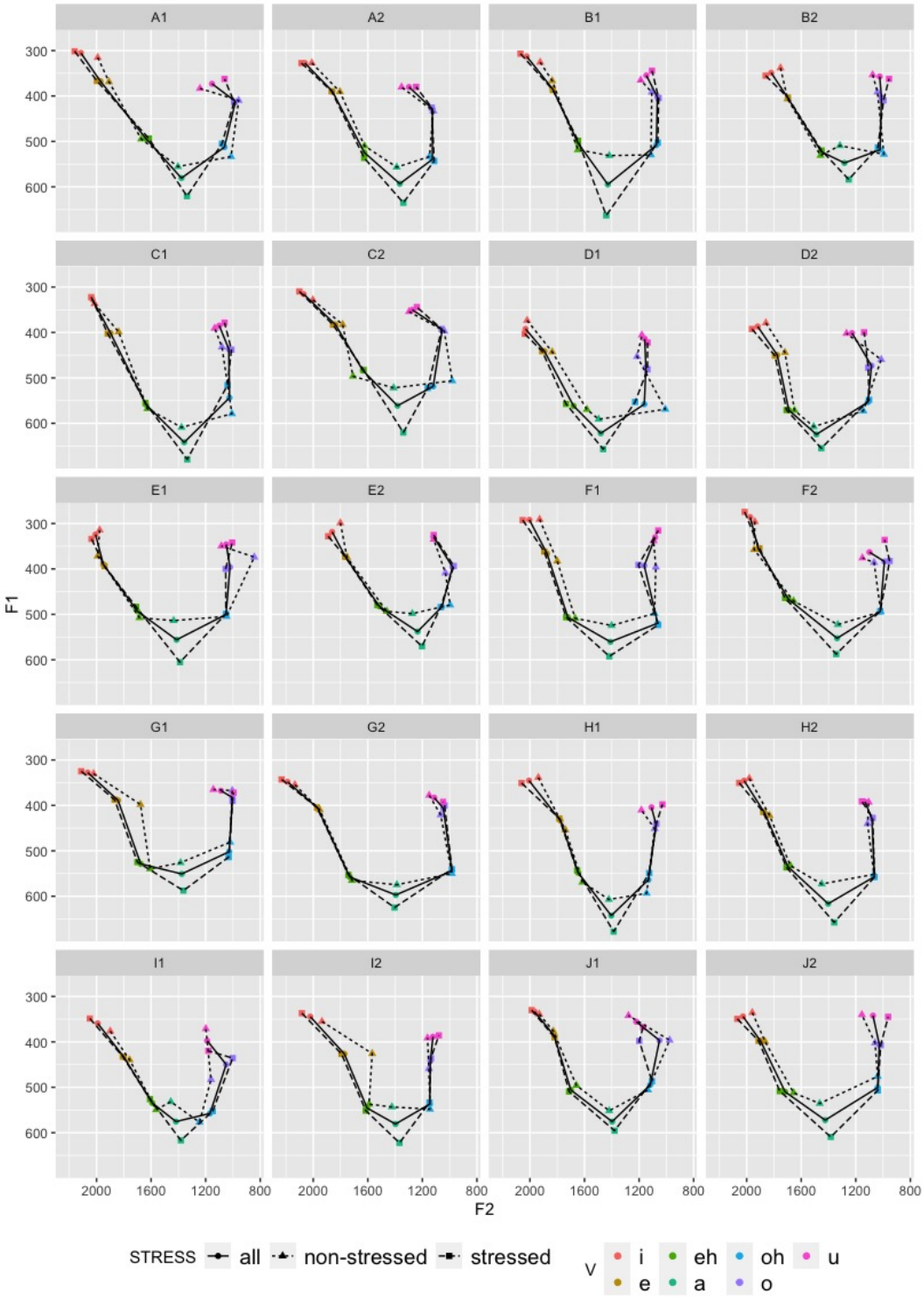
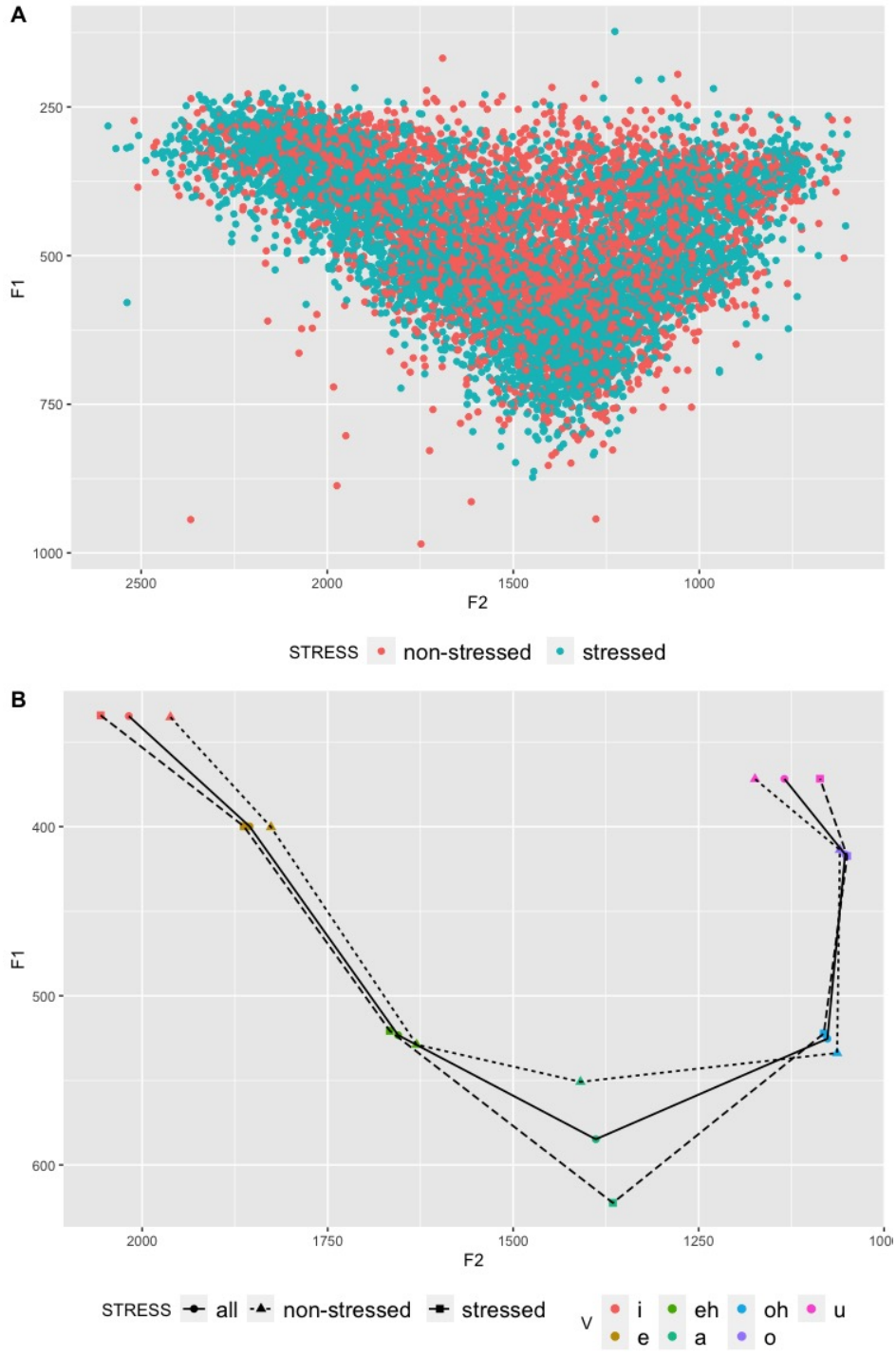


Figure 3.6: **Global representation of stressed and non-stressed vowels' data points in the acoustic space as a function of F1 and F2.**

Figure 3.7: **Global representation of cardinal stressed and non-stressed vowels in the acoustic space as a function of F1 and F2.**



With regard to vowel formant variability, stressed vowels were found more variable than unstressed vowels for F1 and F2, with an observed standard deviation of 128.2 Hz and 403.7 Hz in stressed vowels, respectively, compared to 121.3 Hz and 355.3 Hz in unstressed vowels, respectively. Such difference was also statistically significant when performing the Fligner-Killeen test of homogeneity of variances, suggesting that, in fact, stressed and unstressed vowels display different variances (F1: $\chi^2 = 31.268$, $p < 0.001$; and F2: $\chi^2 = 118.41$, $p < 0.001$).

3.2.5 The discriminatory performance of formant frequencies

We now turn to the reporting of the discriminatory performance of individual formant frequencies as a function of vowel quality and stress. System performance estimates, namely Cllr, EER, and AUC values, are depicted in Tables 3.7, 3.8, and 3.9. Note that the performance estimates reported regards the comparison involving all speakers in the corpus (i.e., 190 inter-speaker comparisons), upon which the discriminatory performance of parameters was assessed. In Tables 3.7 and 3.8 effect sizes are re-evoked to provide a broader overall perspective, including the proportion of differences observed across all individuals in Table 3.7.

By inspecting Tables 3.7 and 3.8, which presents performance estimates as a function of formant frequencies, it can be seen that there was not always a complete agreement concerning different diagnostic/performance estimates. However, such an outcome was already anticipated, mainly due to the fact that such metrics are based on different criteria. As such, a special emphasis will be drawn on those parameters displaying the best relative trade-off between the different assessment approaches.

Despite some level of disagreement observed across metrics, some overall patterns can be noted. As can be verified in Table 3.7, among all tested formant frequencies, F3 was the only parameter which better satisfied most conditions, while also presenting a large effect size on account of the variable *speaker*. Moreover, although a relatively higher binary classification performance was verified for F3 and F4 in relation to lower formant frequencies, along with a higher proportion of inter-speaker differences, F4 was the parameter depicting the highest Cllr and EER values, which is compatible with a lower accuracy when assessed within the likelihood ratio framework.

Regarding the formants' discriminatory performance when considering the variable "*lexical stress*", a similar pattern is suggested. As can be seen in Table 3.8, in which two of the best performing parameters are highlighted, F3 was again the parameter that best satisfied most conditions, with F4 and F2 displaying, overall, the worst performance in terms of accuracy. Note, however, that when considering solely the discriminatory power of formants expressed by the highest AUC values, both F3 and F4 came out, in all cases, as the best performing estimates. Notwithstanding,

between the two high formant frequencies, only F3 was found to display the desired properties, as expressed by lower Cllr and EER values.

Concerning classification performances in terms of ROC analysis, in Figures 3.8, 3.10, and 3.11 ROC curves are presented, in which binary classification performance of formant frequencies are depicted regardless of vowel quality. In Figure 3.8, which concerns intra-twin pair comparisons, a relatively poorer performance can be observed when compared to some of the cross-pair comparisons depicted in Figures 3.10 and 3.11. Overall, it is possible to verify that, as far as AUC values are concerned, the best performing parameters mostly varied between F3 (in blue) and F4 (in red). Notably, some exceptions can be identified.

With respect to the effects of lexical stress on the discriminatory performance of individual vowels (see Table 3.9), some previously mentioned patterns have been corroborated. The first regards to the fact that, when all formant frequencies are fused, a slightly better overall performance could be observed for most vowels (i.e., stressed and unstressed combined). This seemed to be particularly the case when comparing calibrated Cllr and EER values across stressed/unstressed vowels in relation to the combination of both categories (see vowels [a], [e], [ɛ], [ɔ], [o], and [u]). As such, the combination between stressed and unstressed vowels tended to yield an overall better discriminatory performance. The vowel [a] is a clear example of such a trend.

As for vowel quality, by inspecting Table 3.9, it can be easily verified that the central vowel [a] was the segment displaying, overall, the best discriminatory performance, as suggested by the lowest Cllr and EER. In addition, when a comparison between front and back vowels is made, a slightly better performance was suggested for the front segments, as indicated by slightly lower EER values in the former vowel category in comparison to the latter in most cases, mainly when stressed and unstressed vowels are combined (see vowels [i]/[u] and [e]/[o]).

Table 3.7: **Overall discriminatory performance of individual formant frequencies after fusing all vowels and performing a calibration procedure.**

Parameter	$Cllr_{(cal)}$	EER	AUC	% Differences	Effect size	Magnitude
F1	0.40	0.15	0.56	37 %	0.03	Small
F2	0.59	0.25	0.55	17 %	0.02	Small
F3	**0.40**	**0.15**	**0.65**	**39 %**	**0.19**	**Large**
F4	0.74	0.28	0.66	38 %	0.20	Large

Note that, when combining different vowels the discriminatory performance is considerably improved, mainly expressed by a reduction in Cllr values. The same trend is expected to apply for the combination between different specific formants, which was not directly explored here.

Table 3.8: **Discriminatory performance of individual formant frequencies as a function of stress after fusing stressed and unstressed vowels (separately) and performing a calibration procedure.**

Stressed vowels					
Parameter	**Cllr$_{(cal)}$**	**EER**	**AUC**	**Effect size**	**Magnitude**
F1	**0.47**	**0.15**	0.55	0.03	Small
F2	0.69	0.24	0.54	0.02	Small
F3	**0.44**	**0.14**	**0.64**	**0.18**	**Large**
F4	0.74	0.31	**0.66**	**0.21**	**Large**
Unstressed vowels					
Parameter	**Cllr$_{(cal)}$**	**EER**	**AUC**	**Effect size**	**Magnitude**
F1	**0.46**	**0.14**	0.57	0.04	Small
F2	0.78	0.26	0.55	0.02	Small
F3	**0.53**	**0.19**	**0.66**	**0.21**	**Large**
F4	0.77	0.25	**0.65**	**0.18**	**Large**

Table 3.9: **Discriminatory performance of individual vowels after fusing all formants and performing a calibration procedure.**

	Stressed and unstressed						
Performance metric	**i**	**e**	**ɛ**	**a**	**ɔ**	**o**	**u**
Cllr$_{(cal)}$	0.51	0.45	0.52	**0.33**	0.49	0.60	0.57
EER	0.15	0.15	0.17	**0.12**	0.16	0.20	0.25
AUC	0.67	0.66	0.68	**0.67**	0.70	0.66	0.61
	Stressed						
Performance metric	**i**	**e**	**ɛ**	**a**	**ɔ**	**o**	**u**
Cllr$_{(cal)}$	0.56	0.46	0.59	**0.44**	0.46	0.65	0.56
EER	0.25	0.15	0.19	**0.15**	0.16	0.28	0.20
AUC	0.68	0.67	0.69	**0.70**	0.71	0.68	0.65
	Unstressed						
Performance metric	**i**	**e**	**ɛ**	**a**	**ɔ**	**o**	**u**
Cllr$_{(cal)}$	0.56	0.74	0.84	**0.42**	0.61	0.85	0.76
EER	0.22	0.25	0.30	**0.15**	0.21	0.30	0.25
AUC	0.66	0.69	0.68	**0.65**	0.74	0.69	0.59

3.3 Discussion

The present experiment consisted of a survey on the inter-speaker discriminatory power of vowel formant frequencies in genetically-related, namely identical twin pairs, and among all speakers (i.e., cross-pair comparisons). The effect of vowel stress regarding the formant's discriminatory potential was also considered. The main findings are discussed in the following sections.

Figure 3.8: **ROC curves and AUC values: intra-twin pair comparisons I**

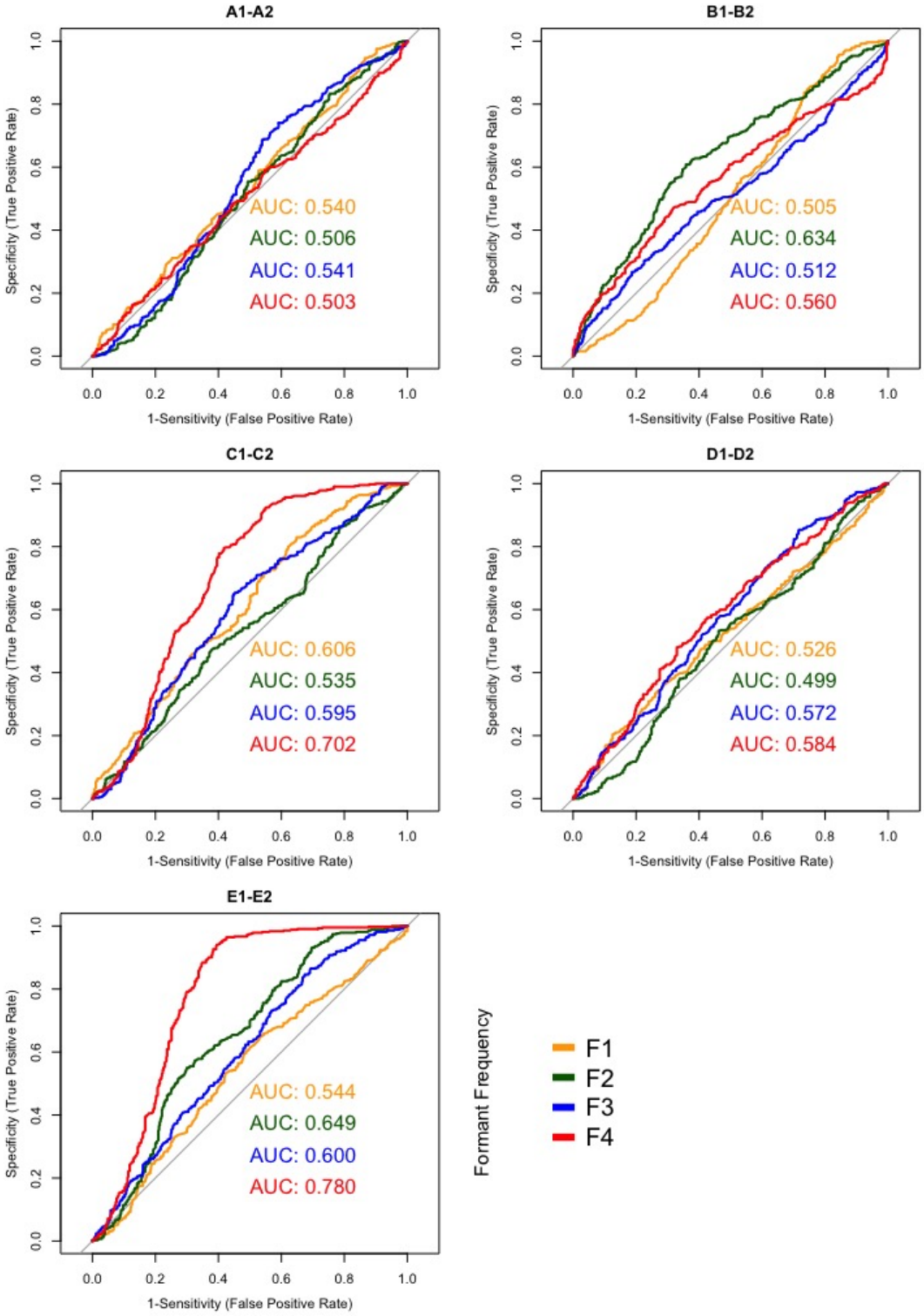


Figure 3.9: **ROC curves and AUC values: intra-twin pair comparisons (II)**

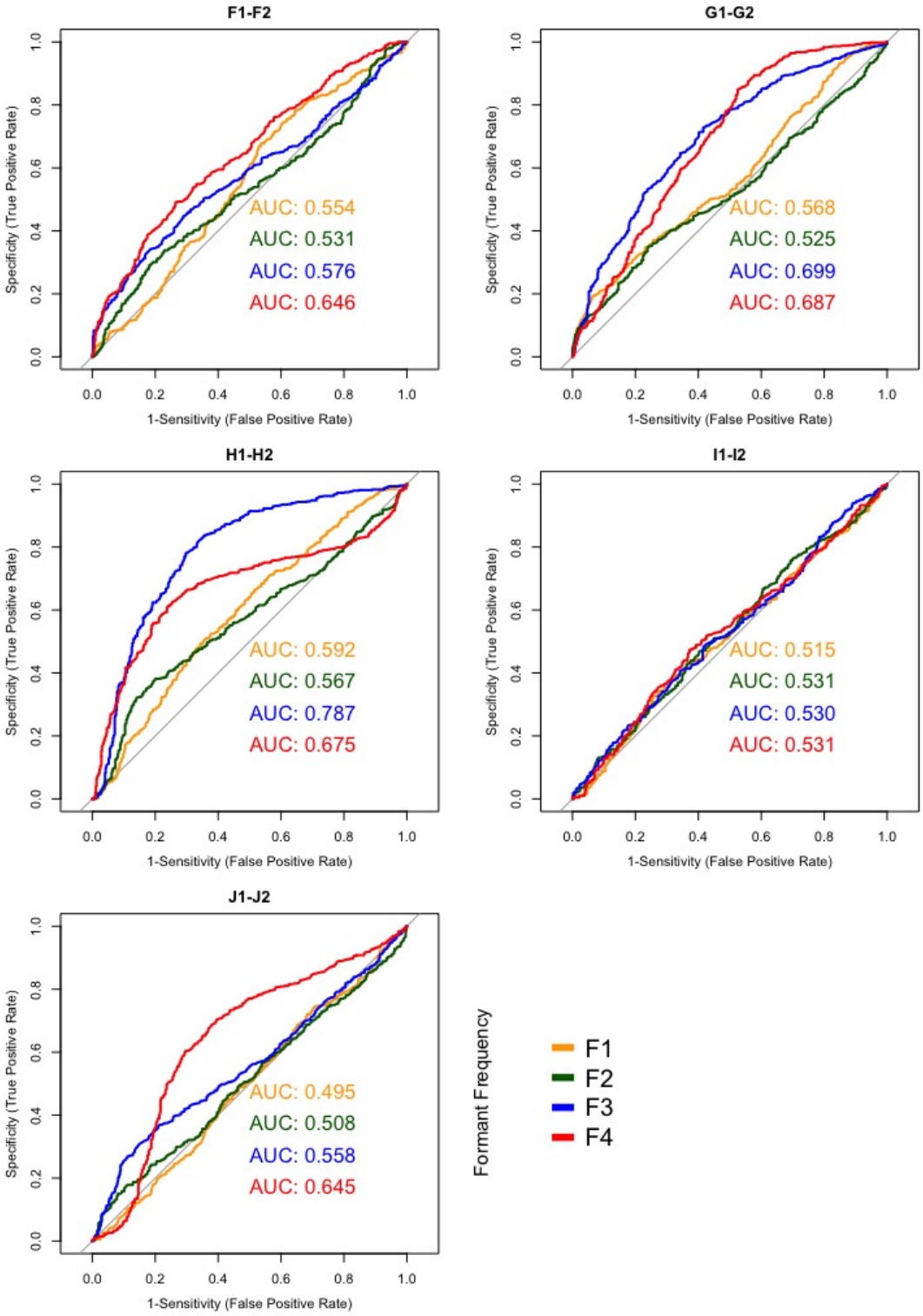


Figure 3.10: **ROC curves and AUC values: cross-pair comparisons (I)**

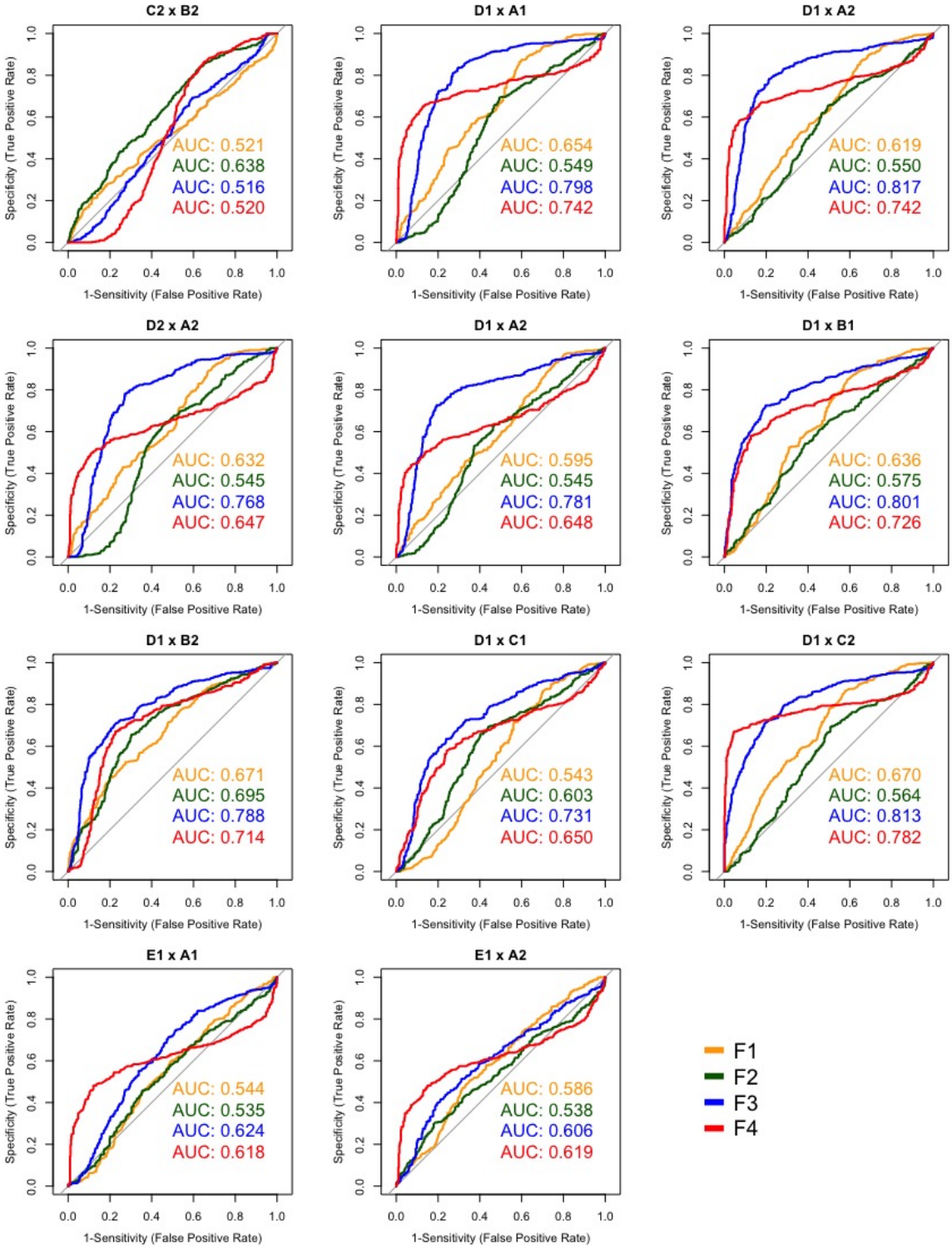
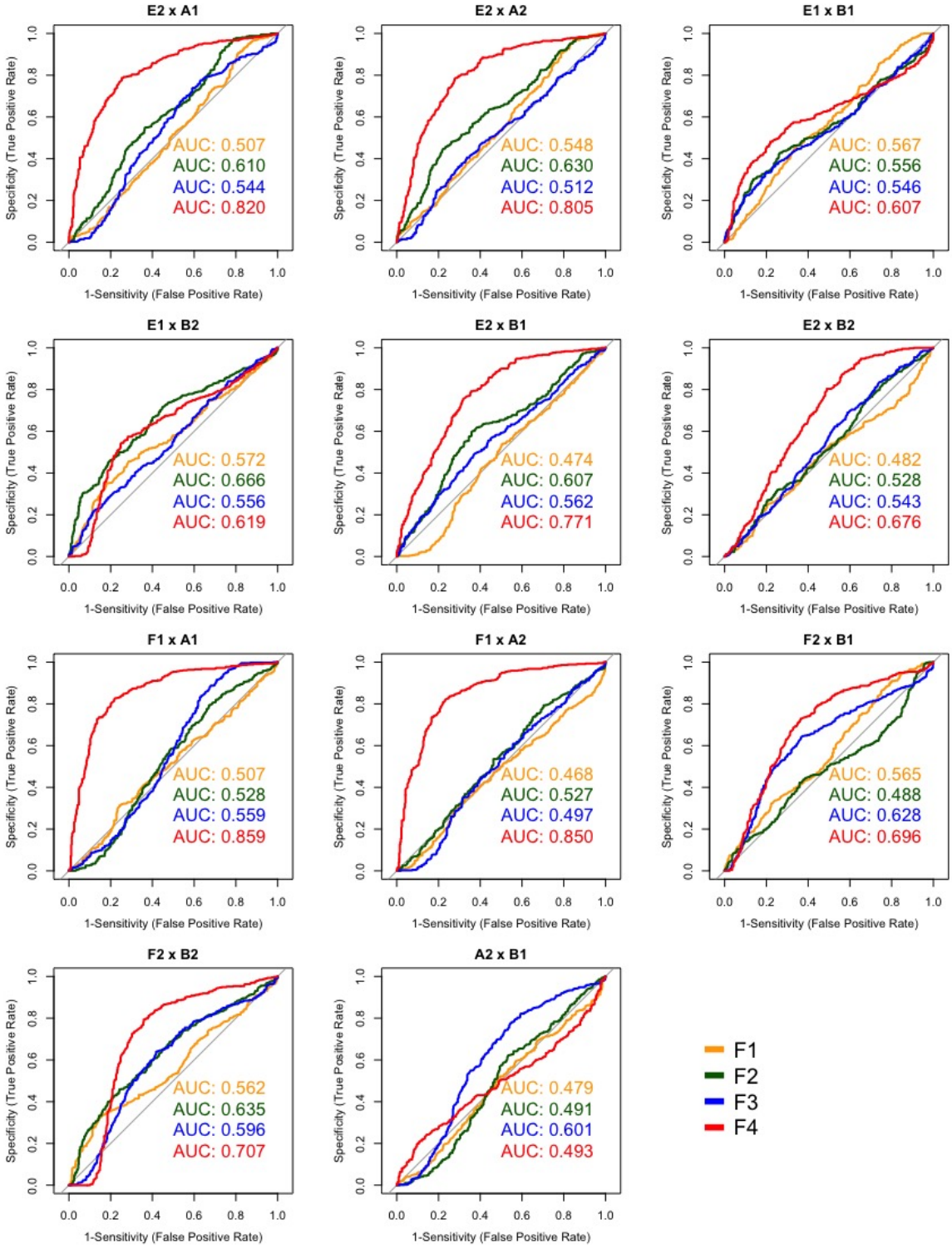


Figure 3.11: **ROC curves and AUC values: cross-pair pair comparisons (II)**



3.3.1 F1-F4 differences in intra-twin and cross-pair comparisons

As observed in the present experiment, for both identical twin pairs and cross-pairs, high-formant frequencies appeared to be more speaker discriminatory in comparison to lower formant frequencies. This finding is in broad agreement with trends reported in the literature regarding experiments with identical twins in both controlled and uncontrolled speech (NOLAN; OH, 1996; WEIRICH, 2010, 2012; LOAKES, 2003); as well as with non-genetically related individuals (CAO; DELLWO, 2019).

Given the control of the linguistic component on the limits of variation allowed for (primary) vowel quality-related formants and the superposition of physiological and linguistic components on identical twin pairs, fewer differences for F1 and F2 compared to higher formants may be presumed. For comparisons carried out across all speakers, however, the impact of the linguistic component, as in the case of the shared dialect, seems to apply, especially for F2. This outcome may be related to a lower variation for F2 observed in back vowels, as shown in Table 3.2. It is worth mentioning that the F2 dimension is importantly related to vowel place of articulation, serving as an acoustic indicator of the constriction position in vowel production (STEVENS; HOUSE, 1955). In this sense, a higher proportion of inter-speaker differences across all individuals for F1 may suggest that variation for such a formant frequency is more tolerated than for F2.

As for higher formants, such as F3 and F4, lower linguistic constraints may be presumed. According to Traunmüller (1984), the position of the higher formants in the spectrum, such as F3, and F4, is largely determined by the vocal tract length. Moreover, in the experiment conducted by Stevens and House (1955), the authors observed that while F2 tended to increase in frequency as the point of constriction moved forward from the glottis, there was only a small increase in F3 as the mouth opening increased in size and became less rounded during the referred movement. In general, the rate of the increase depended mostly on the size of the constriction.

Such an observation invites the hypothesis of high formants' resonance variations associated with vocal tract configurations and individual phonatory settings as possibly more speaker-discriminatory, as justified by the observation of such frequencies as less dependent on the phonetic quality of sounds, which appears to be particularly the case for F4. As observed in the experiment conducted by Takemoto, Adachi, et al. (2006), the fourth formant frequency (F4) seems to be mainly sensitive to laryngeal cavity changes while insensitive to changes in the upper part of the vocal tract. Note, however, that when discriminatory performance is considered on account of the variation observed across all speakers in the present study, F3 was suggested as a more accurate classifier, whereas F4, although being a good classifier, came out as the worst parameter in terms of accuracy.

Two main widely acknowledged factors may be related to the level of phonetic variation observed in identical twins and across all speakers, namely physiological and linguistic factors. The implied relation between these two components has already been addressed by Nolan (1983) when highlighting that physically-related acoustic dimensions, such as fundamental frequency and formant frequencies, are equally exploited by languages and are therefore conflated with linguistic information. In that sense, while the physiological component may establish the limits of physical variation, the linguistic component is responsible for keeping parameters "constant", restraining the variation allowed by the linguistic system. Either the higher or lower superposition of these two dimensions may possibly account for the differences observed across individuals. Different levels of superposition of such dimensions even between identical twins may be observed, given the fact that some identical twin pairs were found to be more similar than others, even when taking part in the same dialogue.

The verification that a lower discrepancy between the discriminatory potential of F1 and F2 was verified for identical twin pairs when compared to the whole group is noteworthy, as revealed by the comparison of the first two rows in Tables 3.1 and 3.2. Such an outcome may be potentially interpreted as the result of a comparable influence of linguistic and structural components on identical twins' linguistic output, yielding a higher similarity in their productions. Moreover, as already acknowledged, it is plausible that such a convergence may be enhanced by the speaking task, i.e., dialogue between twins. The same does not apply, however, to all speakers. The F1 and F2 dimensions seemed to diverge to a greater extent in cross-pair comparisons, where a non-analogous influence of linguistic/environmental and structural factors are not implied.

Notably, studies carried out with identical twins and non-identical twin subjects, as in the case of siblings or unrelated subjects, seem to corroborate this greater phonetic similarity observed regarding identical twin pairs, providing evidence that, considering some exceptions, a higher phonetic convergence is present in such individuals, as verified through the analysis of electromagnetic articulography (WEIRICH; LANCIA; BRUNNER, 2013), acoustic speech (WHITESIDE; RIXON, 2003; WEIRICH, 2012; SAN SEGUNDO, 2014) and voice analysis (SAN SEGUNDO; GÓMEZ VILDA, 2013; SAN SEGUNDO; TSANAS; GÓMEZ-VILDA, 2017; SAN SEGUNDO, 2014).

As observed by Whiteside and Rixon (2003), given the extent of genetic influences on the peripheral structures involved in speech production, such as the vocal tract and the larynx structure, it is expected that higher levels of physical similarity may influence speech characteristics of identical twins. Another critical variable to be considered refers to the implications of the shared linguistic environment on the speech behavior. Therefore, it can be hypothesized that the greater the influence of the linguistic component, the lower is the variability expected regarding a particular physical measure. However, determining the contribution of these two components in the

shaping of speech patterns remains a challenge for phonetic research. In this regard, considering that in Zuo and Mok (2015), identical twins raised apart exhibited the same degree of similarities in their vocalic transitions as identical twins raised together did, suggest a greater influence of physiology over learning.

Moreover, in Weirich, Lancia, and Brunner (2013), MZ twins that were more frequently in contact with each other displayed more comparable levels of similarities than to those that were less frequently in contact. However, it is worth noting that all twins in the latter study were brought up and lived together most part of their lives and as teenagers, which does not allow one to question the environmental influence over the twins, as in the case of the study conducted by Zuo and Mok (2015).

Conversely, evidence pointing to substantial influence of environmental factors on speech patterns have been presented by San Segundo and Yang (2019), while verifying that not only MZ twin pairs but also other siblings were able to deteriorate the performance of a forensic-comparison system based on vowel formant trajectories. According to the authors, factors such as learned variation, individual choice, and the attitude towards one's own sibling seem to play an important role in speech production, and can possibly explain convergences in non-identical individuals.

Finally, the mere fact that identical twins have shown to vary substantially in terms of F3 in the present study, and to a greater extent for F4, suggests that these formants may not be solely dependent on fixed structural features and are as well influenced by dynamic aspects involved in speech production. Differences concerning these formants have been consistently reported by other studies in comparisons involving identical twin pairs (NOLAN; OH, 1996; WEIRICH, 2010, 2012; LOAKES, 2003).

3.3.2 Intra-twin pair comparisons in Bark and Hertz scales

The primary motivation for applying the Bark critical-band scale in the twins' comparisons was to verify whether differences would also be potentially significant when following a psychoacoustic criterion. This verification is especially relevant since variations in anatomic and physiological components involved in speech production could perhaps not be enough to account for all differences observed in identical twin pairs. The verification of convergence in the results between the two scales could imply that the twin pairs may, potentially, be able to perceive such differences, inviting the variable "*choice*" as one possible explanatory component. Notwithstanding, future studies are needed to corroborate this assumption while also estimating the magnitude of the differences observed.

According to Ladefoged (1996), it is possible to categorize speaker-discriminatory variables,

or in the terms used by the author "*speaker-diagnostic*", through two basic distinctions: *organic* versus *acquired/learned*, and within the latter: *individual* versus *group*. As described in Garvin and Ladefoged (1963), the first category has to do with the structure of the vocal organs of speakers and their vocal tract dimensions. Even though such aspects are fundamentally determined by *organic* factors, they are susceptible to suffer modification by learning, as pointed out by the authors.

Concerning the second category— *acquired* or *learned* characteristics— the distinction between *group* versus *individual* is necessary. According to Garvin and Ladefoged (1963), group characteristics are related to social, regional, and cultural conditions, while individual or idiosyncratic features refer to individual variation within a particular group, expressed by patterns that are not predicted from a group perspective. The variable "*choice*" is inserted within this domain.

In regards to speech production, as mentioned in Nolan and Oh (1996), "*choice*" has to do with the selection or adoption of articulatory patterns from available role models or alternative articulatory strategies to satisfy the phonological requirements of a target segment. The level of variation allowed for alternative realizations appears, however, to be determined and regulated by the phonological system, as commented furthermore.

Whether "*choice*" is a conscious or an unconscious process, it may likely require some degree of perceptual processing or mediation. An alternative realization that cannot be perceived by the speaker or that is not linguistically salient may be unlikely to persist. In that sense, some level of auditory feedback might be required.

Moreover, the variable "*choice*" has been widely considered as one of the probable explanations for variations observed between very similar speakers, as in the case of identical twin pairs who had grown up and lived together (NOLAN; OH, 1996; LOAKES, 2008; ZUO; MOK, 2015; SAN SEGUNDO; YANG, 2019). The possible implications of choice as an attempt to establish an individual linguistic identity, especially in contexts where this identity competes or is challenged by another, remains to be explored.

3.3.3 The speaker-discriminatory potential of vowels in BP

Differences within identical twin pairs and across all speakers as a function of vowel quality pointed towards a similar trend for both type of comparisons. This observation is in line with the assumption that differences between groups would be mostly explained in terms of mean formant frequency analysis rather than for individual vowel quality, given the fact that all individuals in the present study spoke the same dialect.

From a qualitative viewpoint, the central vowel [a] appeared to be the most discriminatory segment in both groups, followed by front vowels. Along with the back vowel [u], the central and

front vowels were also the most frequent ones, as previously mentioned.

The observation that the best performing vowels were also the most recurrent ones, as in the case of [a] and [i], may present some statistical implications. It is common knowledge that larger data samples tend to reflect more reliable estimates and are more likely to reveal significant differences in lower alpha values (BERBEN; SEREIKA; ENGBERG, 2012). However, the verification of a higher frequency of occurrence regarding some specific tokens may be considered itself a justification for electing such vowels as viable for the forensic speaker comparison task in contexts where no substantial data is available for analysis. It is worth mentioning that "*availability*" is a crucial factor that comprises one of the main criteria for the parameter selection in the forensic speaker comparison domain. Such a factor refers to the frequency of occurrence of a given element or phenomenon in speech, cf. (NOLAN, 1983).

Despite the difference observed concerning the frequency of occurrence of front and back vowels in the present study, there is reported evidence in the literature of front vowels as being more speaker-specific in Australian English, with F2 of back vowels being considerably homogeneous in comparisons involving similar-sounding speakers (LOAKES, 2004). Also, in Loakes (2008), differences in vowel realization were found where some speakers had more fronted vowels than their twin pairs during spontaneous conversational speech.

According to Recasens and Espinosa (2006), patterns of vowel variability appear to be conditioned differently by contextual and non-contextual factors. In their research with Catalan dialects, front vowels (mostly /i/) were found to be very resistant to context-dependent effects, whereas back vowels varied more along the F2 dimension. Regarding non-contextual variation, i.e., idiosyncratic variation, the opposite trend was observed. Overall, there was more F1 variability for low vs. high vowels and more F2 variability for front vs. back vowels. As pointed out by the authors, "*random*" variability depends inversely on the precision involved in achieving the articulatory target for a given vowel. In the same direction, while analyzing F2 and F3 transitions in American English, Sussman (1990) observed that front vowels were less contextually variable than back vowels.

Overall, the present study's findings appear to be in line with what was reported by Loakes (2004, 2008) and Recasens and Espinosa (2006) while suggesting a higher individual variability for front vowels. Regarding the present study's data, it can be assumed that contextual effects have a comparable influence on the patterns observed across individuals, given the representatives of the material analyzed, which includes a high number of vowel tokens produced in several different phonetic contexts, which is believed to yield a contextual effect minimization. From this point of view, the differences observed are more likely to be mostly related to idiosyncratic phonetic patterns rather than contextual ones.

One possible explanation regarding the differences observed for acoustic distances between

cardinal vowels in this dialect of BP, and perhaps for the higher variability observed for the central and front vowels, is that the perceptual mechanism "knows" that such vowels are usually more variable than the back ones, perhaps due to articulatory working space constraints, performing a perceptual compensation for such discrepancy (ROSE, 2002). From an articulatory viewpoint, as referred by Loakes (2004), speakers may be using different articulatory strategies for front vowels aiming to produce phonologically equivalent vowels. Notwithstanding, this would also imply differences in the articulatory feedback for front, central, and back vowels in BP, as more variation appears to be tolerated for the first and second vowel groups. Conversely, a higher articulatory precision seems to be required for back vowels in BP, given the noticed lower variation for this vowel class, particularly concerning the articulatory horizontal dimension.

Given the fact that the vocalic system of BP is considered relatively symmetric, with seven peripheral oral vowels in a stressed position and the same number of front and back vowels, a homogeneous vowel dispersion could be expected, as expressed by an even spacing of neighboring vowels in the acoustic space. In contrast, a discrepancy in terms of acoustic distances between the entire front and back articulatory dimensions was observed, besides the already reported height asymmetry between these two phonological categories (ESCUDERO et al., 2009). A plausible question concerns whether this acoustic space discrepancy would be related to the lower levels of variation observed for back vowels, given that increased proximity between vowels due to alternative articulatory realizations could imply perceptual difficulties. Another support to an articulatory working space discrepancy may be provided by the fact that in asymmetric phonetic inventories in languages of the world, the number of front vowels is likely to be greater than the number of back vowels, three times more likely in primary systems and two times more in secondary systems (SCHWARTZ et al., 1997a,b).

As mentioned previously, the central vowel [a] was the vowel displaying the best discriminatory performance from the set of vowels analyzed. In terms of vocalic dispersion, this is the only central vowel in the BP system and also the one displaying the highest acoustic distances from its neighbors in the current analysis. The combination of these two factors may place this vowel in a favorable position for a higher acoustic-phonetic variability, as observed in the present study. However, the fact that the vowel [i] displayed the highest proportions of inter-speaker differences within the class of front vowel, even though it was not the one displaying the highest Euclidean distance from its neighbor, may indicate that greater acoustic distance between vowels is not the only factor accounting for the variation in vowel production. In this sense, the concept of "*sufficient contrast*", cf. Lindblom (1990b), meaning the phonetic distance between different vowels may also play a role.

When assessing two performance constraints, namely articulatory simplification, and perceptual

distinctiveness, Lindblom (1990c) states that vowels in vocalic systems of languages in the world seem to have evolved more than anything in response to a *distinctiveness* demand. In this regard, according to his general theory of vowel adaptive dispersion, vowels tend to evolve to provide both sound and feel sufficiently different from each other. This is in line with the claims of Schwartz et al. (1997b) when highlighting that vowel systems should optimize auditory distances in order to enhance contrast and provide as much information as possible about articulatory gestures. According to the authors, if it is true that the interactions between speakers and listeners are responsible for shaping phonological inventories, then, as a result, phonological inventories may provide information about the speaker-listener interaction mechanism.

As was observed in the present study, if it is the case that acoustic distances between neighboring vowels– which have direct implications for perceptual distinctiveness– are in part related to their level of phonetic variation, then it could be justified why the central [a] vowel in BP seems to be the most inter-speaker discriminatory and best performing vowel, as expressed by the lowest Cllr/EER values, and resonably high AUC. The acoustic specification regarding this vowel and its position in the vocalic system would contribute to the higher level of variation observed. Considering that [a] is the only central vowel in BP vocalic system, it may be suggested as less likely to be perceptually confused with other vowels during alternative or imprecise realizations. Notwithstanding, future acoustic and perceptual research are required in order to validate such a hypothesis.

Finally, it is worth mentioning that vocalic systems with larger or smaller vowel inventories and different vocalic dispersions may display different variation patterns, since phonetic content is regarded as dependent on system or inventory size (LINDBLOM, 1990c). In Lindblom (1990c), while referring to consonant systems, the author points out that in small systems, demands for perceptual distinctiveness tend to be less significant than in larger systems. In addition, complex articulatory patterns seem to be required due to a higher intra-system demand for contrast. This trend also appears to apply to vocalic systems, as observed by Schwartz et al. (1997a). By analyzing the structure of 317 primary and 121 secondary systems, as an attempt to identify major trends in vowel inventories, the authors observed that, in general, vowel systems first tend to exploit a "*primary*" system of sounds. However, when systems are found to exceed the inventory size of nine vowels, there is a clear tendency for exploiting at least one new dimension, the so-called "*secondary*" systems, often represented by secondary articulations or duration contrasts.

As for the before-mentioned reasons, forensic-phonetic studies should appraise the language-dependent nature of phonetic variation, especially when outcomes obtained with different language systems are compared.

3.3.4 The lexical stress effect

The analysis of vowel formants' discriminatory power as a function of the lexical stress component in the present experiment was grounded on the assumption that different data distributions may exist concerning distinct vowel groups, i.e., stressed vs. non-stressed. In this regard, different vowel distributions could yield distinct patterns concerning vowel formants' discriminatory power.

In the present study, the lexical stress effects on vowel formants revealed similar trends for intra-twin pair and cross-pair comparisons. The analysis carried based on the combination of stressed and unstressed vowels appeared to be the most explanatory measure in terms of differences observed among speakers, as expressed by the relatively higher number of significant differences, and the reasonably better discriminatory performances: lower Cllr/EER values.

When assessing the stress component separately, by comparing speakers for stressed and unstressed vowels independently, it was found that formant measures obtained through the assessment of stressed vowels appeared more discriminatory than those extracted from unstressed vowels. This finding seems to partially agree with the results reported by Weirich (2012), in which identical twin pairs were found more acoustically similar in the production of unstressed vowels and more distinct for the realization of stressed vowels. According to Weirich (2012), one possible explanation for such an outcome is the assumption of anatomic and physiological aspects as having a more substantial impact on auditorily less salient parameters. The author also observed that MZ twins were more similar in the production of unstressed syllables when compared to DZ twins. Nonetheless, the present study does not corroborate this finding, as the same trend regarding the stress effect was observed for both type of comparisons: intra-twin and cross-pair comparisons. Note that intra-twin pair comparisons are assumed to possess a very low weight on the number of significant differences observed for cross-pair comparisons since they represent only 5.2% of all comparisons performed, namely 10 out of 190 inter-speaker comparisons.

The verification of stressed vowels as being relatively more dispersed in the BP vowel space than unstressed vowels may be considered one plausible factor partially accounting for the discriminatory patterns observed. The assumption of stressed vowels as tending to be more clearly articulated, more prototypical, less reduced, could imply such segments as more acoustically contrasting. In this regard, it may be hypothesized that the presence of the stress component may reflect a better portrayal of individual articulatory adjustments, and consequently, of inter-speaker differences. Conversely, unstressed vowels are suggested as more susceptible to effects such as vocalic reduction, which according to Burzio (2007), may be the consequence of a general principle, namely the realization of the energy downstep necessary for the distinction between stressed–unstressed sequences.

As the results of Fourakis (1991) suggest, although tempo and stress may not have a major influence on the distances of individual vowels from the neutral point, the size of the vowel space overall appears to be susceptible to the effect of these variables. In the study conducted by the researcher, there was a general tendency for larger vowel spaces for the slow stressed condition and smaller for the fast unstressed condition, which may suggest a relationship between precision in speech production and the overall configuration of the vowel space.

Another relevant aspect regards the acoustic salience of segments. Because stressed vowels in BP are referred to as more salient from a perceptual point of view, presenting higher duration, higher $f0$ standard-deviation, and higher spectral emphasis (BARBOSA; ERIKSSON; ÅKESSON, 2013), such segments may be suggested as being potentially more targeted by speakers when implementing alternative realizations. In that regard, the literature signals the critical role of perception when speakers attempt to implement different speech patterns, as in the case of categorical production of pitch variations in imitation tasks (DILLEY; BROWN, 2007).

Duration has been considered to be the most reliable exponent of stress across different languages (GORDON; ROETTGER, 2017). As such, further analysis should also explore the extent to which the higher duration commonly reported for stressed vowels might be related to a higher level of phonetic differentiation among speakers, as the possible result of an increase in the timespan for articulatory differences to emerge.

The present study also found evidence suggesting a vertical reduction of the vowel space in an unstressed condition for all speakers and a horizontal vowel space reduction in some particular cases. This observation is in broad agreement with what was reported by Barbosa (2012) concerning the non-peripheral status of [a] in a post-stressed position in BP. The acoustic outcomes reported by the author revealed the intermediate acoustic nature of the final unstressed [a] regarding its stressed production and the center point in the vocalic space (centroid), which also resulted in a vowel space vertical reduction. This tendency for vowel centralization in unstressed vowels has also been reported for other languages, as in the case of the Spanish vowels /a/ and /o/ (SANTIAGO; MAIRANO, 2018), and the vowels /a/ and /e/ in Hebrew (SILBER-VAROD et al., 2019), as previously mentioned.

Furthermore, one trivial factor may also be related to the discriminatory pattern observed as a function of lexical stress, namely the number of observations for each variable, which is unquestionably a crucial aspect in any statistical inference analysis. In that regard, from the 9,446 vowel points analyzed in the present study, 5,487 (58%) were classified as stressed, while 3,959 (42%) as unstressed, a difference of 16% in the number of data points between the two categories. As such, given the impact of n-size on the statistical strength, a higher number of statistical differences could be presumed for the stressed condition. The same reasoning applies

to the combined stressed and unstressed condition, where an even higher number of data points is present, and a greater number of statistical differences were observed.

A statistical estimate that is able to compensate for such data discrepancy is the analysis of the effect sizes. As observed in this study, stressed vowels tended to display large effect sizes in a higher number of vowels when compared to unstressed vowels. However, when comparing the global inter-speakers differences in effect size's magnitude for both vowel classes, a small difference was observed, which may be due to a considerably higher quantitative effect size estimation for most unstressed vowels, particularly regarding F1. The reason for such a difference remains to be better explored.

3.3.5 Implications for forensic-speaker comparisons

The implications of the present experiment's outcomes on the forensic speaker comparison practice– as of the other experiments carried out in the present thesis– may be qualified as two-fold, as it comprises both a theoretical and a practical demand. The first demand regards the necessity of better understanding the limits of phonetic variation among speakers, including those who display a considerably high level of superposition, and identifying possibly explanatory factors accounting for their speech patterns.

The second demand regards a common goal in forensic phonetics research, namely the identification of relevant and robust parameters for the forensic speaker comparison practice. Here the focus resided on vowel formant frequencies, particularly the most speaker discriminatory ones. Most importantly, such parameters were assessed through an ecologically-valid material comprised of spontaneous speech samples. As Loakes (2004) points out, if inter-speaker differences can be observed in vowel tokens that are not strictly controlled for phonetic context, the potential number of tokens available for forensic analysis is increased. Moreover, the fact that differences between very similar speakers could be found when an "economic" measurement approach, involving an estimate over a single temporal interval in the vowel's midpoint, is also highly relevant from a forensic phonetic perspective.

Regarding the requirements on the selection of forensic phonetic parameters, Nolan (1983) defines some relevant criteria, among which are "availability" and "measurability". The first aspect refers to the frequency of occurrence of a given element, being of particular relevance in real forensic contexts, in which large questioned samples are seldomly available for the experts to base their judgments on. Such aspects are related to the forensic data's representativeness, as remarked by Rose (2002). In this sense, the strength of the evidence is dependent on how well the questioned and known suspect samples reflect their respective sources, which is consequently dependent on

the amount and the quality of the data available.

In the present study, the observation of some specific vowels as being relatively more frequent, aligned with their apparent higher discriminatory potential, as in the case of the central vowel [a] and front vowels, may indicate such units as good candidates for the forensic speaker application in BP, particularly in contexts involving a shortage of data. Notwithstanding, the discriminatory potential of all vowels must be acknowledged in forensic casework whenever possible.

Notably, some crucial limitations may be identified concerning the application of vowel formant analysis in forensic contexts, as in the analysis of telephone-transmitted recordings. The effects of the telephone band-pass on the acoustic signal, commonly referred to as "the telephone effect", cf. Künzel (2001), is mainly characterized by the suppression of higher formant frequencies due to the lower cut-off slope of the telephone band-pass, and the tendency of lower formants to be shifted. As observed by Künzel (2001) in an experiment with German speakers, low formants of vowels produced by males and females tended to move upwards in telephone-transmitted samples compared to direct recordings, resulting in erroneous measurements.

These effects reported by Künzel (2001) have also been corroborated by Passetti (2015) for male speakers in BP, in which the increase of the F1 values in the mobile phone situation caused a global downward displacement of the vowel space. In contrast, the decrease of the F2 values for the front vowels and the increase of this formant's values for back vowels resulted in a vowel space reduction.

Furthermore, Byrne and Foulkes (2007) observed in their experimental study that, due to the filtering effect of the phone transmission, F1 frequencies were found to be higher than their counterparts in the direct recordings for most of the vowels assessed. Additionally, the effects of the mobile phone on F1 were considerably greater than the observed by Künzel (2001), being 29% higher than in the direct speech condition. In general, the authors observed that F2 and F3 measures were generally unaffected by the mobile phone transmission, given some exceptions, as in the case of individuals who presented relatively high F3 values.

Notwithstanding, there may be situations in which the analysis of vowel formants in telephone speech might be relevant, that is, when both questioned and reference materials are telephone-transmitted. As pointed out by Künzel (2001), as long as both questioned and reference material in a forensic case were recorded via telephone, there would not be serious implications, assuming that different telephone channels do not differ substantially in terms of their practical effects.

Finally, it is important to acknowledge that, as pointed out by Cao and Dellwo (2019), voice communication by other means than telephone has become increasingly common, as in the case of cross-platform messaging apps (e.g., WhatsApp and Telegram), allowing users to exchange voice messages. Such technological advances have introduced the possibility of applying higher formant

frequencies in forensic speaker comparisons, as well as other acoustic measures (e.g., voice quality parameters), which certainly demands more experimental work to be carried out. In addition, it should be noted that research on vowel formants are also relevant in the domain of speech technology, providing the experimental background for the enhancement of speech recognition systems.

Given the referred widely reported limitations of employing spectral estimates in forensic speaker comparison analysis, mainly when dealing with telephone transmitted speech, there is a clear demand for assessing the discriminatory patterns of (potentially) more robust and stable parameters, which are less susceptible to system transmission influences. Such a demand is what motivates the experiment conducted on speech timing estimates, as exploited in the following chapter.

Chapter 4

The discriminatory patterns of speech timing parameters

4.1 Introduction

The present chapter[1] concerns the analysis of speech timing features in comparisons involving genetically-related individuals (i.e., identical twins) and across all speakers (cross-pair comparisons). Such a comparison procedure was performed in considering a set 11 speech timing measures, as described in the method's section.

Because it was logical from a practical viewpoint, the studied parameters were classified and grouped into three main categories, following a duration criterion, namely macro, micro, and pause-related temporal parameters. Such a division is expected to help reporting of the outcomes and the discussions.

The first category includes those parameters extracted from units with an average duration superior to that of the phonetic syllable, namely speech rate, articulation rate I, articulation rate II, and stress groups. The second category includes those parameters extracted from units with a mean duration equal to/below the syllable duration, which includes V-V units and vowel segments. Finally, all pause-related parameters were grouped in the same category, namely silent pauses, filled pauses, all pauses, and inter-pausal intervals.

[1]The outcomes deriving from this section of the thesis have been submitted to the open-access journal "Plos One" and was under a peer-reviewing process by the time of its publication.

4.2 Results

Overall, 851 speech chunks were analyzed, an average of 42 chunks and 2.30 min of transcribed material per subject, resulting in an average of 42 speech rate and articulation rate data points per subject. Regarding the total length of transcribed material used in the present study, the experiment carried out by Arantes, Eriksson, and Lima (2018) with different linguistic units (phone, syllable, V-V units, and word) suggests an average stability time for speaking rate parameters of 12.1 seconds, in which vowel-to-vowel intervals (V-V units) was the linguistic unit yielding the shortest stabilization time (9.44s). This outcome supports that the average length of the recordings used in the present study is somewhat representative, at least ten times longer than the specialized literature recommends.

For the sake of a general description, median, mean, standard deviations, and range values are presented in Table 4.1, and the results derived from the statistical analysis summarized in Tables 4.2 and 4.3. Total numbers of data points for each tested parameter are also depicted in Outcomes1.

4.2.1 Intra-twin pair comparisons

As can be seen in Table 4.2, the analysis of speech timing patterns in comparisons involving identical twins revealed a remarkable level of intra-pair similarities. Such an observation applies particularly to the class of macro speech timing parameters, for which the largest effect sizes have been observed here, except for one estimate: SGDUR. As for micro temporal parameters, two pairs out of ten (20%) were found statistically distinct, namely G1-G2 and J1-J2. The same proportion of intra-twin pair differences has also been observed for the class of pause-related estimates, with two pairs diverging significantly: C1-C2 and F1-F2. As observed in Table 4.2, apart from IPI, all temporal parameters pointing to intra-twin pair divergences were found to display "small" effect sizes as a function of the variable "*speaker*". Based on such an observation, a re-test was conducted with down-sampled data points of the respective parameters.

Intra-twin pairs and cross-pair statistically significant differences after the down-sampling procedure are presented in Table 4.3, based on three random samplings. As can be observed, when reducing the number of observations in about 32%, 33%, 15%, 49%, for VVDUR I, VVDUR II, VOWELDUR, and SILPAUSES, respectively, and replicating such a procedure, no consistent differences could be verified for the respective speech timing estimates. Such a lack of consistency[2]

[2]It should be noted that when several additional independent replications are performed, i.e., up to 20 replications, such an inconsistency remains. In fact, twin pairs that could not be contrasted earlier stood out significantly different, as verified for VVDUR I and SILPAUSES in additional tests.

may serve as a clue of a probable n-size or sampling influence on the outcomes.

Finally, when disregarding all reported inconsistent differences, only the twin pair F1-F2 turns out as statistically different for IPI with an AUC of 75%. However, the level of robustness of this parameter for the forensic phonetic application will be described furthermore.

In Figure 4.1, density curves are presented for all temporal parameters and all speakers according to the Kernel density estimate, which may be regarded as a smoothed version of the histogram. Through a close inspection of this figure, it is possible to visualize how similar identical twins were regarding their speech timing patterns, as expressed by the relative overlap of their density curves. It can be seen that they are almost perfectly aligned in terms of their mean values, to a greater extent for micro speech timing estimates and to a lesser extent for pause-related estimates. Conversely, when comparing the density curves across all speakers, some differences can be observed, mainly at a macro speech timing level.

4.2.2 Cross-pair comparisons

All results deriving from the cross-pair comparisons are equally summarized in Table 4.2. In the table, cross-pair significant differences are expressed as total values and percentage values, considering the proportion of differences observed as a function of the number of cross-pair comparisons performed (i.e., 190 cross-pair comparisons).

As can be seen, there was a tendency for those units with a higher frequency of occurrence to display a higher number of inter-speaker differences, as in micro speech timing parameters. The second parameter category displaying the highest proportions of inter-speaker differences was the class of macro speech timing parameters. Finally, parameters pertaining to the class of pause-related estimates have been found to display the lowest proportions of inter-speaker differences, suggesting a higher convergence for such estimates across speakers, especially for silent pauses (SILPAUSES).

When comparing the effect size values presented in Table 4.2, which is an estimate that provides a common metric to compare the direction and strength of the relationship between variables (BERBEN; SEREIKA; ENGBERG, 2012), it is possible to observe how much of the variation for each tested parameter can be explained on the basis of the "*speaker*" variable, as expressed by the effect sizes comparisons across different parameters. Hence, effect sizes can be regarded as an important indication of whether the differences observed are likely to be explained on account of individual differences or better explained by other factors.

Considering the effect sizes' magnitude presented in Table 4.2, two speech timing parameters were found the most explanatory of individual patterns, namely speech rate, and articulation rate

I (i.e., excluding silent pauses). As far as the explanatory potential of the variable "*speaker*" is concerned, these two parameters were virtually identical. However, when considering the proportion of inter-speaker differences, a slightly higher proportion of differences was noted for ARTRATEI in relation to SRATE. Moreover, slightly smaller effect size and proportion of inter-speaker differences were observed for ARTRATE II (i.e., excluding both silent and filled pauses) in relation to SRATE and ARTRATEI. Note that these are the only parameters based on the same number of observations, making their comparison less biased.

Ordering the explanatory potential of the variable "*speaker*" from the largest effect to the smallest effect of the variable on the different speech timing parameters we arrive at the following order:

SRATE = ARTRATE I
> ARTRATE II
> IPI
> FILPAUSES
> SGDUR
> all the other parameters

By comparing overall patterns of the density curves in Figure 4.1 and individual mean values as a function of all parameters, it can be observed how variable speech timing estimates are across speakers, also helping understand their effect size differences.

4.2.3 The speaker-discriminatory performance of speech timing parameters

One of the present study goals was to identify the most suitable speech timing parameters for speaker comparison applications from a forensic perspective, which regards not only how variable estimates are across individuals but also how accurate and consistent they might be. Three different estimates were used to test such a consistency: the Log-likelihood-ratio-cost function (Cllr), Equal Error Rate (EER), and AUC values deriving from multiple ROC analyses. In Table 4.2 performance estimates are presented for each analyzed parameter, including mean Cllr and EER observed values during several tests.

As can be seen in Table 4.2, among all tested parameters, SRATE and ARTRATE I, both from the class of macro temporal parameters, have shown to display the largest AUC values and the lowest EER. These were also the parameters displaying the lowest Cllr values, along with

Table 4.1: **Speech timing parameters' median, mean, standard deviation and range of mean individual values across all speakers.**

Parameter	Category	Median	Mean	Standard deviation	Range (means)
SRATE	Macro	4.6 vv/s	4.6 vv/s	1.3 vv/s	3.5 – 5.7 vv/s
ARTRATE I	Macro	5.5 vv/s	5.4 vv/s	1.1 vv/s	4.7 – 6.2 vv/s
ARTRATE II	Macro	6.0 vv/s	5.9 vv/s	1.0 vv/s	5.2 – 6.6 vv/s
SGDUR	Macro	1.0 s	1.2 s	702 ms	853 - 1.687 ms
VVDUR I	Micro	160 ms	207 ms	199 ms	168 – 264 ms
VVDUR II	Micro	150 ms	163 ms	98 ms	137 – 205 ms
VOWEL DUR	Micro	67 ms	84 ms	67 ms	69 – 104 ms
SILENT PAUSES	Pause-related	480 ms	547 ms	333 ms	398 – 772 ms
FILLED PAUSES	Pause-related	255 ms	298 ms	146 ms	204 – 373 ms
ALL PAUSES	Pause-related	365 ms	449 ms	301 ms	345 – 649 ms
IPI	Pause-related	2.0 s	2.3 s	1.3 s	1.4 – 3.5 s

Table 4.2: **Number of significant differences in comparisons among all speakers and intra-twin pairs for speech timing parameters, followed by effect sizes (η^2), and discriminatory performance estimates.**

Parameter	N	Cross-pair differences	Intra-twin differences	Eff. size	Mag	$Cllr_{raw}$	$Cllr_{cal}$	EER	AUC
SRATE	851	40 (21.0%)	–	15.6%	**Large**	**0.78**	**0.78**	**0.28**	**0.64**
ARTRATE I	851	47 (27.7%)	–	15.5%	**Large**	**0.76**	**0.75**	**0.27**	**0.64**
ARTRATE II	851	26 (13.6%)	–	12.3%	Mod	**0.78**	**0.75**	**0.31**	**0.62**
SGDUR	2.107	42 (22.1%)	–	6.5%	Mod	**0.97**	**0.89**	**0.35**	**0.59**
VVDUR I	12.609	75 (39.4%)	G1-G2	2.2%	Small	**0.82**	**0.81**	**0.33**	**0.55**
VVDUR II	10.495	62 (32.6%)	G1-G2	2.3%	Small	**0.92**	**0.84**	**0.30**	**0.55**
VOWEL DUR	9.447	54 (28.4%)	J1-J2	1.7 %	Small	**0.95**	**0.90**	**0.40**	**0.54**
SIL PAUSES	864	7 (3.6%)	C1-C2	4.6%	Small	**6.06**	**1.00**	**0.55**	**0.58**
FIL PAUSES	560	10 (5.2%)	–	8.3%	Mod	**2.81**	**1.00**	**0.50**	**0.61**
ALL PAUSES	1.424	7 (3.6%)	–	3.3%	Small	**9.97**	**1.00**	**0.50**	**0.56**
IPI	675	21 (11.5%)	F1-F2	11.3%	Mod	**0.88**	**0.88**	**0.43**	**0.63**

Table 4.3: **Number of significant differences in comparisons among all speakers and intra-twin pairs for VVDUR I, VVDUR II, VOWEL DUR, and SIL PAUSES for downsized samples.**

Parameter	N	Random sample I Intra-twins	Random sample II Intra-twins	Random sample III Intra-twins	Random sample I Cross-pairs	Random sample II Cross-pairs	Random sample III Cross-pairs
VVDUR I	8.580	G1-G2 (10%)	–	–	53 (27.8%)	62 (32.6%)	56 (29.4%)
VVDUR II	7.020	–	–	–	51 (26.8%)	47 (24.7%)	41 (21.5%)
VOWEL DUR	8.040	–	–	–	40 (21.0%)	40 (21.0%)	38 (20.0%)
SIL PAUSES	440	–	–	–	1 (0%)	0 (0%)	0 (0%)

ARTRATE II. From such a parameter category, SGDUR presented the worst overall performance with EER around 35%.

Figure 4.1: **Kernel density diagrams for speech rate (A), articulation rate (B), articulation II (C), V-V unit duration I (D), vowel duration (E), and stress group duration (F).**

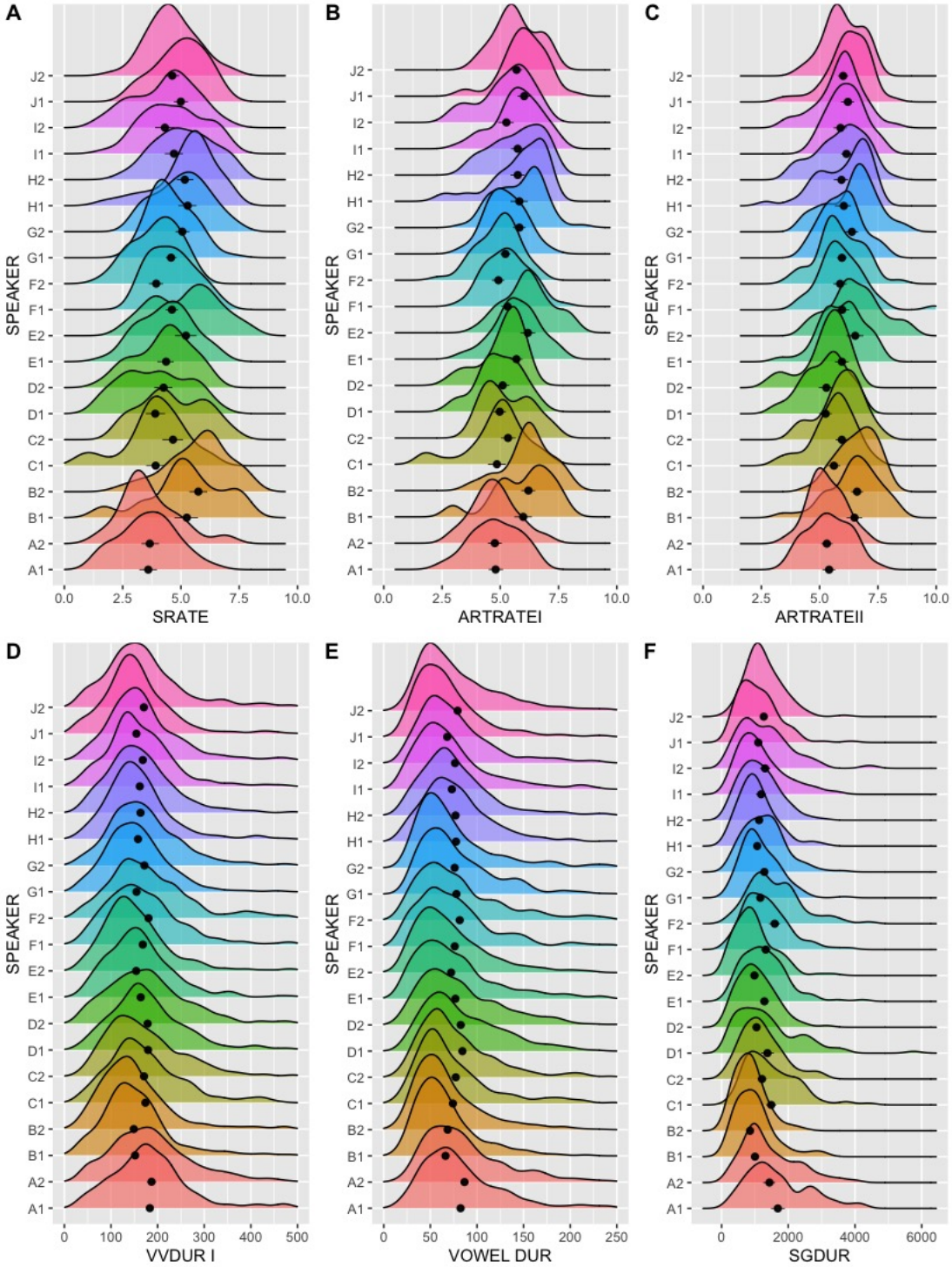
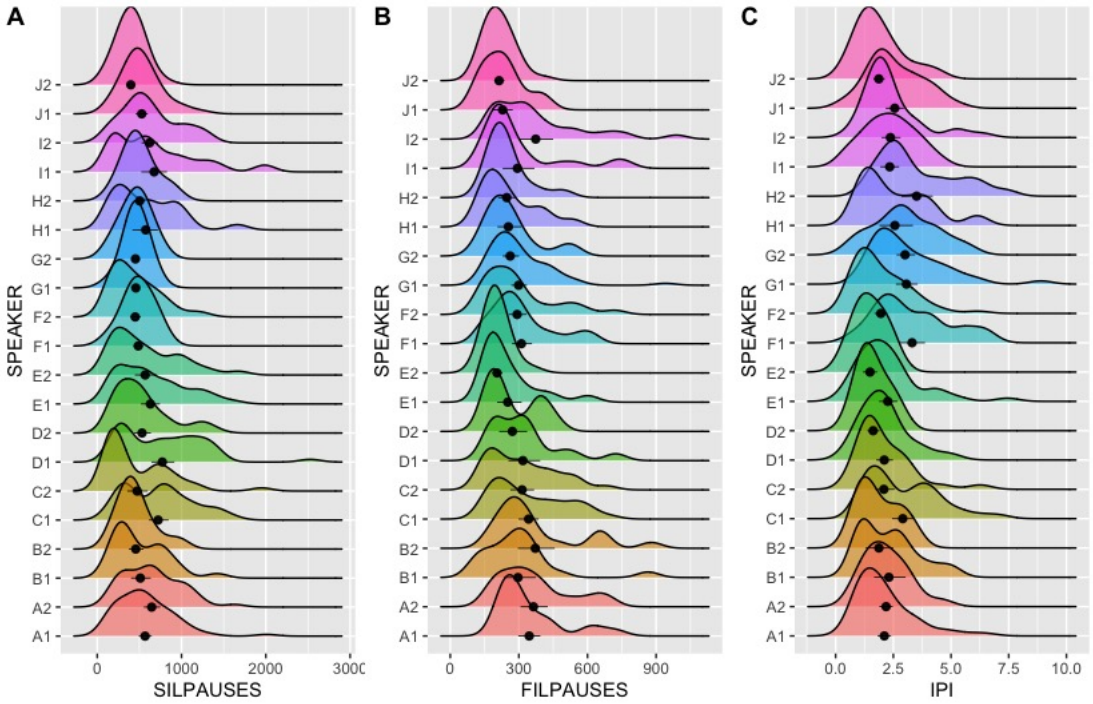


Figure 4.2: **Kernel density diagrams for silent pauses (G), filled pauses (H), and IPI (I).**



Concerning the category of micro temporal parameters, these were found to exhibit the lowest AUC values from all tested estimates, with an overall binary classification performance just above the chance level (54%-55%), being outperformed by all macro temporal estimates. VOWEL DUR has shown the worst speaker-discriminatory performance from this parameter group, expressed by the highest Cllr/EER and the lowest AUC values.

Finally, by inspecting Table 4.2, it can be seen that category of pause-related parameters exhibited the highest Cllr and EER values, with SILPAUSES and ALL PAUSES (i.e., the combination of silent and filled pauses) displaying the worst discriminatory performances, as evidenced by the highest EER and Cllr values among all tested parameters, even when considering calibrated Cllr values[3]. As can be noted, from this parameter class, IPI was found the best performing parameter; however, with equally high EER values: 40%-45%.

Note that, in terms of Cllr, the pause-related parameters category was the one that benefited most from a calibration procedure, as expressed by a considerable reduction between raw and

[3] *Calibration* is a method performed on log-likelihood ratios to reduce the magnitude and incidence of likelihood ratios known to support the incorrect hypothesis, i.e., the contrary-to-fact hypothesis, thereby improving accuracy. As previously mentioned, such a procedure is based on a logistic regression model trained with the same set of data (i.e., self-calibration), cf. Morrison, Zhang, and Rose (2011) and Morrison, Zhang, and Enzinger (2019).

calibrated Cllr values. The other parameters did not benefit as much from such a procedure.

ROC graphs corresponding to intra-twin pair comparisons are displayed in Figures 4.3 and 4.4, while some of the cross-pair comparisons are presented in Figures 4.5 and 4.6. Such plots depict the overall binary classification performance regarding the most discriminatory macro speech timing parameters (i.e., SRATE, ARTRATEI, ARTRATEII) in relation to VVDUR I and VOWEL DUR. For the sake of simplicity, the other parameters were not included; however, their overall performance can be assessed in Table 4.2. Such figures show an important feature that is not represented in Table 4.2.

By inspecting Figure 4.3, it is possible to verify that, as anticipated, intra-twin pair comparisons by means of ROC analysis yielded, overall, a very poor classification performance, expressed by AUC values just above the chance level (50%) for some estimates, and even below the chance level for others. Notwithstanding, regarding cross-pair comparisons, Figures 4.5 and 4.6 show that there was no homogeneous discriminatory performance for any of the parameters assessed across different peers. On the contrary, substantial performance differences were observed on account of the pairs being compared.

Furthermore, when comparing the performance of macro and micro speech timing parameters in Figures 4.5 and 4.6, it is possible to note that, in the large majority of cases, macro speech timing estimates presented overall better performances than micro speech timing estimates. It can also be observed that even across individuals that are not taking part in the same dialogue and who are not genetically related, relatively poor binary classification performances can be observed on account of their speech timing patterns, e.g., D1 x A1; D1 x A2; E2 x B1; E2 x B2. This relatively high variation in classification performance across different pairs of speakers may be regarded as the main reason for such low global AUC values in Table 4.2, since the reported multi-class AUC values consist of the averaging of multiple AUC values, comprehending average, above-average and below-average performances.

Finally, it is worth mentioning that the poor performances observed for intra-twin pair comparisons are expected to bear a very low weight on the global AUC values reported in Table 4.2, once this kind of comparison represents 5.2% of all comparisons performed, namely 10 out of 190 inter-speaker comparisons.

4.3 Discussion

This experiment set out to assess the speaker-discriminatory potential of a set of 11 speech timing parameters in comparisons performed with genetically related speakers (i.e., identical twins)

Figure 4.3: **ROC curves and AUC values for intra-twin pair comparisons.**

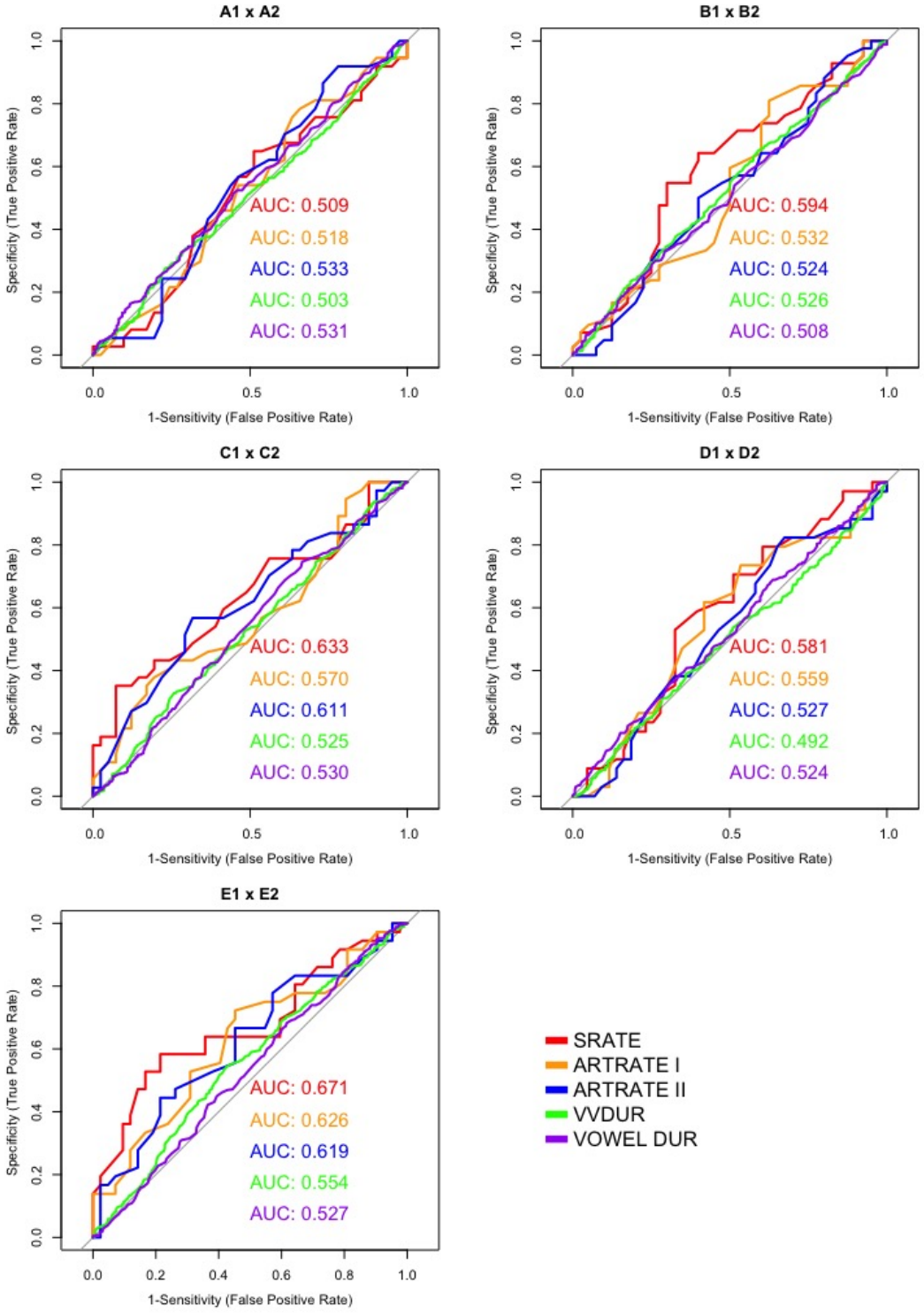


Figure 4.4: **ROC curves and AUC values for intra-twin pair comparisons.**

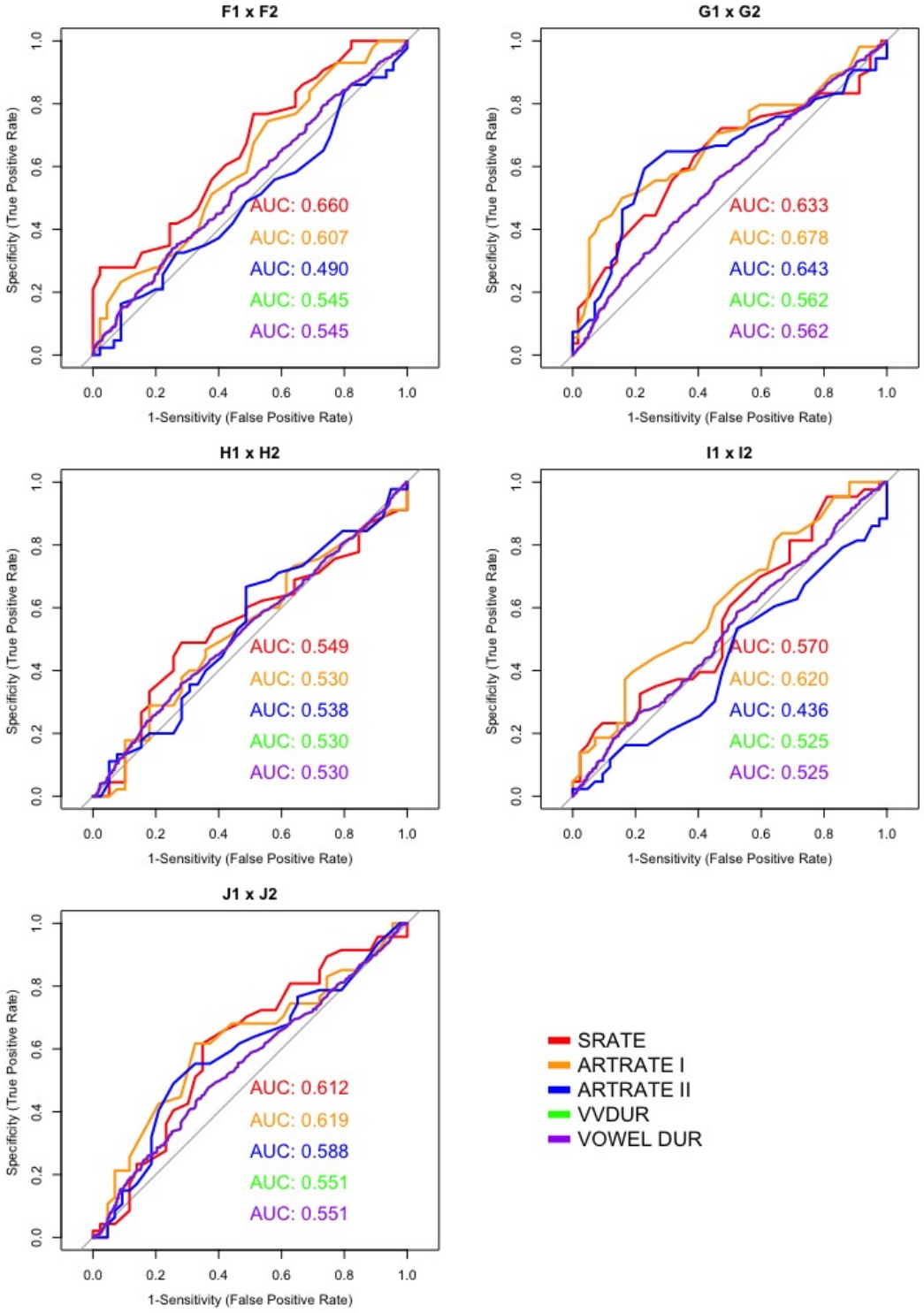


Figure 4.5: **ROC curves and AUC values for cross-pair comparisons (I).**

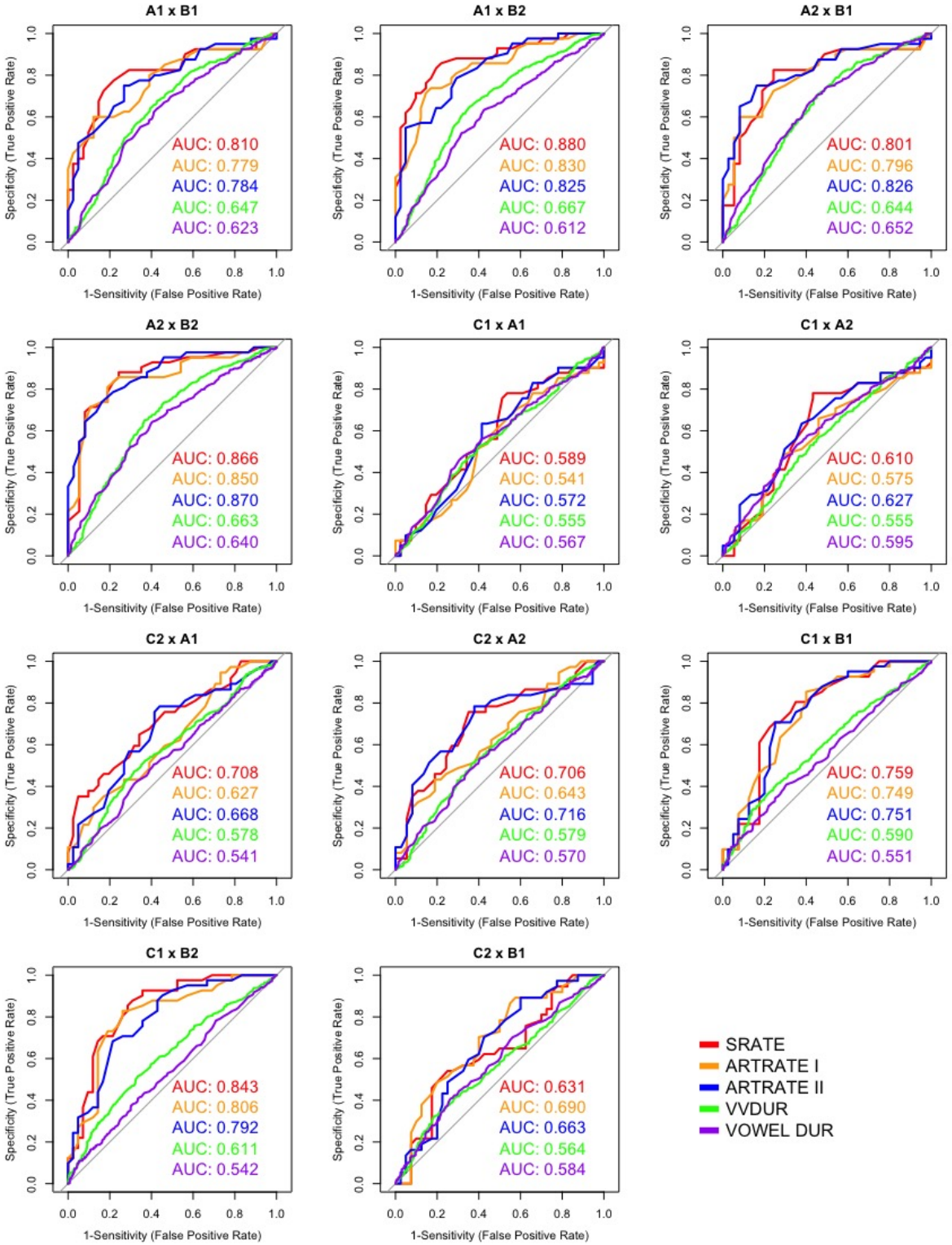
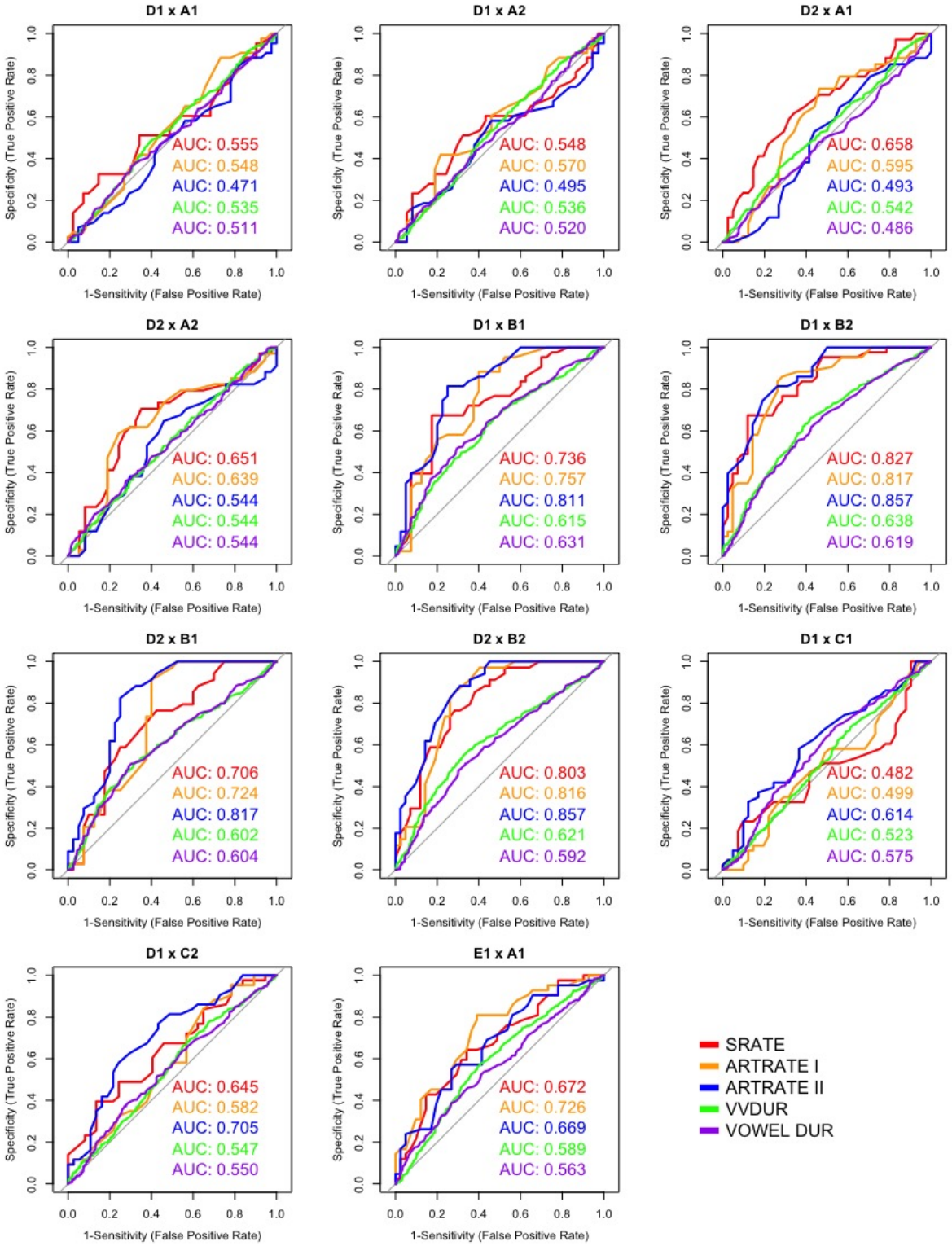


Figure 4.6: **ROC curves and AUC values for cross-pair comparisons (II).**



and cross-pair comparisons. Speech timing estimates pertaining to different dimensions were assessed; namely, macro, micro, and pause-related, having the phonetic syllable duration as the main criterion for contrasting the first two categories. The outcomes are discussed in the following.

4.3.1 The discriminatory performance of macro, micro, and pause-related speech timing parameters

From a general perspective, the present experiment findings suggest the category of macro speech timing parameters as the most reliable estimates when assessed under unscripted speech conditions, mainly SRATE and ARTRATE I. Furthermore, a relatively similar speaker-discriminatory performance has been observed regarding the comparison made between the two estimates, with ARTRATE I displaying a relatively higher number of inter-speaker differences. As for ARTRATE II, which is characterized by the suppression of both silent and filled pauses during its calculation, a slightly lower discriminatory performance was suggested. In addition, SRATE, ARTRATE I, and ARTRATE II were also the estimates presenting the largest effect sizes, which is compatible with a higher explanatory contribution of the variable "*speaker*" in their variation patterns.

However, it is worth noting that, despite being the best performing parameters, the overall performance of SRATE and ARTRATEI was found relatively poor when assessed in isolation, adding some uncertainty whether those parameters would provide enough support for the application in actual forensic conditions, as will be commented on in *Chapter 6*.

As for micro speech timing parameters, despite their considerably high number of cross-pair differences, this was the category displaying the smallest effect sizes concerning all tested estimates as a function of the "*speaker*" factor. Moreover, contrary to our expectations, the reported effect size did not appear to be largely dependent on the V-V units' salience, as defined by means of a duration criterion. Such an outcome invites other explanatory factors regarding micro speech timing parameters' variability, such as factors of a linguistic order.

In this regard, some variables have been systematically reported to significantly affect syllable duration across languages, such as stress and syllables' position in the phrase. In that matter, duration has been acknowledged as the most reliable explanatory factor of stress across different languages (GORDON; ROETTGER, 2017), with stressed vowels (i.e., the nucleus of the syllable) displaying longer duration in comparison to unstressed vowels, as in BP (BARBOSA; ERIKSSON; ÅKESSON, 2013). As for the higher duration of syllables in a phrase final position, the so-called "*phrase-final lengthening effect*", is a widely reported phenomenon occurring to the final syllable rime (SHATTUCK-HUFNAGEL; TURK, 1998). Moreover, studies have shown that, although most of the duration increase seems to occur in the phrase-final syllable rime, significant length-

ening has also been found in the main-stress syllable rime, when this is not the final syllable, as observed by Turk and Shattuck-Hufnagel (2007) for American English. Another potentially relevant explanatory factor regards the presence of pauses within the bounds of V-V units, which is assumed to result in a duration increase; such duration increase is associated with phrasal prominence throughout utterances (BARBOSA, 2007). In this regard, the duration of pauses, either its presence or absence within the bounds of V-V units, may be a better explanatory component regarding these units' variability. Conversely, adding or subtracting such phenomenon from V-V units did not seem to substantially increase the explanatory potential of the factor "*speaker*" on their variation.

As for vowel duration (VOWEL DUR), a low explanatory potential of the variable "*speaker*" has also been observed, which, similarly to the duration of V-V units, may likely be better accounted for by linguistic constraints. As remarked by Lisker (1974), studies of vowel duration have resulted in two well-known general formulations, that is, the duration of a vocalic segment is largely related to the degree of opening of the vowel, resulting in a higher duration for low vowels in comparison to high vowels, and that its duration also depends on the nature of the following consonant, with vowel segments being longer before voiced and shorter before voiceless consonants. Together, the before-mentioned factors may explain, in part, why individual variability seems to display a low explanatory potential concerning micro speech units' duration.

Regarding the specific effects of speaker variability in phone duration, very similar trends have been observed by Heuvel (1996) in an experiment with Dutch speakers using read material. When comparing the effects of three primary sources of variation on vowel duration, namely, vowel identity, context (i.e., phonetic environment), and speaker, the author observed that the vowel identity and consonantal context had a far more substantial influence on vowel duration than speaker idiosyncrasies, with a higher impact of the former in comparison to the latter factor. In addition, long segments, mainly long vowels, presented the largest speaker specificity compared to short vowels and consonants, for instance.

In view of that, the fact that a higher number of inter-speaker differences have been observed for the category of micro speech timing estimates in the present study does not imply that such differences can be attributed to idiosyncratic patterns. Such an outcome may be related to the fact that, as for obvious reasons, vowels and syllables are relatively more frequent than larger speech units (e.g., words, phrases), yielding a greater number of data points, reflecting on the statistical power of the analysis– detecting effects so small which are very likely to be irrelevant in practice. In this regard, as pointed out by Berben, Sereika, and Engberg (2012), while p-values are very influenced by sample size and more likely to be significant when the sample size is large and less likely if the sample is small, effect size estimates, in contrast, are not sensitive to it. The outcomes

of the present study seem to present clear support for this statistical fact.

4.3.2 Some remarks on speaking rate

The observation that ARTRATE I was the parameter displaying the largest proportion of inter-speaker differences within the category of macro speech timing parameters is in broad agreement with the widespread assumption of articulation rate as considerably speaker-specific (KÜNZEL, 1997; JESSEN, 2007). Notwithstanding, it must be noted that different outcomes may be observed depending on the nature of the articulation rate under assessment, as to whether only silent pauses or both silent and filled pauses are excluded. When estimating the articulation rate by considering both silent and filled pauses (i.e., ARTRATE II), a reduction in the proportion of significant differences across speakers was observed. A basis for such a reduction may be suggested, namely the exclusion of idiosyncratic information conveyed by voiced segments, particularly of filled pauses. As pointed out by Künzel (1997), the speaker-specific potential of filled pauses is well known, following the observation that individuals tend to be quite consistent in using 'their' respective personal variant of the hesitation sound. In the present study, filled pauses displayed a moderate effect size, being suggested as more explanatory of individual patterns than silent pauses. It must be recognized that, as pointed out by Gonçalves (2017), the identification and transcription of filled pauses is a somewhat laborious task once it requires several revisions. In some cases, it is uncertain whether one is dealing with a filled pause, an emphasis, or perhaps both, expressed in the form of a lengthened vowel.

Despite the contrasting speaker-discriminatory potential of articulation rate and speech rate, with the direction of the difference depending on the treatment given to silent and filled pauses, a similar discriminatory potential regarding the two parameters has also been suggested in the available literature. Despite observing a higher variance for speech rate in relation to articulation rate, Gonçalves (2017) noted a compatible discriminatory power between speech rate and articulation rate in an experiment with BP speakers. However, the before-mentioned study differs from the present regarding some essential aspects. Firstly, the number of speakers analyzed, 20 in the present in relation to 7 speakers on the other. Secondly, for the articulation rate estimation, not solely silent pauses were excluded from the intervals' total duration, but also filled pauses. As such, the articulation rate reported by the author is somewhat equivalent to ARTRATE II in the present experiment, which appears to be the case when confronting the global (6,19 vv/s) and local (6.20 vv/s) median values reported by the author with the median value observed here (6 vv/s). Thirdly, the minimum silent pause duration threshold in the present study was set at 100 ms in comparison to 130 ms on the other, which, in theory, may reflect different estimates. These

factors combined, adding that different assessment criteria were adopted, may account for possible cross-study differences.

As previously mentioned, the category of macro speech timing parameters was the best explained on account of the variable "*speaker*". This is particularly true for SRATE, ARTRATE I, and ARTRATE II, which are measures extracted from larger temporal windows (see Table 4.2). Such an outcome suggests a rather interesting tendency: the effects of individual variation in speech timing parameters seem more expressive in larger temporal intervals than in small temporal windows. This observation may find support on the observation that while speakers can vary significantly in the proportion of silent or filled pauses, they produce, or even in the proportion of lengthening in word-final segments, as mentioned by Künzel (1997), depending on different factors, such as speaking style, emotional state; the same "freedom" does not seem to apply to smaller units, such as syllable duration or vowel duration, where a great deal of individual variability could have consequences on communication or on the intrinsic rhythm structure. Additional evidence of a higher individual articulatory control on macro over micro temporal units may be provided by the observation of a higher agreement across individuals for macro speech units than micro in the production of synchronous speech (CUMMINS, 2004). It is worth noting that such a speaking condition is regarded as having direct consequences on the prosodic variability, reducing idiosyncratic and expressive variation across individuals (CUMMINS, 2009).

Concerning the degree of speaker specificity in the duration of individual segments, the study conducted by Heuvel (1996) with Dutch speakers found compelling evidence supporting a higher idiosyncratic status of long segments in comparison to shorter, as in the case of long vowels. As observed by the author, such specificity may be accounted for by the effects of individual speaking rate on the duration of such segments, which was further corroborated through a duration normalization procedure. After diminishing the effects of speaking rate on vowel duration, a considerable reduction of speaker specificity could be detected; as a matter of fact, higher than the observed for consonants.

With regard to averages values concerning some of the most commonly studied speech timing estimates, such as speech rate and articulation rate, the average values observed in the present study are in somewhat agreement with values reported in the literature for spontaneous speech. In this regard, average values ranging approximately from 4 to 5 syl/s for speech rate and from 5 to 6 syl/s for articulation rate have been reported across different studies, cf. Künzel (1997), Jessen (2007), Oliveira Jr (2012), Hughes, Brereton, and Gold (2013) and Gonçalves (2017). Such a convergence should not be regarded as arbitrary, as it suggests a regular pattern across different languages (POEPPEL; ASSANEO, 2020). Moreover, the results recently obtained by Assaneo and Poeppel (2018) provide empirical evidence on the cortical levels of what seems to be the

preferred speaking rate in terms of neuron processing. By measuring the synchronization between auditory and speech-motor regions in the brain, while participants listened to synthesized syllables at different rates, it has been found that the auditory-motor *synchrony* was significant only over a restricted range, being enhanced at 4.5 Hz, which, according to the authors, is a value compatible with the mean syllable rate across different languages. According to the authors, these findings suggest that the temporal patterns of speech emerge as a consequence of the intrinsic rhythms of cortical areas, yielding a reliable coupling between acoustic stimuli and auditory cortical activity. In the present study, the observed mean speech rate across speakers lied within this referred interval, where a mean/median speech rate of 4.6 vv's/s and a standard deviation of 1.3 vv's/s were observed.

Several other studies also support the observation that neural activity phase-locks to rhythm, not only in speech but also in music (LUO; POEPPEL, 2007; DOELLING; POEPPEL, 2015; DING et al., 2017; HARDING et al., 2019). In a literature review by Poeppel and Assaneo (2020), the authors explored studies with what they call the "*temporal mesoscale*" of speech, with special attention to regularities in the envelope of the acoustic signal that correlate with syllabic information. It has been observed that the temporal structure of speech at this scale is remarkably stable across languages[4], with a preferred range of rhythmicity of 2–8 Hz. As argued by the authors, this rhythmicity is required by the processes underlying the construction of intelligible speech. The relevance of the referred studies' outcomes in interpreting this study's findings is that they seem to concomitantly signal the limits of variability expected for the rate of speech, suggesting an intertwined relation between production and perception. In that sense, although speakers do tend to vary in their speech temporal patterns, the magnitude of this variation may be seen as under production-oriented and output-oriented constraints, driven by demands of production efficiency on the one hand, and comprehensibility on the other, as thoroughly exploited in Lindblom (1990b).

4.3.3 Differences in pause-related parameters

Notably, within the pause-related parameters category, the inter-pausal intervals (IPI) has shown to display the highest number of significant differences across individuals, as well as the highest effect size (see Table 4.2). These finding are in good agreement with Jacewicz, Fox, and Wei (2010), when mentioning that there is considerable variation in the manner that individuals convey a message, including frequency of pauses which determine the length of inter-pausal chunks of fluent

[4]For a deeper investigation on this matter, see also Varnet et al. (2017), where the speech amplitude-modulation (AM) and frequency-modulation (FM) spectra were analyzed in ten different languages with tendencies to different isochronous units.

speech. In that study, the authors observed a considerable inter- and intra-speaker variation in producing interpausal phrases.

The observation that silent pauses duration yielded the fewest number of differences in relation to all the remaining parameters in the present study seems to suggest a relative regularity of pause-related measures across individuals. This outcome is not all of a surprise when considering the key role played by silent pauses in revealing the prosodic structure of utterances, with its emergence being commonly associated with intonational phrase boundaries (NESPOR; VOGEL, 2007). In that regard, in the study conducted by Krivokapić (2007), aiming to assess the effect of prosodic structure and phrase length on pause duration in read sentences, the author was able to observe a pre- and post- boundary prosodic effect on silent pause duration, in which longer phrases, both before and after the phrasal boundary, yielded longer pauses. Moreover, the author was able to identify a prosodic complexity effect on pause duration, in which medial prosodic boundaries induced shorter pauses in comparison to final boundaries. In the same direction, Smith (2004) was able to observe pause duration-related differences as a function of structural factors concerning the discourse organization. It was also noted that speakers tended to display longer pauses at topic shift than at other discourse boundaries, which also seemed to influence the amount of sentence-final lengthening.

Based on the literature's findings, a point of convergence can be identified across studies, that is, silent pause duration in opposition to its frequency, seems to be largely dependent on intrinsic factors, which may in part suggest a substantially high linguistic control on its variability in communicative contexts. The present study appear to provide evidence for a low inter-speaker variability regarding global measures of pause duration. Additional support for this relative pause duration stability across individuals, may find ground on the observation that frequency of pause has been reported to be more variable than its duration, as suggested by Lane and Grosjean (1973). Regarding such remark, the authors noted that when speakers vary their rate of reading to produce a desired apparent rate, they primarily tend to add or subtract pauses of largely the same duration at strategic syntactic locations, whereas articulation rate and pause duration are much less affected.

Note, however, that different levels of variability in the pausing behaviour can be identified across individuals, when a distinction between sentence medial and sentence final pauses is made. In that regard, Fant, Kruckenberg, and Ferreira (2003) found evidence in Swedish suggesting uniform patterns of pause duration between complete sentences across subjects, whilst pauses within sentences showed large individual variations in reading. Further developments of the present study must also probe this possible pause type-dependent difference in spontaneous speech.

Finally, it should be acknowledged that "silent pause duration" may be efficient when used for

differentiating different speaking styles. In the study conducted by Igras-Cybulska et al. (2016) with Polish speakers, on the application of pauses as a potential source of biometry for automatic speaker recognition, three types of acoustic pauses (silent, filled and breath pauses) and syntactic pauses were analyzed in both spontaneous and read speech. The researchers found that quantity and duration of filled pauses, audible breaths, and correlation between the temporal structure of speech and the syntactic structure were the best performing features for speaker characterization. Silent pause-related features, on the other hand, considerably improved the distinction between read and spontaneous speech with 75% accuracy.

4.3.4 A note on synchronicity in speech production

The present study's findings provide evidence of speech timing patterns as being remarkably similar within twin pairs. Such a convergence was considerably superior to the observed in the previous experiment on vowel formant frequencies (*Chapter 3*), and superior to the observed at the $f0$ dimension, as will be thoroughly described in the next chapter (*Chapter 5*).

Two complementary hypothesis are invited to account for such a striking intra-twin pair similitude regarding the temporal dimension: the sharing of similar mental representation of speech timing features acquired throughout their language acquisition and the possible (overlapping) effects of *prosodic entrainment*.

In this regard, evidence from experimental studies on motor control of articulatory timing at the phoneme (WRETLING; ERIKSSON, 1998) and word level (ERIKSSON; WRETLING, 1997) provide an indication of motor control of timing in those levels as relatively more "hard coded" than the motor patterns involved in the production of spectral components of speech. The experiment conducted in Wretling and Eriksson (1998) with the impersonation of phonetic patterns, including the temporal dimension of speech production, revealed that timing patterns in the imitations were in all cases more similar to those of the speaker's natural production than to the target patterns (i.e. the speech model to be reproduced), which may suggest the speech timing features as more challenging to be deliberately manipulated, and perhaps more stable intra-individually. Such observations are of considerable relevance from a forensic-phonetic perspective, as they suggest a relative intra-speaker stability of speech temporal patterns.

Support for an environmental effect on the establishment of speech temporal patterns has been found under a more controlled situation, i.e., read speech. In the study conducted by Whiteside and Rixon (2013), a pair of male identical twins (T1 and T2) aged 21-year-old and an age- and sex-matched sibling (S), recorded 2 years after, were assessed regarding their speech tempo and fundamental frequency patterns. The authors found evidence for greater intra-twin similarities in

mean $f0$ compared to the control. Conversely, although intra-twin similarities were greater than the sibling, these diminished for both speech tempo and dynamic $f0$ parameters at the sentence, word, and syllable level. According to the authors, such an outcome supports the view that some speech and voice parameters might be under a greater genetic influence, whereas others are shaped more by environmental factors such as accent, dialect, reading style, and speaking style.

It is worth mentioning that the present experiment's methodological design does not allow answering whether such a high intra-twin pair temporal congruence may solely be accounted for by environmental influences, once another equally plausible and perhaps overlapping factor may be on the basis of the patterns observed; *prosodic entrainment*. Notwithstanding, the fact that more differences have been found regarding the same twins' spectral patterns may indicate a possibly higher level of *prosodic entrainment* in speech timing than in vowel formant frequency patterns. In this instance, the experiment conducted by Arantes and Barbosa (2010) with 42 BP native speakers on perceptual entrainment found encouraging evidence supporting the assumption of rhythm perception as a listener-speaker entrainment process, in which duration may be considered the main acoustical feature driving the behavior of listeners. Furthermore, prosodic boundaries (i.e., stress group boundaries) were found to play an important role in such a mechanism, organizing the listener's experience of rhythm.

The concept of abstract clocks responsible for regulating the temporal organization of motor gestures and consequently allowing individuals to enter into a "*synchronization state*" is not exclusive to the realm of speech production. In fact, such mechanism has been observed for other complex forms of timing and motor control, e.g., music playing, typing, cf. Shaffer (1982, 1984). As for speech production, several experimental studies have been carried out in the domain of synchronous speech, with particular attention to the research performed by Cummins (2001, 2004, 2007, 2009) and Cummins, Li, and Wang (2013).

Particularly, in Cummins (2004), while assessing two subjects reading a text in *synchrony*, it has been observed that some prosodic variability is significantly reduced when reading synchronously, where synchronous speech has been demonstrated to exhibit markedly less inessential variability. Moreover, variables associated with global timing, namely major syntactic juncture and phrase length, were found more consistent in the synchronous condition, displaying less variability, while smaller units were not noticeably affected by the synchronous speaking condition, as in the case of stressed and unstressed syllables, as well as the closure to voicing onset (C-V transition). In sum, those variables which were most directly related to macroscopic temporal structure (i.e., phrasing) displayed less variability in synchronous speech. In contrast, the microscopic temporal structure remained stable, suggesting, according to the author, that at a finer timescale, there is little if any change to speakers' timing when speaking in *synchrony*.

Regarding the effects of interpersonal *synchrony* when two or more individuals take part in the same conversation, Buder and Eriksson (1997) points out that conversational speech by one individual must be patterned to somehow "fit" the patterns of the other individuals engaged in the conversation, referring to effective turn-taking as one example of such fit. Moreover, as mentioned by the authors, there are other ways in which speech can be temporally structured during conversational interaction. By analyzing the prosodic cycles of mean speech fundamental frequency and mean voice intensity between subjects while engaging in a conversation, Buder and Eriksson (1997) observed a common occurrence both in Swedish and American English, namely the relative continuance of speakers' "rhythms" across turn-exchanges, in which the period and phase of prosodic cycles initiated by one partner was maintained by the other, across speaker transitions. Furthermore, the analyses suggested that in conversations between same-gender the patterns of rhythmic integration, or "*synchrony*", were substantially similar in American English and Swedish.

As mentioned by Cummins (2004), the answer to synchronicity must lie in the shared knowledge speakers have of what is essential and what is redundant, or optional, in the modulation of the speech organs. Such an observation also concerns speakers of the same dialect, who very likely share similar temporal relations among the discrete speech units and the mechanisms for producing them.

Such synchronicity or "*convergence*" between interlocutors has been observed and corroborated by Cohen Priva, Edelist, and Gleason (2017) at the speech timing level while keeping under control two possible driving forces related to this convergence, namely, conversational factors and the interlocutor baseline speech rate. According to the authors, such a convergence effect may be either a direct or indirect process, in that the interlocutors' different speech rates may affect some intermediate factor (e.g., conversation flow) and thereby affect the speakers' speech rate.

In terms of the "*machinery*" underlying a possibly higher/lower convergence across speakers concerning their rhythm patterns– be it enhanced or not by some level of *prosodic entrainment* during a simultaneous conversation– it could be expressed within the bounds of the dynamic model of speech rhythm production described by Barbosa (2007) and more thoroughly explored in Barbosa (2006).

Such a model comprises two universal oscillators, namely a syllabic oscillator keeping pace with the vowel onset sequence and a phrase stress oscillator that specifies prosodic phrasing and prominence. The referred model acts at three nested temporal domains or three coupling forces to assign segmental duration. The first set of coupling forces is implemented by the probabilistic-controlled coupling between syntax and production constraints controlled by the coupling strength, the second set of coupling forces operates between the phrase stress oscillator and the syllabic

oscillator. The third set of coupling forces stands for the prosody-segments interaction, yielding overt segmental durations (BARBOSA, 2002).

In the present study, this coupled-oscillator rhythmic model is evoked to invite the hypothesis of a higher coupling strength within twin pairs as one possible factor accounting for the substantially higher similarities in their speech timing outcomes. Notwithstanding, given the characteristics of the material analyzed, i.e., spontaneous dialogue between twins, it can not be determined how much of such a higher coupling strength is accounted for by the speech task or possible environmental/learning influences. In order to explore such aspect, analyses of additional speech materials produced by the twin pairs are required.

Finally, with regard to identical twins, a great deal has yet to be explored, mainly concerning the effects of acquiring a language "together", being exposed to very similar models (e.g., same mother, father, and relatives) on the establishment of their "linguistically shared knowledge", and, consequently, on their speech timing patterns. To that end, data from non-identical twins (dizygotic twins) should also be included in future developments of the present study, while also probing the effects of different speaking styles on the speech timing parameters assessed here.

Chapter 5

The discriminatory patterns of fundamental frequency

5.1 Introduction

The present chapter[1] regards the analysis of speaking fundamental frequency ($f0$) descriptors, including centrality, modulation, and variation parameters. Two types of *speech material* were considered for assessment, namely connected speech versus lengthened vowels. The main outcomes are presented and discussed in the following, in light of previous experimental studies carried out with identical twins as well as non-genetically related speakers.

5.2 Results

Average values of $f0$ estimates in connected speech and in lengthened vowels are summarized in Table 5.1. Furthermore, the statistical analysis outcomes are displayed in Table 5.2. In such a Table, the numbers of statistical differences for each tested parameter are expressed as a function of all inter-speaker comparisons (i.e., cross-pair differences) and intra-twin pairs comparisons (i.e., twin pairs differences). The parameters are displayed in decreasing order according to the proportion of inter-speaker differences observed. In addition, effect size estimates regarding inter-speaker comparisons and their respective magnitudes are presented. In Table 5.3, system performance estimates regarding the comparison across all individuals are presented, namely Cllr, EER, and

[1]The outcomes deriving from this section of the thesis have been published in the peer-reviewed open access Journal of Voice in the form of a research article, cf. Cavalcanti, Eriksson, and Barbosa (2021b). The full length research article can be retrieved from https://doi.org/10.1016/j.jvoice.2021.08.013.

AUC.

5.2.1 Intra-twins and cross-pair differences: connected speech

Through the inspection of Table 5.1, it can be observed that centrality measures of $f0$ (e.g., mean and median) fell within a very specific range, varying from 74 to 93 semitones in connected speech and from 76 to 89 in vowel segments for median values. Nevertheless, some inter-speaker differences could be observed within such intervals.

As can be seen in Table 5.2, the analysis of the parameters' discriminatory potential on the basis of the differences observed among all individuals suggested five f0 estimates as being remarkably discriminatory, namely *f0 median*, *f0 mean*, *f0 baseline*, *f0 min*, and *f0 max*, in decreasing order of relevance. Notably, although large effect sizes have been observed for all such measures, centrality measures and $f0$ baseline were the ones displaying the highest scores, suggesting a higher explanatory potential of the variable "*speaker*" on the differences observed. Within this group of parameters, two identical twins out of ten (20%) were found different from each other: E1-E2 and F1-F2, except for $f0$ max that yielded only one difference: F1-F2.

By consulting Table 5.2, it can be seen that $f0$ variability and modulation parameters displayed only moderate to small effect sizes, being accompanied by a lower proportion of differences across speakers. Furthermore, all parameters displaying small effect sizes were the ones corresponding to $f0$ modulation. Regarding such parameters, only one identical twin pair (per parameter) could be differentiated for two out of ten measures, G1-G2 (f0M8) and F1-F2 (f0M5).

In Figures 5.1 and 5.2, Kernel density diagrams (i.e., a smoothed version of the histogram) are presented for the six most speaker-contrasting parameters (Figure 5.1), and the six least speaker-contrasting parameters (Figure 5.2), according to the proportion of inter-speaker (cross-pair) differences observed. Moreover, individual mean values are displayed at the bottom of the distributions, as expressed by rounded points. Through visual inspection of such Figures, general trends regarding the parameters' distributions and variability can be observed, revealing convergences or divergences across subjects.

As can be seen, greater variability in mean values and data distributions can be verified across speakers for those parameters displaying higher proportions of inter-speaker differences, such as $f0$ median, mean, minimum, baseline, and maximum (see Figure 5.1). Conversely, a higher alignment or congruence regarding mean values and distributions can be noted for those parameters displaying a lower discriminatory potential, as observed for the parameters in Figure 5.2. Furthermore, it is noteworthy that either a higher or a lower overlap of individuals' average values directly affects the estimation of effect sizes and is therefore reflected on the estimates. In that regard, greater

divergences in the estimates' average values across speakers tend to yield larger effect sizes and vice-versa, adding valuable information on how variable a metric is.

Table 5.1: **Centrality and dispersion measures of fundamental frequency ($f0$) extracted from connected speech intervals and lengthened vowels (in semitones).**

Connected Speech							
Speaker	$f0$ **med**	$f0$ **mean**	$f0$ **sd**	$f0$ **min**	$f0$ **max**	$f0$ **base**	$f0$ **SAQ**
Mean	83	83	1.8	79	86	80	1.2
Sd	2.2	2.1	0.3	1.5	1.9	1.8	0.2
Range	74–93	77–92	0.1–6.3	71–89	79–96	71–90	0.07–6.70
Lengthened Vowels							
Speaker	$f0$ **med**	$f0$ **mean**	$f0$ **sd**	$f0$ **min**	$f0$ **max**	$f0$ **base**	$f0$ **SAQ**
Mean	81	81	0.4	80	81	80	0.3
Sd	1.8	1.8	0.07	1.7	1.8	1.8	0.06
Range	76–89	76–89	0.03–2.02	73–89	76–90	74–89	0.02–1.85

5.2.2 Intra-twins and cross-pair differences: lengthened vowels

Regarding the analysis of $f0$ parameters extracted from lengthened vowels, as summarized in Table 5.2, a relatively smaller proportion of inter-speaker differences could be observed compared to $f0$ differences in connected speech for the same estimates. Also, relatively similar effect sizes were observed for centrality and $f0$ extreme measures in lengthened vowels. Regarding the comparison of the most speaker-contrasting measures, namely $f0med$, $f0mean$, and $f0base$, a reduction in effect size of approximately 7%-10% was verified between connected speech and lengthened vowels, suggesting a higher explanatory potential of the variable "*speaker*" on $f0$ patterns in the former.

Furthermore, no intra-twin pair differences could be observed when twins were compared for such linguistic material, which is compatible with a higher congruence of their $f0$ patterns in the production of lengthened vowels, in most cases, perceived as filled pauses.

As shown in Table 5.1, $f0$ estimates in lengthened vowels tended to be less variable than in connected speech, as represented by smaller standard deviation and smaller range values for all parameters assessed. Also, the smaller discrepancy between global average $f0$ minimum, median, mean, and maximum values in the production of such phenomena suggest its relative "stationary" status, in contrast to what was verified in connected speech.

Table 5.2: **Number of significant differences for cross-pair and intra-twin pair comparisons for $f0$ descriptors in connected speech and lengthened vowels, followed by effect size estimates (η^2).**

Connected Speech				
Parameter	**Inter-speaker diff**	**Intra-twins diff**	**Effect size (inter)**	**Magnitude**
f0 med	98 (51%)	E1-E2 (2.4 st) F1-F2 (2.6 st) –	50.3%	Large
f0 mean	94 (49%)	E1-E2 (1.9 st) F1-F2 (2.7 st) –	49.4%	Large
f0 base	91 (47%)	E1-E2 (2.4 st) F1-F2 (1.8 st) –	47.4%	Large
f0 min	75 (39%)	E1-E2 (2.6 st) F1-F2 (1.9 st) –	32.1%	Large
f0 max	70 (36%)	F1-F2 (3.5 st)	26.7%	Large
f0M8	37 (19%)	G1-G2	12.7%	Moderate
f0M5	30 (15%)	F1-F2	11.8%	Moderate
f0M7	21 (11%)	None	9.0%	Moderate
f0 sd	20 (10%)	None	9.0%	Moderate
f0M4	18 (9%)	None	8.8%	Moderate
f0M6	14 (7%)	None	5.9%	Small
f0 SAQ	13 (6%)	None	6.1%	Moderate
f0M2	8 (4%)	None	5.1%	Small
f0M1	4 (2%)	None	3.0%	Small
f0M3	1 (0.5%)	None	1.6%	Small
Lengthened Vowels				
Parameter	**Inter-speaker diff**	**Intra-twins diff**	**Effect size (inter)**	**Magnitude**
f0 base	47 (24%)		40.7%	Large
f0 med	47 (24%)		40.0%	Large
f0 mean	46 (24%)		40.0%	Large
f0 min	45 (23%)	None	40.2%	Large
f0 max	45 (23%)		37.8%	Large
f0 sd	1 (0.5%)		1.6%	Small
f0 SAQ	1 (0.5%)		0.9%	Small

Figure 5.1: **Kernel density diagrams and mean individual values of f0med, f0mean, f0min, f0baseline, f0max (in semitones), and f0M8 (peak width).**

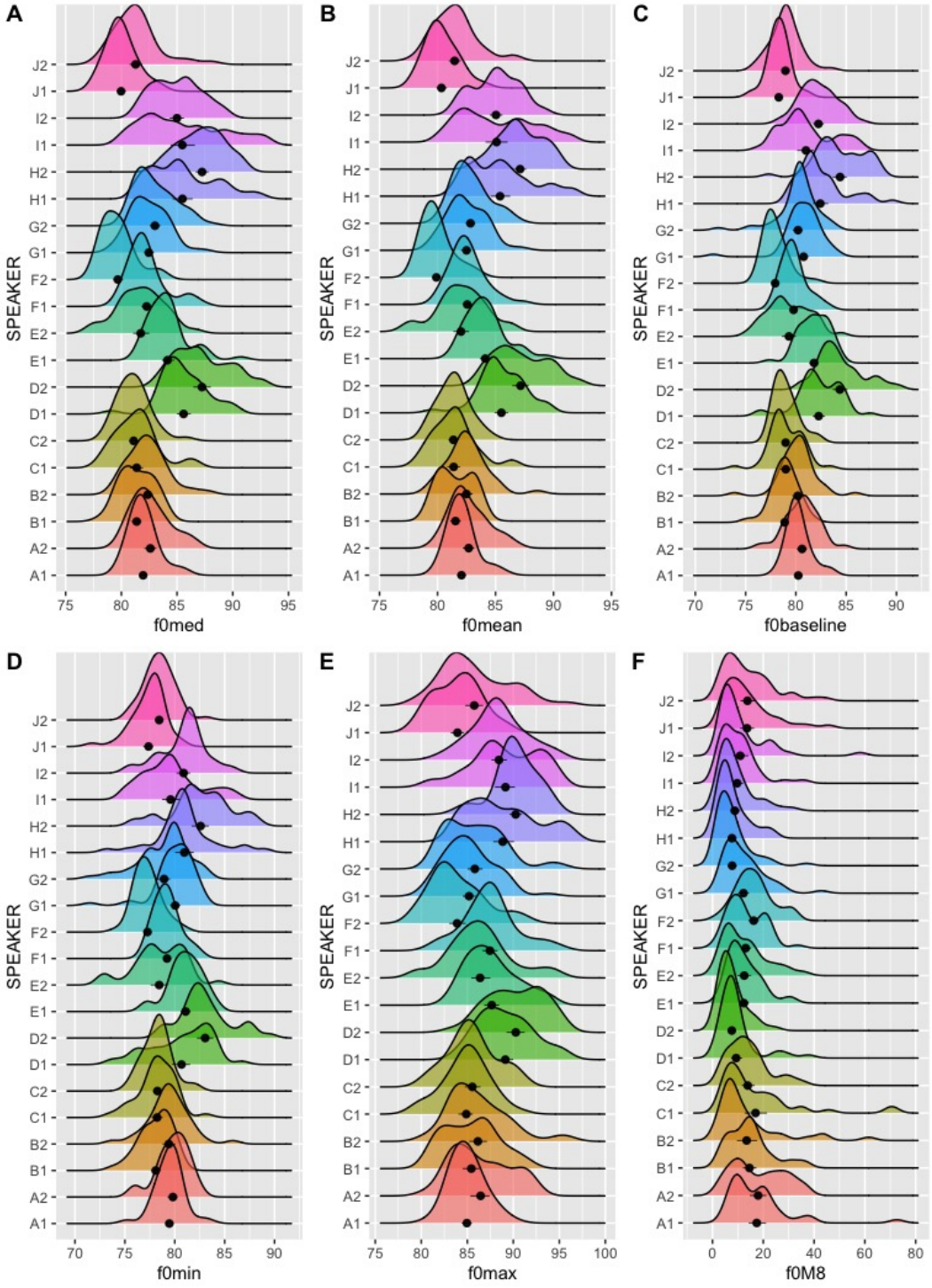
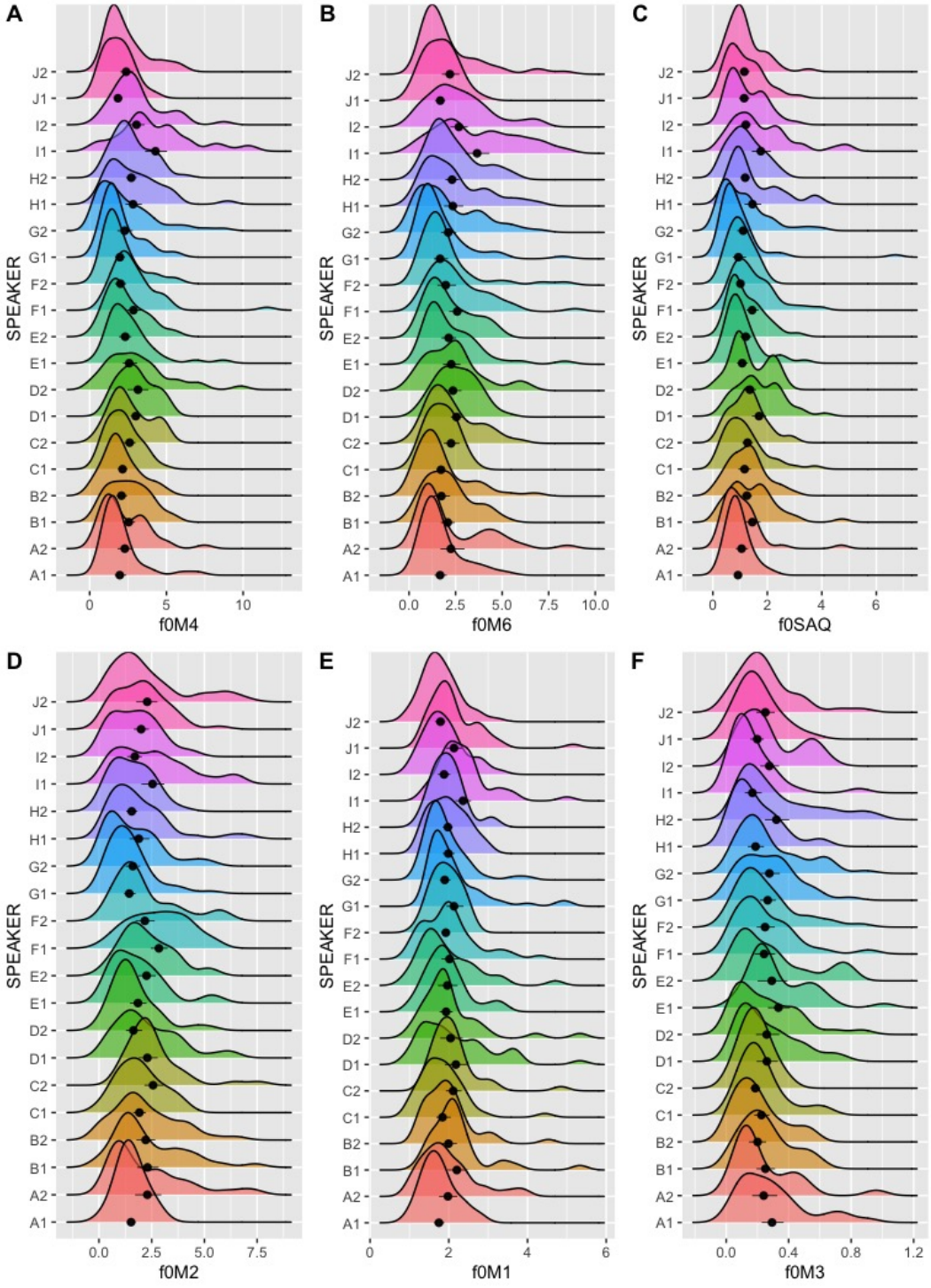


Figure 5.2: **Kernel density diagrams and mean individual values of F0M4 ($f0$ rising rate), F0M6 (sd of $f0$ rising rate), $f0$SAQ ($f0$ dispersion), F0M2 (sd of $f0$ peaks), F0M1 ($f0$ peak rate/s), F0M3 (sd of $f0$ peak's duration), respectively.**



5.2.3 The performance of $f0$ estimates for speaker comparison applications

Through the examination of Table 5.3, which summarizes the overall performance of $f0$ parameters in terms of log-likelihood-ratio costs (Cllr), equal error rates (EER), and AUC values, it seems remarkable that, among all parameters, the median, mean and base value of $f0$ were the ones presenting the best discriminatory performance. This is expressed by the combined lowest Cllr, EER, and highest AUC values observed for such estimates. Furthermore, among all parameters, the base value of $f0$ (i.e., f0base) was the estimate displaying the lowest Cllr and EER value, outperforming all the others in terms of its accuracy. This observation applies mostly to the analysis of $f0$ in connected speech. Note that $f0$ baseline performance estimates in Table 5.3 are highlighted in blue

It is possible to visualize the classification performance of $f0$ estimates for intra-twin pair comparisons in Figures 5.3 and 5.4, which depict ROC curves and their corresponding AUC values. By inspecting the graphs, it can be verified that $f0$ centrality, floor, and ceiling estimates plus $f0$ baseline are, in fact, the ones whose curves most often deviates from the central line towards the upper left point (1.0); hence these are also the estimates presenting the highest AUC values (i.e., the area under the ROC curve). Note that the diagonal line indicates a 50% chance of correct or false acceptance; as such, those parameters closer to this threshold are regarded as less accurate concerning their binary classification performance. Note that, from the plotted estimates, $f0$ standard deviation (f0sd) and dispersion (f0SAQ) are often the ones closer to the diagonal line, suggesting poorer performances. Notwithstanding, some exceptions could be noted, such as for the ROC curves of C1-C2 and I1-I2 in Figure 5.4. It should also be noted that these plotted binary comparisons only consider intra-twin pair comparisons, where the performance of a system is already expected to deteriorate. Moreover, the graphs are not in alphabetic order to draw attention to E1-E2 and F1-F2 in Figure 5.3, i.e., the twin pairs with the highest proportions of intra-twin pair differences.

ROC curves regarding some cross-pair comparisons can be visualized in Figures 5.5 and 5.6, which, basically reflect, in general, a superior performance than the observed for most intra-twin pair comparisons, with some curves reaching AUC values higher than 0.90, suggesting an accuracy of 90% for some peers. Again, note the performance variation on account of the speakers being compared.

Finally, regarding the comparison of a system's performance based on $f0$ estimates extracted from connected speech vs. lengthened vowels samples, relatively better performance is suggested for the former compared to the latter. In Table 5.3, it is possible to verify that EER and particularly

Cllr values were found relatively increased, whereas a slight decrease of AUC values in lengthened vowels is indicated.

Table 5.3: **Log-likelihood-ratio Cost (Cllr), Equal Error Rate (EER) and AUC values for comparisons involving all speakers.**

Connected Speech				
Metric	$Cllr_{(raw)}$	$Cllr_{(cal)}$	EER	AUC
(f0med) $f0$ median	**0.56**	**0.55**	**0.24**	**0.78**
(f0mean) $f0$ mean	**0.54**	**0.54**	**0.20**	**0.77**
(f0base) $f0$ baseline	**0.52**	**0.50**	**0.13**	**0.77**
(f0min) $f0$ mininimum	**0.68**	**0.68**	**0.23**	**0.72**
(f0max) $f0$ maximum	**0.80**	**0.76**	**0.30**	**0.69**
(f0sd) $f0$ standard deviation	**1.03**	**0.95**	**0.45**	**0.61**
(f0SAQ) $f0$ dispersion	**0.95**	**0.94**	**0.46**	**0.59**
(f0M1) smoothed $f0$ peak rate	**3.76**	**1.00**	**0.46**	**0.57**
(f0M2) sd of $f0$ maxima	**1.10**	**0.98**	**0.40**	**0.58**
(f0M3) sd of the $f0$ maxima positions in seconds	**1.25**	**1.00**	**0.45**	**0.57**
(f0M4) $f0$ rising rate in the peaks	**1.00**	**0.95**	**0.45**	**0.61**
(f0M5) $f0$ falling rate in the peaks	**0.98**	**0.91**	**0.40**	**0.62**
(f0M6) sd of $f0$ rising rate	**1.02**	**0.98**	**0.44**	**0.59**
(f0M7) sd of $f0$ falling rate	**0.96**	**0.93**	**0.39**	**0.61**
(f0M8) $f0$ peak width	**1.01**	**0.94**	**0.40**	**0.63**
Lengthened Vowels				
(f0med) $f0$ median	**0.88**	**0.83**	**0.34**	**0.73**
(f0mean) $f0$ mean	**0.87**	**0.83**	**0.32**	**0.73**
(f0base) $f0$ baseline	**0.84**	**0.81**	**0.29**	**0.74**
(f0min) $f0$ minimum	**0.84**	**0.81**	**0.30**	**0.74**
(f0max) maximum	**0.91**	**0.85**	**0.38**	**0.72**
(f0sd) $f0$ standard deviation	**1.09**	**1.00**	**0.49**	**0.57**
(f0SAQ) $f0$ dispersion	**1.13**	**1.00**	**0.54**	**0.56**

5.3 Discussion

The purpose of this experiment was to evaluate the speaker-discriminatory potential of a set of $f0$ parameters in comparison carried out with genetically-related subjects (i.e., identical twins) and in cross-pair comparisons among all speakers. The inter-speaker comparisons were performed on the basis of two different kinds of *speech material*: intervals of continuous speech (i.e., at the level of sentences) vs. lengthened vowels extracted from dialogues which were perceived, in most

Figure 5.3: **ROC plots including AUC estimates for $f0$ median, mean, baseline, standard deviation (sd) and SAQ parameters for E1-E2, F1F2, A1-A2, and B1-B2 twin pairs.**

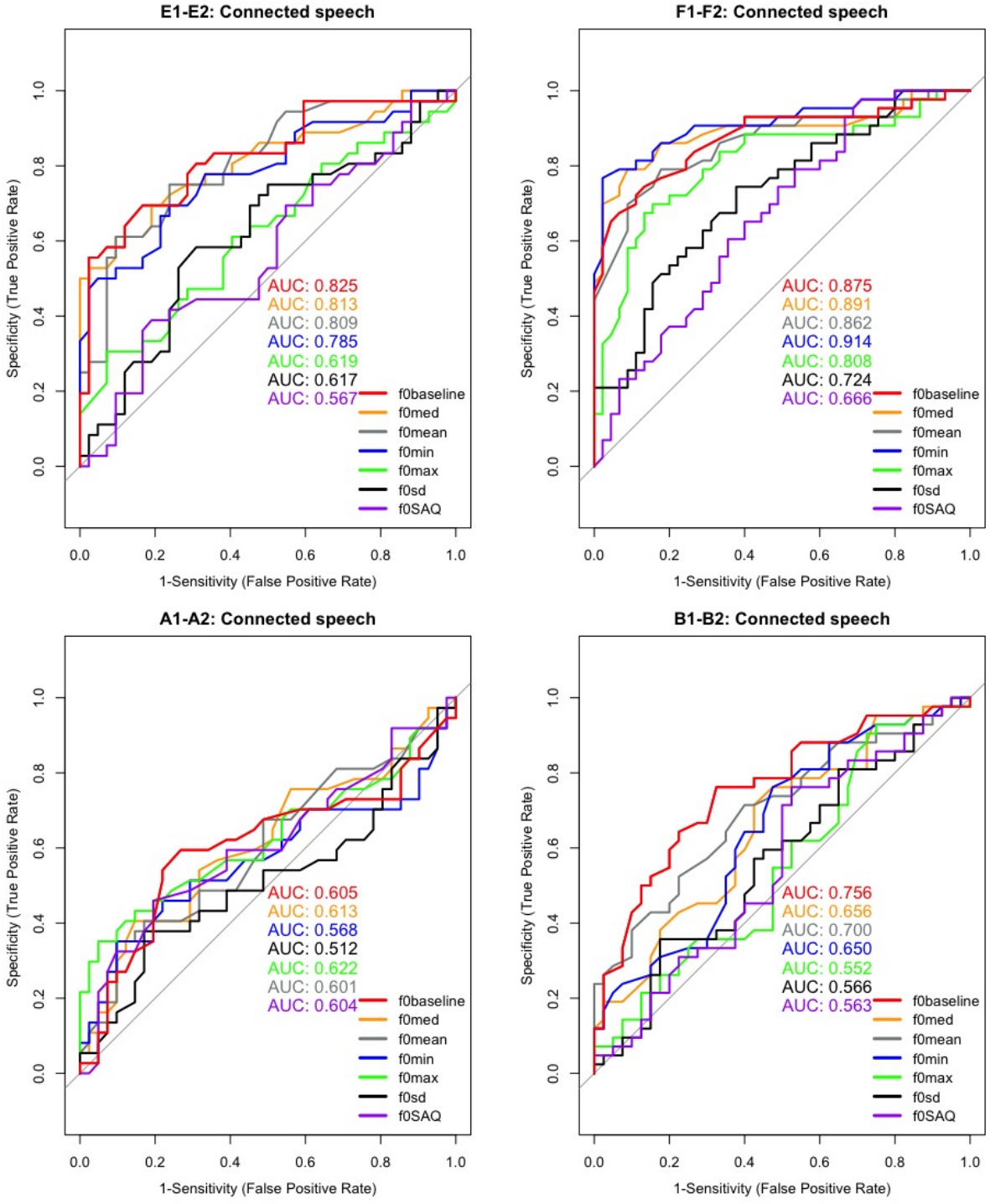


Figure 5.4: **ROC plots including AUC estimates for $f0$ median, mean, baseline, standard deviation (sd) and SAQ parameters for the other twin pairs.**

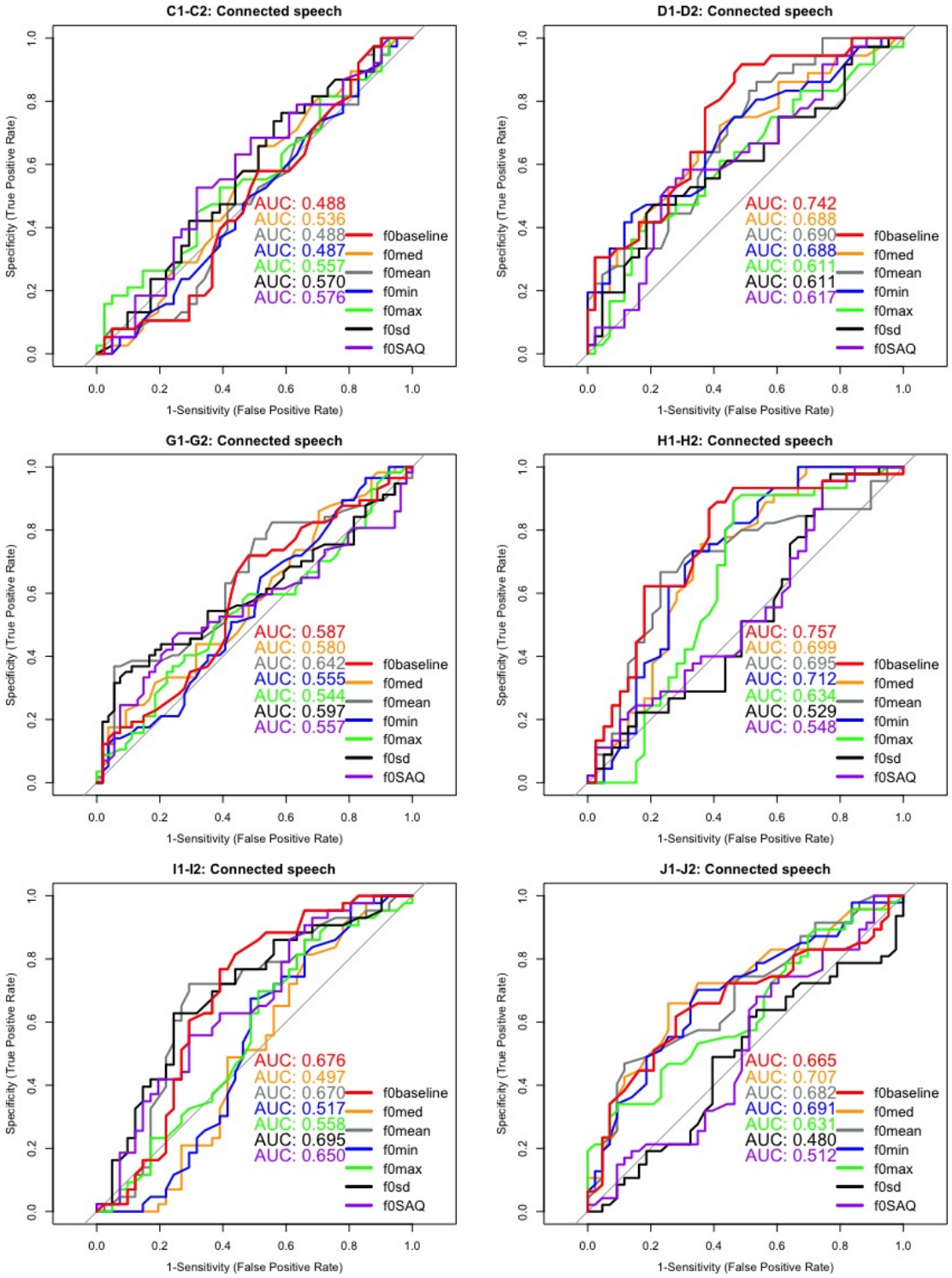


Figure 5.5: **ROC plots including AUC estimates for $f0$ median, mean, baseline for cross-pair comparisons (I).**

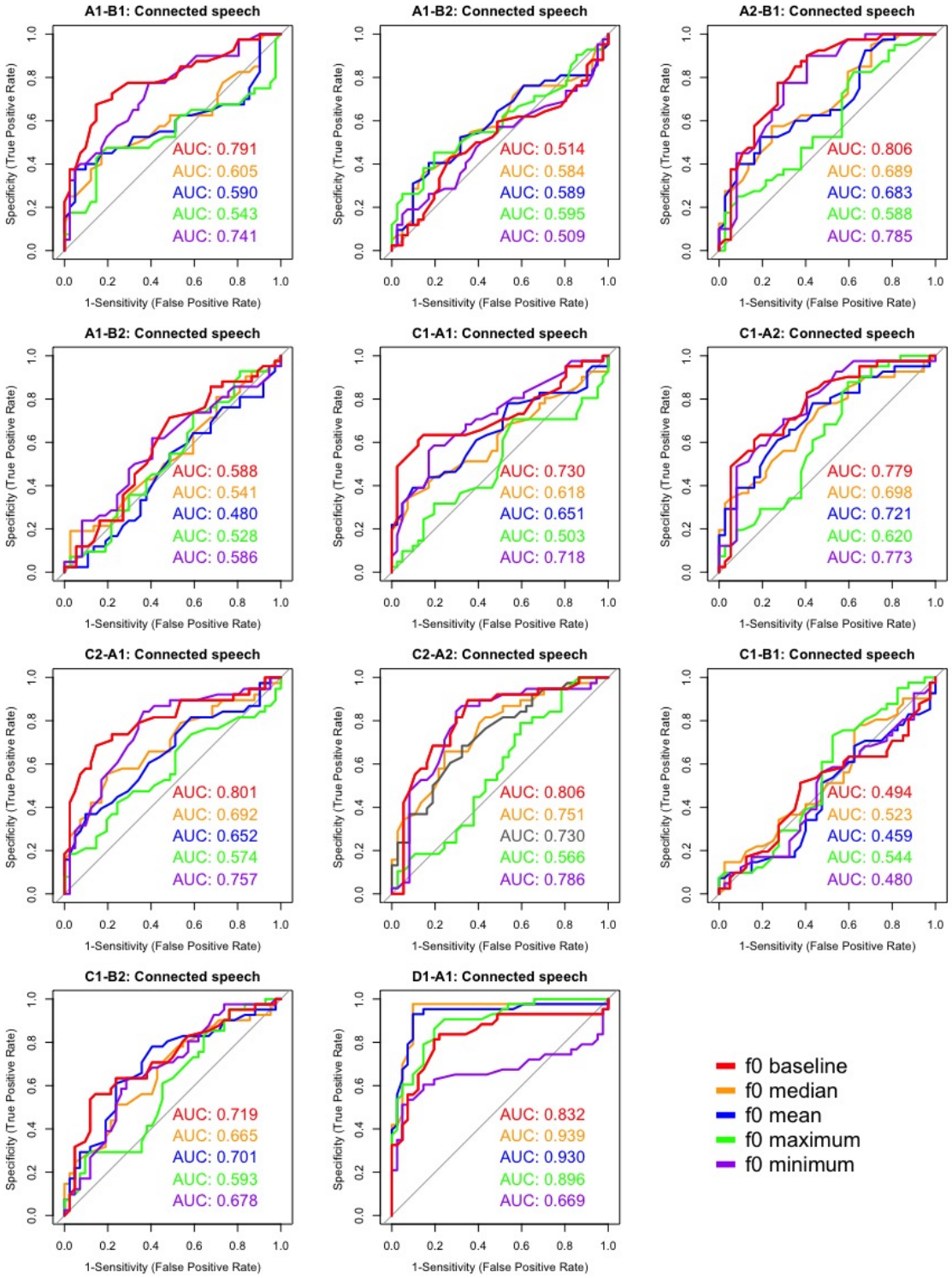
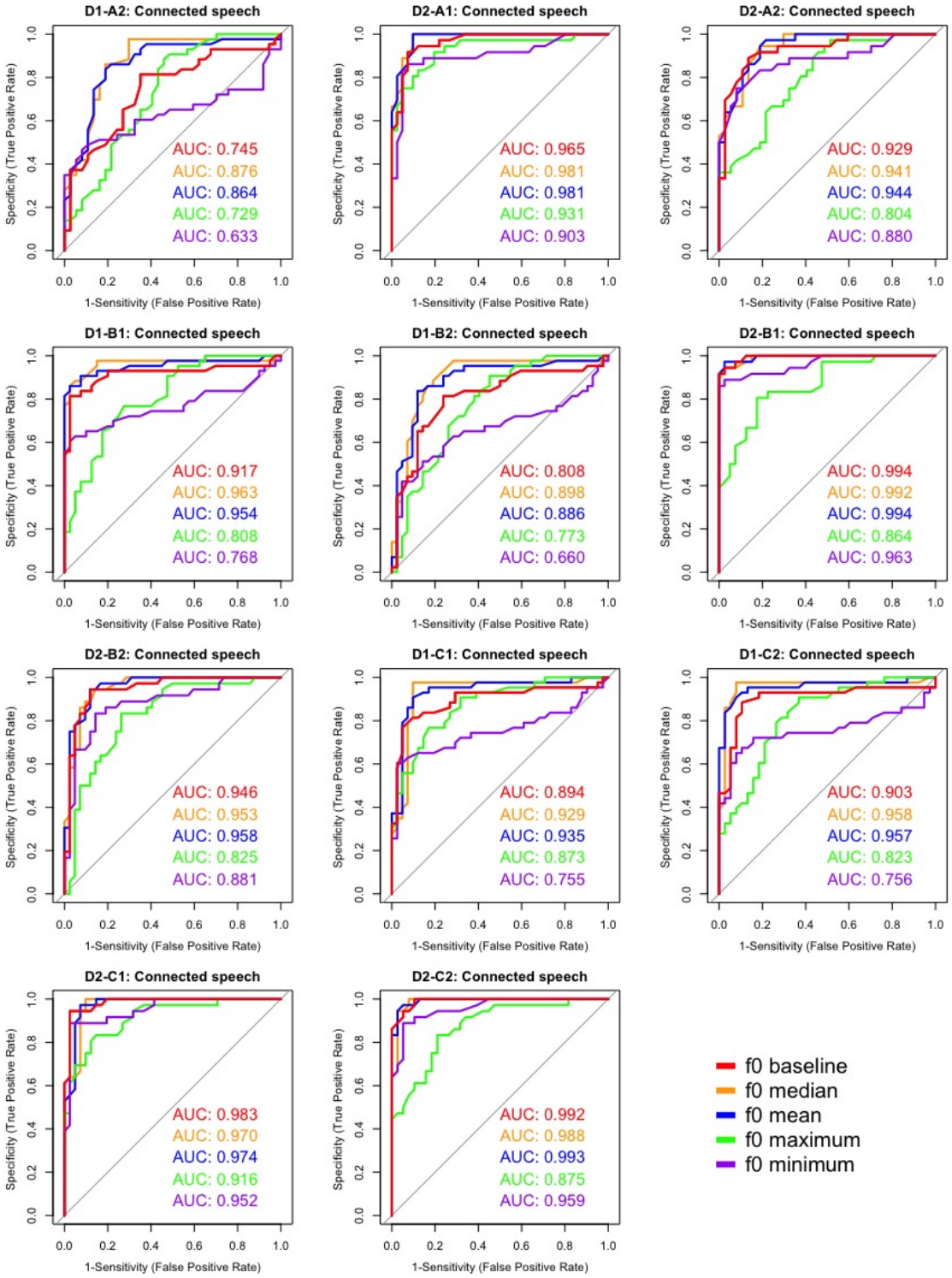


Figure 5.6: **ROC plots including AUC estimates for $f0$ median, mean, baseline for cross-pair comparisons (II).**



cases, as filled pauses.

Some general discriminatory patterns were suggested regarding the variability of $f0$ parameters in intra-twin pair and cross-pair comparisons. Overall, centrality long-term $f0$ estimates plus $f0$ baseline were found the most speaker discriminatory estimates in connected speech as well as in lengthened vowels. The second most contrasting parameters category regards $f0$ floor and ceiling estimates. Such $f0$ descriptors were also the ones displaying the largest effect sizes. Conversely, $f0$ variation and modulation-dependent estimates, such as $f0$ standard deviation and dispersion, were the ones displaying the lowest proportions of inter-speaker differences.

The observation of higher discriminatory potential of long-term $f0$ measurements in more extensive speech intervals is not surprising. As expected, $f0$ excursions on such intervals are much more expressive than in lengthened vowel tokens, which present a relative stationary $f0$ contour pattern.

Such an observation may also have critical perceptual implications, once the listener's discriminatory performance may be potentially better when larger speech stretches are used in auditory discrimination trials, as more individual acoustic information is present.

Although the set of $f0$ parameters assessed in the study conducted by Whiteside and Rixon (2013) was considerably smaller than the set of parameters examined in the present study, the main outcomes converge in the same direction. When comparing siblings, it was observed that the measurement associated with $f0$ variability (i.e., $f0$ standard deviation) was the parameter displaying the greatest level of similarity across the siblings, while mean $f0$ values presented the highest discrepancies, even between an identical twin pair. According to the authors, the results suggest that more dynamic aspects of $f0$ may be under a greater influence of environmental variables, which are responsible for shaping intonation patterns, such as dialect and speaking style. Such a higher environment influence on the $f0$ variability was also previously suggested by Debruyne et al. (2002).

The outcomes of the present study do not disprove the assumption that individuals vary in how they exploit fundamental frequency. However, the observation of lower inter-speaker difference proportions for $f0$ variation/modulation estimates invites the hypothesis of such parameters as varying to a lesser extent in comparisons performed based on the same speaking style, same speaking condition, same dialect, and with a relatively homogeneous group composed only by young male speakers.

Variation in the fundamental frequency dimension has also been referred to as associated with particular characteristics of different emotional states. In the study conducted by Higuchi, Hirai, and Sagisaka (1997), while analyzing $f0$ contours of Japanese sentences on the basis of $f0$ min, $f0$ amplitude at the phrase domain, and $f0$ amplitude at the lexical domain, the authors observed clear

differences when comparing four styles/states, e.g., unmarked, hurried, angry, and gentle. Amongst the analyzed styles/states, more expressive differences were observed for the angry speech, which was characterized by a high $f0$ min and a minimum change in both phrase and lexical amplitudes, yielding reasonable flat $f0$ contours.

Furthermore, the higher convergence among individuals for $f0$ variation estimates in the present study seems to be in broad agreement with Arantes and Eriksson (2019), when observing that speaking style has a significant effect on the shaping of f0 distributions, with a higher or lower variation depending on the speaking style being analyzed. In the study carried out with a multilingual corpus (including the analysis of Brazilian Portuguese, British English, and Swedish) comprised of speech productions in three different speaking styles, the authors observed that interview or sentence reading were the styles in which speakers differed the most in terms of $f0$ distribution shape. Furthermore, encouraging evidence of a remarkable intra-speaker stability of the shape of $f0$ distributions was also found, in which $f0$ contour by the same speaker tended to vary less when speaking in different styles than the contours of different speakers that are speaking in the same style, which was especially the case for the spontaneous speaking style. As suggested by the researchers, this could indicate the suitability of this feature as a good parameter in forensic speaker comparison, as expressed by a low within-speaker variation and a high inter-speaker variability.

5.3.1 Intra-twin pair comparisons

Concerning intra-twin pair comparisons, it was possible to observe that from a total of 11 intra-pair significant differences, only two of those concerned to $f0$ modulation parameters, whereas nine differences derived from $f0$ centrality, baseline, and extreme values, mostly from $f0$ median, mean, baseline and minimum. Moreover, from the group of 10 identical twin pairs assessed, only three pairs could be differentiated according to the present study's approach regarding the connected speech analysis, namely E1-E2, F1-F2, and G1-G2; however, only two of these were more systematically and consistently differentiated: E1-E2 and F1-F2. As for the analysis of lengthened vowel segments, only one twin pair was statistically different: I1-I2.

It is noteworthy that E1-E2, mostly E2, are the only speakers in the present study with a reported smoking habit, whereas both F1 and F2 are the ones who use their voice artistically (i.e., acting). Although, according to E1-E2, smoking is not a shared daily-base habit, such information signals the relevance of further analysis, given the widely reported and well-described effects of smoking on $f0$, characterized mainly by a $f0$ lowering (MURPHY; DOYLE, 1987; GONZALEZ; CARPI, 2004; AYOUB; LARROUY-MAESTRI; MORSOMME, 2019). However, the fact that

such difference could not be observed when considering the analysis of $f0$ estimates in the set of lengthened vowels selected does not allow any inference to be made. In addition, in the analyses conducted on vowel formants in *Chapter 3*, E1-E2 was also found statistically distinct, particularly regarding their fourth vowel formant (F4), revealing that the acoustic-phonetic dissimilarity between E1-E2 is not solely restricted to the $f0$ dimension.

It is noteworthy that F1-F2 was the twin pair displaying the highest proportion of significant differences. Such individuals respond for more than half of the intra-twin pair differences observed. The fact that such speakers have a reported experience in using their voice artistically (which was, in fact, one of their main topics during the dialogues, where vocal quality adjustments for composing a certain character were mentioned) may suggest a higher "self-awareness" regarding how they sound and are perceived by others. In this instance, the factor "choice" can not be disregarded, as suggested and considered by many other twin studies at different acoustic-phonetic levels (NOLAN; OH, 1996; LOAKES, 2008; ZUO; MOK, 2015; SAN SEGUNDO; YANG, 2019).

The observation that no significant differences were observed in the comparison of MZ and DZ twins regarding the $f0$ variation in Debruyne et al. (2002) is a rather interesting finding, once it suggests an unbalanced contribution of genetic and environmental factors on the parameter average values and variation. As remarked by the authors, while the $f0$ variability may be highly determined by behavioral and adaptive elements (i.e., under a strong environmental influence), a comparable influence on both DZ and MZ is presumed. As such, genetic factors may be overwhelmed by external factors, no longer accounting for the difference between MZ and DZ.

In the present study, the verification of a possibly higher intra-twin pair congruence regarding $f0$ variation and modulation parameters solely can not be taken as an indication that such divergences do not occur across twins' speech. Here, the mere fact that identical twins were taking part in a simultaneous dialogue may induce some sort of *prosodic entrainment* or *synchronicity*, which may be on the basis of their higher congruence. Such synchronicity has been explored and corroborated in the experiment conducted by Buder and Eriksson (1997) while analyzing the prosodic cycles of mean speech fundamental frequency and mean voice intensity between subjects while engaging in a conversation.

However, a potential "*synchronicity effect*" can not be generalized to the comparisons performed across all subjects (i.e., cross-pair comparisons), where a higher congruence for $f0$ variation and modulation across speakers has also been suggested. Note that, except for the ten intra-twin pair dialogues in the 190 cross-pair comparisons, 180 comparisons were carried among individuals who were not taking part in the same dialogue (e.g., A1 - B1, C2 - A2, B2 - C1).

It should also be remarked that $f0$ patterns, like any other phonetic dimension, are regulated by the interaction of both intrinsic (system-oriented) and extrinsic (output-oriented) factors, cf.

Lindblom (1990a). Hence, assuming that identical twins were, in all cases, interacting with someone they were very familiar with is also expected to affect the way they exploit $f0$ in their speech. Such an observation adds to the ecological validity of the speech material analyzed, as already acknowledged in several passages of the present thesis.

Finally, of an unquestionable relevance is whether listeners could potentially perceive the discrepancies in $f0$ centrality and extreme values in intra-twin pair comparisons observed. In this regard, while assessing listeners' sensitivity to differences in the amount of change in $F0$ in an experiment with a "forced-choice" sample comparison procedure, where stimulus contained a pitch movement of variable size ranging from 1 to 6 semitones, 't Hart (1981) found evidence suggesting that only differences of more than three semitones seem to play a part in communicative situations. Notwithstanding, there was evidence for good discriminators as being able to perceive $f0$ modulations of about 2.1 - 2.8 semitones, with $f0$ rises being better perceived than $f0$ falls. As verified in the present study, the magnitude of the differences observed for E1-E2 and F1-F2 concerning $f0$ centrality and extreme values in speech were, in some cases, on average, just below three semitones, except F1-F2 for $f0$ maximum. Therefore, it is uncertain whether such differences, for $f0$ only, were large enough to be perceived by the individuals, which certainly demands further analysis.

5.3.2 Connected speech vs. lengthened vowels

Regarding the comparison of $f0$ estimates in connected speech versus lengthened vowels, a reduction in speaker-discriminatory potential was suggested, as expressed by lower proportions of inter-speaker differences, mostly for $f0$ median, mean, and baseline. Besides, smaller effect sizes were observed in lengthened vowels, which is compatible with a higher explanatory potential of the variable "*speaker*" on global patterns of $f0$ in connected speech. From a statistical viewpoint, such effect size reduction may be interpreted as the consequence of a lower variability of $f0$ in lengthened vowels, which tend to present a relatively stationary pattern, in contrast to more extensive speech intervals.

In this regard, different studies within the clinical voice domain (i.e., vocology studies) support the observation that different acoustic outcomes regarding $f0$ may be obtained depending on the task and consequently on the nature of the material under assessment (e.g., comparisons of sentences and sustained vowels), suggesting that for a more reliable assessment of the parameter, material collected in more than one task have to be considered (ZRAICK; SKAGGS; MONTAGUE, 2000; MOON et al., 2012; VIEGAS et al., 2019; SOTOME et al., 2021). Although the differences observed in the present study on account of the comparison of $f0$ parameters in speech intervals and lengthened vowels can not be attributed to the variable "*task*" given that the measurements

were extracted from the same recordings, they can be potentially assigned to differences in the *speech material* under assessment. Overall, the results suggest that not all the acoustic complexity present in speech regarding $f0$ is reflected in the production of lengthened vowels as observed in the present study.

Finally, in this study, a higher intra-twin pair convergence for lengthened vowels comparisons in relation to connected speech comparisons was suggested. According to the present study's analysis approach, no identical twin pair seemed to diverge significantly in the former condition. Such an observation may indicate that a greater inter-speaker variability in long-term $f0$ estimates is present in continuous speech, a condition in which twins can avail of their vocal plasticity and, deliberately or not, diverge from each other.

In summary, the findings of the present experimental study appear to point in the direction that the magnitude of acoustic-phonetic similarity/dissimilarity between identical twins is, in part, dependent on the nature of speech/voice material under analysis, and consequently, on how representative of the speaker's voice/speech behaviour it is in a broader sense.

5.3.3 The discriminatory performance of $f0$ descriptors

The verification of the baseline value of $f0$ (i.e., $f0$ baseline) as the most accurate measure is in broad agreement with previous literature reporting. While conducting a comprehensive study involving different $f0$ parameters with Brazilian Portuguese speakers using a Multivariate Kernel Density (MVKD) analysis, Silva et al. (2016) noted that the $f0$ baseline was the estimate displaying the lowest Equal Error Rate (ERR) in relation to other $f0$ measures. Moreover, the authors noted an even better discriminatory performance when combining $f0$ baseline and $f0$ median. Furthermore, $f0$ baseline has also been reported in the literature as the most stable parameters with regard to different sources of variation, such as speaking style, vocal effort, and recording quality (LINDH; ERIKSSON, 2007).

Additional support for the use of baseline and median values of $f0$ instead, especially when conducting forensic analysis, regards the fact that, as remarked by Lindh and Eriksson (2007), the extraction of $f0$ values from recordings is often done automatically. When manual examinations are carried out after automatic extractions, they may very likely reveal measurement errors which are also dependent on the audio quality, such as octave jumps. It goes without saying that, within a forensic context, audio samples with poor quality are the rule, and good audio recordings the exceptions. Therefore, using median values instead of means would be more reliable as it is less sensitive to possible outliers. In that instance, perhaps the reason why a discrepancy between $f0$ mean and $f0$ median could not be observed in the present study has to do with the fact that only

very high-quality recordings were assessed, not allowing such verification.

Future studies should also explore more in-depth how the patterns observed here vary as a function of different speaking styles and different communicative contexts (e.g., spontaneous dialogue vs. interview with an unfamiliar interlocutor), providing relevant information on the levels of stability and variability of the parameters assessed for forensic speaker comparison purposes.

It is important to emphasize that, although in the present experimental study, $f0$ parameters have been assessed individually, different acoustic-phonetic estimates should be acknowledged in a realistic forensic context. The combination of different metrics from different acoustic-phonetic dimensions (e.g., spectral, temporal) tends to yield better explanatory models and more accurate speaker profiles.

Finally, as remarked by Braun (1995), given the social-cultural influences on how $f0$ is perceived in a language, one should be extremely cautious when considering the use of reference material acquired from one language group to another. This fact justifies the relevance of cross-linguistic studies, with particular consideration to those parameters commonly analyzed in a forensic setting.

Chapter 6

General trends

6.1 Introduction

In the present chapter, an attempt will be made to contextualize some of the main findings deriving from the previous analyses, while also acknowledging the discriminatory power of the different acoustic-phonetic parameters assessed. Furthermore, some brief comments regarding the overall differences observed in intra-twin pair comparisons will also be sketched. Lastly, a summary of the main findings is provided, and some new directions concerning the present research are pointed out.

6.2 General patterns on the acoustic-phonetic variability

As of yet, the analysis performed here has not considered a direct/straightforward comparison of the discriminatory outcomes of all different parameters derived from distinct acoustic dimensions, i.e., formant frequency, speech timing, and $f0$ dimensions. However, such a comparison is valuable given that, in real-world speaker comparison circumstances, not all estimates may be readily available for analysis due to system transmission restrictions, limitations imposed by the integrity/quality of the material assessed, or merely because the extension of the material being analyzed is rather reduced. Hence, knowing the amount of idiosyncratic information that one can derive from different acoustic estimates and understanding their discriminatory capacity and possible limitations is crucial. A brief comparison of the discriminatory and explanatory potential of acoustic-phonetic parameters will be made in view of this.

For such a cross-parameter comparison, a primary focus must be drawn to effect size estimates, which offer us the possibility of comparing acoustic-phonetic parameters without disregarding the

Table 6.1: **Discriminatory performance across parameters from different dimensions, i.e., fundamental frequency, formant frequency, and speech timing dimensions.**

Parameter	Eff. size	Mag	AUC	EER	$\text{Cllr}_{(\text{raw})}$	$\text{Cllr}_{(\text{cal})}$	Twins %	Cross-pairs %
$f0$ **median**	50 %	Large	0.78	0.24	0.56	0.55	20 %	51%
$f0$ **mean**	49 %	Large	0.77	0.20	0.54	0.54	20 %	49%
$f0$ **baseline**	47 %	Large	0.77	0.13	0.52	0.50	20 %	47%
$f0$ **minimum**	32 %	Large	0.72	0.23	0.68	0.68	20 %	39%
F3 [a]	34 %	Large	0.71	0.20	0.72	0.60	30 %	65%
F4 [a]	21 %	Large	0.65	0.24	1.34	0.84	60 %	61%
F1 [a]	10 %	Mod	0.61	0.20	0.70	0.69	30 %	42 %
F2 [a]	9 %	Mod	0.60	0.33	0.94	0.85	20 %	34 %
SRATE	15 %	Large	0.64	0.28	0.78	0.78	None	21 %
ARTRATE I	15 %	Large	0.64	0.27	0.76	0.75	None	27 %
ARTRATE II	12 %	Mod	0.62	0.31	0.78	0.75	None	13 %
IPI	11 %	Mod	0.63	0.42	0.88	0.88	10 %	11 %
SGDUR	6 %	Mod	0.59	0.35	0.97	0.89	None	22 %

impact of n-size on their discriminatory patterns.

In Table 6.1, some of the most relevant parameters from different acoustic-phonetic dimensions are presented. For the sake of consistency, formant frequency outcomes extracted only from vowel [a] will be used, given the impact of vowel quality on the discriminatory patterns of F1 and F2. Also, note that this was the most recurrent vowel in the corpus. Additionally, given the limited number of estimates for this dimension of analysis, all formants are displayed. Finally, for comparability, non-calibrated (raw) and calibrated (cal) Cllr estimates are provided.

By closely inspecting effect size values depicted in Table 6.1, i.e., the magnitude of the variation assigned to the variable "*speaker*", an overall trend is suggested. As can be noted, in general, the explanatory potential of the variable "*speaker*" appears substantially greater for those parameters from the spectral and melodic dimensions than for those of the speech timing domain. The only exception seems to be F1 and F2. However, note that the vowel [a] was suggested as the one most affected by the lexical stress factor, which may be one of the reasons for such an effect size reduction. In addition, when only stressed vowels are considered, F1 and F2's effect sizes turn out to be large: 19% and 16%, respectively.

Although no direct comparison can be done regarding the proportion of inter-speaker differences between formant frequencies and the other dimensions– only for effect sizes and discriminatory performance metrics, it is possible to confront the proportions inter-speaker differences between $f0$ and speech timing estimates since their data originate from the same observations.

As can be observed in Table 6.1, higher proportions of cross-pair differences have been noted

for $f0$ descriptors than for speech timing parameters, twice more in some cases. Such a trend indicates that, when everything else is kept constant, namely, language, dialect, speaking style, speaking condition, sex, higher inter-speaker differences can be observed at the melodic dimension than at the speech timing dimension. In addition, relatively better discriminatory performances were also suggested for $f0$ estimates than for speech timing measures, as expressed by higher AUC and lower Cllr/EER values.

In Figure 6.1, ROC curves are displayed for the twin pair with the greatest dissimilarity defined by the proportion of intra-twin differences: E1-E2. Note that, despite such a tendency, the discriminatory performances of parameters do not seem insensitive to the nature of the *speech material* assessed, given the poorer performance of $f0$ extracted from long vowels in relation to connected speech. Such an observation is of an unquestionable relevance from a forensic phonetics viewpoint, once it signals a possibly weaker strength of evidence when less representative samples of a speaker is elicited or selected for confrontation.

Furthermore, by inspecting the distributions of some of the most discriminatory $f0$ and speech timing parameters in the form of Kernel density curves in Figure 6.2, it is possible to observe that speech rate (SRATE) and articulation rate (ARTRATE I) curves have a relatively more spread distribution than the $f0$ curves, yielding wider distributions with a higher superposition. Conversely, less spread curves can be observed for $f0$ estimates. It can also be noted that the black points at the bottom of distributions, standing for individual mean values, appear less dispersed in speech tempo estimates than for $f0$ estimates. Such a higher/lower superposition bears consequences on the discriminatory performance of the parameters, that is, increasing or reducing the uncertainty around the estimates. These patterns seem to explain the effect sizes and AUC values observed for such parameters.

Some previous research has acknowledged the relatively poorer discriminatory potential of speech timing estimates. While assessing the discriminatory power of speech timing parameters, including speech and articulation rates, Künzel (1997) remarked that, through the verification of their equal error distributions, it is necessary to acknowledge that the discriminating capacity of speech timing measures seemed rather poor in comparison with other acoustic parameters, such as linear predictive coding (LPC) or cepstrum coefficients. The author also emphasizes that, unlike most acoustic parameters, an estimate such as articulation rate is much more appropriate for use under real-world forensic conditions, often involving telephone transmitted speech and non-cooperative speakers. The present study appears to provide further evidence of this uneven discriminatory potential regarding estimates deriving from different acoustic-phonetic dimensions.

Also, in the experiment conducted by Lennon, Plug, and Gold (2019) with 30 English speakers, aiming to compare common speaking rate measures (e.g., rates based on the counting of canon-

Figure 6.1: **ROC curves for E1-E2 intra-twin pair comparison**

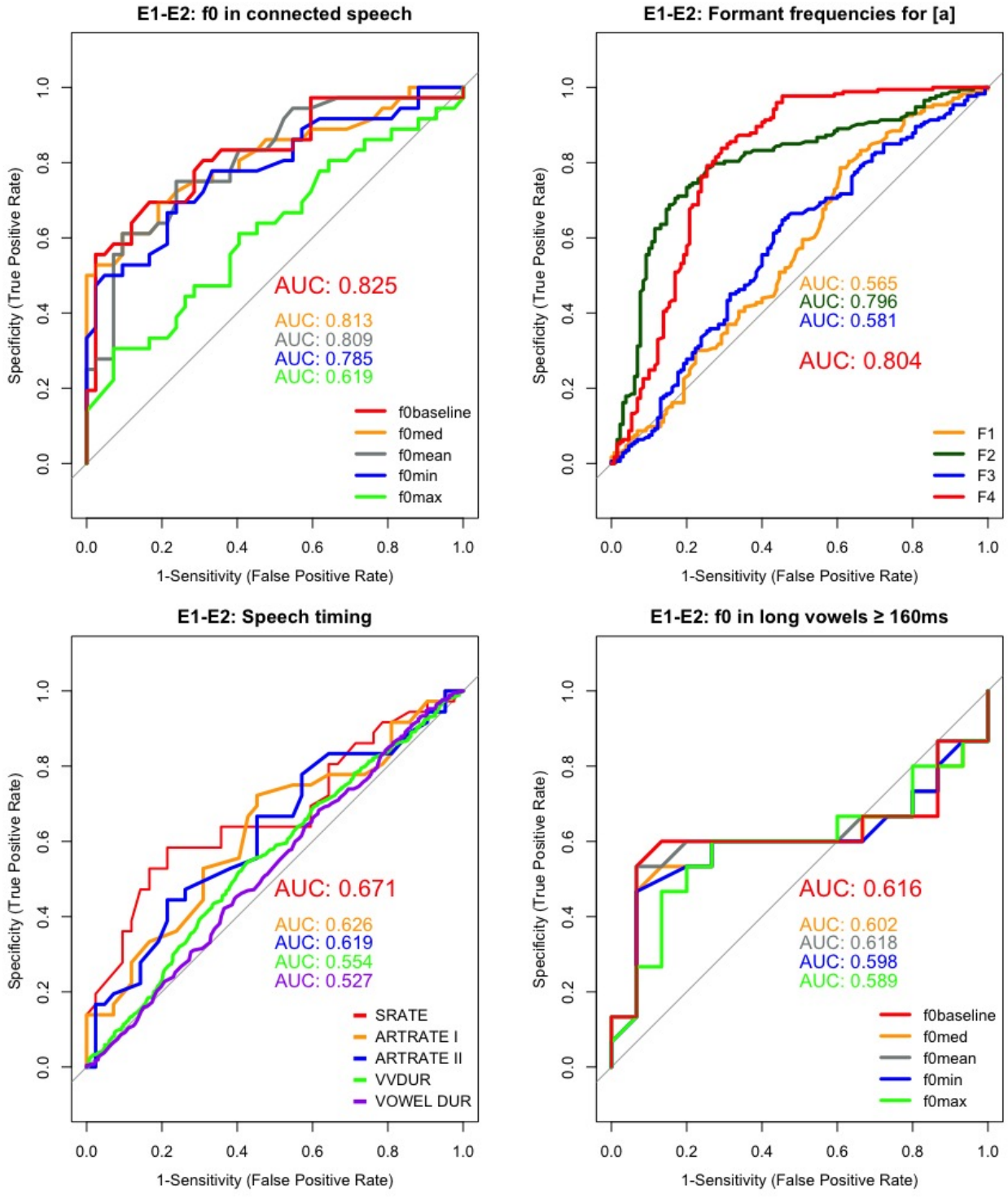
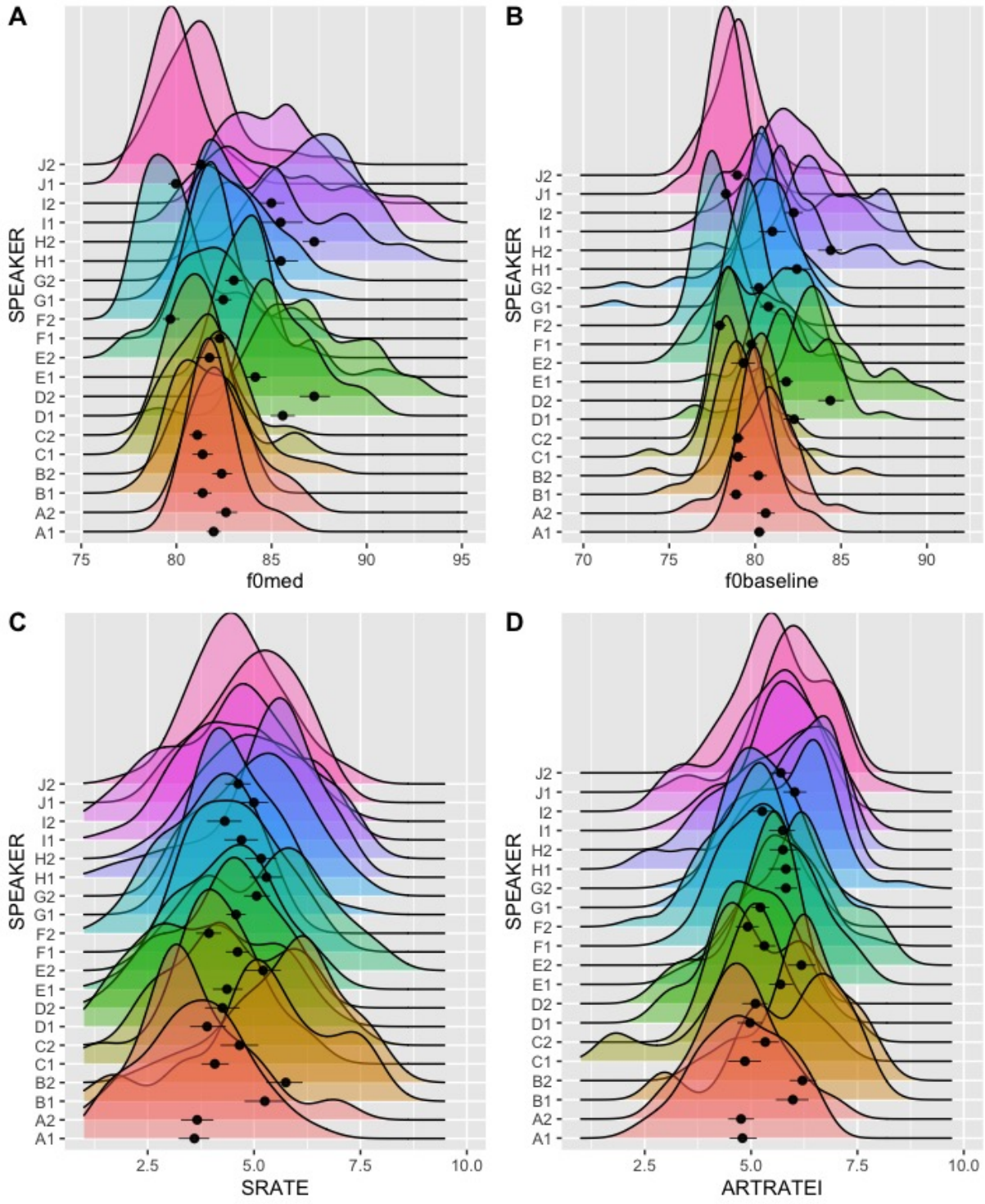


Figure 6.2: **Kernel density curves for $f0$ and speech timing parameters**



ical and surface syllable, phones, and CV segments), it was verified that these rates were closely inter-correlated yielding similar discriminating powers. Notwithstanding, as remarked by the researchers, the results suggested that tempo is a relatively poor speaker discriminant regardless of methodology, as characterized by rather high EERs and Cllrs close to 1. In that study, EER values varied from 0.28 to 0.37, a range that is compatible with the values observed here. As for Cllr values, these varied from 0.88 to 0.89, being slightly higher than the observed for speech rate and articulation rate here.

The analyses carried out by Hughes, Brereton, and Gold (2013) on the implication of reference sample size, and the calculation of numerical LR based on articulation rate revealed the same tendency. In the referred study, both EER and Cllr average values were found relatively high – of 35% and 0.97, respectively– suggesting an overall poor performance of articulation rate for forensic speaker comparison ends. Furthermore, it was verified that the EER estimate tended to remain stable/consistent with the increase in the number of tokens, not presenting important repercussions in terms of categorical system validity. As for Cllr, calibrated LRs were found to be robust to sample size effects, whilst non-calibrated scores displayed much more sensitivity to the amount of reference data used.

6.2.1 Intra-twin pair differences

Now that all three different acoustic-phonetic dimensions have been analyzed, we may now consider inspecting overall differences presented by the twin pairs. In Table 6.2, total numbers of significant differences are depicted as a function of twin pairs, and acoustic-phonetic dimensions analyzed, providing us a more comprehensive picture regarding their similarities and dissimilarities. For the counting of intra-twin pair significant differences at the vowel formant dimension, all formants concerning the combination of stressed and unstressed vowels are considered[1]. For consistency, marginally significant differences were disregarded.

As shown in Table 6.2, among all pairs, E1-E2 and G1-G2 were the ones displaying the highest number of intra-pair differences; consequently, these were the twin pairs displaying the greatest phonetic dissimilarities in terms of absolute numbers. Moreover, as can be noted, E1-E2 appeared to diverge to a greater extent in comparison to G1-G2.

When consulting the personal information provided by E1 and E2 and also checking their responses to qualitative questions, it is not all very clear why this twin pair, in particular, was the one displaying the greatest dissimilarity, as observed in the following.

[1]In order to avoid an inflation in the proportion of intra-twin pair differences for the spectral analysis, significant differences observed by testing stressed and non-stressed vowels separately were not included

Table 6.2: **Number of intra-twin pair significant differences per dimension of analysis. Speaker comparisons performed based on the same statistical approach.**

	Twin pairs									
Dimension	**A1A2**	**B1B2**	**C1C2**	**D1D2**	**E1E2**	**F1F2**	**G1G2**	**H1H1**	**I1I2**	**J1J2**
Formant frequency	0	5	7	4	12	1	11	6	0	1
Speech timing	0	0	0	0	0	1	0	0	0	0
$f0$	0	0	0	0	4	6	1	0	0	0
Sum	0	5	7	4	16	7	12	6	0	1

Regarding some qualitative questions: when questioned how similar they find themselves, E1 responded "similar" while E2 responded "very similar"[2]. When questioned how similar they find their voices, both responded "similar". When questioned whether they like the fact that they are twins– which we assume is a very important and possibly clarifying question, both answered "yes"[3]. When questioned whether they try to sound different from each other, both answered "never"[4]. When questioned whether they try to look different from each other, E1 answered "never", whereas "E2" replied "sometimes", which was the most common answer for this question. When questioned how often they are misidentified or erroneously recognized by other individuals because of their looks, both responded "often", and as a consequence of the way they speak, both responded "sometimes". When questioned whether they share the same social groups, both answered "often".

Notwithstanding, two aspects regarding this twin pair must be considered with attention. The first regards the fact that E2 has reported to smoke (on a weekly basis), while E2 only does it very occasionally. Such a habit also seemed to be accompanied by the alcohol consumption habit, for which E1 answered "sometimes", whereas E2 answered "yes", suggesting a higher frequency. The second aspect regards the fact that E1 is currently attending a higher education course, whereas E2 has did not complete the secondary school ("ensino médio" in Brazil). However, as previously mentioned, it must be noted that $f0$ differences for this pair– which signal some attention on their smoking habit– were restricted to the connected speech domain, not being observed

[2]Possible answers for this and the following question were: "nothing similar", "not so similar", "average", "similar", and "very similar".

[3]Possible answers for this question were: "yes", "no", "indifferent", "very much so", and "not at all".

[4]Possible answers for this and the following questions were: "never", "sometimes", "often", and "very often".

when assessed in prolonged vowels. It should also be noted that differences in $f0$ were equally observed for non-smoker pairs, such as F1-F2, of a very similar magnitude, i.e., just around/below 3 semitones. Hence, it seems premature to infer anything solely on account of this observation. Further investigation is unquestionably needed (e.g., whether such a difference stands out in other speaking conditions, not involving mobile phone transmitted speech).

As for G1 and G2, the second pair with the highest number of significant differences, one qualitative information, in particular, may be a (potential) explanatory factor of their differences, that is: when questioned whether they try to sound different from each other, G1 answered "sometimes", whereas G2 answered "never". Note a non-agreement in their responses. Regarding other relevant qualitative information, a general agreement was observed. When questioned how similar they find themselves, both responded "similar". When questioned how similar they find their voices, both responded "similar". When questioned whether they like the fact that they are twins, one of them answered "yes" and the other "very much so". When questioned whether they try to look different from each other, both replied "sometimes". When questioned how often they are misidentified or erroneously recognized by other individuals because of their looks, both responded "often", and as a consequence of the way they speak, both responded "often". When questioned whether they share the same social groups, both answered "often". Regarding their educational background, both have completed a university degree.

The variable "*age*"[5], which is referred to as a hypothetical factor related to potential differences between identical twins, does not seem a reasonable explanation for differences presented by E1-E2 and G1-G2. Note that, E1-E2 which were 26 years old, were found relatively more dissimilar than G1-G2, aged 29 years old. In addition, such individuals are relatively young. It should also be remarked that J1-J2, the oldest pair in the study, aged 35 at the time of the recording, was among the most similar pairs. Also, it has to be admitted that the age difference among pairs in the present research is considerably restricted for safely considering "age" as an explanatory variable, as done by some previous studies with a more representative age span, cf. Van Lierde et al. (2005). However, such an age "homogeneity" verified in the present study was not unintentional; on the contrary, it reflects a methodological criterion.

Notably, caution must be paid when trying to make a straightforward/direct connection between the twins' responses and the patterns, as it may be often puzzling. Their perception as speakers is not any better than anyone else just because they are twins. As such, questions such as "do you try to sound different from your twin brother?" can be somewhat naive once it disregards the non-reflected reality of linguistic behavior. Even so, we allowed ourselves to pose such

[5]In this regard, identical twins which are considerably older would be expected to be more dissimilar than younger twin pairs, given the influences of environment/learning during their lifetime.

questions. Secondly, one may sometimes see oneself in a way that is not in complete agreement with how others see him/her, perhaps because of a deliberate attempt to emphasize or even reduce an apparent difference/similarity. That, solely, is a source of uncertainty. However, determining the exact reasons underlying intra-twin pair differences is, by far, outside the scope of the present research.

Furthermore, before advancing, it worth noting that, in the present study, identical twins were solely compared regarding the acoustic instance of their speech production. As such, exploring to what extent the acoustic-phonetic patterns observed here may relate to potential perceptual differences or whether such patterns may help or not intra-twin pairs' separability remains a matter for further investigation.

6.2.2 A continuum of inter-speaker similarity

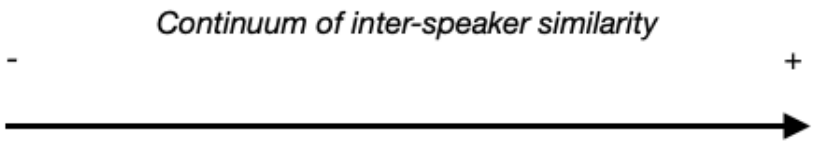
The outcomes of the present research corroborate the observation that, although identical twins appear, in general, similar concerning their acoustic-phonetic patterns– an outcome more or less anticipated, given that they were actively interacting with each other, the magnitude of intra-twin pair similarity varied as a function of the dyad under comparison.

As suggested in Table 6.1, the levels of intra-twin pair similarity, assessed on the basis of their overall significant differences, varied across twin pairs. While some peers did not reveal any significant differences (i.e., A1-A2 and I1-I2), others were found relatively more distinct from each other, as was the case for E1-E2. Based on such an observation, a continuum of inter-speaker similarity is suggested, as sketched in Figure 6.3. Such a higher or lower level of intra-twin similarity in function of the pair being analyzed has also been observed in previous studies for different acoustic-phonetic domains, cf. Loakes (2006), San Segundo, Tsanas, and Gómez-Vilda (2017) and San Segundo and Yang (2019).

Note that such scheme presented in Figure 6.3 is nothing but an oversimplification, a one-dimensional abstract scale that may represent the level o similarity for the comparison of any individuals. On such scale, some pairs would be relatively closer to the right edge of the continuum than others. That is what one would expect for A1-A2 and I1-I2, for instance, as in opposition to E1-E2. Furthermore, identical twins, in general, would likely be much closer to the *(+)* instance when compared to unrelated individuals, given the contributions of different factors from different domains as acknowledged by the literature (e.g., organic, environmental, and their mutual influence).

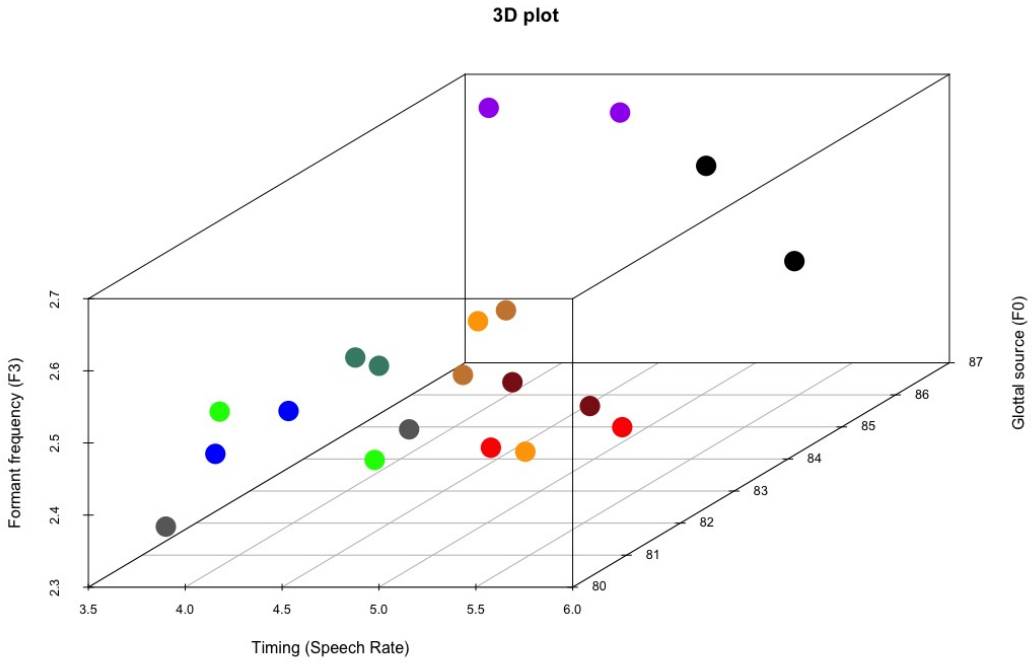
Let us now assume that, instead of using a one-dimensional scale to represent the differences observed across individuals, their outcomes would be depicted in a three-dimensional represen-

Figure 6.3: **One-dimensional representation**



tation. For that, at least three analysis dimensions are required, which we are assuming to be "formant frequency", "timing", and "glottal source" while speakers are represented by points[6], as illustrated in Figure 6.4.

Figure 6.4: **Three-dimensional representation**



If only the glottal source domain was considered, the representation would probably result in

[6]Illustration generated using data from the present study. Same color points indicating identical twin pairs.

a higher overlapping of individual points. That is easy to visualize if we just imagine all points in Figure 6.4 falling at the timing (speech rate) axis. It may be the case that individuals may appear much closer to each other when compared from a one-dimensional approach than when several instances are considered. That is, of course, the cost of neglecting other dimensions of analysis.

That being said, the fact that two out of 10 identical twin pairs assessed in the present study were not found statistically distinct for any of the parameters assessed here does not imply, by any means, they are phonetically identical. More likely, it may simply suggest that a simple/basic model built on one parameter is not explanatory enough of their differences. Furthermore, results could potentially be different if other assessment criteria or other ways of representing their differences were adopted, such as in the form of likelihood ratios instead of a "*match/non-match*" approach. Note that in real-world forensic conditions reporting differences in terms of likelihood ratios is the recommended practice when constructing reports, cf. Rose (2002), Morrison (2009) and Morrison, Zhang, and Enzinger (2019).

Future developments of the present research must consider the implications of different analysis approaches and the possible development of a model that may be regarded as more representative of individual patterns related to speech production. Also, understanding the level of consistency of the parameters assessed here when different speaking styles are compared, particularly those commonly present in a forensic context (e.g., interview versus spontaneous dialogue), is of crucial relevance.

Notably, the level of variation admitted by any of the aforementioned dimensions (in Figure 6.4) is assumed to be regulated by the combination of intrinsic and extrinsic forces, which are expected to reverberate on the level of separability of different acoustic-phonetic parameters. In this instance, according to the H&H theory proposed by Lindblom (1990b), speech production can be understood on the basis of an adaptive organization shaped by general biological processes. In this perspective, speakers are believed to adjust their speech performance according to communicative and situational demands, responding to the interplay between production-oriented factors and output-oriented constraints. Consequently, they are expected to vary their productions along a continuum of hyper- and hypo-speech, towards what the author calls "*sufficient discriminability*". Moreover, according to the author, these adjustments or adaptations would reflect the speaker's awareness of the listener's capability to access sources of information independent of the input (speech signal) and his judgment of the short-term demands for explicit information contained in the signal.

From this perspective, the level of variation observed in different parameters of speech production is far from a random or unregulated process; on the contrary, it seems conditioned to the implications it can bear on the communication process. Thus, the selection of potential estimates

for speaker comparison ends should also consider such intrinsic and extrinsic factors underlying the phonetic variability, and consequently, its language-dependent facet.

6.3 Revisiting some initial hypotheses

Following the specific research questions proposed and presented in *Chapter 1*, some hypotheses have been suggested. Here we draw some very brief comments on them, specifying which ones were or were not corroborated. These are presented as topics in their corresponding order of appearance:

- **Higher formant frequencies as being more speaker discriminatory than lower formant frequencies:** confirmed. However, F3 appeared to outperform F4 when assessed in unscripted speech as far as discriminatory performance is concerned.
- **Cardinal vowels displaying higher Euclidean distances from their neighbors in terms of F1 x F2 as the most speaker discriminatory:** confirmed. The central vowel [a] and front vowels were suggested, in general, as more discriminatory than back vowels. Moreover, the central and front vowels were the ones displaying the largest Euclidean distances from their neighbors.
- **Stressed vowels as being more discriminatory in intra-twin pair comparisons, and unstressed vowels as more discriminatory when considering all speakers:** partly confirmed. Stressed vowels were, in general, more discriminatory in intra-twin pair comparisons and comparisons involving all speakers.
- **Although identical twins may be similar regarding their formant frequency patterns, differences may still be observed between such individuals**: confirmed.
- **Speech timing parameters pertaining to the micro speech timing category as more discriminatory**: not confirmed. Speech timing parameters from the macro speech timing category displayed a better overall discriminatory performance.
- **Although identical twins may be remarkably similar regarding their speech timing patterns, differences may still be observed between such individuals**: inconclusive. Additional data and analyses are required for exploring such a hypothesis more in-depth.
- **The baseline value of $f0$ (i.e., $f0$ baseline) as the best discriminatory parameter among all tested $f0$ estimates:** confirmed.

- $f0$ **centrality measures outperforming** $f0$ **modulation and variation parameters concerning their discriminatory potential when assessed in the same speaking style and dialect:** confirmed.

- **Identical twins as more similar to each other regarding their** $f0$ **dimension than for any other dimension assessed in the present study**: partly confirmed. The type-/nature of the speech material used for the extraction of $f0$ estimates was suggested to influence the outcomes, i.e., whether $f0$ is assessed in stretches of connected speech or lengthened vowels. Here, twins appeared relatively more similar for the former in comparison to the latter type of speech material.

6.4 Final remarks

The present study found evidence suggesting high-formant frequencies, namely F3 and F4, as potentially more speaker discriminatory than low-formant frequencies, as verified by the proportion of significant differences across speakers as well as the comparison of effect sizes. However, between these two formants, F3 has shown to possess the desired properties expressed by the combination of lower Cllr/EER, which is compatible with higher accuracy, and high AUC values, compatible with high discriminatory power. Such an outcome appears to suggest F3 as a potential candidate for speaker comparison ends, particularly when involving unscripted speech.

Regarding the inter-speaker discriminatory potential as a function of vowel quality, evidence was found suggesting the central vowel [a] and front vowels as the most speaker-discriminatory segments. These segments also seemed to display higher Euclidean distances from their neighbors, inviting the hypothesis of a probable relationship between vowel acoustic dispersion and the level of phonetic variation allowed by the phonological system. Furthermore, even though stressed vowels appeared more speaker-discriminatory than unstressed vowels, the combination of both vowel classes seemed to be more explanatory in terms of the observed inter-speaker differences.

Concerning the analysis carried out on the speech timing dimension, evidence was found supporting the category of macro speech timing parameters, mainly speech rate and articulation rate, as the most discriminatory and consistent parameters for speaker comparison applications under unscripted speech conditions. Although very similar outcomes have been observed regarding the comparison of speech rate and articulation rate – with a slightly better performance of articulation rate, different performance outcomes were observed as to whether only silent pauses intervals or both silent and filled pauses were suppressed during the calculation of the parameter. In summary, when only silent pauses intervals were suppressed for the articulation rate estimation, a slightly

better performance was suggested, as expressed by lower EER and higher AUC values along with a higher effect size.

The analysis of speech timing estimates in identical twin pairs, who were taking part in the same dialogue, revealed a remarkable level of intra-pair similarities, substantially higher than the observed for the same speakers' formant frequency patterns. Some explanatory factors, including the overlapping effects of "*prosodic entrainment*" and "shared speech timing patterns/representations" were suggested to account for such a high convergence. Notably, additional analyses are needed.

As for the analysis performed on the $f0$ dimension, the present study allowed identifying a set of potentially relevant speaker-discriminatory $f0$ estimates for speaker comparison purposes. Overall, $f0$ baseline, median, mean, and extreme values were found to display higher proportions of intra-twin pair and cross-pair differences while also presenting the largest effect sizes and best discriminatory performance. Conversely, $f0$ variation and modulation estimates were found relatively more homogeneous across different subjects, inviting the hypothesis of relative control of speaking style and dialect on such metrics. Moreover, $f0$ estimates assessed in connected speech tended to present a better discriminatory potential than lengthened vowels.

The outcomes also suggest that, although identical twins were found very alike regarding their $f0$ patterns, some pairs could still be differentiated acoustically, mainly in connected speech. Whether such differences in $f0$ estimates solely were large enough to be perceived by external listeners is uncertain. However, such findings reinforce the relevance of analyzing long-term $f0$ metrics for forensic purposes, particularly of $f0$ baseline, which displayed the lowest EER values among all tested $f0$ estimates.

Ultimately, a continuum of inter-speaker similarity has been identified, in which some twin pairs were found more acoustic-phonetically similar than others, even when engaging in a conversation. Such a higher or lower convergence regarding identical twin pairs is expected to influence the accuracy with which these speakers can be discriminated.

Finally, the present study's findings suggest that the speaker-discriminatory potential of most acoustic-phonetic parameters is far from uniform across speakers; hence their forensic suitability should be assessed on a case-by-case basis. Further research is encouraged to assess the level of consistency regarding the patterns observed here, including the analysis of different speaking styles, e.g., interview vs. spontaneous dialogue, and non-contemporaneous recordings. Non-identical twin pairs should also be included in future developments of the present research.

References

'T HART, Johan. Differential sensitivity to pitch distance, particularly in speech. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 69, n. 3, p. 811–821, 1981. DOI: https://doi.org/10.1121/1.385592.

AITKEN, Colin GG; LUCY, David. Evaluation of trace evidence in the form of multivariate data. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 53, n. 1, p. 109–122, 2004. Available from: <https://www.jstor.org/stable/3592690>.

ALBERTS, Bruce et al. Molecular biology of the cell. Garland Science, Taylor and Francis Group, 2018.

ARANTES, Pablo; BARBOSA, Plinio A. Production–perception entrainment in speech rhythm. In: SPEECH Prosody 2010-Fifth International Conference. [S.l.: s.n.], 2010. Available from: <https://www.isca-speech.org/archive/sp2010/sp10_221.html>.

ARANTES, Pablo; ERIKSSON, Anders. Quantifying Fundamental Frequency Modulation as a Function of Language, Speaking Style and Speaker. In: INTERSPEECH. [S.l.: s.n.], 2019. p. 1716–1720. DOI: http://dx.doi.org/10.21437/Interspeech.2019-2857.

_____. Temporal stability of long-term measures of fundamental frequency. In: SPEECH Prosody, 2014. [S.l.: s.n.], 2014. p. 1149–1152. Available from: <https://www.isca-speech.org/archive/SpeechProsody_2014/pdfs/224.pdf>.

ARANTES, Pablo; ERIKSSON, Anders; LIMA, Verônica G. Minimum Sample Length for the Estimation of Long-term Speaking Rate. In: PROC. 9th International Conference on Speech Prosody 2018. [S.l.: s.n.], 2018. p. 661–665. DOI: 10.21437/SpeechProsody.2018-134.

ARCURI, Cláudia Fassin et al. Taxa de elocução de fala segundo a gravidade da gagueira. **Pró-Fono Revista de Atualização Cientıfica**, Fró-fono, v. 21, n. 1, p. 45–50, 2009. DOI: https://doi.org/10.1590/S0104-56872009000100008.

ASSANEO, M Florencia; POEPPEL, David. The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. **Science advances**, American Association for the Advancement of Science, v. 4, n. 2, eaao3842, 2018. DOI: 10.1126/sciadv.aao3842.

AWAN, Shaheen N; MUELLER, Peter B. Speaking fundamental frequency characteristics of white, African American, and Hispanic kindergartners. **Journal of Speech, Language, and Hearing Research**, ASHA, v. 39, n. 3, p. 573–577, 1996. DOI: https://doi.org/10.1044/jshr.3903.573.

AYOUB, Marie Reine; LARROUY-MAESTRI, Pauline; MORSOMME, Dominique. The effect of smoking on the fundamental frequency of the speaking voice. **Journal of Voice**, Elsevier, v. 33, n. 5, 802–e11, 2019. DOI: https://doi.org/10.1016/j.jvoice.2018.04.001.

BARBOSA, Plinio A. **AcousticParametersforVowelsExtractor [Praat script]**. 2015. Available from: <https://github.com/pabarbosa/prosody-scripts/blob/master/AcousticParametersforVowelsExtractor/AcousticParametersforVowelsExtractor.psc>.

_____. Do grau de não perifericidade da vogal/a/pós-tônica final. **Revista Diadorim**, v. 12, 2012. DOI: https://doi.org/10.35520/diadorim.2012.v12n0a3973.

_____. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. In: SPEECH Prosody 2002, International Conference. [S.l.: s.n.], 2002.

_____. From syntax to acoustic duration: A dynamical model of speech rhythm production. **Speech Communication**, Elsevier, v. 49, n. 9, p. 725–742, 2007. DOI: https://doi.org/10.1016/j.specom.2007.04.013.

_____. **Incursões em torno do ritmo da fala**. [S.l.]: Pontes, 2006.

_____. **ProsodyDescriptorExtractor [Praat script]**. 2020. Available from: <https://github.com/pabarbosa/prosody-scripts/tree/master/ProsodyDescriptorExtractor>.

BARBOSA, Plinio A; ERIKSSON, Anders; ÅKESSON, Joel. Cross-linguistic similarities and differences of lexical stress realisation in Swedish and Brazilian Portuguese. In: NORDIC Prosody. Proceedings of the XIth conference. Frankfurt am Main: Peter Lang, Tartu. [S.l.: s.n.], 2013. p. 97–106. DOI: https://doi.org/10.3726/978-3-653-03047-1.

BECK, Janet Mackenzie. Organic variation of the vocal apparatus. **The handbook of phonetic sciences**, Blackwell Publishers, Oxford, p. 256–297, 1997.

BEHLAU, Mara et al. Spectrographic analysis of vowels formants in Brazilian Portuguese. **ACTA AWHO**, v. 7, p. 74–85, Jan. 1988.

BERBEN, Lut; SEREIKA, Susan M; ENGBERG, Sandra. Effect size estimation: methods and examples. **International journal of nursing studies**, Elsevier, v. 49, n. 8, p. 1039–1047, 2012. DOI: https://doi.org/10.1016/j.ijnurstu.2012.01.015.

BOERSMA, Paul; WEENINK, David. Praat: doing phonetics by computer [Computer program]. Version 6.0. 37. v. 14, 2018. Available from: <http://www.%20praat.%20org/>.

BONA, Judit. Temporal characteristics of speech: The effect of age and speech style. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 136, n. 2, p. 116–121, 2014. DOI: https://doi.org/10.1121/1.4885482.

BOSKER, Hans Rutger. How our own speech rate influences our perception of others. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, American Psychological Association, v. 43, n. 8, p. 1225, 2017. DOI: 10.1037/xlm0000381.

BRAUN, Angelika. Fundamental frequency: how speaker-specific is it? In: BEITRÄGE zur Phonetik und Linguistik. [S.l.: s.n.], 1995. v. 64. p. 9–23.

BRÜMMER, Niko; DU PREEZ, Johan. Application-independent evaluation of speaker detection. **Computer Speech & Language**, Elsevier, v. 20, n. 2-3, p. 230–275, 2006. DOI: https://doi.org/10.1016/j.csl.2005.08.001.

BUDER, Eugene H; ERIKSSON, Anders. Prosodic cycles and interpersonal synchrony in American English and Swedish. In: FIFTH European Conference on Speech Communication and Technology. [S.l.: s.n.], 1997.

BURZIO, Luigi. Phonology and phonetics of English stress and vowel reduction. **Language Sciences**, Elsevier, v. 29, n. 2-3, p. 154–176, 2007. DOI: https://doi.org/10.1016/j.langsci.2006.12.019.

BYRNE, Catherine; FOULKES, Paul. The'mobile phone effect'on vowel formants. **International Journal of Speech Language and the Law**, v. 11, n. 1, p. 83–102, 2007.

CAO, Honglin; DELLWO, Volker. The role of the first five formants in three vowels of mandarin for forensic voice analysis. University of Zurich, p. 617–621, 2019. DOI: https://doi.org/10.5167/uzh-177494.

CARSON, Cecyle Perry; INGRISANO, Dennis R-S; EGGLESTON, K Donald. The effect of noise on computer-aided measures of voice: a comparison of CSpeechSP and the Multi-Dimensional Voice Program software using the CSL 4300B Module and Multi-Speech for Windows. **Journal of Voice**, Elsevier, v. 17, n. 1, p. 12–20, 2003. DOI: https://doi.org/10.1016/s0892-1997(03)00031-6.

CARTWRIGHT, Lynn R; LASS, Norman J. A psychophysical study of rate of continuous speech stimuli by means of direct magnitude estimation scaling. **Language and Speech**, Sage Publications Sage CA: Thousand Oaks, CA, v. 18, n. 4, p. 358–365, 1975. DOI: https://doi.org/10.1177/002383097501800406.

CAVALCANTI, Julio Cesar; ENGLERT, Marina, et al. Microphone and Audio Compression Effects on Acoustic Voice Analysis: A Pilot Study. **Journal of Voice**, Elsevier, 2021. DOI: https://doi.org/10.1016/j.jvoice.2020.12.005.

CAVALCANTI, Julio Cesar; ERIKSSON, Anders; BARBOSA, Plinio A. Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. **Plos one**, Public Library of Science San Francisco, CA USA, v. 16, n. 2, e0246645, 2021. DOI: https://doi.org/10.1371/journal.pone.0246645.

_____. Multi-parametric analysis of speaking fundamental frequency in genetically related speakers using different speech materials: some forensic implications. **Journal of Voice**, Elsevier, 2021. DOI: https://doi.org/10.1016/j.jvoice.2021.08.013.

CIELO, Carla Aparecida; AGUSTINI, Rosane; FINGER, Leila Susana. Caracterısticas vocais de gêmeos monozigóticos. **Revista CEFAC**, SciELO Brasil, v. 14, n. 6, p. 1234–1241, 2012. DOI: https://doi.org/10.1590/S1516-18462010005000003.

COHEN PRIVA, Uriel; EDELIST, Lee; GLEASON, Emily. Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 141, n. 5, p. 2989–2996, 2017. DOI: https://doi.org/10.1121/1.4982199.

CONRAD, Eric; MISENAR, Seth; FELDMAN, Joshua. **CISSP study guide**. [S.l.]: Newnes, 2012.

CONSTANTINI, Ana Carolina. Caracterização prosódica de sujeitos de diferentes variedades de fala do português brasileiro em diferentes relações sinal-ruıdo. [Doctoral thesis]. [Universidade Estadual de Campinas], p. 1–115, 2014. Available from: <http://repositorio.unicamp.br/jspui/handle/REPOSIP/271138>.

CUMMINS, Fred. Reducing expressive variation in speech with synchronous speech. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 109, n. 5, p. 2416–2417, 2001. DOI: https://doi.org/10.1121/1.4744550.

_____. Rhythm as entrainment: The case of synchronous speech. **Journal of Phonetics**, Elsevier, v. 37, n. 1, p. 16–28, 2009. DOI: 10.1016/j.wocn.2008.08.003.

CUMMINS, Fred. Speech synchronization: Investigating the links between perception and action in speech production. In: INTERNATIONAL Congress of the Phonetic Sciences, Saarbrücken. [S.l.: s.n.], 2007. p. 529–532. Available from: <http://cspeech.ucd.ie/Fred/docs/cumminsICPhS07.pdf>.

_____. Synchronization among speakers reduces macroscopic temporal variability. In: 26. PROCEEDINGS of the Annual Meeting of the Cognitive Science Society. [S.l.: s.n.], 2004. v. 26. Available from: <http://cspeech.ucd.ie/Fred/docs/cummins-cogsci04.pdf>.

CUMMINS, Fred; LI, Chenxia; WANG, Bei. Coupling among speakers during synchronous speaking in English and Mandarin. **Journal of Phonetics**, Elsevier, v. 41, n. 6, p. 432–441, 2013. DOI: https://doi.org/10.1016/j.wocn.2013.07.001.

DEBRUYNE, Frans et al. Speaking fundamental frequency in monozygotic and dizygotic twins. **Journal of Voice**, Elsevier, v. 16, n. 4, p. 466–471, 2002. DOI: https://doi.org/10.1016/S0892-1997(02)00121-2.

DILLEY, Laura C.; BROWN, Meredith. Effects of pitch range variation on f0 extrema in an imitation task. **Journal of Phonetics**, v. 35, n. 4, p. 523–551, 2007. ISSN 00954470. DOI: 10.1016/j.wocn.2007.01.003.

DING, Nai et al. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). **Frontiers in human neuroscience**, Frontiers, v. 11, p. 481, 2017. DOI: https://doi.org/10.3389/fnhum.2017.00481.

DJORDJEVIC, Jelena et al. Genetic and environmental contributions to facial morphological variation: a 3D population-based twin study. **PloS one**, Public Library of Science San Francisco, CA USA, v. 11, n. 9, e0162250, 2016. DOI: https://doi.org/10.1371/journal.pone.0162250.

DOELLING, Keith B; POEPPEL, David. Cortical entrainment to music and its modulation by expertise. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 112, n. 45, e6233–e6242, 2015. DOI: https://doi.org/10.1073/pnas.1508431112.

ERIKSSON, Anders; WRETLING, Pär. How flexible is the human voice? A case study of mimicry. In: FIFTH European Conference on Speech Communication and Technology. [S.l.: s.n.], 1997.

ESCUDERO, Paola et al. A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. **The Journal of the Acoustical Society of America**, v. 126, n. 3, p. 1379–1393, 2009. ISSN 0001-4966. DOI: 10.1121/1.3180321.

FANT, Gunnar; KRUCKENBERG, Anita; FERREIRA, Joana Barbosa. Individual variations in pausing. A study of read speech. In: PROCEEDINGS of Fonetik. [S.l.: s.n.], 2003. p. 193–196.

FAWCETT, Tom. An introduction to ROC analysis. **Pattern Recognition Letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006. DOI: https://doi.org/10.1016/j.patrec.2005.10.010.

FERNÁNDEZ, Eugenia San Segundo. Glottal source parameters for forensic voice comparison: An approach to voice quality in twins' voices. In: INTERNATIONAL Association for Forensic Phonetics and Acoustics Annual Conference. [S.l.: s.n.], 2012.

FIGUEIREDO, Ricardo Molina de. **Identificação de falantes: aspectos teóricos e metodológicos**. [S.l.]: Universidade Estadual de Campinas, 1994. Available from: <http://www.repositorio.unicamp.br/handle/REPOSIP/270642>.

FOURAKIS, Marios. Tempo, stress, and vowel reduction in American English. **The Journal of the Acoustical society of America**, Acoustical Society of America, v. 90, n. 4, p. 1816–1827, 1991. DOI: doi:10.1121/1.401662.

FRIEDOVA, Lucie et al. Slowed articulation rate is associated with information processing speed decline in multiple sclerosis: A pilot study. **Journal of Clinical Neuroscience**, Elsevier Ltd, v. 65, p. 28–33, 2019. ISSN 15322653. DOI: 10.1016/j.jocn.2019.04.018.

FRITZ, Catherine O; MORRIS, Peter E; RICHLER, Jennifer J. Effect size estimates: current use, calculations, and interpretation. **Journal of Experimental Psychology: General**, American Psychological Association, v. 141, n. 1, p. 2, 2012. DOI: https://doi.apa.org/doi/10.1037/a0024338.

FUCHS, Robert; MAXWELL, Olga. The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range. In: ISCA. PROCEEDINGS of Speech Prosody. [S.l.: s.n.], 2016. p. 523–527. Available from: <http://dx.doi.org/10.21437/SpeechProsody.2016-107>.

GARVIN, Paul L; LADEFOGED, Peter. Speaker identification and message identification in speech recognition. **Phonetica**, Karger Publishers, v. 9, n. 4, p. 193–199, 1963. DOI: https://doi.org/10.1159/000258404.

GOLD, Erica; FRENCH, Peter. International practices in forensic speaker comparison. **International Journal of Speech, Language and the Law**, Equinox Publishing Ltd, v. 18, n. 2, p. 293–307, 2011. DOI: https://doi.org/10.1558/ijsll.v18i2.293.

GONÇALVES, Cintia Schivinscki. Taxa de elocução e taxa de articulação em corpus utilizado na perıcia de Comparação de Locutores. **Letras de Hoje**, SciELO Brasil, v. 52, n. 1, p. 15–25, 2017. DOI: https://doi.org/10.15448/1984-7726.2017.1.25540.

GONZALEZ, Julio; CARPI, Amparo. Early effects of smoking on the voice: a multidimensional study. **Medical Science Monitor**, International Scientific Information, Inc., v. 10, n. 12, cr649–cr656, 2004.

GORDON, Matthew; ROETTGER, Timo. Acoustic correlates of word stress: A cross-linguistic survey. **Linguistics Vanguard**, v. 3, n. 1, 2017. ISSN 2199174X. DOI: 10.1515/lingvan-2017-0007.

HAND, David J; TILL, Robert J. A simple generalisation of the area under the ROC curve for multiple class classification problems. **Machine Learning**, Springer, v. 45, n. 2, p. 171–186, 2001. DOI: https://doi.org/10.1023/A\%3A1010920819831.

HARDING, Eleanor E et al. Cortical tracking of rhythm in music and speech. **NeuroImage**, Elsevier, v. 185, p. 96–101, 2019. DOI: https://doi.org/10.1016/j.neuroimage.2018.10.037.

HARMEGNIES, Bernard; POCH-OLIVÉ, Dolors. Formants frequencies variability in French vowels under the effect of various speaking styles. **Le Journal de Physique IV**, EDP sciences, v. 4, n. C5, p. c5–509, 1994. DOI: 0.1051/jp4:19945108.

HEUVEL, Hendrik. **Speaker variability in acoustic properties of Dutch phoneme realisations [Doctoral thesis]**. [S.l.]: [Radboud Universiteit], 1996. Available from: <https://repository.ubn.ru.nl/handle/2066/76416>.

HIGUCHI, Norio; HIRAI, Toshio; SAGISAKA, Yoshinori. Effect of speaking style on parameters of fundamental frequency contour. Springer, p. 417–428, 1997. DOI: https://doi.org/10.1007/978-1-4612-1894-4_33.

HOSOI, HIROSHI et al. Effect of the rate of speech flow on speech intelligibility in normal and hearing-impaired subjects. **Nippon Jibiinkoka Gakkai Kaiho**, The Oto-Rhino-Laryngological Society of Japan, Inc., v. 95, n. 4, p. 517–525, 1992. DOI: 10.3950/jibiinkoka.95.517.

HUGHES, Vincent; BRERETON, Ashley; GOLD, Erica. Reference sample size and the computation of numerical likelihood ratios using articulation rate. **York Papers in Linguistics**, University of York, v. 13, p. 22–46, 2013. ISSN 0307-3238.

IGRAS-CYBULSKA, Magdalena et al. Structure of pauses in speech in the context of speaker verification and classification of speech type. **EURASIP Journal on Audio, Speech, and Music Processing**, Springer, v. 2016, n. 1, p. 18, 2016. DOI: 10.1186/s13636-016-0096-7.

JACEWICZ, Ewa; FOX, Robert Allen; WEI, Lai. Between-speaker and within-speaker variation in speech tempo of American English. **The Journal of the Acoustical Society of America**, v. 128, n. 2, p. 839–850, 2010. ISSN 0001-4966. DOI: 10.1121/1.3459842.

JANNETTS, Stephen et al. Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types. **International Journal of Language & Communication Disorders**, Wiley Online Library, v. 54, n. 2, p. 292–305, 2019. DOI: https://doi.org/10.1111/1460-6984.12457.

JESSEN, Michael. Forensic phonetics and the influence of speaking style on global measures of fundamental frequency. **Formal Linguistics and Law. Berlin: Mouton de Gruyter**, p. 115–139, 2009. DOI: https://doi.org/10.1515/9783110218398.2.115.

_____. Forensic reference data on articulation rate in German. **Science and Justice**, v. 47, n. 2, p. 50–67, 2007. ISSN 13550306. DOI: 10.1016/j.scijus.2007.03.003.

JONG, Gea de et al. The telephone effect on F0. In: IAFPA 2011 conference, Vienna, Austria. [S.l.: s.n.], 2011. v. 11.

KASSAMBARA, Alboukadel. Rstatix: pipe-friendly framework for basic statistical tests. **R package version 0.6. 0**, 2020.

KENT, Raymond D.; VORPERIAN, Houri K. Static measurements of vowel formant frequencies and bandwidths: A review. **Journal of Communication Disorders**, Elsevier, v. 74, June, p. 74–97, 2018. ISSN 18737994. DOI: 10.1016/j.jcomdis.2018.05.004. Available from: <https://doi.org/10.1016/j.jcomdis.2018.05.004>.

KLOFSTAD, Casey A; ANDERSON, Rindy C; NOWICKI, Stephen. Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. **PloS one**, Public Library of Science, v. 10, n. 8, e0133779, 2015. DOI: https://doi.org/10.1371/journal.pone.0133779.

KOREMAN, Jacques. Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. **The Journal of the Acoustical Society of America**, v. 119, n. 1, p. 582–596, 2006. ISSN 0001-4966. DOI: 10.1121/1.2133436.

KORF, Bruce R. Basic genetics. **Primary Care: Clinics in Office Practice**, Elsevier, v. 31, n. 3, p. 461–478, 2004.

KRAUSE, Jean C; BRAIDA, Louis D. Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 112, n. 5, p. 2165–2172, 2002. DOI: https://doi.org/10.1121/1.1509432.

KRIVOKAPIĆ, Jelena. Prosodic planning: Effects of phrasal length and complexity on pause duration. **Journal of phonetics**, Elsevier, v. 35, n. 2, p. 162–179, 2007. DOI: 10.1016/j.wocn.2006.04.001.

KÜNZEL, Hermann J. Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. **Forensic Linguistics**, Citeseer, v. 8, n. 1, p. 80–99, 2001. DOI: 10.1558/ijsll.v8i1.80.

_____. Some general phonetic and forensic aspects of speaking tempo. **International Journal of Speech language and the Law**, v. 4, n. 1, p. 48–83, 1997. DOI: 10.1558/ijsll.v4i1.48.

LABOV, William. **Principles of linguistic change. Volume 3: Cognitive and cultural factors**. [S.l.]: John Wiley & Sons, 2011. v. 36, p. 1–448. ISBN 978-1-444-35146-0.

LADEFOGED, Peter. **Elements of acoustic phonetics**. [S.l.]: University of Chicago Press, 1996.

LADEFOGED, Peter; BROADBENT, Donald Eric. Information conveyed by vowels. **The Journal of the acoustical society of America**, Acoustical Society of America, v. 29, n. 1, p. 98–104, 1957. DOI: https://doi.org/10.1121/1.1908694.

LANE, Harlan; GROSJEAN, François. Perception of reading rate by speakers and listeners. **Journal of Experimental Psychology**, American Psychological Association, v. 97, n. 2, p. 141, 1973. DOI: 10.1037/h0033869.

LENNON, Robert; PLUG, Leendert; GOLD, Erica. A Comparison of Multiple Speech Tempo Measures: Inter-Correlations and Discriminating Power. In: AUSTRALASIAN SPEECH SCIENCE and TECHNOLOGY ASSOCIATION INC. 19TH International Congress of the Phonetic Sciences. [S.l.: s.n.], 2019. p. 785–789. Available from: <https://assta.org/proceedings/ICPhS2019/papers/ICPhS_834.pdf>.

LEWIS, Barbara A; THOMPSON, Lee A. A study of developmental speech and language disorders in twins. **Journal of Speech, Language, and Hearing Research**, ASHA, v. 35, n. 5, p. 1086–1094, 1992. DOI: 10.1044/jshr.3505.1086.

LINDBLOM, Björn. Explaining Phonetic Variation: A Sketch of the H&H Theory. In: SPEECH Production and Speech Modelling. [S.l.]: Springer Netherlands, 1990. p. 403–439. DOI: 10.1007/978-94-009-2037-8_16.

_____. Explaining phonetic variation: A sketch of the H&H theory. In: SPEECH production and speech modelling. [S.l.]: Springer, 1990. p. 403–439. DOI: 10.1007/978-94-009-2037-8_16.

_____. Models of phonetic variation and selection. **Phonetic Experimental Research, Institute of Linguistics, University of Stockholm**, v. 11, p. 65–100, 1990. ISSN 0282-6.

LINDH, Jonas; ERIKSSON, Anders. Robustness of long time measures of fundamental frequency. In: INTERSPEECH. [S.l.: s.n.], 2007. p. 2025–2028. Available from: <https://www.isca-speech.org/archive/archive_papers/interspeech_2007/i07_2025.pdf>.

LISKER, Leigh. On" explaining" vowel duration variation. In: ERIC. PAPER presented at the winter meeting of the Linguistic Society of America. [S.l.: s.n.], 1974. v. 28, p. 225. Available from: <http://www.haskins.yale.edu/sr/SR037/SR037_21.pdf>.

LO, Justing. fvclrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison [R package], version 1.1.1, 2020. Available from: <https://rdrr.io/github/justinjhlo/fvclrr/>.

LOAKES, Deborah. A forensic phonetic investigation into the speech patterns of identical and non-identical twins. In: 15TH International Congress of Phonetic Sciences (ICPhS-15). [S.l.: s.n.], 2003. v. 15, p. 691–694. Available from: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_0691.html>.

_____._____. **International Journal of Speech, Language and the Law**, v. 15, n. 1, p. 97–100, 2008. DOI: https://doi.org/10.1558/ijsll.v15i1.97.

_____. Front Vowels as Speaker-Specific : Some Evidence from Australian English. In: PROCEEDINGS of the 10th Australian International Conference on Speech Science Technology. [S.l.: s.n.], 2004. p. 289–294. Available from: <https://assta.org/proceedings/sst/2004/proceedings/papers/sst2004-375.pdf>.

_____. Variation in long-term fundamental frequency: measurements from vocalic segments in twins' speech. In: UNIVERSITY OF AUCKLAND, NEW ZEALAND. PROCEEDINGS of the 11th Australian International Conference on Speech Science Technology. [S.l.: s.n.], 2006. p. 205–210. Available from: <https://assta.org/proceedings/sst/2006/sst2006-107.pdf>.

LOAKES, Deborah; MCDOUGALL, Kirsty. Individual variation in the frication of voiceless plosives in Australian English: A study of twins' speech. **Australian Journal of Linguistics**, v. 30, n. 2, p. 155–181, 2010. ISSN 07268602. DOI: 10.1080/07268601003678601.

LUO, Huan; POEPPEL, David. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. **Neuron**, Elsevier, v. 54, n. 6, p. 1001–1010, 2007. DOI: 10.1016/j.neuron.2007.06.004.

MAGGIONI, Eleonora et al. Twin MRI studies on genetic and environmental determinants of brain morphology and function in the early lifespan. **Neuroscience & Biobehavioral Reviews**, Elsevier, v. 109, p. 139–149, 2020. DOI: https://doi.org/10.1016/j.neubiorev.2020.01.003.

MANUEL, Sharon Y; KRAKOW, Rena A. Universal and language particular aspects of vowel-to-vowel coarticulation. **Haskins Laboratories Status Report on Speech Research**, ERIC, v. 77, n. 78, p. 69–78, 1984. Available from: <http://www.haskins.yale.edu/SR/SR077/SR077_06.pdf>.

MARYN, Youri et al. Mobile communication devices, ambient noise, and acoustic voice measures. **Journal of Voice**, Elsevier, v. 31, n. 2, 248–e11, 2017. DOI: https://doi.org/10.1016/j.jvoice.2016.07.023.

MOON, Kyung Ray et al. Materials of acoustic analysis: sustained vowel versus sentence. **Journal of Voice**, Elsevier, v. 26, n. 5, p. 563–565, 2012. DOI: https://doi.org/10.1016/j.jvoice.2011.09.007.

MORRISON, Geoffrey Stewart. Forensic voice comparison and the paradigm shift. **Science & Justice**, Elsevier, v. 49, n. 4, p. 298–308, 2009. DOI: https://doi.org/10.1016/j.scijus.2009.09.002.

MORRISON, Geoffrey Stewart; ZHANG, Cuiling; ENZINGER, Ewald. Forensic speech science. **Forensic Science International**, Thomson Reuters, 2019. DOI: https://doi.org/10.1016/j.forsciint.2010.11.001.

MORRISON, Geoffrey Stewart; ZHANG, Cuiling; ROSE, Philip. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. **Forensic science international**, Elsevier, v. 208, n. 1-3, p. 59–65, 2011. DOI: 10.1016/j.forsciint.2010.11.001.

MURPHY, Christopher H; DOYLE, Philip C. The effects of cigarette smoking on voice-fundamental frequency. **Otolaryngology—Head and Neck Surgery**, SAGE Publications Sage CA: Los Angeles, CA, v. 97, n. 4, p. 376–380, 1987. DOI: https://doi.org/10.1177\%2F019459988709700406.

NESPOR, Marina; VOGEL, Irene. **Prosodic phonology: with a new foreword**. [S.l.]: Walter de Gruyter, 2007. v. 28.

NOLAN, FJD. **The phonetic bases of speaker recognition**. [S.l.]: Cambridge University Press, 1983. ISBN 0521244862.

NOLAN, Francis; OH, Tomasina. Identical twins, different voices. **International Journal of Speech, Language and the Law**, v. 3, n. 1, p. 39–49, 1996. DOI: https://doi.org/10.1558/ijsll.v3i1.39.

ÖHMAN, Sven EG. Coarticulation in VCV utterances: Spectrographic measurements. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 39, n. 1, p. 151–168, 1966. DOI: https://doi.org/10.1121/1.1909864.

OLIVEIRA JR, Miguel. A study on speech rate as a prosodic feature in spontaneous narrative. **Alfa: Revista de Linguıstica (São José do Rio Preto)**, SciELO Brasil, v. 56, n. 2, p. 623–651, 2012. DOI: https://doi.org/10.1590/S1981-57942012000200012.

PASSETTI, Renata Regina. Estudo acústico-perceptual do estilo de fala telefônico com implicações para a verificação de locutor em português brasileiro [Doctoral thesis]. Universidade Estadual de Campinas, 2018. Available from: <http://repositorio.unicamp.br/jspui/handle/REPOSIP/332058>.

_____. **O efeito do telefone celular no sinal da fala: uma análise fonético-acústica com implicações para a verificação de locutor em português brasileiro**. [S.l.]: Universidade Estadual de Campinas, 2015. Available from: <http://repositorio.unicamp.br/bitstream/REPOSIP/271133/1/Passetti_RenataRegina_M.pdf>.

POEPPEL, David; ASSANEO, M Florencia. Speech rhythms and their neural foundations. **Nature Reviews Neuroscience**, Nature Publishing Group, p. 1–13, 2020. DOI: https://doi.org/10.1038/s41583-020-0304-4.

POMPINO-MARSCHALL, Bernd. On the psychoacoustic nature of the P-center phenomenon. **Journal of phonetics**, Elsevier, v. 17, n. 3, p. 175–192, 1989. DOI: https://doi.org/10.1016/S0095-4470(19)30428-0.

PROBST, Louise; BRAUN, Angelika. The effects of emotional state on fundamental frequency. In: PROCEEDINGS of the 19th International Congress of Phonetic Sciences, Melbourne, Australia. [S.l.: s.n.], 2019. Available from: <https://icphs2019.org/icphs2019-fullpapers/pdf/full-paper_599.pdf>.

PRZYBYLA, Beata D; HORII, Yoshiyuki; CRAWFORD, Michael H. Vocal fundamental frequency in a twin sample: Looking for a genetic effect. **Journal of Voice**, Elsevier, v. 6, n. 3, p. 261–266, 1992. DOI: https://doi.org/10.1016/S0892-1997(05)80151-1.

QUENÉ, Hugo. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 123, n. 2, p. 1104–1113, 2008. DOI: https://doi.org/10.1121/1.2821762.

_____. On the just noticeable difference for tempo in speech. **Journal of Phonetics**, Elsevier, v. 35, n. 3, p. 353–362, 2007. DOI: https://doi.org/10.1016/j.wocn.2006.09.001.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Available from: <https://www.R-project.org/>.

RECASENS, Daniel; ESPINOSA, Aina. Dispersion and variability of Catalan vowels. **Speech communication**, Elsevier, v. 48, n. 6, p. 645–666, 2006. DOI: https://doi.org/10.1016/j.specom.2005.09.011.

RICE, Mabel L. Causal pathways for specific language impairment: Lessons from studies of twins. **Journal of Speech, Language, and Hearing Research**, ASHA, v. 63, n. 10, p. 3224–3235, 2020. DOI: https://doi.org/10.1044/2020_JSLHR-20-00169.

RILLIARD, Albert et al. Social affect production and perception across languages and cultures–the role of prosody. **Leitura**, v. 2, n. 52, p. 15–41, 2013. DOI: http://www.dx.doi.org/10.28998/2317-9945.2013v2n52p15-41.

ROBIN, Xavier et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. **BMC bioinformatics**, BioMed Central, v. 12, n. 1, p. 1–8, 2011.

ROSE, Philip. **Forensic Speaker Identification**. [S.l.]: Tayler Francis, 2002. v. 3, p. 54–67. ISBN 0415271827. Available from: <http://repositorio.unan.edu.ni/2986/1/5624.pdf>.

SAN SEGUNDO, Eugenia. Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics [Doctoral thesis]. Universidad internacional Menéndez Pelayo, p. 1–318, 2014.

SAN SEGUNDO, Eugenia; GÓMEZ VILDA, Pedro. Matching twin and non-twin siblings from phonation characteristics. **VII Jornadas de Reconocimiento Biométrico de Personas**, Editorial Universidad de Salamanca Zamora, p. 10–17, 2013.

SAN SEGUNDO, Eugenia; TSANAS, Athanasios; GÓMEZ-VILDA, Pedro. Euclidean distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics. **Forensic Science International**, Elsevier, v. 270, p. 25–38, 2017. DOI: https://doi.org/10.1016/j.forsciint.2016.11.020.

SAN SEGUNDO, Eugenia; YANG, Junjie. Formant dynamics of Spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation. **Journal of Phonetics**, The Authors, v. 75, p. 1–26, 2019. ISSN 00954470. DOI: 10.1016/j.wocn.2019.04.001. Available from: <https://doi.org/10.1016/j.wocn.2019.04.001>.

SANTIAGO, Fabian; MAIRANO, Paolo. The role of lexical stress on vowel duration and vowel space in two varieties of Spanish. In: PROC. 9th International Conference on Speech Prosody 2018. [S.l.: s.n.], 2018. p. 453–457. DOI: 10.21437/SpeechProsody.2018-92.

SCHWARTZ, Jean-Luc et al. Major trends in vowel system inventories. **Journal of Phonetics**, v. 25, n. 3, p. 233–253, 1997. ISSN 00954470. DOI: 10.1006/jpho.1997.0044.

SCHWARTZ, Jean-Luc et al. The dispersion-focalization theory of vowel systems. **Journal of phonetics**, Elsevier, v. 25, n. 3, p. 255–286, 1997. DOI: https://doi.org/10.1006/jpho.1997.0043.

SHAFFER, Henry. Rhythm and timing in skill. **Psychological Review**, American Psychological Association, v. 89, n. 2, p. 109, 1982. DOI: https://doi.org/10.1037/0033-295X.89.2.109.

_____. Timing in musical performance. **Annals of the New York Academy of Sciences**, v. 423, p. 420, 1984. DOI: https://doi.org/10.1111/j.1749-6632.1984.tb23450.x.

SHATTUCK-HUFNAGEL, Stefanie; TURK, Alice. The domain of phrase-final lengthening in English. In: THE Sound of the Future: A Global View of Acoustics in the 21st Century, Proceedings of the 16th International Congress on Acoustics and 135th Meeting Acoustical Society of America. [S.l.: s.n.], 1998. p. 1235–1236. DOI: https://doi.org/10.1121/1.421798.

SIEGMAN, ARON WOLFE. The meaning of silent pauses in the initial interview. **The Journal of nervous and mental disease**, v. 166, n. 9, p. 642–654, 1978. DOI: 10.1097/00005053-197809000-00004.

SIGNORELLO, Rosario et al. Vocal Fundamental Frequency and Sound Pressure Level in Charismatic Speech: A Cross-Gender and-Language Study. **Journal of Voice**, Elsevier, v. 34, n. 5, 808–e1, 2020. DOI: http://dx.doi.org/10.1016/j.jvoice.2019.04.007.

SILBER-VAROD, Vered et al. The influence of lexical stress on formant values in spontaneous Hebrew speech. The 19th International Congress of Phonetic Sciences, 2019. Available from: <https://icphs2019.org/icphs2019-fullpapers/pdf/full-paper_154.pdf>.

SILVA, Ronaldo R da et al. Applying base value of fundamental frequency via the multivariate Kernel-Density in Forensic Speaker Comparison. In: IEEE. 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS). [S.l.: s.n.], 2016. p. 1–8. DOI: https://doi.org/10.1109/ICSPCS.2016.7843307.

SMITH, Caroline L. Topic transitions and durational prosody in reading aloud: production and modeling. **Speech Communication**, Elsevier, v. 42, n. 3-4, p. 247–270, 2004. DOI: https://doi.org/10.1016/j.specom.2003.09.004.

SOTOME, Taisuke et al. Analysis of Speech Fundamental Frequencies for Different Tasks in Japanese. **Journal of Voice**, Elsevier, 2021. DOI: https://doi.org/10.1016/j.jvoice.2020.12.021.

STEVENS, Kenneth N; HOUSE, Arthur S. Development of a quantitative description of vowel articulation. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 27, n. 3, p. 484–493, 1955. DOI: https://doi.org/10.1121/1.1907943.

SUNDBERG, Johan. **Ciência da voz: fatos sobre a voz na fala e no canto**. [S.l.]: Editora da Universidade de São Paulo, 2015. ISBN 13: 9788531415104.

SUSSMAN, Harvey M. Acoustic correlates of the front/back vowel distinction: a comparison of transition onset versus "steady state". **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 88, n. 1, p. 87–96, 1990. DOI: https://doi.org/10.1121/1.399848.

TAKEMOTO, Hironori; ADACHI, Seiji, et al. Acoustic roles of the laryngeal cavity in vocal tract resonance. **The Journal of the Acoustical Society of America**, v. 120, n. 4, p. 2228–2238, 2006. ISSN 0001-4966. DOI: 10.1121/1.2261270.

TAKEMOTO, Hironori; MOKHTARI, Parham; KITAMURA, Tatsuya. Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method. **The Journal of the Acoustical Society of America**, v. 128, n. 6, p. 3724–3738, 2010. ISSN 0001-4966. DOI: 10.1121/1.3502470.

THOMPSON, Paul M et al. Genetic influences on brain structure. **Nature neuroscience**, Nature Publishing Group, v. 4, n. 12, p. 1253–1258, 2001. DOI: https://doi.org/10.1038/nn758.

TOMASELLO, Michael. Culture and cognitive development. **Current Directions in Psychological Science**, SAGE Publications Sage CA: Los Angeles, CA, v. 9, n. 2, p. 37–40, 2000. DOI: https://doi.org/10.1111/1467-8721.00056.

TOMCZAK, Maciej; TOMCZAK, Ewa. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. **Trends in Sport Sciences**, Akademia Wychowania Fizycznego w Poznaniu, p. 19–25, 2014. Available from: <http://tss.awf.poznan.pl/files/3_Trends_Vol21_2014__no1_20.pdf>.

TRAUNMÜLLER, Hartmut. Analytical expressions for the tonotopic sensory scale. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 88, n. 1, p. 97–100, 1990. DOI: https://doi.org/10.1121/1.399849.

_____. Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. **Speech Communication**, v. 3, n. 1, p. 49–61, 1984. ISSN 01676393. DOI: 10.1016/0167-6393(84)90008-6.

TRAUNMÜLLER, Hartmut; ERIKSSON, Anders. The frequency range of the voice fundamental in the speech of male and female adults. **Unpublished manuscript**, 1995. Available from: <https://www2.ling.su.se/staff/hartmut/f0_m&f.pdf>.

TSAO, Ying-Chiao; WEISMER, Gary. Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. **Journal of Speech, Language, and Hearing Research**, ASHA, v. 40, n. 4, p. 858–866, 1997. DOI: https://doi.org/10.1044/jslhr.4004.858.

TSAO, Ying-Chiao; WEISMER, Gary; IQBAL, Kamran. The effect of intertalker speech rate variation on acoustic vowel space. **The Journal of the Acoustical Society of America**, v. 119, n. 2, p. 1074, 2006. ISSN 00014966. DOI: 10.1121/1.2149774.

TURK, Alice E; SHATTUCK-HUFNAGEL, Stefanie. Multiple targets of phrase-final lengthening in American English words. **Journal of Phonetics**, Elsevier, v. 35, n. 4, p. 445–472, 2007. DOI: https://doi.org/10.1016/j.wocn.2006.12.001.

VAN DEN BERG, Janwillem. Myoelastic-aerodynamic theory of voice production. **Journal of speech and hearing research**, ASHA, v. 1, n. 3, p. 227–244, 1958. DOI: 10.1044/jshr.0103.227.

VAN LIERDE, Kristiane M et al. Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach. **Journal of Voice**, Elsevier, v. 19, n. 4, p. 511–518, 2005. DOI: 10.1016/j.jvoice.2004.10.005.

VARNET, Léo et al. A cross-linguistic study of speech modulation spectra. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 142, n. 4, p. 1976–1989, 2017. DOI: https://doi.org/10.1121/1.5006179.

VIEGAS, Flávia et al. Comparison of fundamental frequency and formants frequency measurements in two speech tasks. **Revista CEFAC**, SciELO Brasil, v. 21, n. 6, 2019. DOI: https://doi.org/10.1590/1982-0216/201921612819.

VOGEL, F; MOTULSKY, AG. History and Development of Human Cytogenetics. **Human Genetics. Problems and Approaches. Springer-Verlag, Berlin Heidelberg New York Tokyo**, p. 20–24, 1986.

WEINHOLD, Bob. **Epigenetics: the science of change**. [S.l.]: National Institute of Environmental Health Sciences, 2006. DOI: doi:10.1289/ehp.114-a160.

WEIRICH, Melanie. **Articulatory and acoustic inter-speaker variability in the production of German vowels**. [S.l.]: Universitätsbibliothek Johann Christian Senckenberg, 2010. DOI: https://doi.org/10.21248/zaspil.52.2010.381.

_____. The influence of NATURE and NURTURE on speaker-specific parameters in twins' speech: Articulation, acoustics and perception. **Unpublished doctoral dissertation. Humboldt University of Berlin, Berlin, Germany**, 2012.

WEIRICH, Melanie; LANCIA, Leonardo; BRUNNER, Jana. Inter-speaker articulatory variability during vowel-consonant-vowel sequences in twins and unrelated speakers. **The Journal of the Acoustical Society of America**, v. 134, n. 5, p. 3766–3780, 2013. ISSN 0001-4966. DOI: 10.1121/1.4822480.

WEIRICH, Melanie; SIMPSON, Adrian P. Differences in acoustic vowel space and the perception of speech tempo. **Journal of Phonetics**, Elsevier, v. 43, n. 1, p. 1–10, 2014. ISSN 00954470. DOI: http://dx.doi.org/10.1016/j.wocn.2014.01.001.

WHITESIDE, Sandra P; RIXON, Emma. Speech characteristics of monozygotic twins and a same-sex sibling: An acoustic case study of coarticulation patterns in read speech. **Phonetica**, Karger Publishers, v. 60, n. 4, p. 273–297, 2003. DOI: 10.1159/000076377.

WHITESIDE, Sandra P; RIXON, Emma. Speech tempo and fundamental frequency patterns: a case study of male monozygotic twins and an age-and sex-matched sibling. **Logopedics Phoniatrics Vocology**, Taylor & Francis, v. 38, n. 4, p. 173–181, 2013. DOI: 10.3109/14015439.2012.742562.

WILLIAMS, Carl E; STEVENS, Kenneth N. Emotions and speech: Some acoustical correlates. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 52, 4B, p. 1238–1250, 1972. DOI: 10.1121/1.1913238.

WRETLING, Pär; ERIKSSON, Anders. Is articulatory timing speaker specific? evidence from imitated voices. In: PROC. FONETIK. [S.l.: s.n.], 1998. v. 98, p. 48–52.

WUYTS, Floris L et al. The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach. **Journal of Speech, Language, and Hearing Research**, ASHA, v. 43, n. 3, p. 796–809, 2000. DOI: https://doi.org/10.1044/jslhr.4303.796.

ZHANG, Zhaoyan. Mechanics of human voice production and control. **The journal of the acoustical society of america**, Acoustical Society of America, v. 140, n. 4, p. 2614–2635, 2016. DOI: https://doi.org/10.1121/1.4964509.

ZHENG, Weiying et al. Facial morphological characteristics of mouth breathers vs. nasal breathers: A systematic review and meta-analysis of lateral cephalometric data. **Experimental and Therapeutic Medicine**, Spandidos Publications, v. 19, n. 6, p. 3738–3750, 2020. DOI: 10.3892/etm.2020.8611.

ZRAICK, Richard I; SKAGGS, Sheri D; MONTAGUE, James C. The effect of task on determination of habitual pitch. **Journal of Voice**, Elsevier, v. 14, n. 4, p. 484–489, 2000. DOI: https://doi.org/10.1016/S0892-1997(00)80005-3.

ZUO, Donghui; MOK, Peggy Pik Ki. Formant dynamics of bilingual identical twins. **Journal of Phonetics**, Elsevier, v. 52, p. 1–12, 2015. DOI: https://doi.org/10.1016/j.wocn.2015.03.003.

ZWICKER, Eberhard. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 33, n. 2, p. 248–248, 1961. DOI: https://doi.org/10.1121/1.1908630.

ZWICKER, Eberhard; FASTL, Hugo. **Psychoacoustics: Facts and models**. [S.l.]: Springer Science & Business Media, 2013. v. 22. DOI: https://doi-org.ezp.sub.su.se/10.1007/978-3-662-09562-1_6.

Annex I

Ethical approval

Considerações sobre os Termos de apresentação obrigatória:

Foram analisados os seguintes documentos de apresentação obrigatória:

1 - Folha de Rosto Para Pesquisa Envolvendo Seres Humanos: Foi apresentado o documento "folhaRostoDOC.pdf" devidamente preenchido, datado e assinado.

2 - Projeto de Pesquisa: Foram analisados os documentos "PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1171908.pdf" e "ProjetoDoutorado_CEP.pdf". Adequados.

3 - Orçamento financeiro e fontes de financiamento: Informações sobre orçamento financeiro incluídas no documento "PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1171908.pdf". Adequado.

4 - Cronograma: Informações sobre o cronograma incluídas nos documentos "PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1171908.pdf". Adequado.

5 - Termo de Consentimento Livre e Esclarecido: "TCLE_DOC.pdf". Adequado.

6 - Currículo do pesquisador principal e demais colaboradores: Contemplado no cadastro da Plataforma Brasil.

7- Outros documentos que acompanham o Protocolo de Pesquisa:

001.jpg

Recomendações:

Recomendamos incluir no TCLE o tempo de armazenamento do material coletado (no mínimo, 5 anos, segundo a Resolução 510/2016 que regulamenta a pesquisa em Ciências Humanas e Sociais).

Conclusões ou Pendências e Lista de Inadequações:

Consideramos o protocolo de pesquisa aprovado.

Considerações Finais a critério do CEP:

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1171908.pdf	04/07/2018 22:37:15		Aceito
Folha de Rosto	folhaRostoDOC.pdf	04/07/2018 22:36:25	Julio Cesar Cavalcanti de Oliveira	Aceito
Projeto Detalhado	ProjetoDoutorado_CEP.pdf	04/07/2018	Julio Cesar	Aceito

Endereço: Av. Betrand Russell, 801, 2º Piso, Bloco C, Sala 5, Campinas-SP, Brasil.
Bairro: Cidade Universitária "Zeferino Vaz" **CEP:** 13.083-865
UF: SP **Município:** CAMPINAS
Telefone: (19)3521-6836 **E-mail:** epimenta@g.unicamp.br

/ Brochura Investigador	ProjetoDoutorado_CEP.pdf	22:36:07	Cavalcanti de Oliveira	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_DOC.pdf	02/07/2018 14:02:46	Julio Cesar Cavalcanti de Oliveira	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

CAMPINAS, 22 de Agosto de 2018

Assinado por:
Barbara Geraldo de Castro
(Coordenador)

Endereço: Av. Betrand Russell, 801, 2º Piso, Bloco C, Sala 5, Campinas-SP, Brasil.
Bairro: Cidade Universitária "Zeferino Vaz" **CEP:** 13.083-865
UF: SP **Município:** CAMPINAS
Telefone: (19)3521-6836 **E-mail:** epimenta@g.unicamp.br

Annex II

Praat script: *ProsodyTime*

```
# ProsodyTime.psc
# Script implemented by Plinio A. Barbosa (IEL/Univ. of Campinas, Brazil) for
computing
# prosody descriptors from coupled audio/TG files with a syllable-size or
phoneme size unit and TextGrid
# with labels and boundaries
# The TextGrid and Reference-statistics (xy.TableOfReal, where xy = BP, EP, F,
G, or BE) files need
# to be in the same directory!!!
# Copyright (C) 2012, 2014  Barbosa, P. A.
#
#    This program is free software; you can redistribute it and/or modify
#    it under the terms of the GNU General Public License as published by
#    the Free Software Foundation; version 2 of the License.
#    This program is distributed in the hope that it will be useful,
#    but WITHOUT ANY WARRANTY; without even the implied warranty of
#    MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
#    GNU General Public License for more details.
#
# Date: 2012, new version (2.0): Jul, 2014, Jul 2018, Jul, Sept, Oct 2019.
form File acquisition
 word ProsodyOut Parameters.txt
 word PauseOut Pause.txt
 word UnitOut Unit.txt
 word VowelsOut Vowels.txt
 word AudioFileExtension *.wav
 integer VVTier 9
 integer PauseTier 7
 integer TATier 2
 integer VowelTier 4
 integer Ditongo 5
 integer HesitTier 6
 integer F0Thresholdleft 75
 positive MaxFormant 5
 positive Fmax 5000
 integer F0Thresholdright 500
 positive Smthf0Thr 1.5
 positive F0step 0.05
 positive Spectralemphasisthreshold 400
endform
# Read all files in the folder:
Create Strings as file list... list 'audioFileExtension$'
numberOfFiles = Get number of strings
if !numberOfFiles
     exit There are no sound files in the folder!
endif
# Generates the header of the output files:
filedelete 'prosodyOut$'
fileappend 'prosodyOut$' filename uttpart f0mean f0med f0sd f0sk f0min f0max
f0base df0mean df0sd df0sk emph f0r srate artrate hnr slLTAS 'newline$'
filedelete 'pauseOut$'
fileappend 'pauseOut$' filename SilDur 'newline$'
filedelete 'unitOut$'
fileappend 'unitOut$' filename uttname unittype label dur 'newline$'
filedelete 'vowelsOut$'
fileappend 'vowelsOut$' filename uttname label dur F1 F2 F3 F4'newline$'
for ifile from 1 to numberOfFiles
select Strings list
audiofile$ = Get string... ifile
Read from file... 'audiofile$'
filename$ = selected$("Sound")
# Computes the formant trace:
 To Formant (burg)... 0.0 'maxFormant' 'fmax' 0.025 50
tgfile$ = filename$ + ".TextGrid"
Read from file... 'tgfile$'
nintervals = Get number of intervals... 'tATier'
```

```
nintervalsp = Get number of intervals... 'pauseTier'
utterance_number = 0
for iutt from 1 to nintervals
utt$ = Get label of interval... 'tATier' iutt
if utt$ <> ""
 itime = Get start time of interval... 'tATier' iutt
 ftime = Get end time of interval... 'tATier' iutt
 select Sound 'filename$'
 Extract part... itime ftime rectangular 1.0 yes
 uttfilename$ = selected$("Sound")
 totaldur = Get total duration
 begin = Get start time
 end = Get end time
 if totaldur < 0.05
   exitScript: "Intervalo curto para cálculo de HNR em ", 'begin'
 endif
# Cálculo de taxa de elocução
 select TextGrid 'filename$'
 startvv = Get high interval at time... 'vVTier' 'begin'
 endvv = Get high interval at time... 'vVTier' 'end'
 nVV = endvv - startvv - 1
 srate = nVV/(end-begin)
# Tracking de pausas silenciosas
 k = itime
 int = Get high interval at time... 'pauseTier' 'k'
 int = int+1
 tpause = Get start time of interval... 'pauseTier' int
 sdursil = 0
 tant = k
while k < ftime and tpause < ftime
 pause$ = Get label of interval... 'pauseTier' 'int'
 if pause$ == "PS" or pause$ == "P" or pause$ == "p" or pause$ == "#" or pause$
== "-"
    tini = Get start point... 'pauseTier' 'int'
    tfin = Get end point... 'pauseTier' 'int'
    dursil = tfin-tini
     isi = tini - tant
    sdursil = sdursil + dursil
    fileappend 'pauseOut$' 'filename$' 'utt$' 'dursil:3' 'newline$'
 endif
 int = int + 1
 if int > nintervalsp
     k = ftime + 1
 else
     k = Get start point... 'pauseTier' 'int'
 endif
 tant = tini
endwhile
artrate = nVV/(end-begin-sdursil)
# VV units scan
 st = startvv+1
 ed = endvv -1
 for jj from st to ed
  vv$ = Get label of interval... 'vVTier' 'jj'
  tini = Get start point... 'vVTier' 'jj'
  tfin = Get end point... 'vVTier' 'jj'
  durvv = 1000*(tfin - tini)
  fileappend 'unitOut$' 'filename$' 'utt$' VV 'vv$' 'durvv:0' 'newline$'
 endfor
# V units scan
 startv = Get high interval at time... 'vowelTier' 'begin'
 endv = Get high interval at time... 'vowelTier' 'end'
 st = startv+1
 ed = endv -1
 for jj from st to ed
  v$ = Get label of interval... 'vowelTier' 'jj'
```

```
  if v$ <> ""
   tini = Get start point... 'vowelTier' 'jj'
   tfin = Get end point... 'vowelTier' 'jj'
   tmid = (tini+tfin)/2
   select Formant 'filename$'
   f1 = Get value at time... 1 'tmid' Hertz Linear
   f2 = Get value at time... 2 'tmid' Hertz Linear
   f3 = Get value at time... 3 'tmid' Hertz Linear
   f4 = Get value at time... 4 'tmid' Hertz Linear
   durv = 1000*(tfin - tini)
   fileappend 'unitOut$' 'filename$' 'utt$' V 'v$' 'durv:0' 'newline$'
   fileappend 'vowelsOut$' 'filename$' 'utt$' 'v$' 'durv:0' 'f1:0' 'f2:0' 'f3:0'
'f4:0' 'newline$'
 endif
  select TextGrid 'filename$'
 endfor
# Hesitation units scan
 starth = Get high interval at time... 'hesitTier' 'begin'
 endh = Get high interval at time... 'hesitTier' 'end'
 st = starth+1
 ed = endh -1
 for jj from st to ed
  h$ = Get label of interval... 'hesitTier' 'jj'
  if h$ <> ""
   tini = Get start point... 'hesitTier' 'jj'
   tfin = Get end point... 'hesitTier' 'jj'
   durh = 1000*(tfin - tini)
   fileappend 'unitOut$' 'filename$' 'utt$' H 'h$' 'durh:0' 'newline$'
  endif
 endfor
 select Sound 'uttfilename$'
 To Spectrum... yes
 emphasis = Get band energy difference... 0 'spectralemphasisthreshold' 0 0
 select Sound 'uttfilename$'
 To Ltas... 100
 slLTAS = Get slope... 0 1000 1000 4000 dB
 select Sound 'uttfilename$'
 To Harmonicity (ac)... 0.001 120 0.1 4.5
 hnr = Get mean... 0 0
 select Sound 'uttfilename$'
 To Pitch... 0.0 'f0Thresholdleft' 'f0Thresholdright'
 Smooth... 'smthf0Thr'
 f0mean = Get mean... 'begin' 'end' semitones re 1 Hz
 f0median = Get quantile... 'begin' 'end' 0.5 semitones re 1 Hz
 f0sd = Get standard deviation... 'begin' 'end' semitones
 f0skew = (f0mean-f0median)/f0sd
 f0max = Get quantile... 'begin' 'end' 0.99 Hertz
 f0min = Get quantile... 'begin' 'end' 0.01 Hertz
 f0base = f0mean - 1.43*f0sd
 Interpolate
 To Matrix
 To Sound (slice)... 1
 To PointProcess (extrema)... 1 yes no None
 ntones = Get number of points
 tonerate = ntones/totaldur
 select Pitch 'uttfilename$'
############################df0
Down to PitchTier
f0dur = Get total duration
meandf0 = 0
f0ant = Get value at time... 'itime'
i = 1
timef0 = f0step+'itime'
while timef0 <= (f0dur + itime)
f0current = Get value at time... 'timef0'
# Computes f0 discrete derivative, and its cumulative value
```

```
  df0'i' = f0current - f0ant
  meandf0 = meandf0 + df0'i'
  f0ant = f0current
  timef0 = timef0 + f0step
 i=i+1
endwhile
i = i -1
meandf0 = meandf0/i
# Computes f0 discrete derivative standard deviation
sdf0 = 0
for j from 1 to i
 sdf0 = sdf0 + (df0'j' - meandf0)*(df0'j' - meandf0)
endfor
sdf0 = sqrt(sdf0/(i-1))
# Computes f0 discrete derivative skewness
skdf0 = 0
for j from 1 to i
 skdf0 = skdf0 + ((df0'j' - meandf0)/sdf0)^3
endfor
skdf0 = (i/((i-1)*(i-2)))*skdf0
fileappend 'prosodyOut$' 'filename$' 'utt$' 'f0mean:0' 'f0median:0' 'f0sd:2'
'f0skew:2' 'f0min:0' 'f0max:0' 'f0base:0' 'meandf0:2' 'sdf0:2' 'skdf0:2'
'emphasis:1' 'tonerate:2' 'srate:1' 'artrate:1' 'hnr:1' 'slLTAS:1' 'newline$'
endif
select TextGrid 'filename$'
endfor
```

Annex III

Praat script:
ProsodyDescriptorExtractor

```
# ProsodyDescriptorExtractor.psc
# Script implemented by Plinio A. Barbosa (IEL/Univ. of Campinas, Brazil) for
computing
# prosody descriptors from coupled audio/TG files
#
# The TextGrid and Reference-statistics (xy.TableOfReal, where xy = BP, EP, F,
G, or BE) files need
# to be in the same directory if a VV Tier will be used.
# Copyright (C) 2012, 2014  Barbosa, P. A.
#
# The only obligatory tier is the Chunk Tier.
#
#    This program is free software; you can redistribute it and/or modify
#    it under the terms of the GNU General Public License as published by
#    the Free Software Foundation; version 2 of the License.
#    This program is distributed in the hope that it will be useful,
#    but WITHOUT ANY WARRANTY; without even the implied warranty of
#    MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
#    GNU General Public License for more details.
#
#  New functionalities
#
# Date: 2012, 2015, new version (3.0): Jun 2020.
form File acquisition
 word FileOutPar OutPutProsParameters.txt
 word FileOutSil OutPutSil.txt
 word FileOutEff OutPutEff.txt
 word FileOutTones OutPutTones.txt
 word AudiofileExtension *.wav
 boolean HasTonesTier 0
 boolean HasVVTier 0
 boolean HasVowelTier 0
 boolean HasSilTier 0
 boolean InSemitones 1
 integer TonesTier 2
 integer VVTier 9
 integer VowelTier 3
 integer SilTier 7
 integer ChunkTier 1
 integer left_F0Threshold 60
 integer right_F0Threshold 300
 choice Reference: 1
   button BP
   button EP
   button G
   button F
   button BE
   button SP
endform
smthf0Thr = 2
f0step = 0.05
window = 0.03
spectralemphasisthreshold = 400
# Picks all audio files in the folder where the script is
Create Strings as file list... list 'audiofileExtension$'
numberOfFiles = Get number of strings
if !numberOfFiles
     exit There are no sound files in the folder!
endif
filedelete 'fileOutPar$'
# Creates the header of the mandatory output file (includes speech and
articulation rate if there is a VV tier).
if hasVVTier
 if hasSilTier
   fileappend 'fileOutPar$' audiofile chunk f0mean f0med f0sd f0SAQ f0baseline
f0min f0max sdf0peak f0peakwidth f0peak_rate sdtf0peak df0posmean df0negmean
```

```
df0sdpos df0sdneg emph cvint slLTASmed slLTAShigh hnr SPI shimmer jitter sr ar
'newline$'
 else
  fileappend 'fileOutPar$' audiofile chunk f0mean f0med f0sd f0SAQ f0baseline
f0min f0max sdf0peak f0peakwidth f0peak_rate sdtf0peak df0posmean df0negmean
df0sdpos df0sdneg emph cvint slLTASmed slLTAShigh  hnr SPI shimmer jitter sr
'newline$'
endif
# Reads the reference file with the triplets (segment, mean, standard-deviation)
from the
# reference speaker. The variable nseg contains the total number of segments in
the file
 Read from file... 'reference$'.TableOfReal
 nseg = Get number of rows
else
 fileappend 'fileOutPar$' audiofile chunk f0mean f0med f0sd f0SAQ f0baseline
f0min f0max sdf0peak f0peakwidth f0peak_rate sdtf0peak df0posmean df0negmean
df0sdpos df0sdneg emph cvint slLTASmed slLTAShigh  hnr SPI shimmer jitter
'newline$'
endif
# Creates the header of the output file with VQ parameters for vowels
if hasVowelTier
filedelete 'fileOutEff$'
fileappend 'fileOutEff$' audiofile excerpt vowel H1H2 CPP 'newline$'
endif
# Creates the header of the output file for tones
if hasTonesTier
filedelete 'fileOutTones$'
if hasVVTier
 fileappend 'fileOutTones$' audiofile excerpt tonetype time alignVV meanf0VV
'newline$'
else
 fileappend 'fileOutTones$' audiofile excerpt tonetype time 'newline$'
endif
endif
# Creates the header of the output file with pause-related parameters (duration
and Inter Pause Intervals, IPI)
if hasSilTier
filedelete 'fileOutSil$'
fileappend 'fileOutSil$' audiofile type IPI durSIL 'newline$'
endif
##
## Start of all computations for all pairs of audio/TG files
for ifile from 1 to numberOfFiles
select Strings list
audiofile$ = Get string... ifile
Read from file... 'audiofile$'
# filename$ contains the name of the audio file
filename$ = selected$("Sound")
# F0 trace is computed, for the whole audio file
To Pitch... 0.0 'left_F0Threshold' 'right_F0Threshold'
Smooth... 'smthf0Thr'
### Harmonicity
select Sound 'filename$'
To Harmonicity (ac)... 0.01 'left_F0Threshold' 0.1 4.5
# Reads corresponding TextGrid
arq$ = filename$ + ".TextGrid"
Read from file... 'arq$'
begin = Get starting time
end = Get finishing time
###
# Normalized duration computation as in the SG Detector Script (2006)
###
if hasVVTier
# The number of intervals in the VV tier is computed
nselected = Get number of intervals... 'vVTier'
```

```
arqout$ = filename$ + "dur" + ".txt"
filedelete 'arqout$'
arqoutstrgrp$ = filename$ + "SG" + ".txt"
filedelete 'arqoutstrgrp$'
fileappend 'arqout$' audiofile chunk segment duration_ms  z filteredz  boundary
'newline$'
fileappend 'arqoutstrgrp$' audiofile stressgroupduration numberVVunits
finalzfilt 'newline$'
select TextGrid 'filename$'
initialtime = Get starting point... 'vVTier' 2
# VV duration normalisation
kk = 1
 nselected = nselected - 2
 for i from 1 to nselected
  adv = i + 1
  nome$ = Get label of interval... 'vVTier' 'adv'
 itime = Get starting point... 'vVTier' 'adv'
 ftime = Get end point... 'vVTier' 'adv'
 dur = ftime - itime
 dur = dur*1000
 tint = Get starting point... 'vVTier' 'adv'
 call zscorecomp 'nome$' 'dur' 'tint'
 dur'i' = dur
 z'i' = z
 nome'i'$ = nome$
 select TextGrid 'filename$'
 adv = i + 1
endfor
### for i from 1 to nselected
smz1 = (2*z1 + z2)/3
deriv1 = smz1
smz2 = (2*z2 + z1)/3
deriv2 = smz2 - smz1
i = 3
if smz1 < smz2
 minsmz = smz1
 maxsmz = smz2
else
 minsmz = smz2
 maxsmz = smz1
endif
while i <= (nselected-2)
 del1 = i - 1
 del2 = i - 2
 adv1 = i + 1
 adv2 = i + 2
 smz'i' = (5*z'i' + 3*z'del1' + 3*z'adv1' + z'del2' + 1*z'adv2')/13
 deriv'i' = smz'i' - smz'del1'
 if smz'i' < minsmz
  minsmz = smz'i'
 endif
 if smz'i' > maxsmz
  maxsmz = smz'i'
 endif
 i = i + 1
endwhile
tp1 = nselected -1
tp2 = nselected -2
smz'tp1' = (3*z'tp1'+ z'tp2' + z'nselected')/5
deriv'tp1' = smz'tp1' - smz'tp2'
 if smz'tp1' < minsmz
  minsmz = smz'tp1'
 endif
 if smz'tp1' > maxsmz
  maxsmz = smz'tp1'
 endif
```

```
smz'nselected' = (2*z'nselected' + z'tp1')/3
deriv'nselected' = smz'nselected' - smz'tp1'
 if smz'nselected' < minsmz
  minsmz = smz'nselected'
 endif
 if smz'nselected' > maxsmz
  maxsmz = smz'nselected'
 endif
tempfile$ = "temp.TableOfReal"
filedelete 'tempfile$'
fileappend 'tempfile$' File type = "ooTextFile short" 'newline$'
fileappend 'tempfile$' "TableOfReal"  'newline$'
fileappend 'tempfile$'  'newline$'
fileappend 'tempfile$' 2 'newline$'
fileappend 'tempfile$' columnLabels []:  'newline$'
fileappend 'tempfile$' "position" "smoothed z" 'newline$'
tpp = nselected + 2
fileappend 'tempfile$' 'tpp' 'newline$'
time = initialtime
fileappend 'tempfile$' row[1]: "0" 0.0 0.0 'newline$'
boundcount = 0
sdur = 0
sduruns = 0
ssyl = 0
sdurSG = 0
svar = 0
for i from 1 to nselected
 timeinchunk = Get start time of interval... 'vVTier' 'i'
 intervalchunk = Get interval at time... 'chunkTier' 'timeinchunk'
 chunk$ = Get label of interval... 'chunkTier' 'intervalchunk'
 if chunk$ = ""
  chunk$ = "no_label"
 endif
 tempsmz = smz'i'
 tpnome$ = nome'i'$
 adv1 = i + 1
 btime'i' = 0
 time = time + dur'i'/1000
 time'i' = time
 fileappend 'tempfile$' row['adv1']: "'tpnome$'" 'time' 'tempsmz' 'newline$'
 if i <> nselected
  adv1 = i + 1
  if (deriv'i' >= 0) and (deriv'adv1' < 0)
    boundary = 1
    boundcount = boundcount + 1
    btime'i' = time
    bctime'boundcount' = time
    smzbound'boundcount' = smz'i'
  else
    boundary = 0
  endif
 else
  del1 = i -1
  if smz'i' > smz'del1'
     boundary = 1
     boundcount = boundcount + 1
     btime'i' = time
     bctime'boundcount' = time
     smzbound'boundcount' = smz'i'
  else
    boundary = 0
  endif
 endif
 tempz = z'i'
 tempdur = dur'i'
 sdur = sdur + tempdur
```

```
 if boundary == 0
   sduruns = sduruns  + tempdur
 endif
 sdurSG = sdurSG + tempdur
 ssyl = ssyl + 1
 fileappend 'arqout$' 'filename$' 'chunk$' 'tpnome$' 'tempdur:0' 'tempz:2'
'tempsmz:2' 'boundary' 'newline$'
 if boundary == 1
  fileappend 'arqoutstrgrp$' 'filename$' 'sdurSG:0' 'ssyl' 'tempz:2' 'newline$'
   durSG'kk' = sdurSG
   nunits'kk' = ssyl
   zprom'kk' = tempsmz
   kk = kk+1
   sdurSG = 0
   ssyl = 0
   sdurSG = 0
   ssyl = 0
 endif
endfor
### i from 1 to nselected (VV dur norm. computation)
nprom = kk - 1
prate = nprom*1000/sdur
meandur = sdur/nselected
for i from 1 to nselected
 svar = svar + (dur'i' - meandur)^2
endfor
stddevdur = sqrt(svar/(nselected - 1))
tp = i+1
fileappend 'tempfile$' row['tp']: "X" 'end' 0 'newline$'
filedelete temp.TableOfReal
####
#  Write a TextGrid with the stress group boundaries
fileout$ = filename$ + "SG.TextGrid"
filedelete 'fileout$'
fileappend 'fileout$' File type = "ooTextFile short" 'newline$'
fileappend 'fileout$' "TextGrid" 'newline$'
fileappend 'fileout$' 'newline$'
fileappend 'fileout$' 'begin' 'newline$'
fileappend 'fileout$' 'end' 'newline$'
fileappend 'fileout$' <exists> 'newline$'
fileappend 'fileout$' 2 'newline$'
fileappend 'fileout$' "TextTier" 'newline$'
fileappend 'fileout$' "BoundDegree" 'newline$'
fileappend 'fileout$' 'begin' 'newline$'
fileappend 'fileout$' 'end' 'newline$'
fileappend 'fileout$' 'boundcount' 'newline$'
for i from 1 to boundcount
 temp = bctime'i'
 fileappend 'fileout$' 'temp' 'newline$'
 tmpzb = round(100*smzbound'i')/100
 lab$ = string$(tmpzb)
 fileappend 'fileout$' "'lab$'" 'newline$'
endfor
fileappend 'fileout$' "IntervalTier" 'newline$'
fileappend 'fileout$' "StressGroups" 'newline$'
fileappend 'fileout$' 'begin' 'newline$'
fileappend 'fileout$' 'end' 'newline$'
tmp = boundcount + 2
fileappend 'fileout$' 'tmp' 'newline$'
fileappend 'fileout$' 0.00 'newline$'
fileappend 'fileout$' 'initialtime' 'newline$'
fileappend 'fileout$' "" 'newline$'
temp = initialtime
for i from 1 to boundcount
 fileappend 'fileout$' 'temp' 'newline$'
 temp = bctime'i'
```

```
 lab$ = "SG" + string$(i)
 fileappend 'fileout$' 'temp' 'newline$'
 fileappend 'fileout$' "'lab$'" 'newline$'
endfor
fileappend 'fileout$' 'temp' 'newline$'
fileappend 'fileout$' 'end' 'newline$'
fileappend 'fileout$' "" 'newline$'
arqgrid1$ = filename$ + ".TextGrid"
Read from file... 'arqgrid1$'
Read from file... 'fileout$'
plus TextGrid 'filename$'
Merge
Save as text file... 'filename$'Enriched.TextGrid
endif
##
####
if hasSilTier
# Silence sucession descriptors, if the TG has a pause tier (SilTier)
nintersil = Get number of intervals... 'silTier'
tiniant = 0
for i from 2 to nintersil - 1
  label'i'$ = Get label of interval... 'silTier' 'i'
  if label'i'$ <> ""
   type$ = label'i'$
   tini = Get start point... 'silTier' 'i'
   tfin = Get end point... 'silTier' 'i'
   dursil = round(('tfin'-'tini')*1000)
   if tiniant <> 0
    dIPI = tini - tiniant
   else
    dIPI = undefined
   endif
# dIPI contains the duration between the current pause onset and the previous
pause onset, irrespective of pause type
# type if the pause type, marked as a label in the pause tier
# dursil is the duration of the pause
   fileappend 'fileOutSil$' 'filename$' 'type$' 'dIPI:2' 'dursil' 'newline$'
   tiniant = tini
  endif
endfor
endif
### All tones from the Tones Tier is written, together with its time instant
if hasTonesTier
npointstones = Get number of points... 'tonesTier'
for k from 1 to npointstones
 select TextGrid 'filename$'
 label$ = Get label of point... 'tonesTier' k
 time = Get time of point... 'tonesTier' k
 timeinchunk = Get interval at time... 'chunkTier' 'time'
 chunk$ = Get label of interval... 'chunkTier' 'timeinchunk'
 if hasVVTier
  intcurrentVV = Get interval at time... 'vVTier' 'time'
  startinVV = Get start point... 'vVTier' 'intcurrentVV'
  endinVV = Get end point... 'vVTier' 'intcurrentVV'
  alignperc = 100*(time - startinVV)/(endinVV-startinVV)
  select Pitch 'filename$'
  meanf0VV = Get mean... 'startinVV' 'endinVV' Hertz
  fileappend 'fileOutTones$' 'filename$' 'chunk$' 'label$' 'time:3'
'alignperc:0' 'meanf0VV:0' 'newline$'
 else
  fileappend 'fileOutTones$' 'filename$' 'chunk$' 'label$' 'time:3' 'newline$'
 endif
endfor
endif
###
if hasVowelTier
```

```
# H1 - H2 and CPP computation for all open vowels whose intervals and labels
were assigned in the Vowel Tier
select TextGrid 'filename$'
ndesignatedvowels = Get number of intervals... 'vowelTier'
for i from 2 to ndesignatedvowels - 1
 select TextGrid 'filename$'
  label'i'$ = Get label of interval... 'vowelTier' 'i'
  if label'i'$ = "a" or label'i'$ = "A" or label'i'$ = "eh" or label'i'$ = "oh"
   vowel$ = label'i'$
   tini = Get start point... 'vowelTier' 'i'
   tfin = Get end point... 'vowelTier' 'i'
   tmean = (tini+tfin)/2
   timeinchunk = Get interval at time... 'chunkTier' 'tmean'
   chunk$ = Get label of interval... 'chunkTier' 'timeinchunk'
   select Pitch 'filename$'
   f0median = Get quantile... 'tini' 'tfin' 0.5 Hertz
   tleft = tmean - 'window'/2
   tright = tmean + 'window'/2
   select Sound 'filename$'
   Extract part... 'tleft' 'tright' rectangular 1.0 no
   To Spectrum... yes
   spect$ = selected$("Spectrum")
   To PowerCepstrum
   cpp = Get peak prominence... 60 340 Parabolic 0.001 0.0 Straight Robust
   select Spectrum 'spect$'
   To Ltas (1-to-1)
   f0min = 0
   f0max = f0median*1.5
   h1 = Get maximum... 'f0min' 'f0max' None
   f0min = f0max
   f0max = f0median*2.5
   h2 = Get maximum... 'f0min' 'f0max' None
   h1h2 = h1-h2
   fileappend 'fileOutEff$' 'filename$' 'chunk$' 'vowel$' 'h1h2:2' 'cpp:2'
'newline$'
  endif
endfor
endif
###
### All prosodic parameters for each labelled interval in Chunk Tier are
computed:
# f0median f0max sdf0max f0min f0sd tonerate sdpitch meandf0pos meandf0neg
sdf0pos sdf0neg emphasis
###
select TextGrid 'filename$'
nchunks = Get number of intervals... 'chunkTier'
# Spectral emphasis computation
for ichunk from 1 to nchunks
 initime = Get start time of interval... 'chunkTier' ichunk
 endtime = Get end time of interval... 'chunkTier' ichunk
 uttlabel$ = Get label of interval... 'chunkTier' ichunk
 if uttlabel$ <> ""
 select Sound 'filename$'
 Extract part... initime endtime rectangular 1.0 yes
 chunkfilename$ = selected$("Sound")
# Computes the long term spectrum, and gets its standard deviation
 To Ltas... 100
 sltasmedium = Get slope... 0 1000 1000 4000 energy
 sltashigh = Get slope... 0 1000 4000 8000 energy
 select Sound 'chunkfilename$'
 To Intensity... 'left_F0Threshold' 0.0 yes
 mint = Get mean... 0.0 0.0 energy
 sdint = Get standard deviation... 0 0
 cvint = 100*sdint/mint
 select Sound 'chunkfilename$'
 To Spectrum... yes
```

```
 emphasis = Get band energy difference... 0 'spectralemphasisthreshold' 0 0
# f0 descriptors and f0 rate (tonerate) computation per chunk
 select Sound 'chunkfilename$'
 To Pitch... 0.0 'left_F0Threshold' 'right_F0Threshold'
 Smooth... 'smthf0Thr'
 if inSemitones
  f0mean = Get mean... 'initime' 'endtime' semitones re 1 Hz
  f0median = Get quantile... 'initime' 'endtime' 0.5 semitones re 1 Hz
  f0baseline = Get quantile... 'initime' 'endtime' 0.074 semitones re 1 Hz
  f0sd = Get standard deviation... 'initime' 'endtime' semitones
  f0min = Get quantile... 'initime' 'endtime' 0.01 semitones re 1 Hz
  f0max = Get quantile... 'initime' 'endtime' 0.99 semitones re 1 Hz
  f01Q = Get quantile... 'initime' 'endtime' 0.25 semitones re 1 Hz
  f03Q = Get quantile... 'initime' 'endtime' 0.75 semitones re 1 Hz
  f0SAQ = (f03Q-f01Q)/2
 else
  f0median = Get quantile... 'initime' 'endtime' 0.5 Hertz
  f0sd = Get standard deviation... 'initime' 'endtime' Hertz
  f0min = Get quantile... 'initime' 'endtime' 0.01 Hertz
  f0max = Get quantile... 'initime' 'endtime' 0.99 Hertz
  f01Q = Get quantile... 'initime' 'endtime' 0.25 Hertz
  f03Q = Get quantile... 'initime' 'endtime' 0.75 Hertz
  f0SAQ = (f03Q-f01Q)/2
 endif
 Interpolate
 To Matrix
 To Sound (slice)... 1
 Rename... Temp
 To PointProcess (extrema)... 1 yes no None
 ntones = Get number of points
 if ntones <> 0
  initone = Get time from index... 1
  endtone = Get time from index... ntones
  durtones = endtone - initone
  if durtones <> 0
   tonerate = ntones/durtones
  else
   tonerate = undefined
  endif
 else
   tonerate = undefined
 endif
 sdpitch = Get stdev period... 'initime' 'endtime' 0.04 2 1.7
# F0max descriptors (mean and sd)
 meanf0max = 0
 meandrop = 0
 nundefined = 0
 for if0max from 1 to ntones
  tf0max = Get time from index... 'if0max'
  select Pitch 'chunkfilename$'
  if inSemitones
   f0max'if0max' = Get value at time... 'tf0max' "semitones re 1 Hz" Linear
  else
   f0max'if0max' = Get value at time... 'tf0max' "Hertz" Linear
  endif
  tf0left = tf0max - 0.03
  tf0right = tf0max + 0.03
  if inSemitones
   f0maxleft = Get value at time... 'tf0left' "semitones re 1 Hz" Linear
   f0maxright = Get value at time... 'tf0right' "semitones re 1 Hz" Linear
  else
   f0maxleft = Get value at time... 'tf0left' "Hertz" Linear
   f0maxright = Get value at time... 'tf0right' "Hertz" Linear
  endif
  drop = (f0maxleft + f0maxright)/2 - f0max'if0max'
  if drop != undefined
```

```
   meandrop = meandrop + drop
  else
   nundefined = nundefined + 1
  endif
  meanf0max = meanf0max + f0max'if0max'
  select PointProcess Temp
 endfor
 meanf0max = meanf0max/ntones
 meandrop = -1000*meandrop/((ntones - nundefined)*(f0max-f0min))
 sdf0max = 0
 for max from 1 to ntones
  sdf0max = sdf0max + (f0max'max' - meanf0max)*(f0max'max' - meanf0max)
 endfor
 sdf0max = sqrt(sdf0max/(ntones-1))
 select Pitch 'chunkfilename$'
# df0 computations
 Down to PitchTier
 f0dur = Get total duration
 meandf0pos = 0
 meandf0neg = 0
 f0ant = Get value at time... 0
 l = 1
 lneg = 0
 lpos = 0
 timef0 = f0step+'initime'
 while timef0 <= (f0dur + initime)
 f0current = Get value at time... 'timef0'
# Computes f0 discrete derivative, and its cumulative value
  df0'l' = f0current - f0ant
  if df0'l' > 0
   meandf0pos = meandf0pos + df0'l'
   lpos = lpos + 1
   df0pos'lpos' = df0'l'
  else
   meandf0neg = meandf0neg + df0'l'
   lneg = lneg + 1
   df0neg'lneg' = df0'l'
  endif
  f0ant = f0current
  timef0 = timef0 + f0step
  l=l+1
 endwhile
 l = l -1
 meandf0pos = meandf0pos/lpos
 meandf0neg = meandf0neg/lneg
# Computes f0 discrete derivative standard deviation
 sdf0pos = 0
 for j from 1 to lpos
  sdf0pos = sdf0pos + (df0pos'j' - meandf0pos)*(df0pos'j' - meandf0pos)
 endfor
 sdf0pos = sqrt(sdf0pos/(lpos-1))
#
 sdf0neg = 0
 for j from 1 to lneg
  sdf0neg = sdf0neg + (df0neg'j' - meandf0neg)*(df0neg'j' - meandf0neg)
 endfor
 sdf0neg = sqrt(sdf0neg/(lneg-1))
#######
 select Sound 'chunkfilename$'
 To PointProcess (periodic, cc)... 'left_F0Threshold' 'right_F0Threshold'
 plus Sound 'chunkfilename$'
 To Ltas (only harmonics)... 50 0.0001 0.02 1.3
 lowmean = Get mean... 1.4 32 dB
 highmean = Get mean... 32 64.3 dB
 sPI = lowmean - highmean
 select PointProcess 'chunkfilename$'
```

```
 jitter = Get jitter (local)... 0.0 0.0 0.0001 0.02 1.3
 jitter = 100*jitter
 plus Sound 'chunkfilename$'
 shimmer = Get shimmer (local)... 0 0 0.0001 0.02 1.3 1.6
 shimmer = 100*shimmer
 select Harmonicity 'filename$'
 hnr = Get mean... 'initime' 'endtime'
#######
# Speech rate computation per chunk
if hasVVTier
 select TextGrid 'filename$'
 startvv = Get high interval at time... 'vVTier' 'initime'
 endvv = Get high interval at time... 'vVTier' 'endtime'
 nVV = endvv - startvv - 1
 srate = nVV/(endtime-initime)
endif
###
# Tracking of pauses for computing articulation rate
if hasSilTier
 select TextGrid 'filename$'
 int = Get high interval at time... 'silTier' 'initime'
 intfinal = Get low interval at time... 'silTier' 'endtime'
 sdursil = 0
while int <= intfinal
pause$ = Get label of interval... 'silTier' 'int'
 if pause$ <> ""
    tinisil = Get start point... 'silTier' 'int'
    tfinsil = Get end point... 'silTier' 'int'
    sdursil = sdursil + tfinsil-tinisil
 endif
 int = int + 1
endwhile
endif
# Articulation rate computation (requires Pause and VV Tiers)
if hasVVTier & hasSilTier
 artrate = nVV/(endtime-initime-sdursil)
endif
if !hasVVTier
fileappend 'fileOutPar$' 'filename$' 'uttlabel$' 'f0mean:0' 'f0median:0'
'f0sd:2' 'f0SAQ:2' 'f0baseline:0' 'f0min:0' 'f0max:0' 'sdf0max:1' 'meandrop:1'
'tonerate:2' 'sdpitch:2' 'meandf0pos:2' 'meandf0neg:2' 'sdf0pos:2' 'sdf0neg:2'
'emphasis:1' 'cvint:0' 'sltasmedium:1' 'sltashigh:1' 'hnr:1' 'sPI:1' 'shimmer:1'
'jitter:1' 'newline$'
else
 if hasSilTier
  fileappend 'fileOutPar$' 'filename$' 'uttlabel$' 'f0mean:0' 'f0median:0'
'f0sd:2' 'f0SAQ:2' 'f0baseline:0' 'f0min:0' 'f0max:0' 'sdf0max:1' 'meandrop:1'
'tonerate:2' 'sdpitch:2' 'meandf0pos:2' 'meandf0neg:2' 'sdf0pos:2' 'sdf0neg:2'
'emphasis:1' 'cvint:0' 'sltasmedium:1' 'sltashigh:1' 'hnr:1' 'sPI:1' 'shimmer:1'
'jitter:1' 'srate:1' 'artrate:1' 'newline$'
 else
  fileappend 'fileOutPar$' 'filename$' 'uttlabel$' 'f0mean:0' 'f0median:0'
'f0sd:2' 'f0SAQ:2' 'f0baseline:0' 'f0min:0'  'f0max:0' 'sdf0max:1' 'meandrop:1'
'tonerate:2' 'sdpitch:2' 'meandf0pos:2' 'meandf0neg:2' 'sdf0pos:2' 'sdf0neg:2'
'emphasis:1' 'cvint:0' 'sltasmedium:1' 'sltashigh:1' 'hnr:1' 'sPI:1' 'shimmer:1'
'jitter:1' 'srate:1' 'newline$'
 endif
endif
endif
select TextGrid 'filename$'
endfor
endfor
##
procedure zscorecomp nome$ dur tint
 sizeunit = length (nome$)
 sumofmeans = 0
```

```
 sumofvar = 0
 cpt = 1
 while cpt <= sizeunit
  nb = 1
  terminate = 0
  seg$ = mid$(nome$,cpt,1)
  if cpt < sizeunit
#    if phoneticAlphabet$ = "Other"
     if reference$ = "BP" or reference$ = "EP"
      if mid$(nome$,cpt+1,1) == "h"  or mid$(nome$,cpt+1,1) == "N"
         nb = nb + 1
         seg$ = seg$ + mid$(nome$,cpt+1,1)
      endif
      if (cpt+nb <= sizeunit)
       tp$ = mid$(nome$,cpt,1)
       call isvowel 'tp$'
       if ((mid$(nome$,cpt+nb,1) = "I")  or  (mid$(nome$,cpt+nb,1)  = "U"))  and
truevowel
         seg$ = seg$ + mid$(nome$,cpt+nb,1)
         nb= nb+1
       endif
      endif
     endif
     if reference$ = "F"
       if mid$(nome$,cpt+1,1) == "h"  or mid$(nome$,cpt+1,1) == "N"  or mid$
(nome$,cpt+1,1) == "x"
         nb = nb + 1
         seg$ = seg$ + mid$(nome$,cpt+1,1)
      endif
     endif
     endif
    else
      if mid$(nome$,cpt+1,1) == "~"
         nb = nb + 1
         seg$ = seg$ + mid$(nome$,cpt+1,1)
      endif
      if (cpt+nb <= sizeunit)
       tp$ = mid$(nome$,cpt,1)
       call isvowel 'tp$'
       if ((mid$(nome$,cpt+nb,1) = "j")  or  (mid$(nome$,cpt+nb,1)  = "w"))  and
truevowel
         seg$ = seg$ + mid$(nome$,cpt+nb,1)
         nb= nb+1
       endif
      endif
#    endif
  endif
  j = 1
  select all
  tableID = selected ("TableOfReal")
  select 'tableID'
  while (j <= nseg) and  not terminate
     label$ =  Get row label... 'j'
     if seg$ = label$
         terminate = 1
         mean = Get value... 'j' 1
         sd      = Get value... 'j' 2
         sumofmeans = mean + sumofmeans
         sumofvar= sd*sd + sumofvar
     endif
     j = j+1
  endwhile
  if not terminate
   exit Didn't find phone 'seg$' at 'tint'. Pls check the file TableOfReal
  endif
  cpt= cpt+nb
```

```
 endwhile
z = (dur - sumofmeans)/sqrt(sumofvar)
endproc
procedure isvowel temp$
 truevowel = 0
 if temp$ = "i" or temp$ = "e"  or temp$ = "a"  or temp$ = "o"  or temp$ = "u"
or temp$ = "I" or temp$ = "E"
    ...or temp$ = "A"  or temp$ = "y" or temp$ = "O"  or temp$ = "U" or temp$ =
"6"  or temp$ = "@"
    ...or temp$ = "2" or temp$ = "9" or temp$ = "Y"
    truevowel = 1
 endif
endproc
```