UIMP
Universidad Internacional
Menéndez Pelayo

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

TESIS DOCTORAL

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

Eugenia San Segundo Fernández

Máster en Fonética y Fonología
Licenciada en Filología Inglesa
Licenciada en Filología Hispánica

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

2014

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

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

Eugenia San Segundo Fernández

M.A. Phonetics and Phonology
B.A. Hispanic Studies
B.A. English studies

Thesis Supervisors:
Dr. Joaquim Llisterri Boix
Dr. Juana Gil Fernández

AGRADECIMIENTOS

La elaboración de esta tesis ha sido posible gracias a muchas personas, tanto del ámbito académico como ajenos a este. Reservando a los amigos y familiares para el final, debo empezar dando las gracias –dentro del mundo académico, en orden cronológico– a la Dra. Juana Gil por haberme brindado la oportunidad de solicitar una beca FPU en el Laboratorio de Fonética del CSIC. Gracias por dejar que te insistiera en lo importante que era para mí solicitarla ese año, y gracias también por ponerme en contacto con el Dr. Joaquim Llisterri como director de tesis. Gracias, Joaquim, por estar siempre ahí cuando lo he necesitado, pese a la distancia entre Barcelona y Madrid. Eres un ejemplo para mí de rigor y de perfeccionismo. Gracias especialmente por tu apoyo en los últimos días. A ambos, gracias por confiar en mí.

Dos personas muy importantes para mí en la elaboración de esta tesis han sido el Dr. Pedro Gómez Vilda y el Dr. Hermann Künzel. Al primero debo agradecerle todo su tiempo y su paciencia para enseñarme el funcionamiento de *BioMetroSoft* y, en general, para ayudarme a enfocar la tesis. Gracias también por tu amabilidad, por tu humildad y por tu buena disposición para enseñar. Al Dr. Hermann Künzel le doy las gracias por confiar en mí para solicitar conjuntamente una beca a la IAFPA, sin la cual –muy probablemente– no hubiera podido llevar a cabo esta investigación. Gracias por acogerme en la estancia de investigación en Marburg y, sobre todo, por tu impagable ayuda para elaborar el sexto capítulo de esta tesis.

Son muchos los investigadores con los que, a lo largo de estos cuatro años, he podido compartir opiniones y aprender de ellos. Al Dr. G.S. Morrison le agradezco su tiempo y dedicación para descubrirme las relaciones de verosimilitud y el enfoque bayesiano, así como ciertos programas informáticos, desarrollados por él, que utilizo en el capítulo cuatro. Muchas gracias al Dr. Daniel Ramos por su amabilidad para ayudarme a resolver ciertas dudas de ese capítulo y por leer alguna versión preliminar del mismo. Quedo muy agradecida, asimismo, al Dr. A. Tsanas y a la Dra. D. Loakes por haber aceptado realizar sendos informes de esta tesis y por sus muchos comentarios y sugerencias. Mi más sincero agradecimiento a la Dra. Ingrid Hove por su generosidad al aceptar realizar un tercer informe, especialmente con tan corto preaviso.

Gracias a todos los compañeros del Laboratorio de Fonética que han ido yendo y viniendo a lo largo de estos años: contratados, becarios, técnicos, etc., con un agradecimiento especial a la Dra. María José Albalá, por su inagotable amabilidad y por su sonrisa permanente, y a Rocío Peña por haber sufrido conmigo la agonía de encontrar y comprar los micrófonos para grabar a los informantes. Gracias también a mis colegas becarios de Historia del CCHS: por las risas y por los buenos tiempos. Compañera de fatigas en la Universidad de Marburg, Almut Braun, gracias por hacerme la vida más fácil allí, por nuestras excursiones y por soportar con paciencia mis intentos de hablar alemán. Por el último año de tesis, ya en Suiza, quiero agradecer también al Dr.Volker Dellwo que me haya acogido tan bien siempre que he ido a Zúrich: porque tener esperanza en el futuro post-tesis siempre anima a terminar esta con más ganas.

Agradezco su colaboración a todos los gemelos y no gemelos que han prestado su voz para esta tesis, que se ha podido llevar a cabo gracias a la concesión de una beca FPU (AP2008-01524) del Ministerio de Educación. Gracias también al Dr. Antonio Alonso por la prueba de cigosidad.

Finalmente, esta tesis no habría podido realizarse sin el apoyo constante de mi familia. Gracias por enseñarme a ser la persona que soy hoy, con sus cosas buenas y malas. Habéis sufrido la tesis tanto o más que yo. A mi padre, por ser el primero en inculcarme el espíritu científico. A mi madre y a Paula, simplemente gracias por estar ahí, y por los ánimos y abrazos cuando hacían falta. Curro, sin ti nada de esto habría sido posible y, por eso, a ti te dedico esta tesis. Son muchas las cosas que tendría que agradecerte, y alguna se me olvidaría, así que sencillamente: gracias por hacerme reír (de la vida y de mí misma) cada día desde que te conozco.

ACKNOWLEDGEMENTS

The elaboration of this thesis would not have been possible without the help of many people, both within an academic context and outside. I will keep the acknowledgements to my family and friends for the very end. Therefore, I must begin thanking –in chronological order– Dr. Juana Gil, for giving me the chance to apply for a FPU doctoral grant in the Phonetics Laboratory of the CSIC.Thank you very much for letting me insist in how important such a grant was at that moment, and thanks as well for introducing me to Dr. Joaquim Llisterri, my thesis supervisor. Thanks to Joaquim for always being there when needed, despite the distance between Madrid and Barcelona. You are an example for me of rigor and perfectionism. I would like to thank you especially for your support during the last days. To both of you, Joaquim and Juana, thank you for relying on me.

There are two researchers who have played an important role in the elaboration of this thesis. On the one hand, I thank Dr. Gómez Vilda for his patience when teaching me how to use *BioMetroSoft* and, in general, for his friendliness and humility. On the other hand, I am very grateful to Dr. Hermann Künzel for trusting me and having applied with me for an IAFPA grant. Without this, I think I would have never been able to carry out this investigation. Thanks as well for welcoming me in Marburg during my research stay and, above all, for your priceless help in designing and writing the sixth chapter of this thesis.

Throughout these four years, there are many researchers with whom I have had the opportunity to share opinions and I have learnt from them. I am very grateful to Dr. G.S. Morrison for his time and teaching efforts, thanks to which I discovered the Bayes approach and likelihood ratios. Thank you very much also for letting me know about your different software programs, which I use in chapter four. I must especially thank Dr. Daniel Ramos for having so kindly helped me solve some doubts of that chapter and for reading some draft versions of it. I also remain very grateful to Dr. A. Tsanas and to Dr. D. Loakes for having accepted writing a report of my thesis and for their insightful comments, and I extend my grateful thanks to Dr. Ingrid Hove for a further report, particularly when asked on such short notice.

I have to thank all the colleagues in the Phonetics Laboratory who have been there all these years, with a special mention to Dr. María José Albalá, for her affability and permanent smile, and to Rocío Peña for having suffered with me the dying effort to search for the suitable microphones with which the recordings were made. Thank you as well to my colleagues from the History Department: for the laughter and the good old times. Almut Braun, a colleague and a friend during my research stay in Marburg: thanks for making my life there easier, for our excursions and for your patience with my clumsy attempts to speak German. In my final thesis year, already in Switzerland, I would like to thank Dr. Volker Dellwo for having welcomed me so warmly whenever I went to Zürich and for letting me holding out some hope in the post-thesis future.

I extend my gratitude to all the twin and non-twin participants in this thesis, which could be carried out thanks to a FPU grant awarded by the Spanish Ministry of Education (AP2008-01524) and to an IAFPA research grant. Besides, I would like to extend special thanks to Dr. Antonio Alonso for carrying out the zygosity test.

Last but not least, this thesis would have never appeared without the permanent support of my family. Thanks for making me the person I am today. You have "suffered" this thesis as much as me. To my father, because he first instilled in me the passion for science. To my mother and sister, for being there with support and hugs. Curro, without you this would have never been possible. There are so many things I should thank you for. Let me simply say thanks for making me laugh (at life as well as at myself) every single day since I know you.

INDEX

LIST OF ABBREVIATIONS

AAW	Acoustic Average Wave, 157
AMR-WB	Adaptative Multirate Wideband, 58
ASR	Automatic Speaker Recognition, 7, 8, 10, 14, 196, 200, 212, 222
AVI	Amplitude Variability Index, 152
B	Brothers, 20, 21, 28, 29, 30, 41, 53, 60, 65, 71, 72, 73, 74, 85, 90, 98, 103, 104, 111, 112, 113, 114, 115, 116, 117, 119, 120, 122, 123, 124, 125, 128, 175, 179, 187, 189, 191, 193, 194, 197, 203, 205, 211, 213, 216, 217, 218, 219, 226, 227, 228, 229, 230, 231, 240, 242, 244, 245, 246
CDC	Cover Dynamic Component, 157
CSIC	Consejo Superior de Investigaciones Científicas, 79, 232
CV	Consonant-Vowel, 39, 50
CVC	Consonant-Vowel-Consonant, 41
DCT	Discrete Cosine Transform, 62, 89, 90, 95, 96, 97, 98, 99, 100, 109, 110, 111, 118, 122, 216, 221
DFT	Discrete Fourier Transform, 151, 155
DNA	Deoxyribonucleic Acid, 1, 4, 14, 28, 31, 32, 38, 61, 81, 104, 127, 142, 164, 213, 223, 224, 233, 243
DSI	Dysphonia Severity Index, 44, 47
DZ	Dizygotic, 1, 2, 10, 11, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 35, 36, 37, 38, 42, 43, 45, 46, 47, 48, 49, 50, 51, 54, 57, 58, 59, 61, 63, 85, 98, 103, 104, 111, 112, 113, 114, 115, 116, 117, 119, 122, 123, 124, 125, 127, 128, 131, 135, 160, 162, 163,164, 165, 166, 167, 170, 171, 175, 176, 177, 178, 179, 186, 187, 189, 191, 192, 193, 194, 195, 200, 203, 204, 205, 209, 211, 212, 213, 215, 216, 217, 218, 219, 222
EER	Equal Error Rate, 53, 54, 202, 207, 208, 209, 210, 211
EMA	Electromagnetic Articulograph, 48, 50
FA	False Alarms, 200, 202, 209, 210, 211
FFT	Fast Fourier Transform, 46, 151
FR	False Rejections, 200, 202, 209, 210, 211
FSC	Forensic Speaker Comparison, 1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 15, 86, 121, 127, 130, 205, 215, 218, 219
FSI	Forensic Speaker Identification, 6, 58

GMM-UBM
Gaussian Mixture Model-Universal Background Model, 222

GRBAS
Grade, Roughness, Breathiness, Asthenia, Strain, 44, 47

GSM
Global System for Mobile Communications, 58, 197, 199

GUI
Graphical User Interface, 139, 145, 146, 172

HMM
Hidden Markov Model, 55

HNR
Harmonics-to-Noise Ratio, 44, 47, 150

IAFPA
International Association for Forensic Phonetics and Acoustics, 3, 4, 61, 229, 230, 235, 244

IP
Intra-pair, 58, 79, 120, 128, 206, 207, 208, 209, 210, 211, 217, 221

IS
Intra-speaker, 113, 114, 115, 117, 119, 123, 128, 206, 207, 208, 209, 210, 221

LADO
Language Analysis for the Determination of Origin of Asylum Seekers, 3

LDA
Linear Discriminant Analysis, 144, 146

LF
Liljencrants-Fant model, 159

LLR
Log-likelihood Ratio, 103, 115, 117, 119, 123, 175, 186, 187, 188, 189, 191, 192, 193, 217

LPCC
Linear Predictive Coding-Derived Cepstral, 53, 147

LR
Likelihood Ratio, 5, 12, 13, 14, 54, 81, 82, 83, 88, 97, 98, 99, 100, 103, 105, 111, 113, 114, 115, 123, 127, 129, 135, 163, 164, 193, 199, 216

LTAS
Long Term Average Spectrum, 46

LVQ
Learning Vector Quantization, 52

MAE
Mucosal Average Energy, 148, 153, 178, 180, 190, 194

MDS
Multidimensional Scaling, 33

MFCC
Mel-frequency Cepstral Coefficients, 12, 52, 133, 147, 154

MPT
Maximum Phonation Time, 44, 45, 65, 76, 77

MVKD
Multivariate Kernel Density, 90, 97, 98, 99, 102, 109, 121, 122, 129, 138, 162, 222, 238, 239

MW
Mucosal Wave, 148, 149, 157, 161, 178, 180, 182, 190, 194

MWC
Mucosal Wave Correlate, 148, 157, 161, 178, 180, 181, 190

MZ
Monozygotic, 1, 2, 10, 11, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 57, 58, 59, 61, 62, 85, 93, 94, 98, 103, 104, 111, 112, 113, 114, 115, 116, 117, 119, 122, 123, 124, 125, 127, 128,

131, 135, 142, 145, 160, 162, 163, 164, 165, 166, 167, 168, 169, 171, 175, 176, 177, 178, 179, 186, 187, 188, 189, 191, 192, 193, 194, 195, 200, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 215, 216, 217, 218, 219, 222, 224

NHR

Noise-to-Harmonics Ratio, 148, 153

PARCOR

PARtial autoCORelation, 150, 159, 160

PRESEEA

Proyecto para el Estudio Sociolingüístico del Español de España y de América, 75, 242

PSD

Power Spectral Density, 11, 148, 149, 154, 155, 156, 157, 161, 178, 180, 182, 190, 194, 221

RAE

Real Academia Española, 68, 105, 106, 107, 108, 242

SFF

Speaking Fundamental Frequency, 34, 36

SL

Strong Likeness, 177

SOSM

Second Order Statistical Measure, 52

SPL

Sound Pressure Level, 77

SPS

Standard Peninsular Spanish, 68, 69

STR

Short Tandem Repeat, 61, 230

SV

Speaker Verification, 53

UCN

Unconstrained Cohort score Normalisation, 54

US

Unrelated Speakers, 28, 29, 30, 85, 98, 104, 111, 112, 113, 114, 115, 116, 117, 119, 120, 122, 123, 128, 175, 176, 177, 179, 186, 187, 189, 191, 193, 194, 203, 204, 205, 211, 216, 217, 218, 219

VFF

Vocal Fundamental Frequency, 36, 41, 47

VoIP

Voice over Internet Protocol, 58, 197, 199

VOT

Voice Onset Time, 38, 40, 41, 43, 47

VS

Vocalic Sequence, 11, 12, 15, 69, 70, 85, 91, 92, 93, 94, 97, 98, 99, 100, 101, 102, 103, 105, 109, 110, 111, 112, 117, 118, 121, 122, 123, 124, 126, 128, 129, 216, 219, 220, 221, 222, 223

WAV

Waveform Audio File Format, 92

WL

Weak Likeness, 177

LIST OF FIGURES

LIST OF TABLES

Ma liberté
Longtemps je t’ai gardée
Comme une perle rare

1. INTRODUCTION

1.1. Introduction to the investigation

The objective of this thesis is to investigate the phonetic-acoustic similarities and differences in three main speaker groups: monozygotic (MZ) twins, dizygotic (DZ) twins and non-twin siblings. From a forensic-phonetic perspective, the study of this type of speakers is highly relevant, as they represent extreme examples of physical similarity. Distinguishing their voices poses a well-recognized challenge in the forensic realm [see Chapter 2]. Yet, there is an interest in this investigation *per se*, as the study of genetically identical speakers (MZ twins) and their comparison with non-genetically-identical siblings (DZ twins and non-twin siblings), on the one hand, and with a reference population of unrelated speakers, on the other hand, allows gaining insight into the contribution of *nurture* and *nature* in the speech patterns of speakers in general. In other words: to what extent is our voice determined by our DNA and to what extent is it due to educational influences? Besides, this study could be considered the first investigation into the phonetic and acoustic characteristics of Spanish-speaking twins and siblings. According to our research objective, a three-folded approach has been undertaken, as will be described below. For the 54 male speakers recorded *ad hoc* for this study, three different analyses have been carried out. On the one hand, we have labeled and analyzed the F1-F3 formant trajectories of 19 Spanish vocalic sequences. Secondly, several naturally sustained [eː] tokens have been extracted from the speakers' spontaneous vowel fillers and their glottal source characteristics have been analyzed. These two approaches have been complemented with an automatic speaker recognition analysis carried out with the software *Batvox*.

The chapter division of this thesis is as follows: In *chapter one* an introduction is aimed at providing a definition of Forensic Phonetics, describing the main tasks or applications of this discipline (among which Forensic Speaker Comparison is included, and especially focused on in this chapter). Some terminology controversies will be outlined, which will explain our preference for Forensic Speaker Comparison (FSC) over other terms. The current methodologies in this field will be described, with a special emphasis in outlining their advantages and disadvantages. Finally, the methodological approach undertaken for this thesis will be broadly explained, together with a summary of the research objectives.

Chapter two will be devoted to the review of the literature. This refers to the phonetic studies on twins' voices, either from a perceptual, acoustical, articulatory or automatic approach. The studies specifically related to the three types of analyses that we have undertaken (formant trajectories, glottal source and *Batvox* automatic analysis) will be briefly described in their corresponding chapter (chapters four, five and six). The literature review on twins' phonetic studies will be preceded by an introduction to the biological bases of twinning and the twin method, and also by a succinct subsection referring to the forensic relevance of twins' voices.

In *chapter three*, the methodological details for carrying out this thesis are described. This includes a description of the main characteristics (age, dialect, etc.) of the different types of participants recruited: 24 MZ twins, 10 DZ twins, 8 non-twin brothers and 12 unrelated speakers (the latter making up the reference population). An *ad hoc* corpus has been designed and collected for this thesis. It includes five speaking tasks and a vocal control technique. Some details about the recording procedure will be presented in this third chapter: material and technical characteristics of the recording, as well as the data collection set-up. Another section will be devoted to explain the use of a telephone filtering for the recordings[1]. Finally, a description follows to explain the likelihood-ratio approach within which the results of the different analyses are offered.

Chapters four, five and six, correspondingly, will provide all the information related to the three different analyses carried out: analysis of formant trajectories of vocalic sequences, glottal source analysis and automatic analysis. Each of these three chapters is divided in the following sections: 1) Objectives and justification, where the research objectives and corresponding hypotheses are set, and the most relevant studies related to the analysis undertaken are reviewed, as a state-of-the-art background; 2) Speech material, analysis tools and method; 3) Parameters, 4) Results, 5) Discussion, 6) Conclusions. In the case of Chapter 5 (glottal source analysis), the description of a pilot experiment is also included.

Finally, *chapter seven* includes a summary of the results from the previous analyses and the main conclusions drawn from them. The chapter ends with the implications, on the one hand, and directions for future research, on the other, derived from this investigation.

1.2. Defining Forensic Phonetics: current methodologies in Forensic Speaker Comparison.

1.2.1. Towards a definition of Forensic Phonetics: terminology controversies

On numerous occasions, a definition of Forensic Phonetics has been attempted (e.g. Jessen, 2008; Künzel, 1994; Nolan, 1983; 1997; Rose, 2002). What the definitions of all these authors may have in common is that they specify for the discipline of Phonetics the general definition of Forensics as "the application of scientific knowledge to legal problems" (*Merriam Webster Online*). Therefore, Forensic Phonetics would be the application of Phonetics aimed at solving any type of

[1] The telephone filtering was carried out only for the fifth speaking task, as it is explained in Chapter 3.

legal issue. Jessen (2008: 671) probably provides the most accurate and complete definition:

> Forensic phonetics is the application of the knowledge, theories and methods of general phonetics to practical tasks that arise out of a context of police work or the presentation of evidence in court, as well as the development of new, specifically forensic-phonetic, knowledge, theories and methods. (Jessen, 2008: 671)

One of the most typical forensic cases where a phonetic expert may be involved is one in which he has to compare the voice of an offender (i.e. speech samples of an unknown speaker) with the voice of a suspect or several suspects (i.e. speech samples of known origin). When referring to this kind of task we will talk about Forensic Speaker Comparison (FSC from now on). Other possible tasks which a phonetician may be requested to perform for forensic purposes are mentioned, for instance, by French (1994: 169), Rose (2002: 2) or French and Stevens (2013):

- Determination of unclear or contested utterances (e.g. in cases when recordings are of poor quality, when the voice is pathological or when the speaker has a foreign accent). This task, together with the task of *transcription*, belongs to what French and Stevens (2013) call *speech content analysis*[2] (cf. French & Stevens, 2013: 183-185).
- Authenticity examinations of audio recordings (e.g. Cicres, 2011).
- Design and validation of voice line-ups; the equivalent for visual identification parades but in the perceptual domain (e.g. Nolan, 2003).
- Speaker profiling (determining the phonetic profile of an unknown speaker on the basis of his voice; i.e. deriving information about the speaker such as gender, age, dialect, etc.). The Language Analysis for the Determination of Origin of Asylum Seekers (LADO)[3] can be considered an application of speaker profiling (French & Stevens, 2013: 186).

All of the above are important areas of interest for the discipline of Forensic Phonetics. For instance, they are recognized by the IAFPA (International Association for Forensic Phonetics and Acoustics), which states the following in its Code of Practice:

> Recognising the varied array of casework subsumed under the interests of IAFPA (eg. speaker identification/elimination, speaker profiling, voice line-ups, transcription, authentication, signal enhancement, sound propagation at crime scenes), Members should maintain awareness of the

[2] For French and Stevens (2013), the forensic task *Speech Content Analysis* involves any "examination of audio recordings to determine what was said". Nevertheless, within this task, they distinguish between (a) *general transcription* and (b) *questioned/disputed utterance analysis*, "where a very specific section of a recording is in dispute" (French & Stevens, 2013: 185). Yet, they add that "the two tasks occupy two ends of a generality-specificity continuum, and are not qualitatively distinct" (French & Stevens, 2013: 184).

[3] As its name indicates, this application of speaker profiling is useful to "assist immigration authorities with determining the nationality of asylum seekers" (French & Stevens, 2013: 186).

> limits of their knowledge and competencies when agreeing to carry out work. (http://www.iafpa.net/code.htm)

Being FSC the most typical forensic-phonetic task (cf. French & Stevens, 2013), and because it is most closely related with the topic of our study, we will explain its purpose in more detail below.

In FSC, a speech recording, or several, is available for an unknown speaker (the *offender*), which can be associated with a crime. If there also exists recorded material of someone suspected to be the same as the unknown person (the *suspect*) or at least this speech recording can be made, then a voice comparison can be carried out. Some of the most frequently described examples found in Forensic Phonetics which relate to FSC are: bomb hoaxes, fraudulent bank deals (Nolan, 2001), kidnappers making ransom demands over the telephone, drug dealers arranging illegal transactions over a (tapped) telephone, or stalking offenses (Jessen, 2008: 673).[4]

Even though it may seem obvious that the comparison of the unknown speech sample (*offender*) and the known speech sample or samples (*suspect/s*) is always carried out with the eventual goal of *identifying* the offender in a criminal case, some terminology controversies have arisen in recent years which cast doubt on the appropriateness of the use of the term "identification".

It could be established that the publication of Saks and Koehler (2005) somehow triggers the terminology change from "forensic speaker identification" to "forensic speaker comparison", at least in some speech scientists. The authors of this article sustain that all forensic sciences should emulate the approach of DNA typing, where a paradigm shift would already have occurred. González-Rodríguez et al. (2007) and Morrison (2009b) are two representative articles where more detailed information can be found about how FSC could move towards a rigorous framework aimed at meeting current admissibility criteria (González-Rodríguez et al., 2007) and where the main characteristics of the new paradigm are summarized (Morrison, 2009b), such as the importance of constructing "databases of sample characteristics and *using* these databases to support a probabilistic approach to identification" (Saks & Kohler, 2005: 893).

For understanding the terminology controversy arising from the opposition between

[4] Other forensic cases such as those where a victim did not see the offender at the time where the crime was perpetrated but claims to recognize his/her voice are also of interest for Forensic Phonetics but would rather fall under the label "naïve speaker recognition" as opposed to "technical speaker recognition" (Nolan, 1983: 7). While the former refer to the "application of our natural abilities as human language users to the identification of a speaker" (Nolan, 2001: 4), the latter entails the "employment of any trained skill or any technologically-supported procedure in the decision-making process" (Nolan, 2001: 4). See also Künzel (1994), who refers to these two types of forensic-phonetic cases as "speaker recognition by non-experts" and "speaker recognition by experts", respectively.

speaker *identification* and speaker *comparison*, it should be noted that those who support the use of "comparison" over "identification" base their argument of the fact that, according to a likelihood-ratio approach, the specific task of the forensic phonetician in a FSC case is offering an answer to the following question: "*How much more likely the magnitude of the difference between samples is if they came from the same speaker than from different speakers"* (Rose, 2002: 89).[5] The answer to that question would be quantitatively expressed as a *likelihood ratio* (LR)[6]. Although in Chapter 3 we define this concept more exhaustively, it is important to note at this point that a LR is an expression of the probability of obtaining the evidence given same- versus different-origin hypotheses and not the probability of the hypotheses given the evidence: "If the forensic scientist were to present the probability of same-origin versus different-origin [hypotheses] and the evidence were potentially incriminatory, then he would be usurping the *rôle* of the trier of fact" (Morrison, 2009b: 300).[7] The reason for supporting the use of "comparison" over "identification" is more explicitly developed in Morrison (2009b: 300):

> A terminological point which arises from the discussion above is that in the likelihood-ratio framework the forensic scientist does not perform "identification" or "individualization", because these terms imply determining a posteriori probability (see Meuwly (2006) on terminological and logical problems with the use of the terms "identification" and "individualization" in forensic science). A neutral term such as "comparison" is more appropriate. (Morrison, 2009b: 300).

Some prior references to the issue of finding "recognition" and "identification" as misnomers can be found in Rose (2002: 87-90) and Rose (2006a: 164). Since the term

[5] This question has also been formulated as: "How much more likely are the observed properties of the known and questioned samples under the hypothesis that the questioned sample has the same origin as the known sample than under the hypothesis that it has a different origin?" (Morrison, 2009b: 299)

[6] A LR is calculated using the following equation: $LR = \frac{P(E|H_{so})}{P(E|H_{do})}$ where E is the evidence, (i.e, "the measured differences between the samples of known and questioned origin" Morrison, 2009b: 299), H_{so} is the same-origin hypothesis, and H_{do} is the different-origin hypothesis.

[7] In order to calculate the probability of same-origin versus different-origin hypotheses (i.e. *posterior odds*), it is necessary to apply Bayes' Theorem. The odds form of Bayes' Theorem is provided in the following equation (Morrison, 2009b): $\frac{P(H_{so}|E)}{P(H_{do}|E)} = \frac{P(E|H_{so})}{P(E|H_{do})} \times \frac{P(H_{so})}{P(H_{do})}$ → Posterior odds = Likelihood Ratio × Prior odds.

Therefore, for the calculation of the *posterior odds*, which is of interest to the court, it is necessary to know the prior odds, and this information is usually not available for the forensic scientist. See Morrison (2009b: 300) for the different interpretations of *prior odds*: "Under one interpretation of Bayes' Theorem, the prior odds would represent the trier of fact's belief in the relative likelihood of the two hypotheses prior to the evidence being presented. Obviously, when conducting their analysis, the forensic scientist cannot know the trier of fact's prior belief. Under another interpretation pragmatic priors can be calculated, e.g., if the crime were committed on an island and there are known to have been 100 people on the island at the time, then pragmatic prior odds could be 1/99; however this would involve the assumption that each person on the island is equally likely to have committed the crime, and although it may be appropriate for the trier of fact to make such an assumption, it is not appropriate for the forensic scientist to do so." (Morrison, 2009b: 300).

"comparison" seems to be widespread nowadays[8], this is the name that we have adopted in this study. Coulthard and Johnson (2007), French and Harrison (2007), Morrison (2009a), Rose and Morrison (2009a), or French et al. (2010) provide a more thorough discussion about this terminology controversy. It should be noted that, apart from the likelihood-ratio approach, a variety of frameworks exists currently for expressing conclusions in a FSC report. See, for instance, Gold and French (2011) or French and Stevens (2013) for an updated international overview of the present situation.

1.2.2. Current methodologies and most frequently analyzed parameters in Forensic Speaker Comparison

In FSC, the recordings of the offender and the suspect/s can be compared using a wide variety of phonetic or acoustic features, and following several different methods. Indeed, this field is characterized by its lack of consensus, not only over the analysis and comparison techniques to use but also over other matters like how conclusions should be expressed (cf. Section 1.2.1). Both Cambier-Langeveld (2007) and Gold and French (2011) document several international practices in FSC, including most commonly analyzed acoustic-phonetic parameters, comparison methods or reporting strategies. Cambier-Langeveld (2007) analyzes the results of a collection of twelve reports carried out by international experts in FSC. As a collaborative exercise, these anonymous experts were provided with the same audio materials and asked to submit a report on a fake case as if it was a real case, exactly as they would report to a regular costumer (Cambier-Langeveld, 2007: 228). As regards the different methods employed, Cambier-Langeveld (2007) noted the existence of three basic subgroups of methods: auditory-acoustic, semi-automatic, and automatic, even though she admits that this may be an inappropriate simplification[9]. In Gold and French (2011), a different classification of methods is proposed. They distinguish between: 1) auditory phonetic analysis only; 2) acoustic phonetic analysis only; 3) auditory phonetic cum acoustic phonetic analysis; 4) analysis by automatic speaker recognition (ASR) systems; and 5) analysis by automatic speaker recognition systems with human analysis. It is important to note that in the survey carried out by Gold and French (2011), in which thirty-six international FSC experts participated, none of them reported using the fourth type of method: an ASR system alone. The use of an ASR system would be accompanied by what they call *human analysis* (fifth type of FSC method in their own classification), which involves "the use of an automatic system in conjunction

[8] The *Position Statement* in French and Harrison (2007), signed by nine researchers and with several more co-signatories, suggest the replacement of *identification* by *comparison*: "It will be apparent from the arguments developed here that the term FSI should be replaced by FSC" (French & Harrison, 2007: 144).

[9] "Perhaps referring to the field of forensic speaker identification as one discipline in which three subgroups of methods are employed, as I have done so far, is inappropriate; one might equally well speak of several disciplines approaching the same problem (that of forensic speaker identification) from different angles". (Cambier-Langeveld, 2007: 240)

with analysis of the auditory and/or acoustic phonetic kind" (Gold & French, 2011: 296). Several descriptions of the auditory-acoustic phonetic approach can be found, for instance, in Jessen (2008), Künzel (2011), Nolan (1997), or Rose (2002, 2006a).

As far as the most frequent measurements are concerned, Cambier-Langeveld (2007) lists fundamental frequency and formant frequencies as the two most repeated parameters included in the reports – considering the semi-automatic and auditory-acoustic analyses-, the third most frequent measurement being "speech rate", "speaking rate" or "articulation rate". As the author indicates in the conclusions of this collaborative exercise, this is a type of measurement for which different definitions can be found, depending on the expert. Other important questions arise from this study, such as whether pauses should be analyzed separately or as part of "what we perceive as speech rate" (Cambier-Langeveld, 2007: 241). In comparison, the aim of Gold and French (2011) is not only spotting the speech features (linguistic and non-linguistic) that are most widely used in forensic casework but also finding which weighting is attached to those parameters by different experts, i.e. which of those parameters are considered to show the greatest potential for speaker discrimination. Firstly, they distinguish between phonetic features and non-phonetic features. In the first group, a distinction is made between segmental and suprasegmental features. The non-phonetic features would include what they call *higher order linguistic features* (e.g. discourse markers, aspects of turn-taking, lexical features and lexico-gramatical usage) and *non-linguistic features* (e.g. filled pauses, tongue clicking, audible breathing, throat clearing and laughter).

Another common, traditional division of parameters useful for FSC is proposed by some authors who distinguish between *high-level features* and *low-level features* (cf. Kinnunen & Li, 2010; Künzel & Alexander, 2014). The first group would refer to well-known linguistic characteristics of a speaker such as dialect (or 'regional coloring' in Künzel & Alexander, 2014: 244); sociolect, jargon, intonation patterns or pausing behavior, but these high-level features also refer to conversational and lexical aspects such as the use of frequent expressions ("you know", "oh yeah", etc.) or specific word choices. According to Kinnunen and Li (2010: 4), the pioneering investigations on this research line date back to Doddington (2001) who studied how a speaker's idiolect (understood as characteristic vocabulary) could be used to individualize him. In contrast, the second group of parameters (low-level features) would comprise basically short-term acoustic features of the spectrum or cepstrum. In other words, these are essentially the set of cepstral coefficients in which automatic systems are usually based on, and which are characteristic of the resonance behavior of the vocal tract [see Chapter 6].

As regards the most frequent acoustic measures, if we first consider the segmental features, it seems that formant values (mainly F2) are commonly analyzed in the experts' reports, being the centre frequencies of formants in monophthongs more frequently examined than the formant trajectories of diphthongs. Gold and French (2011: 300-301) discuss the percentage of experts examining other type of vocalic measures, like formant bandwidth or formant densities, as well as the most frequent acoustic measurements in relation to consonants. As concerns the suprasegmental features, all experts participating in the above-mentioned survey agree in measuring fundamental frequency, although the specific values (mean, median, mode, etc.) being measured may vary from one expert to another. A great majority of the experts seem to include also voice quality examinations in their FSC procedures, as well as intonation and tempo. As to the question of which parameter was usually more relevant for discriminating speakers, Gold and French (2011: 302) found the following results across all the participating experts:

> For all respondents together, voice quality was reported most often (32%), followed by rhythm (16%). Lexical and grammatical choices, vowel and consonant realizations, phonological processes (e.g. connected speech processes) and fluency were all reported by 13% of the respondents. (Gold & French, 2011: 302)

It can be concluded, then, that nowadays there is a lack of methodological consensus in FSC, and that the acoustic parameters mostly used by international experts are attached different importance by each of them. Despite the fact that the parameters examined usually depend on the specific characteristics (e.g. duration, quality, etc.) of the recordings under comparison, the idea that the more parameters and approaches considered, the more complete the comparison procedure will be has traditionally been accepted in FSC. Delgado (2001), Künzel and González-Rodríguez (2003) or Künzel (2011) are only some of the publications were this hybrid methodology is recommended. Indeed, a combined perspective to FSC seems to be the standard practice in this discipline. A good proof of this is the information related to the distribution of current methods provided by Gold and French (2011: 296). According to their survey, the "Auditory Phonetic cum Acoustic Phonetic Analysis" method is used in most countries, followed by the "Automatic Speaker Recognition System with Human Analysis" method, which we could describe as the most complete one, since it combines the advantages of the ASR method with those of the auditory-acoustic[10] approach, also called "traditional method" by some authors (Künzel, 2011). In next section of this chapter, we present the main advantages and disadvantages of the use of traditional and automatic methods[11].

[10] Actually this method could more accurately be described as "phonetic-acoustic-linguistic", as it also makes use of linguistic cues other that phonetic-acoustic ones. This is the name given by Künzel (2011: 38): "traditional phonetic-acoustic-linguistic method".

[11] For a review of the specific parameters considered and the methodological approach followed by the Forensic Acoustic Laboratory of the Spanish Scientific Police, see Delgado et al. (2009).

A final aspect that we would like to highlight in this section is related to the ideal characteristics that a parameter should have for FSC. Wolf (1972) set out some criteria for selecting a forensic-phonetic parameter and since then, other authors, with more or less variations, have repeated these criteria. In Table 1, we include the six basic criteria already established by Wolf (1972) and redefined by Nolan (1983).

Among all of the aspects that a robust forensic-phonetic parameter should fulfill, the two first in Nolan's (1983) list are probably the most relevant, as their forensic value has been repeated in many publications thereafter[12], the first criterion sometimes with more emphasis than the second:

> The sound properties that a perceptual or acoustic analysis should focus on are those that are known to be subject to large differences between speakers, i.e. have large 'interspeaker variation' (also referred to as between-speaker variation). (Jessen, 2008: 687)
>
> The acoustic properties of speech will be useful forensically to the extent that they have relatively large between-speaker variation and relatively small within-speaker variation. (Morrison, 2010a: 6054)

Table 1

Criteria for selecting a parameter for FSC, according to Wolf (1972) and Nolan (1983)

Wolf (1972: 2044)	Nolan (1983: 11)
"It should vary as much as possible among speakers".	*High between-speaker variability*: "the parameter needs to exhibit a high degree of variation from one speaker to another".
"It should be as consistent as possible for each speaker".	*Low within-speaker variability*: "*it* will have to show consistency throughout the utterances of an individual; and preferably be insensitive to his state of health, emotional condition, or the communicational context".
"It should not change over time or be affected by the speaker's health".	
"It should not be modifiable by conscious effort of the speaker, or, at least, be unlikely to be affected by attempts to disguise the voice".	*Resistance to attempted disguise or mimicry*: "the parameter needs to withstand attempts on the part of the speaker to disguise his voice or mimic that of another, either by virtue of being the acoustic consequence of a physiological characteristic of the speaker which he is not able to alter at will, or by being in some way a "less obvious" attribute of speech which escape his attention during attempts at disguise or mimicry".
"It should occur naturally and frequently in normal speech".	*Availability*: "it is of little use basing speaker recognition on a parameter which occurs only

[12] The fact that the rest of the criteria are not mentioned so often does not imply that they are not equally important for forensic research and practice.

	seldom in speech and therefore necessitates large amount of data in both test and reference corpora”
“It should not be affected by reasonable background noise nor depend on specific transmission characteristics”.	*Robustness in transmission*: “the usefulness of a parameter will be limited if its information is lost or reduced in telephone transmission or tape recording”.
“It should be easily measurable”.	*Measurability*: “the extraction of the parameter in question must not be prohibitively difficult”.

As it will be explained in Chapter 3, all speakers in our study were recorded twice, on non-contemporaneous sessions, separated by a time lapse of two-three weeks. We have thus taken into account intra-speaker variation.

The six basic criteria summarized in Table 1 also appear in the literature specifically related to automatic speaker recognition (ASR). For instance, Kinnunen and Li (2010) refer to the same characteristics for an ideal parameter. Interestingly, they add that “the number of features should be also relatively low” (Kinnunen & Li, 2010: 3), since apparently this would reduce a known problem in ASR, the so-called “curse of dimensionality” (Jain, Duin & Mao, 2000):

> Traditional statistical models such as the Gaussian mixture model (Reynolds, Quatieri & Dunn, 2000; Reynolds & Rose, 1995) cannot handle high-dimensional data. The number of required training samples for reliable density estimation grows exponentially with the number of features. This problem is known as *the curse of dimensionality* (Jain, Duin & Mao, 2000). The computational savings are also obvious with low-dimensional features. (Kinnunen & Li, 2010:3)

1.3. Methodological approach of this investigation

In this thesis we have undertaken a three-folded approach to the voice and speech analysis of the participating speakers. In all the three analyses, after the tokens’ labeling and/or extraction, and after the acoustic measurements, we proceeded to obtain speaker comparisons of the following types: a) intra-speaker, b) intra-pair, and c) inter-speaker. The first type of comparison was carried out for all speakers, comparing their first recording session with the second one. Intra-pair comparisons were made for MZ, DZ and non-twin siblings, while we considered inter-speaker comparisons those which implied the comparison of unrelated speakers. Yet, it should be noted that intra-pair comparisons are *sensu stricto* also inter-speaker comparisons.

Firstly, we carried out an acoustic analysis of all the speakers’ formant trajectories for 19 Spanish vocalic sequences (VS) and for three formants (F1, F2 and F3). As it is detailed in chapter four, after the tokens’ labeling and measurement, the trajectories were fitted using two types of curve fitting methods (parametric representations): polynomials and discrete cosine transforms (DCTs). The coefficients of this curve fitting were used as input for the forensic comparison procedure described above. For this analysis, the speech material was extracted from the second speaking task in the *ad hoc* corpus described in Chapter 3.

In the second type of analysis, several tokens of the Spanish vowel filler [eː] were extracted per speaker and recording session. From these speech units, considered long enough for a glottal analysis, a vector of 68 parameters was created, comprising seven different feature subgroups: 1) f_0 and distortion parameters; 2) cepstral coefficients of the glottal source power spectral density (PSD); 3) singularities of the glottal source PSD; 4) biomechanical estimates of vocal fold mass, tension and losses; 5) time-based glottal source coefficients; 6) glottal gap (closure) coefficients; and 7) tremor (cyclic) coefficients. As in the first analysis described, a forensic comparison followed the voice analysis. In this case, the vowel fillers were extracted from the fifth speaking task, which was especially aimed at eliciting speech with hesitation.

Finally, the third type of analysis was carried out using the automatic speaker recognition system *Batvox 4*, which is based on parameters related to the resonances of the vocal tract, as it will be explained below, and in more detail in Chapter 6. For this analysis, around 120 seconds of net speech[13] were extracted, per speaker and session, also from the fifth speaking task[14].

As we already explained in Section 1.2, there are many different parameters which are used in forensic casework and studied by international researchers nowadays, but no consensus is reached as to which is most useful or robust in FSC. Yet there is some agreement that a single parameter is not usually enough to discriminate between speakers, and that a hybrid approach is the most appropriate methodology. Taking this into account, our investigation is aimed at combining three very different analyses to achieve a thorough understanding of the similarities and differences between the participating MZ, DZ and non-twin siblings. The analyses proposed here basically differ in *where* the speaker-characteristic properties are extracted but also in *how* it is done. In the first approach, the formant trajectories of the vocalic sequences (VS) under study most clearly depend on the vocal tract configuration of the speaker. The second approach is based only on glottal-source information, as the vocal tract decoupling assures that there is no interfering of vocal tract characteristics. Finally, the automatic analysis carried out with *Batvox 4* would be also based on the vocal tract but in a different way as in the case of formant trajectories. In the case of VS formant trajectories, these are expected to reflect not only the morphology of the speaker's vocal tract but also his personal choice for the articulatory achievement of two vocalic

[13] With "net speech" we refer to the valid speech sample after having removed extraneous or undesired noise, like laughter, clicks, etc., according to the recommendations in Künzel (2011: 256).

[14] The remaining speaking tasks will also be described in Chapter 3. Even though they have not been used specifically to extract the speech material analyzed in this thesis, their design and collection were deemed useful in order to obtain a more complete corpus of twins' and non-twins' voices, which could serve as reference material –for instance, the reading of phonetically balanced texts in the third speaking task– and which could be used in future studies.

targets, according to studies like McDougall (2006). An automatic system, on the other hand, extracts a set of features representing the resonance profile (i.e. the Mel Frequency Cepstral Coefficients, MFCCs) but these features are not interpretable in the same sense that formants are. As stated by Jessen (2008: 699-700), "very little is known about how specifics of MFCC information relate to specifics of speech production and perception". In Rose (2002), a more detailed explanation of the differences between formants and cepstral coefficients can be found, as well as a non-technical description of the *cepstrum* and its forensic significance. More information about the cepstral analysis will also be presented in Chapter 6.

Thus, placing this thesis within the hybrid methodology widely extended in FSC, the benefits of undertaking a multifaceted approach are manifold, as we will try to explain in the rest of the chapter.

The main advantage of the traditional acoustic-phonetic-linguistic method is that it makes use of the so-called "natural speech" parameters (Künzel, 1994: 140; Künzel, 2011: 50), such as f_0 (high-pitched and low-pitched voice), or number and duration of silent intervals, *inter alia*. These are easy concepts to explain to non-experts, like lawyers or judges, as they "relate in a direct way to aspects of speech production" (Rose, 2006: 173). In contrast, the features used in automatic systems are abstract and therefore more difficult to explain to non-experts. But it is evident that the traditional method also presents disadvantages, being the most frequently highlighted (Künzel, 2011: 39) the following ones: 1) the *subjectivity* implicit in the scientist's choice of parameters due to his educational background and professional experience; 2) the *working time* which each comparison case requires, in view of the necessity to analyze many phonetic features (not to mention the time required by the careful listening of recordings for the auditory –e.g. voice-quality– analysis).

In relation to the automatic systems, we should first note that, although in this kind of methods the results of the comparisons are expressed in LRs, the advantages of these systems should not be confused with the advantages of using a LR-based approach. As explained in Künzel (2011: 41), the general acoustic principle of the automatic method has important advantages on its own. For instance, it allows the comparison of the voices of speakers regardless of the language they speak. This is very useful in real forensic casework where multilingual speakers are involved, for instance in organized crime or terrorism. Künzel (2011: 41-42) describes an example of such a case. As compared with the traditional method, the automatic method offers here a clear advantage, as the forensic expert needn't have a high level of proficiency in the language of the speakers being compared. Besides, automatic methods are very important, if not really necessary, in cases involving a great amount of recordings or speakers. The use of this kind of methods does not require as much time as the traditional method would[15].

The advantages of using a LR-based approach are also manifold: 1) it implies that the scientist does not give a verdict on identification (as explained in Section 1.2.1, the *a posteriori* probability should not be expressed by the forensic expert), since that responsibility is the role of the trier of fact; 2) another important advantage of using an LR-based approach is that the LR from the voice comparison can be combined with the LRs resulting from other analyses or evidences from the same forensic case, like footwear evidence, fingerprints, etc. It is essential to highlight that LRs are not only yielded by automatic systems but can also be obtained from traditional parameters, as explained by Rose (2006a: 173). Hence the importance of separating the advantages of a LR-based approach from the advantages of automatic systems.

Nevertheless, methods based on LRs are not free of problems. First of all, there is no agreement as to whether verbal scales should be included in forensic reports accompanying LR numeric values. Champod and Evett (2000: 240) propose some verbal equivalents for LRs. For instance, if a LR is between 100 and 1000, it would be described as giving "moderately strong" support for the prosecution hypothesis[16]. However, as explained by Rose (2006a: 167), "neither the verbal equivalents nor their use is universal; for Royall (2000: 760), for example, LRs of 8 and 32 count as 'fairly strong' and 'very strong', respectively." Actually, the grades proposed by Evett (1998: 201) are as follows:

1 < LR > 10 → Limited support
10 < LR > 100 → Moderate support
100 < LR > 1000 → Strong support
LR > 1000 → Very strong support

In Künzel (2010: 274, note 5) we find that "Evett (1998) recommends that in all forensic disciplines except DNA 'we must use linguistic qualifiers' to indicate to the court the level of support that a LR gives to the stated propositions". However, not all the scientists agree in using such scales. For instance, Künzel (2010: 275) considers that on the basis of his personal experience with ASR systems, he "would not consent to use such a scale generally, that is in all cases, until a more quantifiable and objective way to assess the fitness of a reference population to a given case has been found". In the case of Rose (2006a: 167), he elaborates on another

[15] As explained in Künzel (2011: 49), "assuming that the technical requirements are met (e.g. having appropriate reference populations), an automatic system takes less than one minute to calculate a result" (our translation).

[16] Yet, in Evett (1998: 201), a LR between 100 and 1000 would indicate "strong support".

criticism received by the opponents of the verbal equivalents for the LRs, namely the circularity of its use:

> Their use (of LRs) can be criticized as circular: in response to the claim that the evidence gives 'strong support' to the hypothesis it can be enquired what is meant by 'strong support', the only real response to which involves reference to the original LR (Rose, 2003: 2055, in Rose, 2006: 167).

Other shortcomings associated with the use of LRs and the Bayesian evaluations of evidence are: 1) the difficulty for accepting or understanding that a LR indicating "strong support" for a hypothesis can be overturned when the prior odds[17] are considered (cf. Rose, 2006); 2) the lack of a straightforward understanding of Bayesian inference by the court. To this, Rose (2006a: 168) adds the disadvantage that "LRs are all too easily transposed into probabilities of hypothesis given the evidence". Yet, again in Rose (2006a: 168), some references are mentioned of investigations showing that "human minds are capable of Bayesian evaluation, providing that the wording is carefully chosen and refers to incidence ("out of 100 people, 3 will have this disease") rather than probability ("there is a 3% probability of this disease")" (Gigerenzer, 2003; Gigerenzer & Hoffrage, 1995; Pinker, 1997; in Rose, 2006: 168).

In general, from what has been explained in this section, we can conclude that both approaches (traditional and automatic) complement each other, as the advantages of one add to the advantages of the other, and at the same time, their respective strengths compensates their weaknesses. For instance, the easy interpretability of the parameters in the traditional approach can somehow compensate for its limited discriminant power (as compared with the features used in automatic systems), especially if a hybrid approach is adopted, where also an automatic system is complementarily used. Rose (2006a: 173) explains that trade-off between both approaches as follows:

> The distinction between traditional and automatic features is important, since it reflects a tension between interpretability and discriminant power: traditional features have much greater interpretability – more *Anschaulichkeit*- which is a bonus for explanations and justifying methodology in court. Automatic features, on the other hand, are very much powerful as evidence: they will, on average, yield likelihood ratios that deviate much more from unity. (Rose, 2006: 173)

We have tried in this section to show the advantages and disadvantages of the traditional and automatic methods in FSC, as our investigation draws on both types. The first approach, relying on the formant dynamic characteristics of VS, is the one which most clearly fits in the traditional method, both in the time-consuming procedure of extraction and acoustic analysis, and

[17] Nevertheless, as pointed out by Rose (2006: 168), "it is in fact sometimes the case that the prior odds are ignored by the court –whether by commission or omission is not clear." (Rose, 2006: 168).

in the self-explanatory nature of these parameters. In the case of the glottal analysis, it stands at the crossroads of the traditional and automatic methods, as it combines more *traditional* parameters, such as jitter and shimmer, with more abstract ones, such as the cepstral coefficients of the glottal source power spectral density. Finally, the third type of analysis, which clearly belongs to the automatic methods, complements the other analyses, giving thus a complete picture of what the voice of the participating speakers look like from different angles.

1.4. Outline of research objectives

The *general objective* of this thesis is investigating the phonetic characteristics of three main speaker groups: MZ, DZ and non-twin siblings. Their phonetic-acoustic similarities and differences would be studied in relation to a reference population of unrelated speakers. The reasons why this global objective is of research relevance from a forensic perspective have to do with the fact that the above-mentioned speakers are considered very similar as far as their voice and speech patterns are concerned. Research on this type of speakers allows therefore testing the performance of a forensic-comparison system, since a robust system –and by extension, the parameters in which such system is based– should be able to distinguish between very similar speakers. This is why the study of twins has traditionally been deemed to challenge this discipline in a constructive way [see Chapter 2 for a literature review of twins' studies].

Nonetheless, the interest of this investigation is not limited to the forensic realm, as stated at the beginning of this introduction [see Section 1.1]. It is not uncommon in the literature to consider the comparative study of MZ and DZ twins –to which we have added non-twin siblings and unrelated speakers– as a valuable resource for understanding the interplay of genetic factors and non-genetic factors influencing certain features. This does not only apply to voice; on the contrary, this type of comparative studies exists in various other disciplines to test the relative weighting of *nature* and *nurture* in several possible features. We are referring to the use of the classical twin method, which will be described in more detail in the second chapter [see Section 2.2].

According to what has been just explained, the *first* and *second* objectives of this investigation –derived from the general objective: investigating the phonetic characteristics of MZ, DZ and non-twin siblings– could be outlined as:

1) Investigating the robustness of certain voice features for forensic purposes by means of comparing their performance in MZ twins, DZ twins, non-twin siblings and unrelated speakers.

2) Investigating –for the proposed voice parameters– how the comparison results vary according to the type of speakers considered: MZ twins, DZ twins, non-twin siblings and unrelated speakers. In other words, investigating the degree of genetical influence of the analyzed voice features.

As will become apparent when explaining the hypotheses in Chapter 3, objectives one and two are strongly linked together since we consider that a parameter or set of parameters found to be genetically related will be robust for speaker comparison. We split them in two different objectives, as the first one is more clearly related to the forensic application of this study while the second one would simply refer to the nature-nurture dichotomy –trying to shed light on this topic– and it is of interest *per se,* regardless of its forensic usefulness.

Taking into account that this investigation is based on three types of analyses (Chapters 4, 5 and 6, correspondingly), the general objectives that we have just mentioned could be specified as follows:

- *Testing the forensic validity of formant trajectories extracted from Spanish vocalic sequences* [see Chapter 4]. This general objective can be split into the following specific or secondary objectives:

  [Chapter 4 – Objective 1]: *Testing whether there is higher intra-pair similarity for this kind of parameters in MZ twins than in other speaker comparisons (DZ, B or US). This would imply that the parameters are genetically influenced and would therefore be useful in a typical forensic context.*

  [Chapter 4 – Objective 2]: *Testing whether the fusion of the scores obtained for all the vocalic sequences (VS) outperform the individual systems based on single VS.*

  [Chapter 4 – Objective 3]: *Testing whether certain procedures for parameter curve fitting of the formant trajectories outperform the others.*

- *Testing the discriminatory power of a series of glottal features extracted from Spanish vowel fillers* [see Chapter 5]. This general objective can be split into the following specific or secondary objectives:

  [Chapter 5 – Objective 1]: *Testing whether there is higher intra-pair similarity for this kind of parameters in MZ twins than in other speaker comparisons.*

  [Chapter 5 – Objective 2]: *Testing whether some glottal parameters yield better identification results than others.*

- *Testing the performance of the automatic system Batvox 4 for discriminating MZ, DZ and non-twin siblings* [see Chapter 6]. This general objective can be formulated as:

  [Chapter 5 – Objective 1]: *Testing whether there is higher intra-pair similarity for this kind of parameters in MZ twins than in other speaker comparisons.*

2. REVIEW OF THE LITERATURE: TWIN STUDIES

2.1. Introduction

The structure of this chapter is as follows: First, we introduce the basic types of human twins and their most relevant genetic and environmental aspects. Besides, the validity of including also non-twin siblings in our study is considered, as regards the nature-nurture dichotomy. In a second section, we will try to acknowledge the importance of research on twins' voices for forensic purposes. We do so, on the one hand, by reviewing the existing phonetic studies which, not being twins its main research goal, do mention at some point the singularity of twins' voices and its potential forensic relevance. On the other hand, we gather together some twin-related forensic casework in which national and international voice experts have been involved. Thirdly, a more thorough literature review is undertaken in which several studies on twins' voices are not only summarized but also critically described. On this occasion, the criterion for publications' selection has been that the studies should focus on twins' voices as the main goal of the research[18]. Since we are especially interested in the forensic application of twins' voice research (hence, adult twins), studies related to child language acquisition have not been reviewed.

2.2. The biological bases of twinning and the twin method

Roughly speaking, there are two basic types of twins (see Figure 1):

- **Monozygotic (MZ) twins**, also called identical twins, occur when a single ovum is fertilized by a sperm cell to form one zygote, which then divides in two. The members of MZ twin pairs share all their genes in common (Segal, 1990:612) and, with rare exceptions (Vandenberg 1966; Dallapicolla et al., 1985; in Segal, 1990: 612), they are always of the same sex.
- **Dizygotic (DZ) twins**, non-identical twins, or fraternal twins, occur when two separate ova in the same menstrual cycle are independently fertilized by two different sperm cells: "The result is two zygotes, each of which develops its own placenta and amniotic sac" (Stromswold, 2006:

[18] The works under review are therefore quite diverse, since the issue of twins' voices is often tackled from very different perspectives. Actually, the study of voice itself is considered from different points of view and by varied experts (speech therapists, engineers, phoneticians and so on). Therefore, due to the multidisciplinarity of voice research in general, some terminology discrepancies may exist between different studies for same or similar parameters, like "pitch" and "f_0", or "intensity" and "loudness", which might be used by different authors to refer to the same concept.

336-337). DZ twins, as well as full siblings, share 50% of their genes, on average, by descent (Segal 1990: 612). [19]

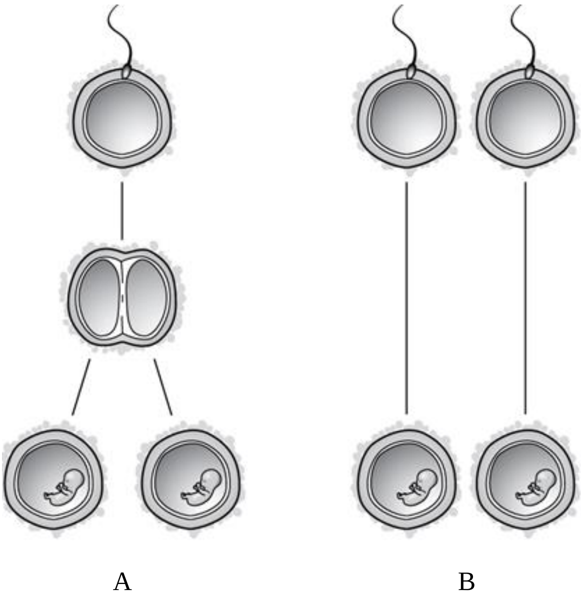


Figure 1. Basic classification of twins in (A) MZ twins and (B) DZ twins. Adapted from "Dizygotic Twins – Twins, Triplets, and More". Netplaces.com. Retrieved 30 October 2013. (Fierro, 2013)

A more accurate classification of twins would take into account that there are a few subgroups within MZ twins. While DZ twins are always dichorionic-diamniotic (i.e. each zygote develops its own placenta[20] and amniotic sac), MZ twins may be of three types (see Figure 2):

- **Dichorionic-diamniotic MZ twins** have two different placentas and two different amniotic sacs. This type of MZ twinning accounts for about 20-25% of all MZ twins. They occur when the zygote splits during the first three days following fertilization (Stromswold, 2006: 337).

- **Monochorionic-diamnotic MZ twins** have one shared placenta and two different amniotic sacs. This type of MZ twinning is the most common: it accounts for about 70-75% of MZ twins. According to Stromswold (2006: 237), "they occur when the inner cell mass splits after blastocyst formation but before the formation of the amniotic sac (at 8 days after fertilization)".

- **Monochorionic-monoamniotic MZ twins** have one shared placenta and one shared amniotic sac. This is the rarest form of twins, accounting for only 1-5% of all MZ twins. They result

[19] The theoretical range of shared genes for same-sex pairs is 0%-100%, but a more realistic range is 25%-75% (Pakstis et al., 1972; in Segal, 1990: 612).

[20] Note however that, according to Stromswold (2006: 338), DZ may have fused placentas (0-5% of DZ twins) or unfused placentas (95-100% of DZ twins). See Figure 2.

when "division occurs after the formation of the amniotic sac, but before the establishment of the embryonic axis (at about 15 days after fertilization)" (Stromswold, 2006: 337).

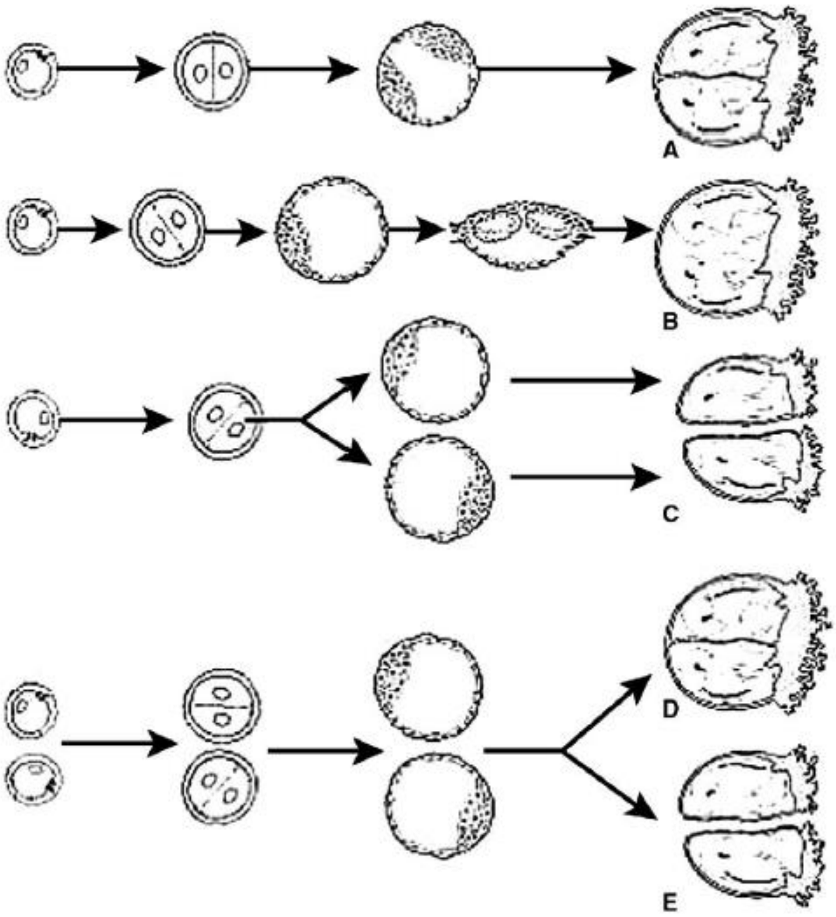


Figure 2. Extended classification of twins (cf. Stromswold 2006). Letters A, B and C represent MZ types and letters D and E represent DZ types: (A) monochorionic, diamnotic MZ twins, (B) monochorionic, monoamniotic MZ twins, (C) dichorionic, diamnotic MZ twins with separate placentas, (D) DZ twins with fused placentas, and (E) DZ twins with unfused placentas.

This classification is of utmost importance since the existence of distinct types of MZ twins may have *genetic* and (perinatal)[21] *environmental* implications (Stromswold, 2006: 337). When studying twins, two basic concepts appear, usually as opposite terms: genes and environment. Indeed they are rather intermingled. The *twin method*, used in most twin studies, tries to "provide a useful indication of the relative contribution of genetic and environmental factors on individual differences in measured traits" (Haworth, Asbury, Dale & Plomin, 2011: 1).[22] These authors provide a well-known definition of the twin method: "The twin method uses

[21] The term *perinatal* is used by Stromswold (2006) to refer to "the period that begins with the implantation of the embryo and ends at 44 weeks gestation", while *postnata*l refers to any time after that period.

[22] Hence, scientists sometimes refer to this as the "nature-nurture dichotomy", first outlined by Sir Francis Galton in 1875: "It is, that their history affords means of distinguishing between the effects of tendencies received at birth, and those that were imposed by the circumstances of their after lives; in other words, between the effects of nature and nurture" (Galton, 1875, in Segal 1993: 45). In the phonetic realm, this

MZ and DZ twin intraclass correlations to dissect phenotypic variance into genetic and environmental sources” (Plomin, DeFries, McClearn & McGuffin, 2008, in Haworth, Asbury, Dale & Plomin, 2011: 1). The twin research methodology offers several design variations of the classic twin method (see Table 2).

Table 2

Twin Research Designs

Classic Twin Study: MZ and DZ Twins Reared Together
Cotwin Control Studies
Singleton Twins
DZ Twin Studies
Longitudinal Twin Studies
The Twin-Family Design
Twins as Couples
Twins and Nontwins
Partially Reared Apart Twins
Twins Reared Apart

Note. Classification extracted from Segal (1990: 612-3).

The experimental design in our thesis essentially follows the classic method but it also draws on the following designs: “Twins as Couples”, “Twins and Nontwins” and “Partially Reared Apart Twins”.

- **Classic Twin Study (MZ and DZ Twins Reared Together).** The classic twin method compares the resemblance within identical twin pairs to the resemblance within fraternal twin pairs, assuming equal environment influences for both types of twins. Hence, “greater resemblance within MZ twin pairs, relative to DZ twin pairs, is consistent with (although not proof of) a genetic explanation for the trait under investigation” (Segal, 1990: 613). The rationale of the classic twin method underlies the different twin methodologies: “Differences within MZ twin pairs are explained by environmental effects because all genetic inheritance is commonly shared. In contrast, differences within

would be translated as the “organic-learned dichotomy”, as stated by Nolan (1996: 39): “The differences between voices are often broadly categorized as either ‘organic’ or ‘learned’. This definition implies that some aspects of personal voice quality are determined by our anatomical inheritance and others by what we either copy from people around us or choose arbitrarily in order to mark our personality. The organic-learned dichotomy needs elaboration if a full understanding of the bases of speaker differences are to be understood but it is nonetheless a useful conceptual starting point”.

DZ twin pairs are associated with both genetic and environmental influences because these twins share half their genes, on average, by descent" (Segal, 1990: 613). The discovery of the twin method has traditionally been credited to Francis Galton's 1875 article on twins, entitled "The History of Twins, as a Criterion of the Relative Powers of Nature and Nurture" (Galton, 1875). However, he would have never actually proposed that the resemblance of identical twins be compared to the resemblance of fraternal pairs in order to assess genetic influence[23], which is the essence of the classic twin method (Rende, Plomin & Vandenberg, 1990). According to these authors, the first description of the method appeared 50 years after Galton's paper, in Merriman (1924) and Siemens (1924).

- **Twins as Couples.** To a certain extent, our thesis also follows this design insofar as we have taken into account the *couple effect*, defined as "the varying functional roles assumed by the members of MZ twinships because of their social interdependence" (Zazzo, 1978, in Segal, 1990: 613). This means that, since some twins behave differently when interacting with their cotwins, as compared with acting alone, we have recorded cotwins speaking to each other in different communication situations as well as reading out alone and holding a conversation with the researcher.

- **Twins and Nontwins.** This variation of the classic twin method consists in the comparison between twins and non-twins across different measures. Its interest lies in the fact that these comparisons "may highlight the effects of the unique biological and psychological aspects of twinship" (Segal, 1990: 614). The term *non-twins* may include the singleton siblings of twins, or either just sibling pairs who are close in age or pairs of unrelated, age-matched singletons. For our study, besides twins we have recruited both siblings and singletons.

- **Partially Reared Apart Twins.** This kind of design involves comparisons of twins who have lived apart for a certain period of time with twins who have always lived together. For our thesis, examples of twins of both types are available for comparison, especially since both older and younger twins have participated in our study: "Twins of various ages may be selected to determine if the living situation has a more important impact on resemblance during the early or later years" (Segal, 1990: 614).

So far we have outlined the basics of the twin method, used widely in different disciplines like Medicine or Psychology. However, twin designs are scarcely used in disciplines like

[23] Yet he first acknowledged the existence of two different types of twins (Rende, Plomin & Vandenberg, 1990).

Linguistics, except for the study of some speech disorders (e.g. Grigorenko, 2009) or in speech acquisition research. Besides, in this last field it is usually agreed that "the speech of MZ and DZ twins has never been analyzed qualitatively and systematically compared" (Locke, 1989: 554). This author's study departs from the hypothesis that "greater concordance of speech articulation is expected between MZ than DZ twins" (Locke, 1989: 555) and bases his hypothesis on the following reasoning:

> Since the morphology of these physical systems [*e.g. vocal tract and respiratory system*] is genetically transmitted, it is reasonable to hypothesize that specific sound production patterns, which reflect the physical characteristics of these production systems, are more alike in children whose genetic endowment is similar (Locke, 1989: 555).

Stromswold (2006:334) explains the logic of twin studies to investigate the impact of genetic factors on language as follows:

> Identical (monozygotic, MZ) twin pairs and fraternal (dizygotic, DZ) twin pairs share essentially the same pre- and postnatal environment, whereas MZ twins share 100% of their alleles and, on average, DZ twins share only 50% of alleles. Therefore, if MZ twin pairs' linguistic abilities are more similar than DZ twin pairs', this suggests that genetic factors play a role in language. One way to determine whether MZ twins are linguistically more similar than DZ twins is to compare the concordance rates for language disorders in MZ and DZ twin pairs.

The heritability[24] estimates obtained by Stromswold indicate that MZ cotwins are more linguistically similar than DZ cotwins. However, estimates of the role of genetic factors rarely exceed 60% and some MZ cotwins have measurably different linguistic abilities. Having noted that some MZ twin pairs are discordant for language impairments, Stromwold wondered why this happens if they really have the same genetic and environmental endowments. She discusses how genetic, epigenetic[25] and perinatal environmental factors can lower heritability estimates for language causing MZ twins to be linguistically discordant.

So, even though it is agreed that MZ twins share the 100% of their genes while DZ twins share only half their genetic information, some genetic differences may exist between the different types of MZ twins. This is important since "the validity of heritability estimates obtained from

[24] For a definition of the concept of "heritability" see Tomblin and Buckwalter (1998: 188-189): "If a continuous trait such as language achievement or a dichotomous trait such as developmental language disorder (DLD) is genetically influenced, there should be a greater similarity on the trait between MZ twins than DZ twins. This elevated similarity in MZs over DZs for a continuous trait is often reported in terms of *heritability* (h^2). Heritability refers to the proportion of phenotypic variance that can be attributed to genetic variance. High heritability values reflect greater genetic contribution to the trait. For those traits that are qualitative in nature, such as the presence or absence of disease, the similarity of twins has been reported in terms of *concordance*. If the disease is the product of a genetic etiology, the concordance rate of the MZ twins should be greater than the concordance rate for DZ twins".

[25] The concept of "epigenetics" is discussed further below in this section.

twin studies is predicated on MZ cotwins having identical genotypes. If they do not, heritability estimates will be lowered" (Stromswold, 2006: 337):

> A number of mechanisms can cause MZ cotwins to have different genotypes. Although the vast majority of MZ twins are karyotypically identical (i.e., the number and general morphology of the cotwins' chromosomes are the same), if chromosomal non-disjunction occurs just before or at the time of twinning, MZ twins will have different karyotypes and are said to be heterokaryotic (Lejeune, 1963). […] A more subtle way that MZ twins may have different genotypes is if a spontaneous mutation occurs either before or after the zygote has split. […] When mutation occurs earlier (i.e. in dichorionic MZ twins), there is a greater chance that MZ cotwins will have different spontaneous mutations. Thus, dichorionic MZ twins are more likely to differ genetically than monochorionic MZ twins. (Stromswold, 2006: 337).

Besides MZ and DZ twins, non-twin siblings have also participated in this study. It is important to take into account that full siblings (i.e. of the same father and the same mother) are genetically the same as DZ twins: they share 50% of their genes (yet, see arguments supporting that DZ cotwins are genetically more similar to one another than non-twin full siblings)[26]. The inclusion of siblings in the twin research methodology is rare (see Table 2). However, it is not uncommon in phonetic studies to consider also the recruitment of siblings for the experiments (e.g. Whiteside & Rixon, 2004; Kinga, 2007; Feiser, 2009; Feiser & Kleber, 2012; see Section 2.4). In forensic studies, the most widespread reasons to investigate siblings, alone or together with the study of twins, are pinpointed by Feiser (2009):

> Siblings' voices are of high importance in forensic speaker identification/comparison. For example, not uncommonly the question is posed in court whether a given unknown recording could have been spoken by the subject's brother(s) instead of the subject himself. Other than being a possible legal strategy, this question suggests itself because siblings often have similar sounding voices (including speech patterns in general). (Feiser, 2009:1)

We have detailed above the specific genetic load shared by MZ cotwins, DZ cotwins and full siblings. Some explanation about what we understand as environmental influences seems also necessary at this point.

Firstly, it is widely repeated that "given the genetic identity in MZ cotwins, behavioral and physical differences between them are associated with differences in their environments" (Segal, 1990: 612), but where can such environmental differences occur? The first level where

[26] According to Stromswold (2006), "twin-derived heritability estimates will also be skewed if DZ cotwins share more (or less) than 50% of their alleles"; she mentions the case of transplant surgeons, who "have known for decades that the incidence of graft rejection is lower between DZ cotwins than between non-twin full siblings, and this clinical observation has been used to argue that DZ cotwins are genetically more similar to one another than non-twin full siblings (see Geschwind, 1983)" (Stromswold, 2006: 338-9).

differences may occur is the *prenatal* level (e.g. fetal transfusion), then at the *perinatal* level, some circumstances may affect the twins' initial similarity (e.g. order of delivery), and at the *postnatal* level, many are the possible factors which could trigger twin differentiation, like accident or injury to one cotwin (see Segal, 1990 and Stromswold, 2006).

Secondly, it has been already mentioned that any "excess" of similarity in MZs over DZs refers to "the proportion of phenotypic variation that can be attributed to genetic variance" (Tomblin & Buckwalter, 1998: 189). This being the essence of the twin design, it requires, like all research designs, that an important assumption be made: the *equal environment assumptio*n, i.e. it is assumed that the two twin types have similar environmental experiences. Although some authors (Lewontin, Rose & Kamin, 1984) have questioned the validity of this assumption, others have provided empirical support for it (Scarr & Carter-Saltzman, 1979; Vandenberg & Wilson, 1979, in Tomblin & Buckwalter, 1998: 189).

Thirdly, if we understand "environment" from a sociolinguistic point of view, it is clear that the family[27] exerts important effects on the linguistic output of individuals, although this field has not been thoroughly investigated so far (Hazen, 2001). Acknowledging the family as one of the most basic units in society (Benson & Deal, 1995), Hazen (2001) focuses on the family's effects on language variation, a research line that would be at the intersection of language acquisition and language change. It seems that so far there are more open research questions than clear answers to those questions (e.g. could the children in a family have exactly the same language variation patterns as the parents?).[28] These issues go beyond the purposes of our thesis. Yet they allow us to raise awareness on the difficulty for separating the effects of genetic factors from external (ambiance) factors, and more importantly, they point to unsolved questions about inter-speaker similarities in very close speakers (for the case of our study, affecting not only MZ and DZ twins but also siblings). Actually, influences between siblings may have opposite directions: accommodating or distancing influences, as it has been specially noted for twins (see below). Considering siblings in general, it is clear that the language acquisition process is strongly related with the transfer of language variation patterns:

> Most often, parents provide the stimulus triggering language acquisition […] *Nevertheless* a safe hypothesis is that no child copies exactly the language variation patterns of the parents. Neither is

[27] In variationist studies like Hazen (2001), the term *family* is used to refer to any modern instantiation of the family: "There may very well be differences in sociolinguistic variation as factors of types of families (e.g. single parent vs. two parent families; gay parents vs. straight), but discovering this first requires general assessments about the influence of any family on sociolinguistic variation" (Hazen, 2001: 520).

[28] A further example of opposite opinions about these issues is found in Koeppen-Schomerus, Spinath and Plomin (2003: 97): "Although theories of socialization assume that environments are doled out on a family-by-family basis, behavioral genetic research shows that, after controlling for genetic resemblance, growing up in the same family does not make children similar in personality or psychopathology." (Harris, 1998; Plomin & Daniels, 1987)

> there a radical break from the parents: no child creates a separate language from the language(s) of the parents (Hazen, 2001: 503).

As well as the close ties between children and parents may influence speech patterns, we can infer that the higher or lower closeness between siblings (which for our thesis was measured by means of the questionnaire described in Chapter 3 and available in Appendix A2), may be at the core of the environmental factors affecting voice for the phonetic parameters under study per sibling pair. As stated by Hazen (2001: 506), "the fields of discourse analysis and language and gender studies illustrate that the family is an influential context for construction of social identities". Equally important is the peer group, sometimes competing with the family linguistic model[29]. Therefore we have also gathered information in our questionnaire about the siblings' leisure (shared or not) activities, group of friends, etc.

If we specifically focus on twins, several authors before have tackled the question of how environmental factors affect their similarity, although there is no agreement in their experimental results. For instance, in Newman, Freeman and Holzinger (1937), twins reared apart and twins reared together were equally similar, but in Shields (1962) twins reared apart were more similar than twins reared together. Although this last finding may sound surprising, it has been suggested that "twins reared together may 'create' differences between themselves in an attempt at differentiation from the twin" (Segal, 1990: 615). As regards the question of whether MZ twins have a closer relationship than DZ twins, it seems that "an impressive body of experimental, clinical, and observational data suggests that MZ twins share a more intimate social bond, relative to DZ twins (Burlingham, 1952; Mowrer, 1954; Parker, 1964; Smith, Renshaw & Renshaw, 1968; Loehlin & Nichols, 1976, Paluszny et al., 1977; Segal, 1984; In Segal, 1990: 619).[30] This fact

[29] In Hazen (2001: 506), we find a description of the family as a *CofP (Community of Practice)*, whose norms could be in competition with those of other CofPs: "Innovative scholarship bearing on the sociolinguistics of the family has come about from Community of Practice (CofP) theorists (Holmes 1999). A CofP is defined as "an aggregate of people who come together around mutual engagement in an endeavor" (Eckert & McConnell-Ginet 1992: 464), and Holmes and Meyerhoff (1999: 174) label the family as a type of CofP. One working assumption of the CofP model is that becoming a "core member" of a CofP involves the acquisition of sociolinguistic competence (Ochs & Schieffelin 1983, Romaine 1984); the implication is that family members do follow the sociolinguistic patterns of their families (cf. Daly 1983). But family members are also going to be members of other CofPs – groups of friends, clubs, sports teams- and the sociolinguistic norms of the family may compete with those of other CofPs" (Hazen, 2001: 506).

[30] It may not be related to that "more intimate social bond" which, according to Segal (1990:619), MZ twins share, relative to DZ twins. Yet it is also of interest the difference in competition (rivalry) behavior between MZ and DZ twins: "Several studies have tested the hypothesis that social-interactional processes and outcomes may differ between genetically identical individuals (MZ twin pairs) and genetically non-identical individuals (DZ twin pairs). Von Bracken (1934) demonstrated that young MZ twins tended to maintain equality during work activities performed in the company of the cotwin. DZ twins, in contrast, either behaved competitively (if matched in skill) or disinterestedly (if unmatched in skill)" (Segal, 1990: 619). The question of whether this different behavior can influence speech is open for research. For our thesis, we have designed certain speaking tasks [see Chapter 3] which imply interaction and collaboration between cotwins where the influence on speech of the aspects noted by Segal could be tested in future studies.

could be at the base of what Debruyne, Decoster, Van Gysel, and Vercammen (2002) call “intratwin mimetism”.

Having explained the two best-known aspects affecting (dis)similarities between twins (i.e. genes and environment), a third element comes into play: *epigenetics*, which is the study of the changes in gene expression caused by mechanisms other than changes in the underlying DNA sequence. Epigenetics can explain why identical twins, born with the same DNA, may become completely different as they grow up. This developing scientific field reveals how certain factors, like stress or food habits, can cause divergence in twins by altering the expression of specific genes, as Miller (2012) explained to the general public in a recent article of the National Geographic Magazine:

> One way the study of epigenetics is revolutionizing our understanding of biology is by revealing a mechanism by which the environment directly impacts genes. […] a particular epigenetic process called DNA methylation […] is known to make the expression of genes weaker or stronger. (Miller, 2012: 4)

Even though this discipline is still in its infancy, many researchers (e.g. Danielle Reed, specialist in genetics; in Miller, 2012: 5) suggest that epigenetics could be in the origin of many differences between cotwins. In line with this, Reed evoked the following powerful metaphor:

> What I like to say is that Mother Nature writes some things in pencil and some things in pen. Things written in pen you can’t change. That’s DNA. But things written in pencil you can. That’s epigenetics. Now that we’re actually able to look at the DNA and see where the pencil writings are, it’s sort of a whole new world (Reed, in Miller, 2012: 5).

According to what is known so far about twin and non-twin siblings, their genetic endowment and the environmental influences possibly affecting their voice and speech, our starting point for research begins with the scheme in Figure 3. This shows how much genetic influence and environmental influence is expected per speaker-type pair. Accepting the *equal environment assumption* described above, MZ and DZ cotwins are expected to share the same environmental influence, but DZ twins will share half the genetic information than MZ twins. Male siblings (i.e. brothers – B-) will share the same genetic endowment as DZ twins but, on average, less environmental factors, mainly because of the age gap between them. Finally, unrelated speakers (US), who have also participated in this study [see Chapter 3], will neither share nature nor nurture.

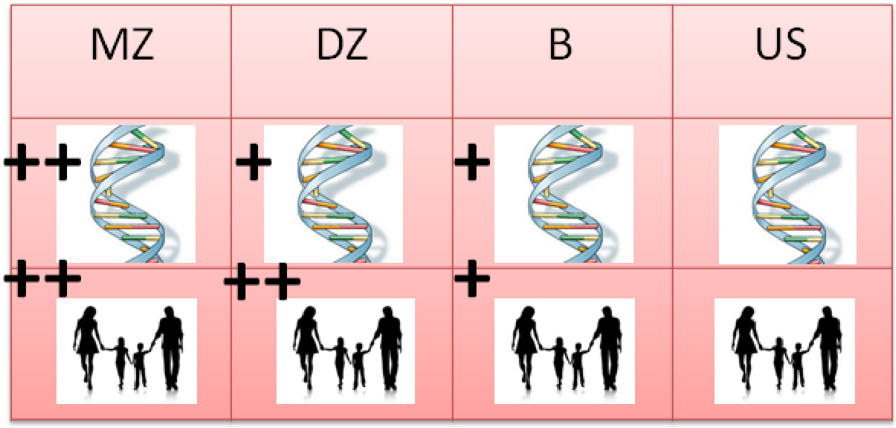


Figure 3. Outline of the genetic load and environmental influences expected to be shared by the four speaker types in our study: monozygotic twin pairs (MZ), dizygotic twin pairs (DZ), brother pairs (B), and unrelated speakers (US). The symbol (++) means more than (+) but not necessarily double. The absence of symbols means "neither genetic nor environmental factors shared".

In line with this scheme, five working hypotheses have been established for our thesis (see Table 3). Firstly, we assume that a speaker's voice would be similar to itself, i.e. from one recording session to another. This assumption is made for all speaker types (H_1). Secondly, accepting that MZ twin pairs are the most similar speakers that can exist (because of their shared genes and shared environmental influences), we hypothesize (H_2) that MZ intra-pair comparisons will yield matching scores similar to those obtained in intra-speaker comparisons. The third hypothesis (H_3) implies that DZ intra-pair comparisons will yield relatively large matching scores but not as large as in the case of MZ twins (the genetic load shared by DZ cotwins is still important and they shared the same environmental characteristics as MZ cotwins). In the fourth hypothesis (H_4), we state that the intra-pair comparisons in the case of brothers will yield matching scores over the *background baseline* (i.e. the values obtained by the reference population, namely the unrelated speakers). That means that brothers should be more similar than unrelated speakers because they share 50% of their genes, exactly the same as DZ twins (see Figure 3), and they usually have environmental influences in common, although less than DZ twins. Finally, we hypothesize (H_5) that a background baseline should exist for the matching scores obtained by the unrelated speakers.

These hypotheses will be further developed in the following chapters, where the phonetic parameters under study will be described. It seems though useful to outline them in this section, as they result from the nature-nurture differences acknowledged after the literature review carried out in this chapter.

Table 3

General Research Hypotheses

H_1: It is expected that intra-speaker comparisons will yield large matching scores for all type of speakers (MZ, DZ, B and US).

H_2: MZ intra-pair comparisons will also yield large matching scores.

H_3: DZ intra-pair comparisons will yield large matching scores but not as large as MZ intra-pair comparisons.

H_4: B intra-pair comparisons will yield matching scores over the background baseline.

H_5: US inter-speaker comparisons will yield matching scores aligned with the background baseline.

Note. MZ means monozygotic pairs; DZ is used for dizygotic pairs; B for brothers, and US for unrelated speakers.

2.3. Forensic relevance of twins' voices

We find several references to the speech of twins or siblings not only in introductory works about Forensic Phonetics (Rose, 2002; Rose, 2006a) but also in certain chapters dealing with the forensic application of Phonetics within more general books about voice. For example, Kreiman and Sidtis (2011) encompasses very diverse aspects of voice, including a whole chapter about voice identification in a forensic setting, and within this, a section devoted to the importance of twins' voices. Here the authors claim that identical twins' voices are a challenge for forensic speaker identification, since this sort of twins are genetically identical and usually raised in the same circumstances. Therefore, their voices are highly confusable. In line with the etiology references found in the rest of their book, the authors make a comparison between human twins and other kind of twins in the animal kingdom:

> [...] the cries of twin noctule bat pups are more similar than are the calls of unrelated pups, and remain so as the pups grow, although the cries of unrelated individuals grow more distinct with age. (Knörnschild, von Helversen, & Mayer, 2007; In Kreiman & Sidtis 2011: 244-245)

Rose (n.d.) gives special importance to the forensic comparison of twins' voices when, in order to provide support for the combined use of traditional and automatic approaches in Forensic Phonetics, he draws on an (apparently) hypothetical instance of forensic casework which involves a twin pair.

> Consider, for example, a case where two samples are from different speakers who have very similar global acoustics, like some identical twins, but where one twin consistently uses a funny 'r' sound (technically a labio-dental approximant). It is likely that a global automatic approach, which cannot focus on single speech sounds, will evaluate the difference between the two samples as more probable assuming they have come from the same speaker. A traditional approach would not make this mistake. (Rose, n.d.: 4).

In a second example, a real case is mentioned by Rose (2002) very early in the introduction of his best known book. A telephone conversation between two brothers was intercepted in Australia. One of the brothers was charged with drug-related offences, despite his defense lawyer's claim that the brothers' voices were so similar that the incriminating recordings could not be attributed to the suspect. Nevertheless, thanks to a forensic phonetic analysis it was shown that the brothers' voices could be distinguished: "Although their voices were indeed acoustically very similar in many respects, they still differed in others, and in particular they both had different ways of saying their 'r' sound" (Rose 2002: 1-2).

In another publication, in which this same author describes in detail what Forensic Speaker Recognition is, Rose (2006a) includes a reference to twins' voices when he writes about the "forensically much-neglected indexical function of language". As concerns this issue, he brings up the case of siblings (especially identical twins), explaining that these speakers have a very similar vocal tract and yet they are able to exploit the plasticity of their vocal anatomy very differently. As a result, they may sound different. For example, Rose (2006a) mentions that each twin can have different allophones for the same phoneme (Nolan & Oh 1996, in Rose 2006a) or they may tend to use systematically different articulatory settings.

We can also find cases of offences or crimes that involve the participation of siblings or twins in Spain. In one case, a telephone conversation between two brothers was intercepted and each telephone-end conversation transcribed. Due to the confusability of the brothers' voices, each end of the line was wrongly transcribed. An expert (Hellín, 2010) noticed this previously made error only when he looked at the exact times when the telephone call took place. The speech fragment wrongly attributed to one of the brothers could have never been uttered by him since his call wasn't intercepted until he picked the telephone up. The disputed speech fragment belonged to his brother, the one who made the phone call, and who happened to be speaking (saying the disputed utterance) while the phone was still trying to communicate, that is to say, before his brother had time to pick up the receiver. Another case (Hellín, 2010) involves a twin pair, one of whom had previous charges and could be imprisoned due to a new offence. His cotwin, free of accusation, decided to declare himself culprit so that his twin (the real offender) wouldn't go to jail.

In a recent piece of news (Mora, 2013) we could read about the arrest of two MZ twins in France charged of six rapes and sexual assaults. Even though the victims claimed that the aggression took place by only one person, the police could not determine, on the basis of the DNA found in the sperm cell, which one of the two twins committed the crime. The reason set out by the police was that the DNA is the same for identical twins[31].

[31] Nevertheless, the newspaper article specified, as described in San Segundo (2013a: 61) "that the French police would need to pay around one million Euros for a very complex and specific DNA test which would reveal the identity of the offender". However, according to other studies, no DNA test seems to allow the

To sum up, it may not be extremely common to find twins as criminals, this being a usual criticism of twin research in Forensic Phonetics. However, real forensic casework related to twin and non-twin siblings is not as unlikely as one may a priori think. In Ma (2011), more than ten famous twin-crime stories are related. Describing them in detail goes beyond the purposes of this thesis. The key purpose of carrying out research on twins' voices is that it could shed light on the limits of between-speaker (inter-speaker) and within-speaker (intra-speaker) variation, since twins represent the most extreme physical similarity in human beings. As this resemblance also applies to their voice, distinguishing them is still a challenge in speaker identification.

2.4. Phonetic studies on twins

We have considered a four-perspective approach to the literature review on twins' voices. First, we will deal with studies focusing on the perception of twins' vocal productions. Secondly, several acoustic parameters will be reviewed, as they have been previously described in the literature about twins. Thirdly, we will mention a couple of articulatory studies which investigate twins' voices. Finally, the few studies which have considered this issue under an automatic approach will also be reviewed. Of course, some works offer combined perspectives. In appendix F we include a table which classifies the twin studies in chronological order, where information about the speaker sample size of each study can be found.

2.4.1. Perceptual studies

It could be useful to begin the section devoted to perceptual studies by mentioning the (probably) most frequently cited work about twins' *self-perception* (in other words, twins' attempts to correctly identify their own voices aurally). Indeed, the study by Gedda, Fiori-Ratti and Bruno (1960) is usually brought up by other authors (Alpert, Kurtzberg, Pilot, & Friedhoff, 1963; Decoster, van Gysel, Vercammen, & Debruyne, 2000; Debruyne et al., 2002) to exemplify the difficulty of identifying twins' voices. This classic study consisted in a perceptual test in which each member of a twin pair was presented randomly his voice and that of his co-twin. The results showed that most monozygotic (MZ) twins were not able to distinguish who was talking at each time, while the opposite happened with dizygotic (DZ) twins, who could mostly tell apart their own voice from that of his co-twin. The age range of the twins recruited for this experiment was 8-16 years.

distinction of identical twins, at least in cases such as the one reported in Künzel (2011: 274). Hence the importance of carrying out research in disciplines like forensic phonetics, which relies on pieces of evidence other than DNA, like voice samples, which have not been proved to be totally identical in MZ twins.

Luchsinger and Arnold (1965) describe a similar phenomenon, now referring to telephone-transmitted voices: “[…] *MZ pairs* whose voice and speech patterns were so similar that not even their other sister could tell them apart on the telephone” (In Ryalls, Shaw & Simon 2004: 166). Likewise, in a study basically aimed at pointing out the importance of designing an adequate voice line-up, Yarmey, Yarmey, Yarmey, and Parliament (2001) include a brief mention to the challenge imposed by familiar voices in identification tasks. One of the participants in the perceptual experiment carried out by these researchers, after having listened to the voice of his identical twin, stated: “I am positive that I have never heard that voice before in my life” (Yarmey et al., 2001: 298).

A typical approach to the study of twin voices under a perceptual point of view consists in carrying out a listening experiment in order to investigate the feasibility of twins’ identification; in other words, whether listeners[32] are able to detect differences between twins. For instance, Johnson and Azara (2000) found that listeners’ performance was greater than chance level in three different experimental conditions: In the first one, half of the stimuli were words spoken by MZ twins and the other half were stimuli from the same person repeating a word twice. The second condition was identical to the first one with the additional characteristic that some pairs of stimuli were repetitions of the same word by two unrelated twins. In these two conditions, the listeners knew that they would be hearing twins in some of the trials. Finally, in the third condition, the listeners were not told that some of the stimuli would have twins paired with each other.

A multidimensional scaling analysis of perceived speaker similarity was carried out with the results of the third experiment condition. The pairs of stimuli presented to the listeners were judged as belonging to the “same speakers” or “different speakers”. Thus, considering that the proportion of “same speaker” judgments would be a good estimate of the perceived similarity of two speakers, these data were entered in a multidimensional scaling (MDS) analysis in order to create a perceptual map of the speakers. Interestingly, the results of this analysis showed that “sometimes twins are perceived to be as different from each other as are unrelated speakers” (Johnson & Azara, 2000: 2).

In a similar line of research, Decoster et al. (2001) study on MZ twins also tried to answer the question of whether twins can be perceptually identified. In both this article and Johnson and Azara’s (2000), it seems that one of the reasons for them to perform a perceptual experiment on twins’ voices is to justify a subsequent acoustic analysis, which actually is only undertaken in Decoster et al. (2000):

[32] In this case, with “listeners” we do not mean the twins themselves, as in the studies just reviewed, but any other listeners, whether familiar or unfamiliar with the twins.

> One rationale for this experiment is as a pretest for a phonetic study of twin's speech. If listeners are able to detect differences between twins then we are encouraged to look more closely at acoustic and articulatory records to determine the ways in which the twins' speech differs, but if listeners are unable to detect differences between twins we wouldn't expect to find any meaningful differences in acoustic or articulatory studies. (Johnson & Azara, 2000: 7)

As well as Johnson and Azara (2000: 2) concluded that "listener's sensitivity to twin differences is greater than chance level", Decoster et al. (2000: 54) found that "perceptually twin voices of young adult men and women are identifiable when presented in random order with a genetically unrelated voice". This result is what triggers the subsequent acoustic analysis. It may be worth mentioning that identification performance was better when the perceptual stimuli were sentences than when they were sustained vowels. The most plausible reason for this is that sentences yield more voice information and prosodic cues about the speaker, which are available for the listener in order to make his identification judgment. This is in line with previous findings in Phonetics (Goldstein, Knight, Bailis & Conover, 1981; Kreiman & Sidtis, 2011). Not only in general phonetic studies has it been found that listeners perform better at identification when listening to larger stretches of speech than to shorter ones, but the same results were also obtained in studies specifically dealing with siblings' voices (San Segundo, 2014).

As we have previously explained, in Decoster et al. (2009), the authors first investigate listeners' ability to identify MZ voices and then look for an acoustic explanation for the perceptual results. Concerning the acoustic results, they hypothesized that two parameters related to voice pitch (fundamental frequency, i.e., f_0, in sustained vowels, and speaking fundamental frequency, i.e., SFF, in sentences) could have been the cues used by the listeners in the perceptual identification of twins. However, these acoustic aspects as well as other combined approaches (like the one undertaken by Homayounpour and Chollet (1995), embracing perceptual, acoustical and automatic methods) will be described in next sections, as this section covers only perceptual studies.

In this sense, Decoster et al. (2009) is not the only study including a perceptual test and a further acoustic evaluation. Whiteside and Rixon (2000) follow the same experimental schema. They first designed a perceptual experiment in which familiar listeners[33] had to distinguish a single pair of twins from their own recorded voices. Then, an acoustic analysis followed, using 23 parameters. Since this is a combined approach, we will review in detail this study in the following section (2.4.2.), together with Whiteside and Rixon (2001). Furthermore, this procedure seems more coherent in order not to break the authors' chronological order of publications.

[33] The listeners who carried out this perceptual experiment were all familiar to the twins: the twins' brother, two housemates, three close friends and the twins themselves.

Some other phonetic studies that have compared MZ and DZ twins within a perceptual approach are Weirich and Lancia (2011) and Weirich (2011). The latter showed that unrelated speakers were significantly easier to distinguish than twins, but zygosity had no effect on perceived similarity. In the former, the results of an AX discrimination test showed that stressed syllables did not differ in terms of inter-speaker variability between MZ and DZ twins.

It seems appropriate to include some brief comments about siblings' voices, which are by far less researched than twins'. Feiser and Kleber (2012) tested voice similarity among brothers (3-5 years of age difference between them) in a perceptual experiment and found that correct identifications of brothers by their voices are significantly above chance. Besides, the authors indicated that a particular pair of brothers was significantly better identified than the other pairs in the experiment. From the authors' point of view, some plausible explanation for this finding could be that the brothers better identified showed more pronounced dialect features.

Recapitulation

The main conclusions that we can draw from all these studies are:

(1) Twins' voices are highly confusable, making the task of twin voices' identification a difficult one, as it has been shown in experiments where even twins were not able to distinguish their own voice from that of their cotwin. (Gedda et al., 1960; Yarmey et al., 2001)

(2) However, listeners can tell twins apart by their voice above chance level (Decoster et al., 1999; Johnson &Azara, 2001), and the same happens with siblings' voices (Feiser & Kleber, 2012); this makes us hypothesize that there must be some acoustical parameters which allow for speaker identification even in challenging conditions, such as in comparisons between very similar speakers.

(3) Moreover, MZ twin pairs can be as different between them as non-twin speakers usually are (Johnson & Azara, 1999), which has made many scientists wonder what proportion of inter-speaker variation is due to genetic factors and which to environmental reasons. More information about the twin method applied to Phonetics can be found in Section 2.2.

2.4.2. Acoustical perspectives

We have briefly mentioned above a study analyzing FUNDAMENTAL FREQUENCY (f_0)[34] (Decoster et al., 2000) in which the authors concluded that this could have been the acoustic cue used by listeners to identify twins. Other authors have also investigated this parameter in twins' voices, although with a different purpose. Wondering if this feature could constitute a vocal phenotype[35], and acknowledging that at present the effects of anatomical difference on voice parameters is unknown, Przybyla, Horri and Crawford (1992) analyzed the *average VFF (Vocal Fundamental Frequency)* of 50 pairs of female twins and 12 pairs of male twins, comprising both MZ and DZ pairs. Their results indicate that both MZ and DZ twins are significantly correlated for this parameter. Furthermore, the larger discrepancies in VFF found in DZ twins would suggest the presence of a genetic component in the variation of this voice parameter. Nevertheless, this thesis is not very strongly supported, since it could be difficult to disentangle genetic from environmental factors affecting voice, as the authors state:

> The twin method applied in the present study (twins reared together) may be insufficient for genetic study for disentangling the specific environmental and genetic components in the human voice. The possibility that genetic effect is completely masked by common environment cannot be excluded. Since voice is a feature dependent upon environmental influences, other family study methods such as adoption studies (especially study of twins reared apart) may be more informative. (Przybyla et al., 1992: 265)

Debruyne et al. (2002) is a follow-up of the study carried out in Decoster et al. (2001). Now the authors focus on the following parameters in twins' voices: (1) *average fundamental frequency in Hz* (they refer to this parameter as Speaking Fundamental Frequency, SFF*)* and (2) the *standard deviation of the fundamental frequency in Hz* (further called varSFF, or intra-individual variation of the SFF). It is important to note that the analysis is based on read speech. Besides, the aim of this study is twofold. On the one hand, the authors want to test whether twins' voices are similar considering these parameters. On the other hand, the study investigates to what extent the observed similarity is genetically determined, by comparing the results in MZ twins with the results in DZ twins. They found that, for this parameter, it was not possible to discern the influence of genes from the influence of shared environment. Whereas in SFF there was a

[34] Throughout this chapter, we will mark in small capitals the different parameters analyzed in twins' studies. We will use italics to signal all the possible subtypes of features encompassed by the term previously marked in small capitals.

[35] Przybyla et al. (1992) define "phenotype" and "vocal phenotype" as follows: "The initial step in a genetic analysis of quantitative traits is the recognition and precise definition of the phenotype. The term phenotype denotes a measurable trait in an individual characteristic that can be observed and measured. A phenotype is determined by the interaction of genes and environment. Such quantitative features can be measured and described as attributes of a given individual. Within a population, the variation of a quantitative phenotype tends to have a normal shaped distribution. The normal distribution usually suggests that such a trait is the result of the interaction of genes and environment." (Przybyla et al., 1992: 262)

difference between MZ and DZ twins, being the correlations higher in the former, in varSFF there was not such difference.

All in all, we can draw some relevant information from this study. First, the well-known fact that voice fundamental frequency “is dependent on various organic elements such as the length and mass of the vocal folds, while personality and character traits also play a role” (Debruyne et al., 2002: 469) makes us consider the use of a method to extract biometric characteristics of the glottal source (i.e. length and mass of the vocal folds, among others – see Chapter 5) which could better inform of speaker idiosyncrasies at the glottal source level. It is also worth pointing out that Debruyne et al. (2002) take into consideration intra-speaker variation, which is an aspect usually neglected in other voices studies, and it is actually important in Forensic Phonetics [see Chapter 1]. Finally, among the authors who describe the environmental factors influencing the development of voice since birth, Debruyne et al. (2002) are, to our knowledge, the first ones who introduce, in phonetic studies, the concept of “intratwin mimetism” [see Section 2.2.]

Loakes (2006b) focused on *long-term f_0* (fundamental frequency sampled at various points throughout a speech sample) and found that, even though general speakers tend to fall within a specific f_0 range, “twins have a more similar mean long-term f_0 than what has previously been reported for unrelated pairs of speakers”. The main asset of this study is that, unlike other investigations on twins, two non-contemporaneous sessions are recorded per speaker, which allows for intra-speaker comparison.

Some other publications exist which have considered the analysis of f_0 or related parameters in the study of twins. However, either they constitute multiparametric approaches to voices (Forrai & Gordos, 1983; Whiteside & Rixon, 2000) or should rather be considered together with other clinical studies (Fuchs et al., 2000). In any case, we will review these two studies later on in this chapter, when it seems more appropriate, since f_0 is not the focus of the analysis in them but one parameter more, among many others.

Besides fundamental frequency, some other acoustic parameters in which MZ twins have been found to be similar are the following ones: SPECTRUM SHAPE, NUMBER OF HARMONICS and FREQUENCY PEAKS (Gedda, Fiori-Ratti & Bruno 1960; Flach, Schwickardi, & Steinert, 1968; Cornut, 1971; Forrai & Gordos, 1983), although we must point out that the first and the third study actually analyze infant or children voices.

If we go back to pioneer works on twins’ voices from a spectral point of view, discarding the above-mentioned Gedda, Fiori-Ratti and Bruno (1960) and Cornut (1971), who analyzed pre-adolescent and infant twins’ voices (respectively), we should mention the study of Alpert et al.

(1963). These authors found that the voices of MZ twins are more similar than those of DZ twins, but only at HIGHER SPECTRAL POINTS, not at lower spectral points. They suggest that this could be due to the fact that “lower frequencies are more sensitive to transitory factors while the higher frequencies are more likely to reflect anatomic similarity” (Alpert et al., 1963: 340). Their explanation seems logical. Besides, it has been reported elsewhere in the phonetic literature (cf. Rose 2002, to limit ourselves to Forensic Phonetics).

According to Decoster et al. (2001), whose literature review on twins’ voices is quite comprehensive, the studies on twins which have considered INTENSITY MEASUREMENTS have not yielded so concordant results as those studies focusing on spectral features: “While Flach et al. (1968) reported on very comparable voice intensities among MZ voices, Cornut (1971) found higher speaking intensity levels in the most “dominant” co-twin” (Decoster et al., 2001: 50).

The study of Forrai and Gordos (1983) is notably singular, since the authors analyze a great amount of acoustic parameters with the purpose of creating a system capable to discriminate between MZ and DZ twins. In other words, as well as there are blood-group tests to determine twin zygosity (the most widespread method together with DNA analyses), these authors try to do the same with voice samples; a technique which would save, according to them, time and money expenses. For this purpose they create a learning algorithm based on 14 acoustic parameters, being one of their intentions “to determine the minimum number of parameters necessary for a safe zygosity discrimination” (Forrai & Gordos, 1983: 317). Among other parameters, basically spectral ones, the authors also include VOICE ONSET TIME (VOT) and, again (as in previously mentioned studies) f_0 average and standard deviation. The results obtained under this approach to twins’ voices suggest a “perfect fit between the zygosity diagnoses obtained independently by the phonetic and the blood-grouping method” (Forrai & Gordos, 1983: 319).

As regards VOT, noting a lack of twin studies focusing on this parameter, Ryalls et al. (2004) carried out an experiment aimed at investigating whether VOT is more similar in MZ pairs who live together (shared linguistic environment) than in MZ pairs who live separated (i.e. they stopped sharing the same linguistic environment at some point in their lives). The shortcoming of this study is that only one MZ female twin pair for each of the above mentioned conditions participated in the experiment. The pair who shared a household was 21 years old, while the other twin pair was 70 years old at the time of the recording and they had been living for 45 years in two different linguistic regions of the United States. The results of this study showed that VOT averages were similar for the younger twins who still lived together, but they were different for the older twins living in separated geographical regions.

COARTICULATION PATTERNS have also been studied in twins’ voices. Nolan and Oh (1998) examined the production of /l/ and /r/ phonemes in three pairs of MZ twins to see if they exhibited between-speaker differences in coarticulation. The results showed that there were no strong differences within twin pairs as regards the following parameters: (1) the mean frequency of the first four formants in the initial consonants /l/ and /r/; and (2) the mean frequency of the first four formants in the vowel following the previously measured consonant. Notwithstanding the lack of coarticulatory differentiation, some spectral differences in the consonants under consideration were found for some twin pairs, and these differences were perceptually salient. The authors' explanation for this fact is that "twins avail themselves of the same freedom as other speakers to opt for alternative realizations of a phoneme, at least where these are sanctioned by the relevant speech community" (Nolan & Oh, 1998: 48).

Whiteside and Rixon (2004) found that coarticulation patterns in CV sequences in a variety of phonetic contexts were not so different in MZ twins as in siblings who were gender- and sex-matched. In order to measure the degree of coarticulation the authors used F2 locus equations, which "parameterize the relationship between F2 mid and F2 onset values of vowels in consonant-vowel sequences […], and provide an indirect representation of the dynamics of lingual gestures which are involved in the production of consonant-vowel sequences" (Whiteside & Rixon, 2004: 6).

It seems logical to review first Whiteside and Rixon (2004) since we have just discussed another study analyzing coarticulation in twins. However, these authors have published previous studies about twins, as we will next describe. In Whiteside and Rixon (2004), the authors examined a range of acoustic parameters including F2 locus equations in order to gauge the extent of between twin similarities in differences across their coarticulatory patterns, distinguishing different places of articulation. Indeed, for bilabial, alveolar, velar and glottal place of articulation, it was found that twins displayed similar coarticulation patterns. Again, a shortcoming of that study is that only a pair of MZ twins participated in it.

A first study on twins signed by Whiteside and Rixon dates back to 2000, and it will be reviewed a year afterwards by van Dommelen (2001) with some suggestions for data reevaluation. The experiment of speaker identification carried out by Whiteside and Rixon (2000) includes a perceptive test and a further acoustic evaluation. As we briefly introduced in the previous section, the perceptual experiment is designed so that familiar listeners[36] have to identify a single pair of twins from their own recorded voices. The stimuli used in this test are of two types: (1) pure monosyllables, i.e. syllables produced originally by one twin, and (2) hybrid syllables consisting

[36] Note that, as explained in Section 2.4.1, by "familiar" listeners the authors meant the twins' brother, two housemates, three close friends and the twins themselves.

of two fused demisyllables, i.e. the first half of the syllable is produced by one twin and the second half of the syllable is produced by his co-twin.

The results showed that the identification scores for the pure syllables were above chance (one twin was correctly identified in 71.45% of the cases and the other twin in 72.9% of the cases)[37] while identification scores for the hybrid syllables did not differ significantly from chance (44.4% of correct identifications for one twin and 54% of correct identifications for the other). In discussing these latter results, the authors seem to find a confirmation to their expectations that listeners (even familiar with the twins) would find difficult the task of assigning the stimuli to a particular twin.

As regards the acoustic analysis of this study, Whiteside and Rixon tried to find the relationship between the listeners' responses and the acoustic parameters investigated by means of four sets of Pearson product-moment correlation coefficients. Out of the 23 acoustic parameters studied, 13 showed significant between-twin differences, and five of these 13 parameters were fundamental frequency parameters. Of the three temporal measures investigated (word duration, vowel duration and VOT), only the first one did not show significant between-twin differences.

The originality of the approach used in this article -especially the perceptual experiment- is unquestionable and it probably is a good technique to challenge the ability of familiar listeners to identify a voice. However, for forensic purposes, the relevance of its results has to be called into question. The concept of hybrid syllables is foreign to the realm of Forensic Phonetics, so the results of this study should be regarded as an interesting theoretical approach to the similarity of twins' voices but not as a way to shed light on the reality of twins' voices in natural speech. The article received some criticism by van Dommelen (2001), concerning the statistical method used in order to correlate identification scores and acoustic parameters. Considering that the results obtained by Whiteside and Rixon (2000: 8) were "contradictory and inconclusive", van Dommelen (2001) suggested a re-evaluation of their data using multiple regression techniques:

> The pure syllables were analysed acoustically to obtain a set of 23 acoustic parameters. To investigate which cues the listeners used to identify the speakers, correct raw identification scores were correlated with each of the acoustic parameters. The results of these calculations are counterintuitive and hard to interpret. (Van Dommelen, 2001: 8)
>
> [...] it might seem doubtful whether the results of Whiteside and Rixon's experiment allow reliable conclusions at all. However, a possible solution to explaining the seemingly inconclusive results could be to re-evaluate the data by performing a multiple regression analysis including all the 23 acoustic parameters measured in production. (Van Dommelen, 2001: 9)

[37] The twins' responses to their own speech were excluded from the analysis.

In the case of Whiteside and Rixon (2001), they undertook a study on the same MZ twin pair as in their previous study (Whiteside & Rixon, 2000) but this time only within an acoustic approach. While in Whiteside and Rixon (2000), 23 acoustic parameters were examined to assess the extent of the similarities and differences in the twins' voices, on this later occasion, 21 acoustic parameters were investigated, including temporal, frequency and temporal-frequency measures. The only two parameters analyzed in Whiteside and Rixon (2000) which are left aside in Whiteside and Rixon (2001) are jitter and shimmer values for the vowels. However, no reason for the absence of these parameters is offered in their second publication. Actually no reference to Whiteside and Rixon (2000) appears in the latter. Of all the parameters analyzed in Whiteside and Rixon (2001), nine displayed significant differences for between-twin comparisons: *vowel duration, timelag* ("the temporal midpoint values taken from the onset of the vowel to the midpoint of the vowel"), *VOT, f_0 onset, f_0 mid, f_0 offset, F1 onset, F1 mid, and F2 mid*.

Of the four TEMPORAL PARAMETERS considered (word duration, vowel duration, timelag values and VOT), only word duration did not indicate a significant between-twin difference. According to the authors, "this suggests that at the macro level of the word, the twins read the CVC words with similar temporal gestalts" (Whiteside & Rixon, 2001: 18). Noteworthy of this study is that, unlike most of the publications reviewed so far, these authors acknowledge the importance of having more than one non-contemporaneous speech sample of the twins, which is actually strongly recommended in forensic research. They state:

> The potential effects of environmental influences such as health and physical condition, fatigue, lifestyle, stress and mood on VFF should not be underestimated. The extent of these effects could have been gauged more accurately for twins B and D, had multiple sets of speech samples collected over several recording sessions been available for analysis. (Whiteside & Rixon, 2001: 18)

The studies by Whiteside and Rixon (2000, 2001) are very relevant for Forensic Phonetics, since they appear to be almost the only references that take into account temporal parameters in twins' voices. Its relevance lies in the fact that the use of temporal information of the acoustic signal has indeed yielded very good results for speaker identification in general (Dellwo, Kolly & Leemann, 2012) but its application to challenging conditions, like comparison of very similar speakers (i.e. twins and siblings) is yet open to further research. Only one pilot study exist so far (Leemann, Dellwo & Kolly, 2012), in which one MZ twin pair and 7 unrelated speakers of Zurich German were compared in relation to one temporal measure that had proven very effective for speaker discrimination in previous non-twin studies, ie. the percentage over which speech is vocalic (%V).

Acoustic parameters have certainly been the features most frequently used to characterize twins' voices. Let us now briefly describe the works by Loakes (2006a) and Weirich (2011), two

PhD dissertations which analyze very diverse features of the acoustic signal. After this, a final section will be devoted to a particular type of acoustic parameters: those traditionally related to voice quality and specially linked to the study of the voice source.

Loakes (2006a) focused on the FORMANT PATTERNS of twins, showing that F3 is the most speaker-specific formant frequency. Another important finding of this study is that lax vowels are more speaker-specific than tense vowels, at least in the Melbourne variety of Australian English that she analyzes. The great asset of her research lies in having granted special attention to forensically realistic conditions for the data collection. Thus, she analyses (1) telephone-transmitted, (2) non-contemporaneous, and (3) spontaneous speech samples, which are three major characteristics of speech data in real forensic casework. However, telephone-transmitted speech data are available only for a part of the total of twins analyzed[38]. Her work also shows that the extent to which twins' speech can be discriminated highly depends on the specific twin pair being compared, apart from the parameters under analysis. As concerns the consonant production of twins' voices, she found that consistent FRICATION OF /k/ AND /p/ is speaker-specific, being thus potentially useful for speaker identification. These specific findings are explored more fully in Loakes and McDougall (2010).

As regards the acoustic analysis performed by Weirich (2011) in her study of German twins, she distinguishes between ACOUSTIC *TARGETS* FOR PARTICULAR PHONEMES (formant frequencies for vowels /a/, /iː/ and /uː/, as well as /s/ and /ʃ/) and ACOUSTIC *TRANSITIONS* BETWEEN TARGETS (formant transitions in sequences like /ʃə/). The main finding of Weirich's (2011) study is that, for the purpose of distinguishing MZ and DZ twins, acoustic transitions as well as articulatory gestures [see Section 2.4.3] are more important than acoustic targets[39].

> Thus, MZ twins are assumed to show fewer differences than DZ twins in dynamic speech patterns (like TRANSITIONS and GESTURES) but not necessarily in static ones (like TARGETS). These findings are in line with the suggestions from Nolan et al. (2006), Kühnert & Nolan (1999) and Rose (2002), who propose that TARGETS are linguistically determined and influenced by the learned and shared language system, while TRANSITIONS and coarticulatory strategies are organically determined and idiosyncratic. (Weirich, 2011: 234)

Despite the fact that the acoustic transitions were not found to be auditorily salient in the perceptual test, in the acoustic analysis sibilant-schwa transitions did turn out to be more similar

[38] Firstly, she analyzed the speech of eight twins (three male MZ twin pairs and one male DZ twin pair) in high-quality conditions. Afterwards, she offered the analysis of 10 more speakers (four male MZ twin pairs and one male DZ twin pair) recorded via telephone.

[39] However, the results of the perceptual test carried out [see Section 2.4.1] suggested that the investigated transitions are not relevant for perceptual speaker identification.

in MZ twins than in DZ twins. Fundamental frequency measures and voice quality parameters were also analyzed by Weirich (2011), although only in the word “wasche” /^{1}vaʃə/, which was used as a stimulus in the perceptual test conducted in order to measure perceived similarity and acoustic correlates in twins’ voices [see Section 2.4.1].

From a predominantly CLINICAL POINT OF VIEW, Fuchs et al. (2000) compared MZ twins with non-related speakers and found that the voices of the former are significantly more similar than those of the latter regarding the following features: *vocal range, highest and lowest vocal fundamental frequency, fundamental speaking frequency, maximum voice intensity, number of partials and vibrato of intensity*. The most remarkable aspect of this study is that the authors compare the “suitability of the voices of MZ twins for professions with a high demand on voice”. They actually find that it is highly common that both members of a twin pair share not only the same leisure activities but also the same profession. That is why the authors suggest that the voice problems diagnosed to a co-twin, should be taken into account also by the other cotwin, especially if these voice issues are of particular importance for an adequate and long-lasting vocal performance, in cases of jobs which imply an intensive use of voice. Fuchs et al. (2000) also hypothesized that there would be a strong difference in intra-pair correlation between older and younger twin pairs, being the former more different than the latter. If this would happen, it could show the effect of an exogenous influence. However, they could not prove that hypothesis.

Other researchers who have unsuccessfully tried to test this hypothesis are Johnson and Azara (2000), although they used a perceptual approach. They expected that older twins would be more easily distinguished from each other than younger twins “because of their divergent linguistic experience during adulthood” (Johnson & Azara, 2000: 16). However, their perceptual results did not show this effect. Indeed, the twins who were the most similar to each other were the oldest ones, aged 67, and the youngest ones, aged 20. The next oldest twin pair (aged 43) and the next youngest twin pair (aged 21) were the most different from each other. Therefore, according to this study “there is no tendency for twins to differ increasingly from each other with increasing age” (Johnson & Azara, 2000: 16). We should link this result with the findings of Ryalls et al. (2004), which were already explained above when we dealt with the parameter VOT. In Ryalls et al. (2004) it was found that VOT averages were similar for a young MZ twin pair who lived together, but they were different for old MZ twins living in different geographical regions. In this case, age and dialect overlap and it is difficult to separate the influence of one variable from the other.

Within a clinical context, Luchsinger and Arnold (1965) provide the first review of the early literature on speech and language studies in twins. However, according to Ryalls et al. (2004: 165), these authors (Luchsinger and Arnold, 1965) actually “attribute Seeman (1937) with the introduction of twin research to the field of clinical linguistics and speech-language pathology”. This type of studies lies beyond the scope of our research, since they mainly focus on language delayed development, speech and language disorders or poor language skills among twins. They do not show, therefore, an interest in the comparison of twins’ voices. We will limit our literature review of twins in the clinical realm to studies focusing on voice quality, comprising more specifically voice source features, apart from fundamental frequency, whose importance regarding twins’ voices has been described above.

Weirich and Lancia (2011) found auditory similarity in intra-pair twin comparisons but also in the intra-pair comparisons of unrelated speakers, which, according to them, could be explained by the acoustic parameters F_0, shimmer, jitter and Harmonics-to-Noise Ratio (HNR). All these features refer to the voice source, which has been hypothesized by some authors as being more relevant than vocal tract features for speaker identification (Gil & San Segundo, 2013; Alves et al., 2012). However, this is an important question which has not been completely solved. Some other authors (Lavner, Gath, & Rosenhouse, 2000: 9) sustain, according to their experimental results, that “on average, the contribution of the vocal tract features to the identification process is more important than that of the glottal source features”. Despite this, they allow a variation margin, acknowledging speaker idiosyncrasy: “large individual differences exist between speakers, suggesting that each speaker has a different personal combination of acoustic features that cues his identity” (Lavner et al., 2000: 25).

Van Lierde et al. (2005) focus on voice quality (they termed it “vocal quality”) and vocal performance only in MZ twins, understanding voice quality in a broad sense. Their analysis implies a multiparameter approach, comprising: *perceptual evaluations (GRBAS scale)*[40], *aerodynamic aspects*, *acoustic analysis* and *Dysphonia Severity Index (DSI) measurements.* These last measurements are described as follows:

> […] the DSI is based on the weighted combination of the following selected set of voice measurements: highest frequency (f_0-high in Hertz), lowest intensity (I-low in decibels), maximum phonation time (MPT in seconds), and jitter (in percent). The DSI ranges from +5 to -5, respectively, in healthy and severely dysphonic voices. The more negative the person’s index, the worse the vocal quality. (Van Lierde et al., 2005: 512)

While the perceptual rating of the twins’ voices was assessed during connected speech, the acoustic parameters f_0, *jitter* and *shimmer* were obtained using the central segment of a sustained /a/ produced at the habitual loudness and pitch of the speaker, and embedeed in a carrier sentence (Van Lierde et al., 2005). The authors could not find significant correlation coefficients

[40] In the GRBAS scale (Hirano, 1981), five parameters are examined: G (Grade), which refers to the overall voice quality, R (Roughness), B (Breathiness), A (Aesthenicity) and S (Strain, or vocal tension).

in the voice production of the cotwins for jitter and shimmer alone. The explanation put forward for this pattern was that “the presence of frequency and amplitude perturbation parameters may be influenced by many factors (eg., environment, state of health, anxiety, tension)” (Van Lierde et al., 2005: 517). However, they found high correlations in aerodynamic measures of Maximum Phonation Time (MPT), in acoustic parameters (f_0 and intensity) as well as in the perceptual evaluations. Therefore, they conclude that these voice characteristics must be genetically determined:

> As a result of the high correlations found […], we are inclined to think that muscular contractions and vocal fold vibratory amplitude motion (to achieve higher pitch) and subglottal pressure (for intensity control) are also genetically transmitted in healthy MZ. (Van Lierde et al., 2005: 517)

We should note that in this study the classical “twin method” has not been followed, since only MZ twins have been recruited for the experiments. The absence of results for DZ twins to compare with the MZ twins’ results might diminish the strength of this statement. However, the fact that further investigation is needed to sustain their conclusions (although not mentioning explicitly the inclusion of DZ twins) is already acknowledged by the authors.

The research carried out by Cielo, Agustini and Finger (2010) belongs to the studies on voice quality. Although the twin sample is quite small (two MZ pairs, one per gender), their analysis is interesting as far as they tackle some features that have not been considered in twins’ studies before, namely *vocal onset* and resonance characterization, measured in *number of harmonics*. Other of the parameters analyzed include f_0, loudness and MPT, which have been actually analyzed in previous studies on twins, as we have mentioned above. A perceptual assessment of the voices was also carried out, for instance, to evaluate the twins’ breathing types. No significant differences were found between twins as regards vocal onset, f_0 or loudness, while the results for MPT showed significant differences. This could be explained, at least for one twin pair, in terms of exogenous influences: one member of the twin pair was sportier and has trained his voice as a journalist.

As in Section 2.4.1, it seems appropriate to finish with the review of the scarce literature on siblings’ voices. Besides Whiteside and Rixon (2004), who studied both twin pairs and sex-matched siblings (therefore we mentioned this study above), there are a couple of further studies on siblings which deserve some attention. In an attempt to characterize numerically the similarity or dissimilarity of the voices of sisters and identical female twins, Kinga (2007) analyzes the following parameters in three female 21-22-year-old MZ twin pairs and in both members of three female 20-24-year-old pairs of sisters: *average pitch*, *first three formants and formant bandwidths* of nine vowels, *duration of words, vowels and alveolar fricatives, word intensity* and *Fast Fourier Transform (FFT) spectra of vowels.* Despite concluding that complete discrimination cannot be accomplished by a single parameter, the results of her study show that intra-pair differences between MZ twins are lower than those between sisters.

Feiser (2009) based her study on a larger database than Kinga (2007), namely, 10 German-speaking male and 10 female siblings in the age range between 19 and 27 years. Under a clear forensic-phonetic perspective, the acoustic analysis carried out in this study consists of an examination of the following parameters: (a) *fundamental frequency*, (b) the frequencies of *formants F1 to F4* in selected vowels, and (c) *Long Term Average Spectra (LTAS)*. The results show that "acoustic similarities between pairs of unrelated speakers (sibling-external comparisons) are often stronger than similarities between siblings (sibling-internal comparisons)" (Feiser, 2009: 1). This was especially evident when considering f_0, what would suggest that in real forensic casework involving siblings this parameter could reveal differences between them despite their being very similar-sounding.

Finally, the results of a preliminary acoustic study of three Spanish-speaking brothers (San Segundo, 2010a) showed that F3 and F4 yielded more significant differences than F1 and F2 in the inter-speaker comparison of these siblings. However, it was suggested that the relevance of low formants (F1-F2) should not be underestimated, as the specific comparison of two of the three brothers resulted in significant differences for the F1 mean values in all the five vowels examined. The conclusion that more studies would be necessary to understand voice (dis)similarities in related speakers was clearly derived from this investigation, as only three siblings were recruited and the speech material from where the vowels were extracted consisted in carrier sentences.

Recapitulation

The main conclusions that we can draw from all these studies are:

(1) First of all, twins are not necessarily identical as their voice concerns. Furthermore, despite being genetically identical (MZ twins) or very similar (DZ twins and siblings), they can be distinguished (maybe among other possible ways) "by making use of the leeway allowed them by the phonological system of their language", as Nolan and Oh (1998) put it.

(2) The degree of similarities and dissimilarities in twins' speech is not uniform across twin pairs (Loakes, 2006a).

(3) Many different acoustic parameters have been studied to assess twins' (dis)similarities. Our literature review comprises the following ones: a) fundamental frequency (f_0) and related parameters, like average VFF (Vocal Fundamental Frequency) and long-term f_0; b) spectral

parameters, like spectrum form, number of harmonics and frequency peaks; c) intensity measurements; d) Voice Onset Time (VOT); e) temporal parameters; f) coarticulation patterns; g) formant patterns; h) formant transitions; i) plosives frication; j) shimmer, jitter and HNR; k) voice quality features, like GRBAS-scale perceptual evaluations and Dysphonia Severity Index (DSI) measurements.

(4) Among the studies reviewed, only a few of them analyze a sufficiently large number of speakers. The average number of participants is 26 twin pairs[41] (twin participation ranges from 1 pair to 202 pairs). Being the standard deviation 41 and the mode 1 (we find only exceptionally large number of participants, like 202, 117, 62 or 60 [see Appendix E]), we calculated the median (12 twins), which depicts more accurately the usual number of participants recruited for these studies. Some of the studies reviewed only recruit MZ twins as participants, while most of them interest themselves in both MZ and DZ pairs; only a minority focuses on siblings.

(5) Among those studies related to siblings, apart from those works approached from a perceptual point of view [see Section 2.4.1], only three are tackled from an acoustic perspective: both Whiteside and Rixon (2004) and Kinga (2007) studied MZ twin pairs and sex-matched siblings (the former studied coarticulation patterns while the latter undertook a multiparameter-approach). Feiser (2009: 1) focused only on siblings, finding that "acoustic similarities between pairs of unrelated speakers (sibling-external comparisons) are often stronger than similarities between siblings (sibling-internal comparisons)". It is noteworthy that no study has considered so far the study of both types of twins (MZ and DZ) together with siblings.

(6) Of all the studies reviewed in this section, only Nolan and Oh (1998), Loakes (2006a, 2006b) and Feiser (2009) stem clearly from a forensic point of view, as they state it. However, it seems that only Loakes (2006a) takes into account forensic realistic conditions such as channel-mismatch, non-contemporaneous speech samples and different speaking styles, including spontaneous speech.

(7) The main research objectives of these studies are either a) trying to find a genetic component in the variation of certain voice parameter by searching differences between MZ and DZ twin pairs (Alpert et al., 1963; Przybyla et al., 1992; Debruyne et al., 2002) or b) creating a system capable to discriminate between MZ and DZ twins through the analysis of a large number of acoustic parameters (Forrai & Gordos, 1983) or else, in a forensic context, c) trying to test whether it is possible to identify a speaker, distinguishing him/her from his/her co-twin (Loakes, 2006a, 2006b; Nolan & Oh, 1998).

[41] For this count, we have considered all the studies about twins, not distinguishing between those related to perception, acoustics or articulation.

(8) Among the main results obtained in these studies, some seem to point out to the difficulty or impossibility to discern, for certain voice parameters, the influence of genetic factors from the influence of shared environment (e.g. f_0: Debruyne et al., 2002). Some results may apply only to the language spoken by the twins, such as the findings related to formant transitions in sequences like /ʃə/ in German (Weirich, 2011) while other results are supposed to be less language-dependant, for instance, the finding that F3 is the most speaker-specific formant frequency (Loakes, 2006a). Of course, it would be necessary to have studies in other languages in order to confirm this. Caution should be taken when comparing the same voice parameters from the results obtained by different studies with a divergent number of speakers, or analyzing very different speaking styles.

2.4.3. Articulatory studies

There are few phonetic studies which use articulatory techniques to measure similarities and differences between twins, at least comparatively fewer than studies dealing with acoustic parameters or with perceptual tests. This is understandable taken into account the intrusiveness of these techniques, as well as their expensiveness and (sometimes consequently) their scarcity in most phonetic laboratories. Although the value of this approach for describing twins' similarity is clear, the forensic applicability of these techniques is more doubtful, being the main reason that palatography or EMA (electromagnetic articulograph) techniques could hardly be used in real forensic casework. At least, these articulatory data are not to be found in the case of the offender's recording and thus, would not be comparable with the suspect's data.

The first account of differences in twins as regards the physiological and anatomical structures used for speech production probably dates back to mid-twentieth century with the study of Lundström (1948). His study is, nevertheless, included in the discipline of Dentistry, not in Phonetics. Therefore, this article is more interesting from a physiological point of view, rather than from an articulatory approach *per se*. This author showed that "natural variations between MZ twins are much smaller than between DZ twins in terms of a variety parameters related to the jaw and teeth, such as size, breadth, position and inclination of teeth, overbite and others" (Lundström, 1948, in Künzel, 2010: 252).[42]

[42] As cited in Künzel (2011: 252), "on the whole, it appears that genetic factors play an important part, in any case equally important as environmental factors" (Lundström 1948: 187, in Künzel 2011: 252). Yet, it seems that when extreme malocclusions happen, heredity is the most important factor (Lundström 1948: 187, in Künzel 2011: 252).

Likewise, the interest of Spielman et al. (2011) lies in the resemblance of tongue anatomy in twins (6 MZ pairs and 3 DZ pairs). After 30 subjects were asked to match the photographs of the tongues from twins, it was found that, "based on visual assessment, monozygotic twins have highly similar tongues (60% matches); similarly, dizygotic twins were matched 31% of the time, which is a higher probability than would be expected from random selection" (Spielman et al., 2011: 277). We reckon that this article does not belong properly to the group of articulatory studies, notwithstanding its relevance to characterize twins' different physiological structures. We are interested in reviewing the results found in studies dealing with speech articulation in twins. In this respect, the first research where it is actually stated that there is a great overlap in the articulation skills of twins is Whiteside and Rixon (2004). They also refer to three possible factors influencing this overlap, which is larger in MZ twins than in DZ twins:

> Both morphological (Locke & Mather, 1989), cognitive and neuromuscular factors (Matheny & Bruggemann, 1973) have been proposed as explanations for the greater overlap in the articulation skills of MZ twins when compared to dizygotic (DZ) twins. These suggestions are supported by recent evidence which suggests that those brain structures which subserve speech and language input and output processing (e.g., sensorimotor cortex, linguistic cortices such as Broca's and Wernicke's areas as well as frontal brain regions) are also influenced genetically, and that MZ twins display very high levels of similarity in these brain regions (Thompson, Cannon, Narr, van Erp, Poutanen, Huttunen, Lönnqvist, Standertskjöld- Nordenstam, Kaprio, Khaledy, Dail, Zoumalan & Toga, 2001; Plomin and Kosslyn, 2001; as cited in Whiteside & Rixon, 2004: 3-4).

Weirich (2011) approaches the study of twins' voices from a threefold (perceptual-acoustic-articulatory) perspective. As we have already dealt with the two first, let's now turn to the main results found under the articulatory perspective. We have explained that this author makes a basic distinction between *static* and *dynamic* approaches to voice parameters. While under the acoustic approach dynamic aspects of the acoustic signal (formant transitions) were found to be more relevant than static parameters (formant frequencies of particular speech sounds) to discriminate between twins, similarly, in terms of articulatory techniques the /aka/ gesture was found to show fewer differences in MZ twins than in DZ twin. This would show that coarticulatory strategies, measured by means of palatography, are organically determined and thus idiosyncratic.

Weirich also conducted articulatory measurements of the fricative /ʃ/ and the affricate /tʃ/ in CV syllables. The articulatory recordings were carried out using a 2-D electromagnetic articulograph. Using this technique, she also found that individual physiology shapes articulation more in sibilants than in vowels and that there is a difference in inter-speaker variability between

MZ and DZ twins in their articulatory strategy to realize the phoneme contrast between /s/ and /ʃ/ in German:

> The amount of distance and especially the relation of the vertical to the horizontal distance between the two target tongue positions were more similar in MZ than in DZ twins. Thus, the precise realization of the phoneme contrast is influenced by the individual palatal shape and hence NATURE. (Weirich, 2011: 234)

Before this study, the same author had previously analyzed the articulatory behavior of twins in Weirich (2010). Following the same methodology (i.e. EMA), her results suggest that "MZ twins […] are more similar in their articulatory targets (vertical and horizontal tongue positions) […] of the vowel /i/ when a velar consonant precedes the vowel". Similarly, she found supporting evidence for the existence of an interaction between physiology and syllable stress:

> Physiology seems to have a stronger influence on the production of /i/ when produced in an unstressed syllable. Both DZ twins revealed more differences in formants in the unstressed condition, and the 2 female MZ twins with the remarkable similar palatal shape showed more differences in formants in the stressed condition. In their articulatory targets the MZ pairs revealed no inter-speaker variability in the unstressed condition, but one of the DZ twins did. (Weirich, 2011: 234)

As with any other study with a small sample, some caution should be taken not to extrapolate these results, since only 5 subjects participated in the experiment: 3 MZ twin pairs (2 female and 1 male pair) and 2 DZ female twin pairs.

Recapitulation

The main conclusions that we can draw from all these studies are:
(1) The phonetic studies on twins which undertake an articulatory perspective are clearly minoritary and quite recent (Weirich, 2010; Weirich, 2011).

(2) Some studies can be found which focus on twins' physiology of, for instance, the teeth and the jaw (Lundström, 1948) or the tongue (Spielman et al., 2012) but they lack a phonetic point of view and limit themselves to the description of the size and form of the above-mentioned anatomic structures.

(3) The main findings from the scarce existing studies are: (a) that MZ are more similar than DZ twins in their articulatory strategy (measured as the tongue position, both vertically and horizontally) to realize the phoneme contrast between /s/ and /ʃ/ in German (Weirich, 2011), and

(b) that both the syllable stress and the presence of a velar consonant before a vowel intensify the impact of the identical physiology of the vocal apparatus of MZ twins (Weirich, 2010).

(4) A possible reason for the almost complete absence of articulatory approaches to twins' voices could be the non-linear relationship between articulation and acoustics (Stevens, 1972): "whereas small differences in articulation can result in large differences in acoustics, some differences in articulation do not necessarily result in differences in the acoustic output" (Weirich, 2011: 6). This could be the explanation for the lack of investigations, especially in Forensic Phonetics, focusing on articulatory analysis in general, and on twins' voices in particular. In real casework, the police officers are almost certainly not allowed to place intrusive devices in the suspects' oral cavity and, in case they could, the lack of articulatory data for the offender –together with the above mentioned non-linear relationship between acoustics and articulation– makes the use of these articulatory techniques unfeasible in terms of forensic applicability, notwithstanding its clear research interest.

2.4.4. Automatic approaches

In this section, we will review in chronological order the different studies on twins' voices that have been carried out so far in the field of automatic speaker recognition and verification.

Homayounpour and Chollet (1995) adopted a three-folded approach (perceptual, acoustic and automatic). Basically, they compared the results obtained by listeners (familiar and unfamiliar to twins) in a perceptual study with the results provided by two automatic systems. First, a listening test was designed in order to assess the difficulties involved in the discrimination of twins and also to verify if there was a large difference in speaker verification performance when the listeners were familiar with the twins, as compared with listeners not familiar with them. This approach resulted in smaller recognition error rates when a listener was familiar with the twin voices. In a further analysis step, the acoustic approach based on Long-term Spectra (LTS) showed that twins have very close LTS but only when they were recorded over the same telephone line. Since LTS was very different when twins were recorded over different telephone lines or handsets, LTS was rejected as a relevant feature to distinguish between twins. As regards the third and last approach, two automatic systems were developed: one based on a LVQ3[43] supervised neural net algorithm and another one based on a SOSM (Second Order Statistical) measure. The main results of this study were that both automatic speaker verification systems discriminate the voices of identical twins worse than listeners familiar with them. However, these automatic

[43] This concept is not defined in the article but is supposed to mean Learning Vector Quantization (LVQ).

systems and listeners not familiar with the twins have about the same ability to discriminate between identical twins. A possible explanation for this is offered by the researchers:

> Automatic speaker verification systems use only low level features which are related to the acoustic aspects of speech. The spectral representations of speech such as Cepstrum and delta Cepstrum parameters cannot capture the behavioural differences between the twins. But, of course, it should be noticed that twin relatives and friends have received much more speech material for training than our automatic system. (Homayounpour & Chollet, 1995: 301)

This links to a crucial issue in the study of twins, namely the nature-nurture dichotomy, which is only superficially tackled by the authors at this point, when they highlight the difficulty of automatic systems to capture behavioral differences between twins. However, an important question, which is indeed raised by the same authors, remains: Are the results of speaker verification by human listeners comparable to those of automatic systems on a twin database?

Scheffer et al. (2004) performed two kinds of experiments. Both were carried out using the LIA (*Laboratoire d'Informatique d'Avignon*) speaker recognition platform AMIRAL, developed by the Laboratory of Computer Sciences in Avignon, France. In the first experiment, aimed at identifying the twin of a certain speaker, their automatic system was able to actually identify a twin with an acceptable performance (85% of good identification). The second experiment was a classical speaker verification task, for which they obtained 6% False Acceptances (False Positive) for 0% False Rejections and 53% False Rejections for 0% False Acceptances. The database was made up of recordings from 17 MZ twins (10 female and 7 male). Each one had read a text passage of around 40-70 seconds. For the parameterization of the acoustic signal the method used was MFCC (Mel Frequency Cepstrum Coefficients).

As regards the first experiment, it serves the authors to confirm the great similarity of voice between MZ twins. However, 4 out of 34 twins (approximately 15% error) were not detected correctly as the twins of their actual twins, which would suggest that "the twin of a speaker is not necessarily the most difficult impostor for an automatic speaker recognition system" (Scheffer et al., 2004: 2).

Furthermore, some other interesting results were found by these researchers: (1) one of the speakers (Speaker A) had a cold at the time of the recording and this person was not identified as the twin of his actual twin (Speaker B). Likewise, Speaker B was not identified as the twin of Speaker A; (2) However, the twin (Speaker C) of another speaker (Speaker D) had just undergone surgery in one of his vocal folds and was yet correctly identified as the twin of Speaker D. (3) Even more surprising is the case of two other twin pairs (Speakers E and F, and Speakers G and H), of whom only one twin member in each pair (Speaker E in one pair, and Speaker G in another pair), not having any particular voice problems, were not identified by the automatic system as the twin of this actual twin. On the contrary, their cotwins (Speaker F in one pair, and Speaker H in the other pair) were correctly identified.

For the second experiment, i.e. the automatic speaker verification, a UBM (Universal Background Model) is created in order to comply with the use of Likelihood Ratios, under a Bayesian approach. Besides, impostors are added to the system. However these do not pose a problem to the verification, since the voice similarity between co-twins is larger than the similarity between the impostors.

Ariyaeeinia et al. (2008) define speaker verification as a "principal subclass of speaker recognition (voice biometrics)" and begin by acknowledging how important the study of MZ twins is in this field:

> An important issue in the field of automatic speaker verification (SV) is the potential challenge posed by identical (monozygotic) twins. The expectation of this challenge is due to the general concept that monozygotic twins should be highly similar in every respect including their voices. (Ariyaeeinia et al., 2008: 182)

Ariyaeeinia et al. (2008) based their study on a database consisting of speech from 98 verified MZ twins (40 female pairs and 9 male pairs), so it was very dominant in female gender. As regards the characteristics of the speech data, every speaker was recorded first reading a poem (approximately 60 seconds in duration) and secondly saying their date of birth, spoken as digits (approximately 5 seconds in duration). The former type of speech constitutes the "long test data" of a speaker and the latter the "short test data". For the automatic system used, LPCC (Linear Predictive Coding-Derived Cepstral) parameters were extracted and the speaker representation was based on the use of adapted Gaussian Mixture Models (GMMs). The results, presented in terms of Equal Error Rates (EER %) showed that the use of long test utterances lead to smaller error rates than the use of short test utterances. According to the author's own explanation, "this is a clear indication of the non-genetic (extraneous) factors influencing the characteristics of the voices of each pair of the twins" (Ariyaeeinia et al., 2008: 185).

The authors resolved to use unconstrained cohort score normalisation (UCN) since this is supposed to lead to dissimilarities between the non-genetic characteristics of the monozygotic twins. Using this approach, it was possible to exploit the non-genetic characteristics of the twins' voices for the benefit of increasing the discrimination capability of speaker verification: EER were reduced from over 2.8% to around 0.5% in the case of short test utterances (about 5 seconds in duration).

Kim (2009) studied 22 Korean female twin pairs (17 MZ, including 1 triplet and 5 DZ) using *Agnitio's Batvox 3.0*, an automatic speaker recognition program. Two different speaking styles –text reading and spontaneous interview– were used. The results showed that every twin speaker was correctly identified in the same speaking style condition (when models and test files were "read" speech). According to the author, this would suggest that, at least in automatic speaker recognition, the same speaking style setting should be provided in order to get more confident results. Noteworthy of this study is also that in 9 out of 22 pairs, intra-twin LRs in the same speaking style condition were higher than intra-speaker LRs in different speaking style condition. This situation is highly undesirable in a forensic context, where inter-speaker variation should be larger than intra-speaker variation (Wolf, 1972).

Künzel (2010) is the most recent study on automatic speaker recognition, in which a Bayes-based system (*Batvox 3.1*) was used to calculate LR distributions for inter-speaker, intra-pair and intra-speaker comparisons. A total of 35 MZ pairs (26 female and 9 male) participated in this study and two different tests were designed. In the first one, both target voices consisted of the same read text, while in the second one the speaker models were built from spontaneous speech samples but read speech samples were used as targets. The results showed that in the first experiment the automatic system allowed a perfect distinction of each member of a male twin pair (i.e., 0% of Equal-Error Rate, EER) and 0.5% EER for female twin pairs. In the second experiment, the EER rose to 11% for male twin pairs and 4.4% for female twin pairs. These values represent the crossover point in the Tippett plot for the inter speaker / intra speaker LR distributions. However, the results for female twins are worse when considering intra-pair / intra-speaker distributions (19% EER in the first experiment and 48% EER for the second experiment). Therefore, the performance of the system was clearly superior for male than for female voices. The author's explanation for this phenomenon is that "as a consequence of the higher fundamental frequency of female voices the spacing of the harmonics is less dense than for male voices, which in turn yields less speech sound- and speaker information in the spectrum" (Künzel, 2010: 270). This becomes clearer if we bear in mind that the spectrum is used for the extraction of the MFCCs (mel-frequency cepstrum coefficients), which are the features in which the automatic system used is based upon.

We have not found in the literature review any reference to the similarity of siblings' voices analyzed under an automatic speaker recognition approach. The closer account of this phenomenon could be that of Charlet and Peral (2007), who tested France Telecom R&D speaker recognition system with voices in a family context. For this aim, 33 families were recorded pronouncing their complete name. The reason which triggered this study was that "the genetic links that exist between parents and children as well as the same cultural and geographical contexts they share may hamper speaker recognition" (Charlet & Peral, 2007: 93). The text-dependent speaker recognition system developed by France Telecom R&D relies basically on HMM modeling of cepstral features and, in this case, an open-set speaker identification task was carried out. Of each type of speaker (father, mother, daughter/s and son/s) per family, their ability to defeat the system and their "fragility" to impostor attempts (Doddington et al., 1998, in Charlet & Peral, 2007) was evaluated. For the purpose of our dissertation topic, the most relevant results of this study lie in the performance of the system with siblings' voices, particularly in the case of brothers:

> The son, as an impostor, has a low success rate on his father, a high level on his mother and brother and a very high level on his sister (the difference between brother and sister success rate might not be significant because of the small number of attempts in the case of 2 brothers attempts). As a target, he is moderately fragile against his parents and highly against his sister. (Charlet & Peral, 2007: 99)

Recapitulation

The main conclusions that we can draw from all these studies are:

(1) The automatic approach to twins' voices is clearly the less developed so far, only after the articulatory perspective, in number of studies. This is probably due to the fact that the automatic systems used for speaker identification are created (and mostly used) by engineers, who are only a subgroup of professionals interested in the study of voice.

(2) The main objectives of the studies reviewed in this section have been one of the following ones: a) comparing the performance of an automatic system with the ability of familiar and non-familiar listeners to discriminate twins' voices, b) testing if an automatic system was able to detect correctly which speaker was the twin of which speaker; c) in general, testing the intra-speaker, inter-speaker and intra-pair similarity of twins, for example in terms of Likelihood-Ratios.

(3) According to the above-mentioned objectives, the main results obtained so far are: a) the automatic system in Homayounpour and Chollet (1995) discriminates the voices of MZ twins worse than listeners familiar with them. However, its performance would equal non-familiar listeners in the "ability" to discriminate between identical twins. It is nevertheless doubtful that the results of speaker verification by human listeners are comparable to those of an automatic system; b) Not every twin was correctly matched with his twin in Scheffer et al. (2004). Although an approximately 15% of error would suggest that the twin of a speaker is not necessarily the most difficult impostor for an automatic speaker recognition system, in general twins apperar to be good voice impostors, which justifies, in the forensic field, the importance of finding certain parameters in which (even) they can be distinguished. The specific twin pairs for whom a match error happened should be studied further, for instance through a detailed acoustic analysis –like in our study looking at biometric parameters– in order to gain some knowledge of the possible

causes of mismatch; c) Kim (2009) and Künzel (2010) are similar in that they use the same Bayes-based automatic speaker recognition system to calculate inter-speaker, intra-pair and intra-speaker distributions: *Agnitio's Batvox.* In both cases, the twins are recorded in two different speaking styles: text reading and spontaneous interview. The study of Kim (2009) proves that there can exist more intra-speaker variation than inter-speaker variation in the case of twins. While Kim (2009) focuses on female twins, Künzel (2010) considers both male and female twins, finding an interesting result: the performance of the system was clearly superior for male than for female voices. This phenomenon finds an explanation in the different spacing of harmonics in male and female voices: narrower in the former and broader in the latter. Since the automatic system used is based on MFCCs, the portion of the spectrum from which these MFCCs are extracted has a much higher amount of speaker-related information when a narrower spacing of the harmonics exists.

(4) Finally, it is important to note that, although most of these studies analyze a large number of twins, these are not recruited on a dialectal basis, as is the case in our study, where all the twins speak the same language variety.

(5) References to siblings' voices within an automatic approach are almost inexistent except for the study of Charlet and Peral (2007), whose results are especially interesting for the purpose of our dissertation in that, in a text-dependent speaker recognition system tested with 33 families, the son was highly confused with his brother. This implies that someone could be a good impostor of his brother's voice, making this type of speakers especially relevant in forensic studies and thus justifying not only the study of twins but also of non-twin siblings.

3. METHOD

3.1. Introduction

In this chapter we will first describe the subjects selected for recording, detailing how they are distributed in four groups, depending on whether they are MZ twins, DZ twins, brothers or unrelated speakers making up the reference population. We will also explain the reasons for having selected only male speakers (cf. Section 3.2). Secondly, we will thoroughly describe the procedure for the corpus elaboration, consisting on five speaking tasks and a vocal control technique (cf. Section 3.3). Thirdly, we will focus on the recording procedure, which will include an account of the materials and technical characteristics of the recordings followed by a description of the data collection set-up (cf. Section 3.4). Next section is devoted to explain how the telephone filtering was carried out (cf. Section 3.5) and, finally, the last section will focus on the likelihood-ratio approach within which the results are offered (cf. Section 3.6). The speech material used in this thesis and the different acoustic analyses performed will be described in detail in the corresponding section, according to the type of analysis (Chapters 4 to 6).

3.2. Participants

For this study we have recruited 54 speakers, distributed in four different groups:

1. Monozygotic twins: 24 speakers (12 pairs)
2. Dizygotic twins: 10 speakers (5 pairs)
3. Full brothers: 8 speakers (4 pairs)
4. Unrelated speakers (friends or work colleagues) as reference population: 12 speakers (6 pairs)

The speakers had to come in pairs for the recordings, as we will explain later in Section 3.4. We have already referred to the existence of two main types of twins. Monozygotic twins (also called identical) develop from one zygote that splits and forms two embryos, while dizygotic(also called "fraternal") develop from two separate eggs that are fertilized by two separate sperm cells (Abril et al., 2009: 90). Full brothers are male siblings with the same father and the same mother. The importance of these three first speaker groups has been acknowledged elsewhere [see Chapter 2].

Friends or work colleagues –the fourth speaker group– were recruited in order to create a reference population, whose relevance for Likelihood-Ratio-based forensic studies will be explained in Section 3.6. Besides fulfilling the age and dialect criteria, the only requisite for their participation in this study was that they had to come in pairs either with a friend or work colleague. The importance of this requisite lies in the search for a speaking style similar, and thus comparable, to that found in the conversations between twins, usually characterized by their spontaneity due to a close long-term relationship.

The ages of the speakers recruited for this study ranged between 18 and 52 years old (median age: 28.96). The age difference between the siblings in each pair varied between four and eleven years (see Table 6).In all cases they had an adult voice, neither presbiphonic nor adolescent. One of the requirements for recruiting the subjects of our study was that they had to be male speakers. The reasons for establishing this criterion were:

a) In real forensic cases, there is a higher incidence of crimes committed by men[44].
b) For the study of one sex group, it is necessary (according to the method we have adopted) to record not only twins and brothers but also a reference population of normophonic speakers of the same sex[45]. Therefore, the inclusion of female speakers in this study would have implied the consideration of a further variable which would have doubled the number of speakers necessary: MZ and DZ twins, sisters and a group of normophonic female speakers. Consequently, this study is limited to a single sex.
c) Female voices are more difficult to study not only in a forensic context but generally in any phonetic analysis involving the harmonics. As explained in Section 2.4.4, Künzel (2010) found that the performance of the automatic system used in his study was clearly superior for male than for female voices due to the higher f_0 of the latter: "as a consequence of the higher fundamental frequency of female voices the spacing of the harmonics is less dense than for male voices, which in turn yields less speech sound- and speaker information in the spectrum" (Künzel, 2010: 270) Besides, since the telephone channel is a band-pass filter which cuts off frequencies below 300 Hz and above 3,400 Hz[46], some formant frequencies are more affected than others, especially higher formants (see, for example, Künzel, 2001 for the effect of telephone transmission on the measurement of formant frequencies).

[44] According to the statistical annual directory of the Spanish *Ministerio del Interior* (Ministerio del Interior, 2011), 92.5% of the Spanish prisoners in 2011 were male. Besides, Rose (n.d.) notes that "the vast majority of crimes committed where FSI is required (armed robbery, blackmail threats, bomb threats, murder threats, drug offences) are by males".

[45] In order to account for typicality, see Section 3.6.

[46] The frequency cut-off depends on whether the recordings are made via landline telephone, GSM (Global System for Mobile communications) or the cutting edge wideband codecs for VoIP (Voice over IP) like AMR-WB (Adaptative Multi Rate Wideband) which offers broader passband.

As concerns the dialectal aspects, the language variety spoken by all the subjects was North-Central Peninsular Spanish[47] (for a description of this variety, see Hualde 2005). They all were native speakers of this variety coming from different regions in Spain. However, the majority of them were born and lived in either Madrid or Castile-Leon (see Tables 4-7).

Regarding the speakers' recruitment method, there were various strategies for getting enough number of participants, which was especially difficult in the case of twins due to their low incidence.[48] In our case, it was more difficult finding DZ twins than MZ twins since the former may be both same-sex pairs and female-male pairs, while the latter are always of the same sex. Noting a lack of twin associations in Spain[49], some of the methods for recruiting participants were:

- Mailing lists in several universities and research centres.
- Posting information on notice boards in different places like public libraries.
- Creation of *Facebook* events.
- Sending *Twitter* messages.
- Snowball method[50].

While the sample size of the different speaker groups is uneven in this study (12 MZ pairs, 5 DZ pairs, 4 non-twin sibling pairs and 6 unrelated speaker pairs), it should be acknowledged that this situation is very common in the literature (see Appendix F for a summary of twin studies in chronological order), with the ratio MZ:DZ being sometimes really disproportionate. See the ratio 20:4 in Gedda, Fiori and Bruno (1960), 53:9 in Przybyla, Horii and Crawford (1992) or 17:5 in Kim (2009). Only in exceptional cases we find balanced distributions

[47] Speakers 19 and 20 (see Table 5) were born in Cáceres and speaker 32 (see Table 7) was born in Albacete. However, their accent was considered predominantly North-Central Peninsular Spanish as they no longer lived in their hometowns at the time of the recordings but in Madrid.

[48] "The rate of identical twins is constant at approximately four per thousand. It is remarkable that the incidence of identical twins remains the same no matter where a person lives, and it has remained the same throughout history. The rate of fraternal twins, on the other hand, can change depending on where a person lives, the mother's age, etc. Fraternals account for the differences in the twin rate, the fraternal rate being approximately 22.8 per thousand in the world." (National Organization of Mothers of Twins Clubs, Inc., n.d.)

[49] Actually, Boomsma (1998: 35) in his overview of twin registers in Europe, acknowledge that "unfortunately, twin registers from Southern European countries are currently underrepresented".

[50] This is the method used in Decoster, Van Gysel, Vercammen and Debruyne (2001) and described elsewhere as follows:

> Researchers use this sampling method if the sample for the study is very rare or is limited to a very small subgroup of the population. This type of sampling technique works like chain referral. After observing the initial subject, the researcher asks for assistance from the subject to help identify people with a similar trait of interest.
>
> The process of snowball sampling is much like asking your subjects to nominate another person with the same trait as your next subject. The researcher then observes the nominated subjects and continues in the same way until the obtaining sufficient number of subjects. (Explorable.com, 2009).

of MZ and DZ pairs like 100:100 (Lundström, 1948) or 30:30 (Debruyne, Decoster, Van Gysel, & Vercammen, 2002). Yet, it should be noted that the data collection method in those cases implied the use of previously collected twin registries, which undoubtly makes the search for participants easier. While the existence of balanced distributions of MZ and DZ twins is advantageous for statistical purposes, it should be highlighted that the uneven sample size in our study has always been borne in mind in the discussion of the results for all the three voice analyses considered in this investigation[51].

The candidates for participation in this study first had to fill in an online questionnaire (see Appendix A1) which was designed 'ad hoc' in order to assess their suitability as regards the age and language criteria, besides gathering some other useful information like possible voice pathologies. This questionnaire was created through *www.e-encuesta.com.*[52]

The subjects who were finally selected for participating in this study had to fill in a more complete questionnaire at the day of the recording (See appendix A2). There were actually two recording sessions on different days, due to the importance of accounting for intra-speaker variability, as explained in Chapter 1. In the first recording session, the speakers had to fill in a questionnaire of 12 pages in the case of siblings (twins and brothers) and 9 in the case of the reference population. In the second recording session, all the speakers had to fill in another questionnaire, this time shorter (4 pages in all cases). This second questionnaire was aimed at evaluating any possible change in the health condition of the speaker in the time elapsed between one session and the other. The two sessions are separated by 2-4 weeks (mean: 22 days), so they are non-contemporaneous sessions.

The main questionnaire (i.e. the first one) includes questions about *personal data* (name, surname, birth date and contact details); *linguistic data* (residence places, in Spain and abroad, and duration; mother tongue, languages spoken and proficiency level; residence places and languages spoken by father, mother and partner (if any); *health*[53] (voice and speech pathologies, hearing difficulties, smoking habits, etc.); *other data* (studies, profession, leisure activities involving voice use and abuse, etc.). This is part A of the questionnaire, which is the same for all speakers. However, for twins and brothers a further Section B is included in the questionnaire. In this section, there are questions about members of the family (number, age and sex), whether the siblings participating in the study share leisure activities and friends, whether they went to the

[51] Note that Debruyne et al. (2002: 467) point out that "in the literature, especially DZ are often poorly represented". Time and budget constraints are the main reasons for not having considered a larger sample of non-twin siblings and unrelated speakers in our study who could balance the number of MZ speakers.

[52] The full link where the online questionnaire could be found is the following one: http://www.e-encuesta.com/answer.do?testid=hoovkxCrN4g=&chk=1

[53] For the creation of the questions in this section we asked for the guidance of an otorhinolaryngologist.

same school and the same classroom, how often they see each other, how often they talk and other questions about their relationship (if they like having a twin or brother, how close their relationship is – in a 1-5 Likert scale–, who is more confident, if they think they are different or similar and if they think they speak similarly or differently).

In this questionnaire, some questions are related to their voice resemblance and similar soundingness, following Loakes (2006a). In our study, the twins were asked to check (via official documents) whether they were MZ or DZ twins, in case they weren't sure. Besides, in the case of one MZ twin pair we collected saliva samples of them through mucosal scraping in order to have a DNA testing done. The reasons for having this test done were the following ones:

1) This twin pair was the only one who was unsure of their zygosity (whether they were MZ or DZ twins), although they thought they were MZ twins.

2) Their answers in the questionnaire about similar-soundingness were very discordant with the other MZ twin pairs: they said they were seldom confused aurally.

3) In a previous study (San Segundo, 2012) this twin pair obtained very low LRs (see Section 3.6) in a voice quality analysis, in comparison with other MZ twin pairs, which would cast doubt on their true zygosity.

This DNA test was carried out by the *Instituto Nacional de Toxicología y Ciencias Forenses* (Spanish National Toxicology and Forensic Sciences Institute) and it exactly consisted in a DNA profiling of 16 short tandem repeat (STR) loci, which was performed by multiplex microsatellite typing using the AmpflSTR NGM Select PCR Amplification Kit (Applied Biosystems) according to the manufacturer's protocol. Both samples (that of twin MML and that of twin PML) yielded the same DNA profile with a likelihood ratio of 1.19E23, meaning that they actually were MZ twins.

The participants could be compensated for their participation thanks to a grant awarded to Prof. Dr. Künzel (University of Marburg, Germany) and to the author by the International Association for Forensic Phonetics and Acoustics (IAFPA). Prior to the execution of the first recording, the speakers had to sign an informed consent, including the following sections: general description of the research, participant selection criteria, characteristics and duration of the speaking tasks, risks (i.e. absence of risk associated with the research), confidentiality, right to refuse or withdraw, benefits –economic compensation– and contact email of the researcher for future communication.

Table 4

Datasheet for the MZ twins

Speaker pair	Speaker initials	Date 1st recording session	Date 2nd recording session	Birthdate	Birthplace
01-02	APJ	03/01/2012	06/02/2012	1991	Ávila
	RPJ	03/01/2012	06/02/2012	1991	Ávila
03-04	CGP	11/01/2012	10/02/2012	1984	Madrid
	AGP	11/01/2012	10/02/2012	1984	Madrid
05-06	EMG	01/02/2012	20/02/2012	1992	Madrid
	AMG	01/02/2012	20/02/2012	1992	Madrid
07-08	PAS	07/02/2012	29/02/2012	1976	Madrid
	CAS	07/02/2012	29/02/2012	1976	Madrid
09-10	JCT	16/02/2012	06/03/2012	1992	Madrid
	DCT	16/02/2012	06/03/2012	1992	Madrid
11-12	PML	13/02/2012	29/02/2012	1979	Madrid
	MML	13/02/2012	29/02/2012	1979	Madrid
33-34	RSM	11/07/2012	02/08/2012	1994	Madrid
	ASM	11/07/2012	02/08/2012	1994	Madrid
35-36	MHB	31/08/2012	21/09/2012	1982	Valladolid
	JHB	31/08/2012	21/09/2012	1982	Valladolid
37-38	CSD	13/09/2012	17/10/2012	1979	Ávila
	DSD	13/09/2012	17/10/2012	1979	Ávila
39-40	DSA	02/10/2012	07/11/2012	1976	Madrid
	ISA	02/10/2012	07/11/2012	1976	Madrid
41-42	ARJ	05/10/2012	23/10/2012	1993	Madrid
	JRJ	05/10/2012	23/10/2012	1993	Madrid
43-44	SGF	11/12/2012	20/12/2012	1984	Salamanca
	AGF	11/12/2012	20/12/2012	1984	Salamanca

Table 5

Datasheet for the DZ twins

Speaker pair	Speaker initials	Date 1st recording session	Date 2nd recording session	Birthdate	Birthplace
13-14	JRJ	05/01/2012	27/01/2012	1976	Madrid
	MRJ	05/01/2012	27/01/2012	1976	Madrid
15-16	IPG	13/01/2012	03/02/2012	1978	Madrid
	MPG	13/01/2012	03/02/2012	1978	Madrid
17-18	DZL	15/03/2012	29/03/2012	1994	Madrid
	PZL	15/03/2012	29/03/2012	1994	Madrid
19-20	PCL	09/03/2012	28/03/2012	1985	Cáceres
	ACL	09/03/2012	28/03/2012	1985	Cáceres
45-46	SSB	17/12/2012	02/01/2013	1988	Madrid
	VSB	17/12/2012	02/01/2013	1988	Madrid

Table 6

Datasheet for the brothers

Speaker pair	Speaker initials	1st recording session	2nd recording session	Birthdate	Birthplace
21-22	JCM	09/02/2012	27/02/2012	1960	Madrid
	FCM	09/02/2012	27/02/2012	1971	Madrid
23-24	NJM	01/03/2012	20/03/2012	1986	Burgos
	MJM	01/03/2012	20/03/2012	1979	Burgos
47-48	RPR	12/09/2012	10/10/2012	1985	Madrid
	DPR	12/09/2012	10/10/2012	1993	Madrid
49-50	IFC	29/08/2012	03/10/2012	1984	Madrid
	DFC	29/08/2012	03/10/2012	1988	Madrid

Table 7

Datasheet for the reference population

Speaker pair	Speaker initials	1st recording session	2nd recording session	Birthdate	Birthplace
25-26	FAM	23/12/2011	11/01/2012	1967	Guadalajara
	JPP	23/12/2011	11/01/2012	1984	San Sebastián
27-28	RDP	26/12/2011	19/01/2012	1983	Madrid
	DPC	26/12/2011	19/01/2012	1980	Madrid
29-30	DSF	08/02/2012	02/03/2012	1985	Madrid
	JAA	08/02/2012	02/03/2012	1965	Navarra
31-32	ESB	24/02/2012	16/03/2012	1978	Burgos
	AIP	24/02/2012	16/03/2012	1983	Albacete
51-52	CSM	07/11/2012	03/12/2012	1976	Santander
	RAG	07/11/2012	03/12/2012	1985	Santander
53-54	FVV	08/11/2012	28/11/2012	1984	Valladolid
	PCR	08/11/2012	28/11/2012	1982	Valladolid

3.3. Corpus Design

We have created an 'ad hoc' corpus for this thesis dissertation. In its design, we have distinguished five types of tasks[54] to be carried out by the speakers participating in this study:

1. Semi-structured spontaneous conversation
2. Fax exchange to elicit specific vocalic sequences
3. Reading of two phonetically-balanced texts
4. Mathematical calculations aimed at eliciting hesitation marks
5. Informal interview with the researcher

The objective of these tasks is to elicit certain speaking styles. Besides, the corpus includes the recording of a vocal control technique. It consists in asking the speakers to sustain both the vowel [a] and the consonant [s] –independently– as long as possible. This was repeated three times. Thanks to this vocal technique, two measures could be calculated: *Maximum Phonation Time (MPT)* and *s/a Ratio*, which are "two traditionally popular indirect clinical measures of respiratory integrity and laryngeal valving efficiency" (Aronson & Bless, 2009: 148). These measures are especially interesting for the posterior analysis of voice quality parameters.

In the following pages, we will describe each of the tasks which made up our corpus, explaining how they were performed by the speakers and which is the final end pursued with each of them. Then, the two vocal control techniques will be explained.

3.3.1. First task: semi-structured spontaneous conversation

In this first task, each pair of speakers, whether they come in pairs as twins, as brothers or as friends, had to hold a telephone conversation (see Section 3.4.2) of about 10 minutes. This task is called "semistructured" since the topics of the conversation are suggested by the researcher. Prior to the first recording session, the speakers have to read in silence a brief text about twin anecdotes and, then, during the telephone conversation, either they discuss whether they have similar funny anecdotes or not (in the case of the twins) or they talk about some twin pairs they have ever met (in the case of non-twin speakers, i.e. brothers and friends). In the case of the second recording session, several topics for conversation were suggested to the speakers (see Appendix B for the original instructions in Spanish): a) Speak with your partner about a situation in your life when

[54] Although not all the five speaking tasks have eventually involved an acoustic analysis for this thesis, they are still described here, as an important contribution of this thesis is the corpus design itself. The speaking tasks which have not been considered for this thesis could be useful for future investigations.

you felt you were in serious danger of death; b) What would you do if you had all the money in the world?; c) Speak with your partner about your favorite holidays.

The ‘danger of death’ question comes from the sociolinguistic research tradition (Labov, 1972). It is believed that when a speaker narrates such a dreadful situation, he does not pay attention to the way he is speaking but concentrates on what he is talking about. Similarly, the other conversation topics suggested are intended to have the same aim. They are adapted from the questions proposed in Loakes (2006a) to elicit spontaneous conversation.

3.3.2. Second task: Fax exchange to elicit specific vocalic sequences

The purpose of the second task is that, throughout an exchange of fax sheets, the speakers use certain words, whose interest lies in their inclusion of specific vocalic sequences. We will now describe the procedure followed to create the fax sheets in which this task is based.

1. *Methodology for the search of words containing the vocalic sequences of interest*

First of all, we should emphasize that the scope of this research is limited to the following 19 Spanish vocalic sequences:

ae, ao, ea, eo, oa, oe, ai, ei, oi, au, eu, ia, ie, io, ua, ue, uo, iu, ui.

The first step consisted in finding a sufficient number of words containing the above-mentioned vocalic sequences. For the search of such words we used *BuFón, Buscador de Patrones Fonológicos* (Alves, Rico & Roca, 2010). This tool allows the user to insert the desired search elements and displays the results found with those characteristics, both in a corpus of texts from the press and in dictionaries.

As concerns the search criteria of this tool, the following options were selected:

a) Search mode: phonological
This type of search was deemed more appropriate since, if the search term “ao” is entered, the system displays both orthographic correspondences (e.g. *b**ao**bab* and *bacal**ao***) and phonological correspondences (e.g. **aho**rrador and zan**aho**ria)

b) Database: press and proper names (first name, surname and/or place)
We preferred to consult only the press database and not the dictionary database since most of the words found in both databases were the same. Using both implied finding redundant information. Furthermore, only the press database contains details about word frequency and this is an aspect that we wanted to take into account for selecting

the words which would eventually make up the faxes of the second task in our corpus. Likewise we opted for selecting the search option “proper names” because they were useful for the subsequent creation of the fax sheets. Besides, certain vocalic sequences were found almost exclusively in surnames: e.g. “áe” (*Sáez, Herráez, Arráez, Peláez*).

Concerning the search syntax, these were the terms entered:

a) General search:
ae, ao, ea, eo, oa, oe, ai, ei, oi, au, eu, ia, ie, io, ua, ¬[qg]ue, uo, iu, ¬[qg]ui

This is a first approach to the search of vocalic sequences we are interested in, without specifying whether they should be stressed or not. In the case of “ue” and “ui”, we just specify that the program display the examples in which these sequences were not preceded by “q” or “g”, since in those cases the written <u> does not have a phonetic manifestation (e.g *que* [ke]).

b) Specific search for unstressed vocalic sequences:
ae. 'S, ao.'S, ea.'S, eo.'S, oa.'S, oe.'S, ai.'S, ei.'S, oi.'S, au.'S, eu.'S, ia.'S, ie.'S, io.'S, ua.'S, ¬[qg]ue.'S, uo.'S, iu.'S, ¬[qg]ui.'S

Here we simply add *“.'S”* to the search terms above. This means that the stress should be in the syllable following the vocalic sequence.

c) Specific search for vocalic sequences with the stress in the first vowel[55]:
áe, áo, éa, éo, óa, éo, ái, éi, ói, áu, éu, ía, íe, ío, úa, úe, úo, íu, úi

d) Specific search for vocalic sequences with the stress in the second vowel[56]:
aé, aó, eá, eó, oá, eó, aí, eí, oí, aú, eú, iá, ié, ió, uá, ué, uó, iú, uí

In a first search, we found that there were almost no words with the combination –ou-; only the compound word *estadounidense*. The rest were foreign loanwords (e.g. *glamour, soul, country, boutique*) which have not been considered for this study. These results for –ou- agree with the description for this diphthong in our literature review [see Chapter 4]. In Aguilar (2010: 45), we find that “only the diphthong /ou/ has been considered rare in Spanish, as no Latin-origin

[55] Not every word with the stress in the first vowel was found after entering these specific search terms. Many words whose first vowel of the sequence was stressed were found in the general search. That is why this is the first search which was carried out for all the vocalic sequences.

[56] Just as in the case of words with the stress in the second vowel of the sequence, many of the words with the stress in the second vowel of the sequence were found in the general search, in which no restriction regarding the stress was specified.

words contain it" (*author's translation*). This vocalic sequence was therefore discarded from our corpus.

As concerns the rest of vocal combinations, we distinguished between unstressed sequences (e.g. *israelí*) and stressed sequences. In these latter, we made a further distinction: stress in the first vowel (e.g. *Sáez*) or in the second one (e.g. *Rafael*). For each subtype, the possible largest number of words containing such sequence was found. With the results of the search, the tables in Appendix C were created.

2. *Word selection for their inclusion in the fax task.*

Once the word search for each of the vocalic sequences was carried out, we proceeded to select only two examples per type. If we remember that there are 3 subtypes (unstressed, with the stress in the first vowel and with the stress in the second vowel) for each of the 19 vocalic sequences, the total number of words making up the corpus would be 114 (19 sequences x 3 types x 2 examples). However, eight words of the corpus contain two vocalic sequences of our interest (*Beng**oe**ch**ea***, *p**oesía***, *c**ue**st**ió**n*, *fis**io**terap**eu**ta*, *d**ie**cis**éi**s*, *c**eu**t**íe**s*, *j**ui**c**io**so*) and there is also one compound (*jal**ea** r**ea**l*), considered one item, with two vocalic sequences. Therefore, the total amount of words in our corpus is 106. The fact that some words contained two vocalic sequences was considered very convenient in order to reduce the total number of words which should be produced by each speaker and so that the second task would not be very long and tedious for the speakers.

Besides the 106 words making up the corpus (see Table 8), we considered worthy the inclusion of 12 further words (see Table 9) since they present certain particularities of interest, basically due to the variability in their pronunciation. For example, we are interested in words beginning with *h-* plus *–ue* and *–ie* (*huevera, huevo, hueso, hielo* and *hierro*) since from an orthological point of view, at least in the case of *hue-*, the articulatory support of this group resulting in a plosive consonant is inadmissible (Aguilar, 2010). However, the phonetic reality of Standard Peninsular Spanish (SPS)[57] yields different degrees of plosive support before this vocalic sequence.[58] In contrast, there is no clear agreement as to which is the behavior of the group *hie-* but this is usually considered an orthographic equivalent to *ye-*, since it can be pronounced the

[57] In Aguilar (2010), the term Standard Peninsular Spanish (SPS) is preferred over North-Central Peninsular Spanish. However, at least for the aspects that we describe in this section, they can be considered equivalents.

[58] The factors contributing to the consonatization processes of vocalic sequences *ue* and *ie* in words like *huevo, huésped, hierro* or *hierba* are described in RAE (2011: 351) where a thorough description of the phonetic manifestations of such consonantization can be found.

same way. From this point of view, the words *hierro* and *yerro* would be homophonic in SPS, although not in other varieties.

The additional words incorporated to the corpus and containing the sequence *ua* are interesting for several reasons. On the one hand, in *Atahualpa* we intend to investigate whether there is between-speaker variation in the "h" pronunciation, as it happens in *huevo* y *hueso*. In words like *tatuaje* and *suave*, two trends have been observed (Navarro Tomás, 1918: 158-159)[59] in SPS: one consists in the heterosyllabic pronunciation of the vocalic sequence (e.g. /tatuˈaxe/), that is, a hiatus; while the other consists in the homosyllabic pronunciation of both vowels (e.g. /taˈtu̯axe/), thus a diphthongized pronunciation. The same phenomenon occurs in the group *ia*, for example in words like *viaje*, *confianza*, *mundial* and *oficial*, whose pronunciation varies between diphthong and hiatus.

In sum, as shown in Tables 8 and 9, the corpus is made up of 118 words: 106 containing the 19 types of vocalic combinations, plus the 12 extra words which represent specific study cases for certain phenomena deemed to favor between-speaker variation.

Table 8

Words making up the corpus for the second speaking task

Vocalic Sequence (VS)	Unstressed VS		VS with the stress in the first vowel		VS with the stress in the second vowel	
ae	*israelí*	*Aeróbic*	*Herráez*	*Sáez*	*maestro*	*Rafael*
ao	*baobab*	*Ahorrador*	*bacalao*	*Laos*	*Paola*	*zanahoria*
ea	*argéntea*	*Bronceador*	*jalea (real)*	*Bengoechea*	*teatro*	*(jalea) real*
eo	*espontáneo*	*Leonés*	*boxeo*	*feo*	*león*	*gaseosa*
oa	*Joaquín*	*Toallero*	*anchoa*	*Balboa*	*croata*	*almohada*
oe	*poesía*	*Bengoechea*	*aloe (vera)*	*Villarroel*	*bohemio*	*soez*

[59] *[Grupos con acento, interiores de palabra, con i, u como elemento secundario] "Cualquiera que sea la vocal que lleve el acento, estos grupos se pronuncian generalmente en una sola sílaba cuando el elemento más débil del conjunto vocálico se halla constituido por los sonidos i, u. Cada grupo forma un diptongo o triptongo: aire, gaita, llamaís, aciago, vaciáis, despreciáis, causa, flauta, guapo, [...] En ciertos casos, sin embargo, la tendencia fonética a reducir los grupos de vocales a una sola sílaba lucha con influencias etimológicas o analógicas, siendo posible pronunciar una misma palabra con reducción o sin reducción. El lenguaje lento, el acento enfático y la posición final favorecen en dichos casos el hiato. La pronunciación rápida y el tono corriente y familiar dan preferencia a la sinéresis"* (Navarro Tomás, 1972: 158-159).

ai	*faisán*	*Vainilla*	*bonsái*	*káiser*	*bilbaíno*	*Países (Bajos)*
ei	*aceituna*	*Voleibol*	*béisbol*	*dieciséis*	*increíble*	*seísmo*
oi	*Moisés*	*Boicot*	*hoy*	*Zoila*	*Eloísa*	*egoísta*
au	*auténtico*	*Paulina*	*Paula*	*flauta*	*Saúl*	*ataúd*
eu	*ceutíes*	*Mileurista*	*Ceuta*	*fisioterapeuta*	*transeúnte*	*feúcho*
ea	*historia*	*Asociación*	*poesía*	*policía*	*estudiante*	*piano*
ie	*dieciséis*	*Ansiedad*	*ceutíes*	*Díez*	*siete*	*viernes*
io	*fisioterapeuta*	*Funcionario*	*vacío*	*Ríos*	*juicioso*	*cuestión*
ua	*lengua*	*Puntuación*	*cacatúa*	*ganzúa*	*donjuán*	*guapa*
ue	*cuestión*	*Pueril*	*bambúes*	*tabúes*	*sueco*	*cruel*
uo	*antiguo*	*Mutuo*	*búho*	*flúor*	*Fructuoso*	*cuota*
iu	*ciudad*	*Diurético*	*triunfo*	*viudez*	*viuda*	*diurno*
ui	*ruiseñor*	*Juicioso*	*buitre*	*fortuito*	*suizo*	*genuino*

Note. The following color legend is used:

Blue: Words for which the searched vocalic sequence did not exist. These have been then replaced by the following ones:

oe → *Villarroel* (VS with the stress in the second vowel) since there are no more cases, besides *aloe*, with the stress in the first vowel of the VS *oe*.

iu → *triunfo* (VS with the stress in the second vowel) and *viudez* (unstressed VS) since no word exists with the stress in the first vowel of the VS *iu*.

ui → *buitre* y *fortuito* (VS with the stress in the second vowel) since no word exists with the stress in the first vowel of the VS *ui*.

Red: Words with two vocalic sequences.

Table 9

Extra words which make up the corpus for the second speaking task

Vocalic Sequence (VS)	Words
ia	*mundial, oficial, viaje, confianza*
ie	*hielo, hierro*
ua	*tatuaje, suave, Atahualpa*
ue	*huevera, huevo, hueso*

The criteria for the selection of the 106 words making up the corpus, without taking into account the 12 extra cases are the following ones (described in order of importance):

1) Most frequent word

2) Different within-word location of the vocalic sequence from one case to another (for the same group). Example: *espontáneo* (post-tonic position) and *leonés* (pre-tonic position).

3) Different consonantal context between words in the same group. Example: *boxeo* y *feo*. On the one hand, it was impossible to find the same consonantal context for all the 19 VS and thus to control this variable. On the other hand, choosing varied consonantal contexts seemed more forensically realistic.

4) Semantic suitability for the context of the faxes (see Section a: *Procedure for the creation of the fax sheets*).

5) Consonantal contexts conveying less coarticulation (according to Marrero et al., 2008). Preference for voiceless plosives [p,t,k], fricatives [s,f,z], rhotic [r] and affricate [tʃ], instead of nasals, voiced plosives and approximants. However, these are also represented in our corpus, although in minority: *maestro, jalea, león*.

a. Procedure for the creation of the fax sheets

Once the words chosen for the corpus were selected, several fax cover templates were searched to create 6 types of fax sheets to be used in the second task. These templates were found in *FaxCoverSheets.net* (http://www.faxcoversheets.net/samples.htm). After some modifications, such as the translation from English into Spanish and some other format and structure changes, the 6 fax sheets created are those available in Appendix D.

For the creation of these, several context settings were devised which served to make more realistic the exchange of information carried out in the second task. In other words, we aimed to create fax sheets which could have been written in real life. For instance, in the first one, there is an opening with the reason for having sent the fax: the Human Resources Department of a firm has prepared a database with the names of several candidates for different jobs in other collaborating enterprises. With this database, the firm creates a table with the following information: name of the candidate (e.g. *Paula Sáez* or *Moisés Díez*), occupation or profession (e.g. *estudiante de Historia*, *fisioterapeuta*), work shift (e.g. *diurno*) and day in which the interview will be held (e.g. *miércoles cuatro* de *junio*). The creation of this context allows the “realistic” insertion of 18 corpus words in this first fax sheet.

The last fax sheet presents some differences from the others. Despite the intent to include all the corpus words in faxes fulfilling the criterion of shared semantic fields, for some words this was not plausible and these were thus inserted in the sixth and last fax sheet. In this, the sender sends a crossword which should be published in the newspaper where both the sender and the addressee work. Then, this fax sheet is the only case where the setting created does not need that the words share a semantic field. The crosswords were created with a tool available online: http://www.genempire.com/generador-de-crucigramas.This is an automatic generator of crosswords which uses as input the words selected by the user and displays them in the form of a crossword.

Finally, two copies per fax sheet are created: one per speaker. These two copies are not exactly the same. While some words cannot be read properly in one copy, other words are illegible in the second copy. The effect of illegibility was made using *Adobe Photoshop*. In order to understand the aim of this artificial creation of illegibility, we describe in the following section how the second task was carried out.

b. Execution of the fax task by the speakers

This type of task is an adaptation, with several modifications, of the fax task described in Morrison, Rose and Zhang (2011). As in the Map Task (Anderson et al., 1991) –used by Aguilar (1999), for instance- the aim of this taks is to create a realistic context for the interaction between the participants; they have to gather information from each other so as to accomplish a common goal. For the execution of this task, each speaker is in a different room and must follow the following instructions (see the Appendix B for the original instructions in Spanish), which appear written in a card that the researcher has given to him:

> *You will find some fax sheets on the table. Their quality is not very good and some of the information on them is difficult to read.*
>
> *Your sibling*[60] *has also received these fax sheets. Maybe theirs have a better quality than yours. Dial 2839*[61] *and ask him to give you the information that you cannot read properly in your fax sheet.*
>
> 1) *Write down this information in your fax sheet and read it aloud while your do it.*
> 2) *When you have finished asking your brother for the missing information, check that you have not misheard anything.*
>
> *Your brother will do the same with the information that is missing in his fax sheet. Help him telling him the information that he needs.*

[60] In the case of brothers and twins, the instructions say "your brother". In the case of friends, the instructions for this task are identical but instead of "your brother", we have written "your friend".

[61] In the instructions of the other brother or friend a little modification has been made in these instructions: "Wait for your brother's /friend's call and ask him for the information that you cannot read properly in your fax sheet".

The goal of eliciting the reading aloud of the information that each speaker has to write down (i.e. the missing information in his fax sheet) at the same time that he is writing is obtaining a hyperarticulated pronunciation of the words. This is expected to inform of the syllable separation strategy of each speaker.

The aim pursued with the confirmation task is that the same corpus words appear in the recording of one speaker and his conversation partner. If only one of them asked for certain words and his partner only answered, we would not have comparable words in both recordings.

Besides, in a pilot experiment that we have carried out with some speakers before testing this task with the subjects of our study, we have observed two trends: The first one, which is the fastest way to carry out this task, consisted in one speaker enumerating as a list the words asked by his conversational partner. For instance, A says "*I need information about the first person listed in the table, Amalia García. What I specifically need is information about her profession, her work shift and the day she will be doing the interview*" and then B answers: "*teacher, day shift, Wednesday the 4th June*". The second trend consisted in framing the missing words in a sentence. For instance, with the same question asked by A, B answers little by little, giving A time to write down the information: "*Amalia García is a teacher, she works in a day shift and she will do the interview on Wednesday the 4th of June*." In order to avoid the "list effect" implied in the first trend described, we indicated the speakers that they should try to avoid answering as if they were listing the words.

Before carrying out this task, the two speakers were informed of what the task would be like and the fax sheets were given to them so that they could read them and understand the context described in each sheet. They had time to ask any question they might have and eventually they began the fax task with a dummy fax sheet (see the first fax sample in the Appendix D) which will be recorded but not used for this study. The execution of this dummy fax was aimed at evaluating whether the subjects had understood how to carry out the whole task.

Example of the performance of the fax task:

A: I would need information about the first person in the table, Amalia García. The details that I'm missing are: her profession or educational background, her work shift and the day she has to do the interview.

B: Amalia García works as a teacher…

A: (writes it down at the same time) ma-es-tra (English: Teacher)

B: …she has a day work shift

A: (writes it down at the same time) di-ur-no (English: day work shift)

B: … and her interview is on Wednesday, the 4th of June

A: (writes it down at the same time) miér-co-les cua-tro de ju-nio (English: Wednesday, the 4th of June)

[The conversation follows successively with the rest of illegible words]

A: Ok, I will double check what I have just written. Amalia García works as teacher in a day shift and she has the interview on Wednesday, the fourth of June.

B: Excellent. Now it's my turn. I need information about the person who has the interview today 16th December. Can you please tell me his or her name, profession and work shift?

A: …

3.3.3. Third task: Reading of two phonetically-balanced texts

In the third task, we asked the subjects to read two phonetically balanced texts. The first one, designed by Ortega, González and Marrero (2000) contains 179 words (712 phonemes) while the second one, created by Bruyninckx, Harmegnies, Llisterri and Poch (1994) contains 103 words (440 phonemes). Both can be found in Appendix B.

The speakers had time to familiarize with the texts before reading them. The instructions given to the participants were that they had to read both texts at their natural speaking rate and with the loudness they felt comfortable. They were instructed to repeat any sentence where they got tangled up.

Since disguised speech is highly frequent in a forensic context, we instructed the speakers to read one of the texts (Bruyninckx et al., 1994) while pinching his nose. They did this after having first read the two texts in an undisguised way. The main advantage of having the speakers read the same text has been often acknowledged in Forensic Phonetics, namely obtaining comparable units (Nolan, McDougall, de Jong, & Hudson, 2009).

3.3.4. Fourth task: Mathematical calculations

This task is aimed at eliciting filled pauses or hesitation markers in speakers. Although there are several ways in which speakers can hesitate while organizing their discourse, the main goal is obtaining enough tokens of the most common hesitation mark in Spanish which is "eh" (Gil, 2007: 299), although this unit can also be found in elongations of a vowel at final position (e.g. "Hemos estado hablando deee…"). These long vowels will be useful for the posterior analysis of glottal parameters. Actually, the software *BioMet®Soft* requires for this type of analysis the use of vowels longer than the vowels usually found in connected speech [see Chapter 5].

This task is also carried out on the telephone. Each speaker has a sheet with some mathematical calculations. He will have to ask his brother or friend for the solution to these calculations, and afterwards he will answer aloud to the mathematical questions that his partner asks him. Of course, the calculations that each speaker has written in his paper are different, but of the same difficulty. Furthermore, the subjects are required to perform this task as quickly as possible so that the possibilities of producing hesitation speech increase (hesitation in this case would be used to gain time to think the correct answer).

3.3.5. Fifth task: Informal interview with the researcher

This task is also carried out on the telephone. The researcher is at one end of the telephone and one member of each speaker pair at a time is at the other end of the telephone. Meanwhile, the other conversational partner who is not participating in the interview must fill in the questionnaire form (see Section 3.2). When the interview finishes with one speaker, this will fill in the questionnaire and the other one, who has been waiting, is then also interviewed.

The purpose of this task is obtaining speaking samples in a more formal context, since the speakers are not familiar with the interviewer. A different speaking style is then supposed to be elicited, as compared with the previous tasks, where the two interlocutors were always either brothers or friends.

In this interview, the researcher asks the speakers about any of the topics that they had been discussing with each other in the first task. Since the topics raised in both the first task and the fifth one are the same, we obtain comparable phonetic units in two speaking styles. This task lasts around 5 and 10 minutes, approximately. We have described above (see Section 3.3.1) the most usual topics which the speakers could choose to talk about. For specially sparing speakers, other possible topics were raised, following the *PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y de América)* methodology specified in the guidelines of the project (PRESEEA, 2003). Besides, in order to avoid or minimize the "observer's paradox" (Labov, 1972), we have followed the indications in Moreno (2011) who, among other strategies, suggests the use of "icebreakers" as conversational starting points.

3.3.6. Vocal Control Techniques

Maximum Phonation Time

The Maximum Phonation Time (MPT) is an easy and quick measure of a speaker’s glottal efficiency. While the glottal efficiency implies the capacity to close the vocal folds in an efficient way making them vibrate through fast opening and closing cycles, the glottal insufficiency describes the inadequate closing of vocal folds and thus, the inefficient vibration of the vocal folds (Aronson & Bless, 2009).

The MPT is defined as “the maximum duration of a sustained vowel after maximum inhalation and is typically averaged across multiple trials” (Aronson & Bless, 2009: 148). The MPT is measured in seconds and, being completely non-intrusive, is a quick and simple clinical technique which has been traditionally used by voice clinicians.

Following the method suggested by Eckel and Boone (1981), the subjects were asked to do this test twice with a sustained /a/. The longest duration of the two trials was considered for further analyses. For the interpretation of this measure, the usual standard for an adult man with no laryngeal pathologies is 24-35 seconds of sustaining a vowel. In those cases where the speaker suffers from laryngeal pathology, the MPT decreases considerably.

We have to point out that the MPT itself does not allow the diagnosis of a laryngeal pathology. It would be necessary to do a laryngoscopy exam to determine possible organic or functional damage. Nevertheless, the MPT is still considered a useful indicator of a potential pathology.

The s/a Ratio

The second vocal control technique used for this study consists in measuring the s/a ratio of each speaker. The use of this phonation ratio as an indicator of laryngeal pathologies was proposed for the first time by Boone (1977). Originally, to calculate this ratio, we have to measure, on the one hand, the time a person can sustain the sound [s] and, on the other hand, the time this same speaker can sustain the sound [z]. Afterwards, the first measure is divided by the second one in order to obtain a numerical ratio. The highest the resulting number, the higher the possibility that that person has phonation difficulties. In other words, “the principle underlying this measure is an assumption that maximum glottal efficiency will result in equal duration for both the /s/ and /z/ fricatives, yielding a theoretical ratio of 1.0” (Aronson & Bless, 2009: 148).

Since in Spanish, the phoneme /z/ does not exist, we calculated the ratio measuring the duration of /s/ and that of a sustained /a/. This vocalic sound, substituting /z/, is also a voiced sound (i.e. with vibration of the vocal folds), as opposed to /s/, which is voiceless. The replacement of /z/ with /a/ is common practice in speech therapy clinics in Spain (Núñez & Suárez, 1998: 65). Most speakers without any laryngeal difficulty are able to sustain both sounds for the same time. Consequently, if an adult man can sustain an /s/ as well as the vowel /a/ for 25 seconds each, his s/a ratio will be 25/25 = 1.0. On the contrary, 95% of the patients suffering from any difficulty which involves the movement of the vocal folds obtain an s/a ratio above 1.40 (Eckel & Boone, 1981). These people are not able to sustain the voiced and the voiceless sound for the same time. This is due to the fact that the former, unlike the latter, requires vibration of the vocal folds and any lesion or pathology in them would interfere in their vibration cycle, thus reducing the time of phonation.[62]

Just as we have mentioned before for the measure MPT, the s/a ratio should not be considered the only technique to diagnose a laryngeal pathology. A ratio of 1.4 or above does not guarantee the presence of pathology. Indeed, this measure is used as a first test for the early identification of laryngeal difficulties, as well as a tool to control the evolution of certain treatments.

3.4. Recording Procedure

3.4.1. Materials and Technical Characteristics of the Recording

For all the recordings we used the same recording material (microphones, soundcard and software) with the characteristics specified below. Besides, the recordings are always made by the same researcher –the author – as a way to control that all the recordings were carried out using the same protocol. Since the participants came to the recording sessions in pairs, two microphones were needed to record them at the same time. The microphones chosen for the recordings, following the recommendation of Morrison, Rose and Zhang (2012), were two identical *Countryman E6i Earset* microphones. These are omnidirectional condenser microphones especially suitable for this research since they are very small, thin and light, being thus unobtrusive. On the one hand, this helps the speaker to forget that he is being recorded, which is advisable in order to obtain spontaneous speech. On the other hand, since it is an earset device which is held close to the mouth, undesirable noise in the recordings is avoided. In addition, it ensures that the distance from the mouth to the microphone is always fixed. This microphone has a flat frequency response (20 Hz to 20 kHz), a sensitivity of 2.0 mV/Pascal, Equivalent Acoustic Noise 29 dBA SPL and Overload Sound Level 130 dB SPL. See Figure 4 for an illustration of

[62] From: http://www.speech-therapy-information-and-resources.com/sz-ratio.html. Date of retrieval: 26th May 2013. See also http://www.sltinfo.com/sz-ratio/. Date of retrieval: 27th September 2014.

this microphone and how it was used in the recordings of this thesis. Figure 5 and 6 offer information about the polar and the frequency response of the microphone, respectively.

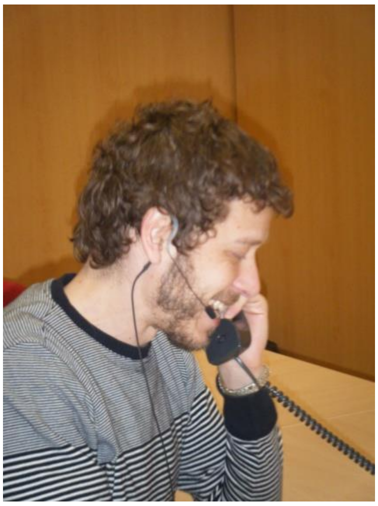


Figure 4. Participant: microphone detail.

The microphones were connected to a soundcard through two long cables, each one to one channel. The soundcard was a *Cakewalk by Roland UA-25EX USB AudioCapture* and the specifications selected for the recording were the following ones:

- Sample rate: 44,100 Hz
- Resolution: 16 bits
- Channel: Mono

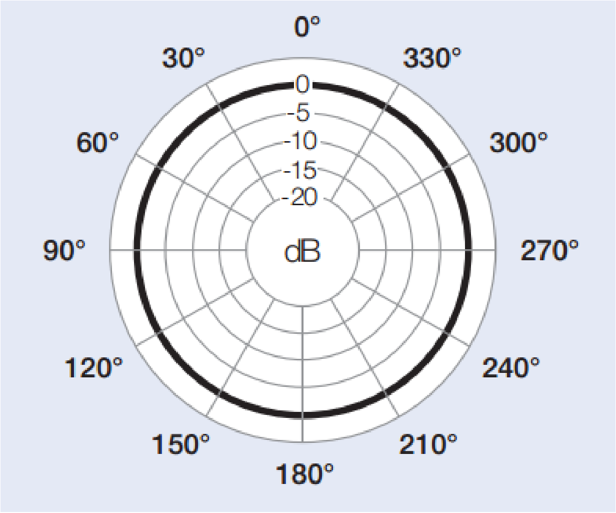


Figure 5. 1kHz Polar Response of the *Countryman E6i Earset* microphone.

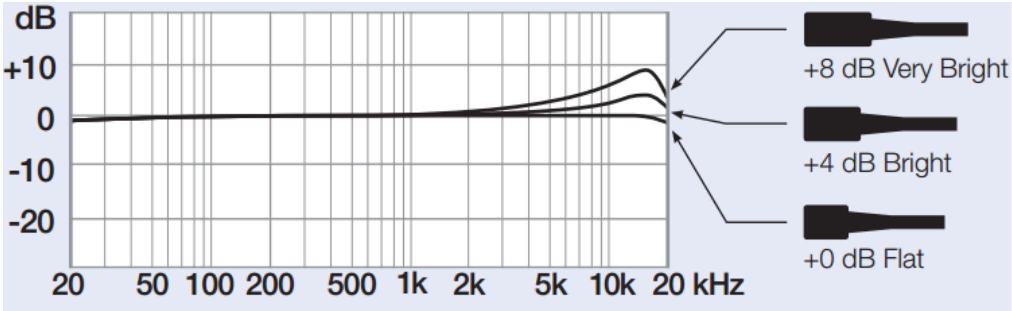


Figure 6. Frequency response of the *Countryman E6i Earset* microphone measured at 15.24 cm with different caps. The cap we used for our recordings was the +0 dB.

The software used for the recordings was *Adobe Audition CS5.5* and the operating system of the computer used was *Microsoft Windows XP Professional, Version 2002, Service Pack 3.* The telephone used for the communication between twins and between the researcher and the twins was a *Cisco IP Phone 7912 Series (Cisco Systems)*[63]. The model of the headphones used by the researcher to monitor the recordings was *AKG K240 Studio.*

3.4.2. Data Collection Set-up

The recordings took place always in the same setting. As we said previously, the speakers came in pairs to the Phonetics Laboratory of the CSIC. Here, they were first gathered together in the same room to receive the instructions on how to carry out the vocal control technique and the first speaking task. They were informed that the instructions for the rest of the task would be given later on. This way they did not have to remember at a time all the instructions for all the tasks and they could concentrate only on one.

They were them separated in two quiet almost identical rooms where the recordings took place and they were adjusted the microphones (Figures 4 and 7). Then, they were given the questionnaire appropriate for each recording session (see Appendix A2 and A3) so that they could fill it when they did not have to speak (because his conversation partner would be involved in an individual task). They were also instructed not to provoke noise which could be undesirable for the proper recording of the acoustic signal, for example, playing with the pen (necessary for the second speaking task), moving the papers noisily or tapping the table. During all this set-up process –and of course, also when the rest of the instructions were given to them– the participants were able to ask any question they might have. The instructions to carry out the different speaking

[63] These telephones were only used by the participants to communicate with each other. To simulate real-condition telephone interceptions, we applied a telephone filter to the high-quality recordings, as specified in Section *3.5*.

tasks as well as the material necessary for some tasks (e.g. text passages) can be found in Appendix B.

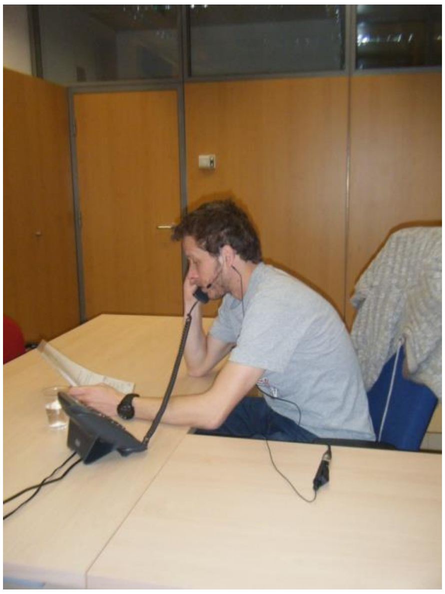


Figure 7. Participant carrying out one of the corpus tasks through the telephone.

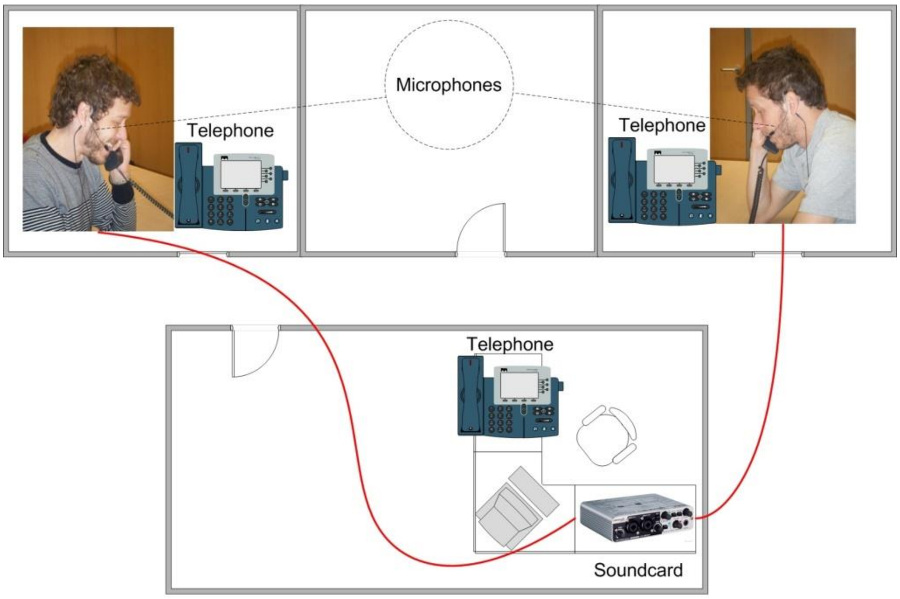


Figure 8. Data collection set-up. The speakers were separated in two different (acoustically isolated) rooms and they communicated via telephone. The red lines represent the cables from each of the two microphones to the soundcard, which was in turn connected to the computer of the researcher. This was in a third room, from where she monitored the recordings and communicated via telephone with the participants, when necessary.

3.5. Telephone Filtering Procedure

One of the main characteristics of the recordings found in a forensic setting is that they are telephone-degraded. That is the reason why for this thesis we have applied a telephone filter to the recordings previously made in high-quality conditions. The procedure for doing so was as follows. First, the audio recordings in WAV format are played through a laptop connected to a loudspeaker *FOSTEX 6301B*. The signal is captured by a landline telephone (*ALCATEL*) and a mobile telephone *HTC Desire* –at at different times– and a call is established with another telephone device. All this setting takes place in an isolated recording booth to avoid extraneous noise. Outside the booth, a second landline telephone is receiving the call, and then the audio signal is saved with the new telephone quality in another computer.

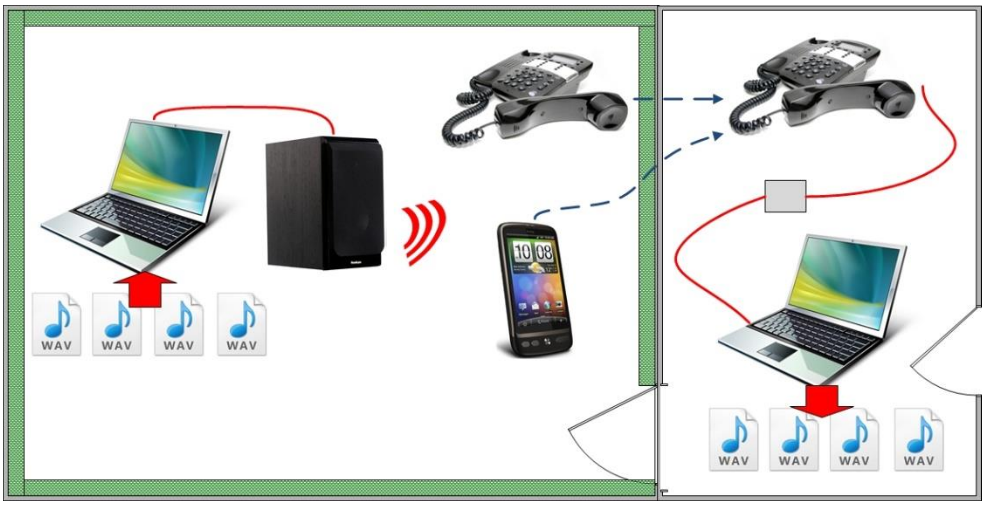


Figure 9. Set-up for the filtering of the audio signal.

3.6. The likelihood-ratio approach

The likelihood-ratio (LR) approach is a framework for the evaluation of forensic evidence which is described by some authors (Saks & Koehler, 2005) as a *paradigm shift* which would have already occurred in DNA profile comparison and which other forensic sciences "could and should emulate" (Saks & Koehler, 2005: 893). Specifically, they refer to these two aspects:

> Each subfield must construct databases of sample characteristics and use these databases to support a probabilistic approach to identification. […] A second data collection effort that would strengthen the scientific foundation of the forensic sciences involves estimating error rates. (Saks & Koehler, 2005: 893).

For more accurate descriptions of this framework, see Berger, Robertson & Vignaux (2010) for a thorough review; Roberts (2004) for the LR-based approach applied to DNA-profile comparison; and Ramos-Castro (2007) or Morrison (2010a) for a description of this *paradigm* *shift* from the perspective of voice comparison. As stated in Morrison (2010a: 2051): "Forensic voice comparison is one branch of forensic science in which this shift is now well underway but in which it is still far from reaching universal acceptance among researchers and practitioners". In this section we will summarize the main aspects that characterize the likelihood-ratio framework, which we have adopted in the acoustic analyses carried out for this thesis.

a) What is the task of the forensic scientist under the LR-framework?

The forensic scientist –in this case, the forensic phonetician– who presents the results of his investigation within the LR framework, should provide the court with an answer to the question:

> How much more likely are the observed differences between the known and questioned samples to occur under the hypothesis that the questioned sample has the same origin as the known sample than under the hypothesis that it has a different origin? (Morrison, 2010a: 3052)

b) What is the formula used to calculate a LR?

As Morrison (2010a: 3052) states, the answer to the previous question (i.e. the answer that should be provided to the court) is a statement of the "strength of the evidence" and is quantitatively expressed as a LR, calculated using the following formula[64]:

$$LR = \frac{p(E|H_{so})}{p(E|H_{do})} \qquad (1)$$

where *LR* is the likelihood ratio; *E* is the evidence ("the measured differences between the samples of known and question origin" Morrison, 2010a: 3052); *p(E|H)* is "probability of *E* given *H*", H_{so} is the same-origin hypothesis, and H_{do} is the different-origin hypothesis. In the case of forensic voice comparison H_{so} could be represented as H_{ss} (same-speaker hypothesis) and H_{do} as H_{ds} (different-speaker hypothesis).

c) How should be the size of a LR interpreted?

Likelihood ratios above 1 mean that the evidence is more likely to occur under the same-origin hypothesis than under the different-origin hypothesis, whereas likelihood ratios below 1 indicate that the evidence is more likely to occur under the different-origin hypothesis than under the same-origin hypothesis. The size of the LR shows the strength of the evidence with respect to the

[64] As can be seen in the formula, a LR has a numerator and a denominator: "The numerator of the LR can be considered a *similarity* term, and the denominator a *typicality* term. In calculating the strength of evidence, the forensic scientist must consider not only the degree of similarity between the samples, but also their degree of typicality with respect to the relevant population. In fictional television shows, forensic scientists are often portrayed comparing two objects, finding no measurable differences between them, and shouting: "It's a match!" Similarity alone, however, does not lead to strong support for the same-origin hypothesis" (Morrison, 2010a: 3052).

competing hypotheses. In other words, a LR=100 means that one would be 100 times more likely to observe the differences between the known and questioned samples under the same-origin hypothesis than under the different-origin hypothesis. Likewise, a LR= -100 (this is the same as LR=1/100) means that one would be 100 times more likely to observe the evidence under the different-origin hypothesis than under the same-origin hypothesis (examples taken from Morrison, 2010a: 3052).

Morrison (2010a: 3052-3053) provides an answer to the question "Why the forensic scientist must present the probability of the evidence, and must not present the probability of hypotheses" and describes the formula to be used by the forensic scientist if he wants to calculate the probability of same-origin versus different-origin hypotheses (the odds form of Bayes' Theorem), although this author considers inappropriate for the forensic scientist to present the posterior odds, as these "include information and assumptions from sources other than a scientific evaluation of the known and questioned samples" (Morrison, 2010a:3053).

d) What is a background population and why it is needed in a LR-based approach?

A background population is a "database representative of the relevant population" (Morrison, 2010a: 3054) to which the offender belongs:

> In forensic voice comparison, this [population] can usually be at least restricted to speakers of the same sex and general age speaking the same language and dialect as can be inferred from the questioned speaker on the basis of the questioned-voice recording. (Morrison, 2010a: 3054)[65]

Since a LR contains a *similarity* term and a *typicality* term, corresponding to the numerator and denominator of the LR formula, respectively (see note 64 and Morrison 2010a: 3054), within a LR-based approach, a background population is necessary for the quantitative estimation of the typicality of the known and questioned samples.

[65] "The exact nature of the relevant population is, however, dependent on the exact nature of the different-speaker hypothesis advanced by the defence" (Morrison, 2010a: 3055).

4. ANALYSIS OF FORMANT TRAJECTORIES

4.1. Objectives and justification

In this section, we will establish our research objectives and hypotheses for the analysis of formant trajectories. We will set these objectives against the background of the state-of-the-art (i.e. the specific studies under this line of research). For this purpose, Section 4.1.2 will include a brief literature review of the main studies which have investigated the formant trajectories of vocalic sequences (VS) with forensic purposes.

4.1.1. Objectives

The main objective of the analysis presented in this chapter is testing the forensic validity of formant trajectories extracted from Spanish vocalic sequences. This general objective can be split into the following specific or secondary objectives:

O1: Testing whether there is higher intra-pair similarity for this kind of parameters in MZ twins than in other speaker comparisons (DZ, B or US). This would imply that the parameters are genetically influenced and would therefore be useful in a typical forensic context.

O2: Testing whether the fusion of the scores obtained for all the vocalic sequences (VS) outperform the individual systems based on single VS.

O3: Testing whether certain procedures for parameter curve fitting of the formant trajectories outperform the others.

For the above-mentioned objectives, we propose the following hypotheses:

H1: Formant trajectories in the vocalic sequences under study will be somehow genetically influenced: higher similarity values will be found in MZ twins than in DZ twins, siblings or in the reference population. This is in agreement with the 5 basic hypotheses established for this thesis [see Table 3 and Chapter 2].

H2: A forensic-comparison system based on all the VS fused together will yield better performance than individual systems, each based on a single VS.

H3: According to previous preliminary studies (San Segundo, 2010a), identification results will not be much better with one parameter curve fitting method as compared with the other.

4.1.2. Justification

Formant frequencies have been traditionally used in FSC since they are one of the clearest acoustic correlates of the vocal tract resonances[66]. There are several ways in which formant frequencies can be approached from a forensic point of view, being the most classic perspective the study of the central values in the four first formants of vowels (F1-F4). Our literature review will not focus on this kind of studies, since the acoustic analysis carried out for our thesis is based on a different approach to formant frequencies, namely the consideration of the formant trajectories of vocalic sequences. Therefore, in the next pages we will review the main publications which have found that the "dynamic" properties of a sound (i.e. formant trajectories) are more speaker-idiosyncratic than their "static" central values[67].

The use of the term "dynamic" for the analysis of vowels implies the consideration of temporal characteristics that are disregarded when studying these sounds at a single point. In the case of vocalic sequences, like diphthongs, the existence of two vowels leaves the speaker some leeway in his vocal tract (e.g. in the movement of his articulators) to achieve the transition between the two acoustic targets. It is a kind of movement flexibility that McDougall (2006) compares with other human motor activities:

> […] future research should pay more attention to dynamic as opposed to static properties of speech on the basis that time-varying features reflect the movement of a speaker's articulators and, just as people exhibit their own personal styles in carrying out skilled motor activities such as walking and running, they use their articulators for speech in an individual manner. Properties of the acoustic signal 'in between' the moments at which phonetic targets are achieved might therefore be expected to exhibit greater between-speaker variation than static measures at a single time-slice. Formant frequencies were chosen to investigate this idea because these acoustic features reflect both differences in the morphology of each person's vocal tract and differences in the choice of articulatory gestures made by each individual to satisfy the targets of the phonetic plan. (McDougall, 2006: 121)

The pioneering studies of this research line probably come back to Goldstein (1976), who examined the speaker-identifying potential of the formant structure of three diphthongs, four tense

[66] In next section we will briefly consider the main acoustic aspects of these parameters (formant frequencies).

[67] The terminology "dynamic" vs. "static" is mainly widespread by the work of McDougall (2004 and following). However, the suggestion that the study of the center of a vowel leaves much information unexplored for forensic usage is already found in Goldstein (1976: 176): "The use of formant information in speaker identification systems have been limited almost exclusively to the measurement of formant frequencies inside a single window at the center of a vowel, leaving much of the formant structure unexplored (Wolf, J. 1972; Sambur, M.R. 1975)".

vowels, and three retroflex sounds in American English. Her interest in these phonetic units lay in the fact that there is a large dialect variation shown by those sounds in American English (Goldstein, 1976: 179). The results of this approximation pointed to the prominence of mid vowels for speaker identification: "Since mid vowels are not produced with an extreme high or low tongue position, they may be subject to more individual variation than [i] or [u]". However, the diphthongs in this study did not rank above other parameters in Goldstein's list of the ten best features for speaker identification.[68]

Greisbach, Esser and Weinstock (1995) and Ingram, Prandolini and Ong (1996) are two later studies approaching the same issue from different perspectives. In the first study, the authors compare the discriminatory performance of F1 and F2 measured at a single central point and these same formants measured at five different points of the same phonetic unit. The specific sounds explored in this study are the five long vowels of German and the diphthong [æ]. Two non-contemporaneous sessions are considered for this study. The five-point measurement outperformed the single-point measurement in the percentage of correct identifications (using Euclidean distances as a measure to identify speakers). The improvement of the five-point measurement (vs. the single-point method) was more evident for the diphthong (78% vs. 39%) than for the monophthongs (92% vs. 72%). As far as the methodology is concerned, the study of Ingram et al. (1996) is very different since the authors use phonologically matched sonorant segments of speech without specifying which vocalic sequences may contribute better to speaker discrimination. Instead, the authors found nine best discriminating segments ranging from 2 to 3 seconds of speech in what they call *formant trajectory matching*, consisting in "using phonetically controlled acoustic segments of sonorant speech spanning one or two vocalic nuclei" (Ingram et al., 1996: 143).

Most recent approaches to the study of formant frequencies in Forensic Phonetics suggest the use of parametric functions to fit the formant trajectories or contours of vowels, either monophthongs (McDougall & Nolan, 2007) or diphthongs. For the purpose of this thesis, we will only focus on the latter. McDougall (2004) first explored the possibility of characterizing the formant dynamic properties of diphthongs for speaker classification. For this purpose, she divided each diphthong in 10 intervals of the same duration, thus obtained 9 points from which she extracted the first three formants. Based on Discriminant Analysis, she found that Australian English /aɪ/ diphthongs followed by /k/ showed promising results for speaker identification, as speakers displayed individual differences in the formant contours for all the three formants analyzed. In McDougall (2005; 2006) she designed a different method for capturing the "most

[68] The method used for statistical evaluation consisted in computing the ratio of between-speaker variance to within-speaker variance, known as an *F ratio*.

defining aspects of the curvature of the contours" (McDougall, 2006: 102), namely linear regression techniques –this method is explained in Section 4.2.2. McDougall (2005, 2006) tested the effectiveness of this parameterization (fitting polynomial equations to the F1-F3 formant contours of each speaker) for distinguishing speakers. However, in her study the cubic-based approach did not always outperform the quadratic-based one:

> Differences among speakers in the formant dynamics of /aik/ can be captured effectively by parameterizing the curves, but it is not clear whether a quadratic-based or cubic-based approach is more appropriate: in most cases the quadratic coefficients provided sufficient information to enable the speakers to be distinguished. (McDougall, 2006: 110)

Several authors have followed McDougall's research line on formant trajectories, like Eriksson and Sullivan (2008), López-Escobedo (2010) and Thaitechawat and Foulkes (2011), whose interest lay in testing the forensic effectiveness of formant trajectories in languages other than Australian English, namely Swedish (Eriksson & Sullivan, 2008); Mexican Spanish (López-Escobedo, 2010), and Thai (Thaitechawat & Foulkes, 2011). Scarce methodological innovations exist in these studies, except for Eriksson and Sullivan (2008), who suggest a leave-one-out method intended to crosscheck the validity of discriminant functions and which is supposed to be "less prone to over-estimating the discriminatory ability of the functions than the method used by McDougall" (Eriksson & Sullivan, 2008: 55). In the case of López-Escobedo (2010) and Thaitechawat and Foulkes (2011), they analyzed additional parameters like f_0 or tone data, respectively.

However, McDougall's studies also gave rise to other investigations, which either suggested new methods for curve fitting, or incorporated the application of a LR-based approach as a methodological novelty, or both things. The first study which tried to explore the forensic discriminability of diphthongs from a LR-based perspective is Rose (2006b). Using the Bernard corpus (Bernard, 1967), which provides information on the F-pattern (F1-F3) of 11 monophthongal vowels as well as 7 diphthongal phonemes of 170 male Australian speakers, Rose (2006b) examined the following diphthongs: /aɪ/, /ɛɪ/, /ɐʉ/, /ɪə/ and /ɛə/. As a result of his investigation, Rose (2006b: 69) concluded that "it is clear that diphthongs have considerable potential in forensic speaker recognition, and need to be researched more". In this line, Kinoshita and Osinai (2006) recorded 27 male adult native speakers of Australian English in order to examine the use of the F2 slope in the glide of the Australian English diphthong /aɪ/. They found that this feature produced as good results as the F2 of the first and the second targets of that vocalic sequence. Rose, Kinoshita and Alderman (2006) carried out a discrimination experiment using the same diphthong (in read speech) and the same database of Australian English speakers but using only two targets of the F-pattern in this diphthong.

The studies reviewed so far considered only one recording session per speaker. In González-Rodríguez et al. (2007) non-contemporaneous speech samples are used for the first time. Highlighting the importance of calibration and exemplifying it with LRs from Australian English diphthongal F-pattern, these authors carry out two experiments:

> The first experiment uses the F-pattern of four diphthongs /ai ei oi ou/ to show that traditional features do indeed have forensic discriminatory potential under relatively clean conditions. The second experiment, with just one diphthong /ai/, takes this further and shows that the discriminability extends to non-contemporaneous data. The experiments are taken from Rose, Kinoshita and Alderman (2006) and Rose (2006b). (Gónzalez-Rodríguez, 2007: 2108)

In Morrison and Kinoshita (2008), who focused on Australian English phoneme /o/[69], speakers were also recorded on two occasions, separated by approximately two weeks. These authors fitted different parametric curves (quadratic and cubic polynomial functions as well as the first three and four coefficients derived from discrete cosine transforms, DCT) to the F1-F3 formant trajectories of this vowel and calculated cross-validated likelihood ratios using the multivariate kernel density formula developed by Aitken and Lucy (2004) and implemented in Morrison (2007). Using calibration techniques, their results showed that "for both the three and two-formant analyses, the best performance was achieved using third-degree polynomials fitted to linear-hertz-scaled equalized-duration formant trajectories" (Morrison & Kinoshita, 2008: 4).

The work in Morrison (2008) is based on the same speaker data used by Kinoshita and Osanai (2006) and Rose et al. (2006) with the aim of exploring whether a new parametric-curve model of formant trajectories could outperform the dual-target model carried out by the former authors in the study of the /aɪ/ diphthong. As a result of the study of Morrison (2008), it is concluded that:

> Using the same set of recordings and measuring acoustic properties of the same set of /ai/ tokens with those recordings, substantially better performance was obtained from likelihood-ratio forensic-voice-comparison analyses which made use of polynomial curves fitted to formant trajectories than from analyses which made use of initial and final formant values. (Morrison, 2008: 260)

This study also concluded that the model based on cubic polynomials fitted to duration-equalized trajectories outperforms models based on quadratic polynomials and models fitted to raw formant trajectories (at least for the /aɪ/ diphthong under examination).

[69] They specify "often transcribed as /oʊ/" (Morrison & Kinoshita, 2008). That is why we include this article in the literature review, since we have specified before that we would only deal with studies on the formant trajectories of vocalic sequences, not on monophthongs.

With a similar methodology (parametric curves fitted to formant trajectories), Morrison (2009c) compared non-contemporaneous speech samples from 27 male speakers of Australian English. On this occasion, the diphthongs considered were /aɪ/, /eɪ/, /oʊ/, /aʊ/ and /ɔɪ/. The likelihood ratios resulting from the analysis were calibrated and also fused using logistic regression fusion.

Using the kind of parametric representations suggested in Morrison's diverse studies, other publications have appeared which have applied in several languages the methodology already tested for Australian English diphthongs. This is the case of Enzinger (2010), San Segundo (2010a), and Zhang, Morrison and Thiruvaran (2011).

Enzinger (2010) explored the discriminatory capacity of the Viennese German diphthong /aɛ/. Apart from polynomial and DCT fuctions, the parametric representations of formant trajectories that he used also included B-splines[70] and Bent-cable models[71]. Both polynomials and B-splines were found to display low error rates, as well as DCT-based methods. However, Bent-cable coefficients showed much higher error rates than the other methods.

San Segundo (2010a) has been, to our knowledge, the first approach to the study of formant trajectories in Castilian Spanish diphthongs from a forensic point of view. Following Morrison (2009c), non-contemporaneous speech samples from 30 Spanish male speakers were compared within a likelihood-ratio framework. Certain parametric curves (polynomials and DCT) were fitted to the formant trajectories of the vocalic sequences [u̯e], [i̯e] (diphthongs) and [ia] and [ai] (hiatuses). The estimated coefficient values from the parametric curves were used as input to a MVKD[72] formula (Aitken & Lucy 2004; Morrison, 2007) for calculating likelihood ratios. The results of this study showed that there were no large differences between using polynomials and DCT. However, due to the small sample size and the methodology used, strong conclusions could not be drawn as regards which vocalic sequences yielded better discrimination.

Finally, Zhang et al. (2011) investigated formant trajectories in the Standard Chinese triphthong /iau/ (extracted from female recordings) with the aim of incorporating this information to a "generic automatic forensic-voice-comparison system, which did not itself exploit acoustic-phonetic information" (Zhang, Morrison, & Thiruvaran, 2011: 2280). The results of this

[70] B-splines are defined (Enzinger, 2010: 48) as "pairwise polynomials […] a generalization of Bézier curves […] advantageous for numerical reasons, as they are locally linearly independent and numerically stable, meaning that small changes in the coefficients result in small changes to the respective spline function and vice versa".

[71] Bent-cable models are defined (Enzinger, 2010:48) as "an extension of so-colled broken stick piecewise-linear models".

[72] The MVKD (Multivariate-kernel-density) formula is described in Section 4.2.2 (cf. likelihood-ratio calculation).

investigation showed that in doing so there was a substantial improvement in system validity but a decline in system reliability.

4.2. Speech material, analysis tools and method

4.2.1. Speech material

For the analysis of the formant trajectories, the speech material consisted of 11,773 phonetic units (i.e.VS). This number results from:

$$54\ speakers\ \times 2\ recording\ sessions \times 19\ types\ of\ VS^{\text{[73]}}\ \times 3\ stress\ variations^{\text{[74]}} \\ \times 2\ examples\ of\ each\ stress\ condition$$

The product should be 12.312 but some tokens had to be discarded for one of the following reasons:

- Non-modal phonation, like creak (in most cases) or whisper.
- Overlap of the phonetic unit of interest with extraneous noises.
- Hyperarticulation, due mainly to an emphatic pronunciation of the stressed vowel[75].

The application of these exclusion criteria resulted in the selection of only homogenous tokens. This fact did not prevent that at least one example per stress condition and type of VS was selected. As explained in Chapter 3 (cf. *Corpus elaboration*) the VS which make up the speech material for this analysis were extracted from the second speaking task (fax exchange).

Speech material extraction

The procedure for the analysis of formant transitions implied a previous extraction, and subsequent labelling, of the vocalic sequences of interest. The procedure for extraction was as follows (see Figure 10):

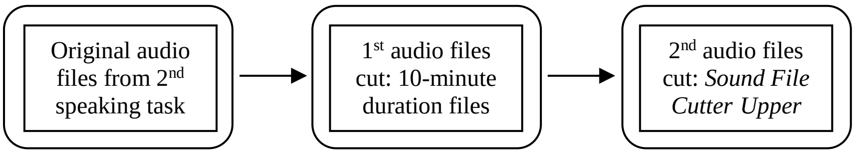

Original audio files from 2nd speaking task
1st audio files cut: 10-minute duration files
2nd audio files cut: *Sound File Cutter Upper*

Figure 10. Procedure for speech material extraction.

[73] Vocalic sequences in Spanish are twenty. Yet we did not consider [ou] for the reasons explained in Chapter 3.

[74] The three types of stress variations are: unstressed VS, stressed in the first vowel and stressed in the second vowel (See the section on corpus elaboration in Chapter 3).

[75] This could be due to the fact that some speakers wanted to highlight where the stressed vowel was so that his interlocutor could correctly place the accent mark in the paper.

Firstly, the sound files obtained in the second speaking task (around 20-30 minutes duration) were cut into smaller files (10 minutes duration)[76], which were then cut using the software *Sound File Cutter Upper* (Morrison 2010b)[77]. The purpose of this software is cutting up a sound file, saving the non-silent portions as a series of short WAV files. This is useful for more easily dealing with a number of short sound files in a later labeling step. Briefly summarizing how this software operates, it calculates the running amplitude of the sound file selected, displaying it against time and indicating a default cutoff value: "portions of the sound file above the threshold (plus one padding before and after) will be saved as separate files" (Morrison 2010b: 3). The threshold can be optionally adjusted. This software seemed useful first to discard the silent portions. As the audio files came from a conversation between two speakers, many silences were expected, corresponding to the moments where each speaker was listening to his interlocutor. Besides, the software was also found useful in order to divide the long audio files of each speaker's recording, which would enable the subsequent labeling.

Speech material labeling

Once the audio files were reduced to short files, we proceeded to use the software *SoundLabeller: Ergonomically designed software for marking and labelling portions of sound files*, developed by Morrison (2012)[78]. As other programs which allow the labelling of sound files, like the *TextGrid* function in Praat (Boersma & Weenink, 2012), this software displays the waveform and spectrogram of a sound file, enabling the user to mark the beginning and end of certain parts of the recording and to use labels for the selected fragments. Figure 11 shows an example of labelling for a VS of a speaker from our corpus.

[76] This first step was necessary since the software *Sound File Cutter Upper* requires sound files of around 10-minute duration. This was made with a simple Matlab script that reads in batch mode all the WAV files of interest and performs a split based on the number of samples given a fixed sample rate.

[77] Software release 2010-12-02. Stable URL: http://geoff-morrison.net/#CutUp

[78] Software release 2012-07-30. Stable URL: http://geoff-morrison.net/#SndLbl

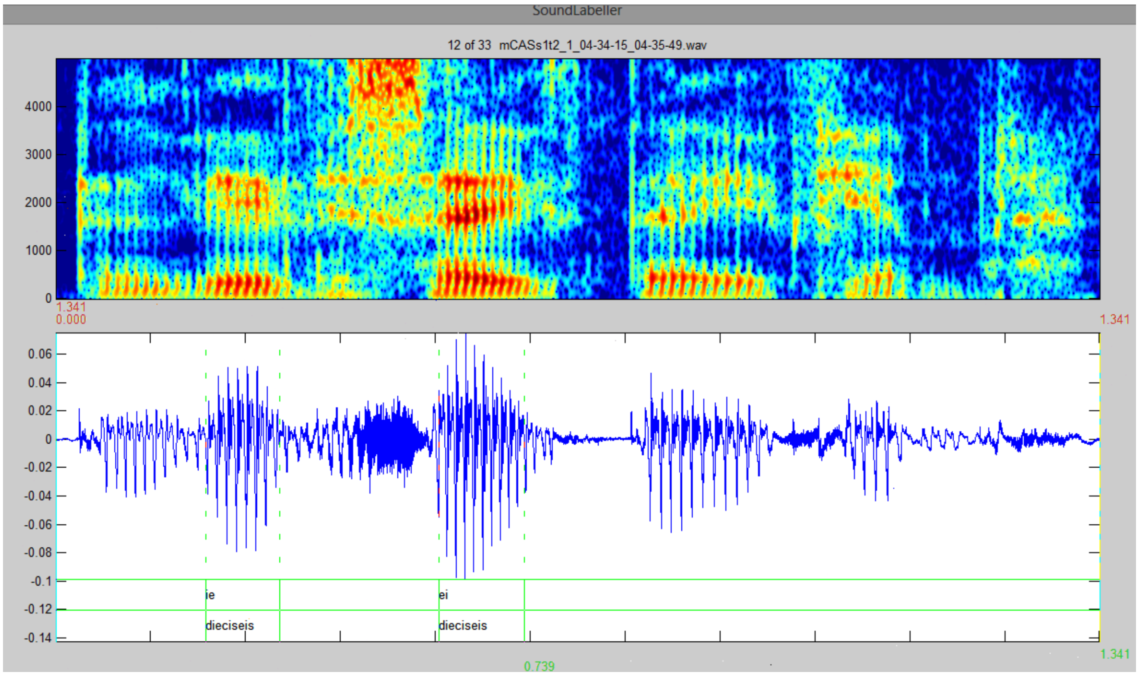


Figure 11. Example of labeling: in row 1 the VS [i̯e] and [ei] are labeled, while row 2 is used for the labelling of the words where those VS appear. In this case, both VS appear in the word *dieciséis*. This labelling belongs to the MZ speaker CAS, first session, second speaking task.

4.2.2. Analysis tools and method

Acoustic analysis

The VS obtained following the steps described in previous section, were then analyzed with *FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories*, developed by Morrison and Nearey (2011).[79] This software measures formant trajectories of the specified sequences using the formant tracking procedure outlined in Nearey, Assmann and Hillenbrand (2002). As specified in the software manual (Morrison & Nearey 2011: 3), “the software measures formant trajectories using a range of parameters for linear-predictive-coding (LPC)[80], runs some heuristics to attempt to identify the best track for each of the first three formants (F1, F2, F3)[81], and presents the results to a human for checking”[82].

[79] Software release 2011-05-26. Stable URL: http://geoff-morrison.net/#FrmMes

[80] The number of linear-predictive-coding coefficients is fixed at nine (according to Morrison & Neary, 2011: 3, “this will find four peaks out of which the three best formant candidates will be selected”), the sampling frequency is fixed at 10kHz, and the cutoff frequency is roved: “the idea is to put the cutoff frequency between F3 and F4 so that there are exactly three formants below the cutoff frequency, and (hopefully) produce good formant tracking results for F1, F2, and F3” (Morrison & Nearey, 2011: 3).

[81] The formant trajectories are extracted using the algorithm described in Markel & Gray (1976).

[82] The formants are tracked eight times using eight different cutoff values for F3 (2500-4000 Hz). As specified in Morrison (2008: 252): “Each of the eight formant-track sets are visually displayed overlain on a spectrogram. The measured intensity, fundamental frequency, and formant frequencies are also used to

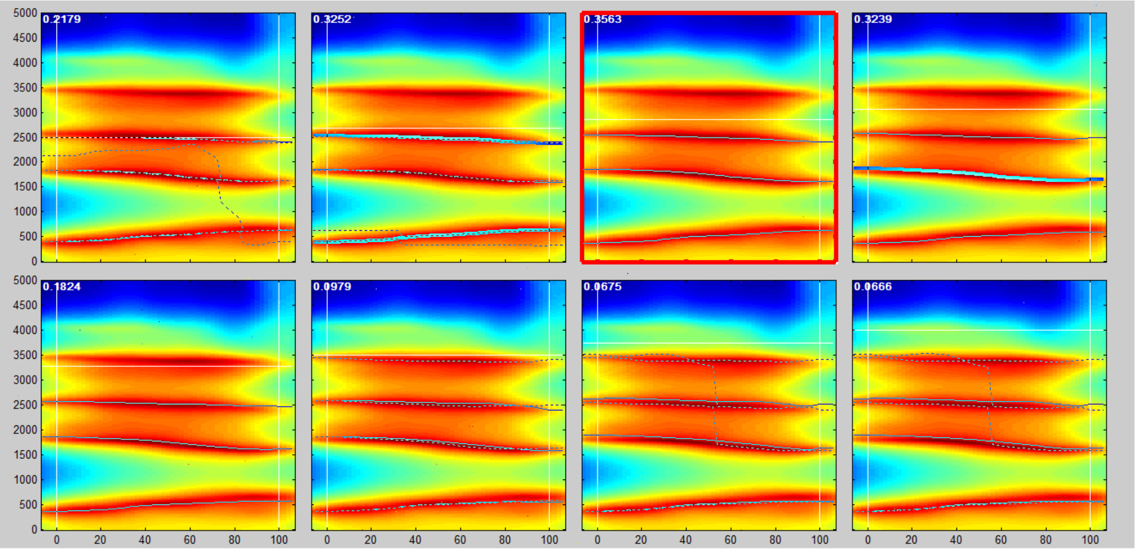


Figure 12. Example of selection of formant tracks with *FormantMeasurer*. This figure shows the VS [i̯a] for MZ speaker AGP, second session, second speaking task.

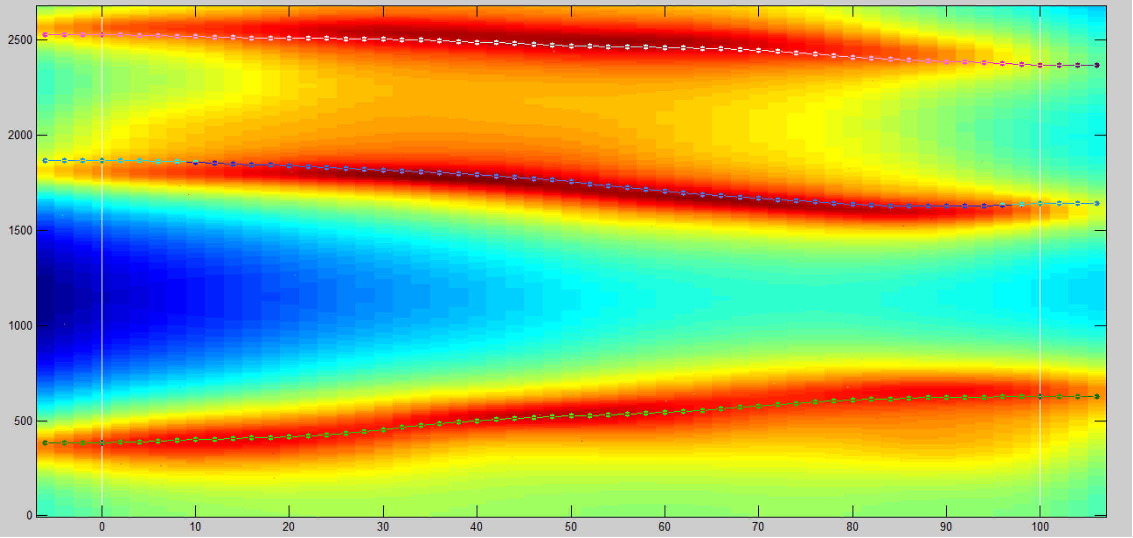


Figure 13. Best formant-track set for one of the VS [i̯a] of MZ speaker AGP, second session, second speaking task.

As can be seen in Figure 12, eight tracksets are displayed per VS. These tracksets correspond to eight different F3-F4 cutoff values. Solid lines are used for the tracks from three-formants-below-the-cutoff, while the tracks from four-formants-below-the-cutoff appear as dotted lines. The F1-F3 tracks with thick lines are those determined to be the best on the basis of the heuristics (Morrison & Nearey, 2012:12). If the researcher does not agree with the selected best track, he can choose other tracks. Once the selection of the best formant-track set is done,

synthesise a vowel. The researcher can listen to the original vowel and a synthesized vowel based on any desired track set. On the basis of visual and auditory inspection, the researcher *can* select what he judge to be the best formant-track set. The researcher also *has* the option of manually editing formant tracks, and of adjusting parameters for fundamental frequency measurement” this separately appears in another window (Figure 13) with the option of manually editing the formants tracks.

Curve fitting

Once the F1- F3 trajectories of each vocalic sequence were obtained, different parametric curves were fitted to each trajectory. Replicating the procedure followed by Morrison (2009c) or Enzinger (2010), we used two types of curve fitting: first, second and third order polynomials and also first- through third-order Discrete Cosine Transforms (DCT), which will be briefly described below. The coefficients of the parametric curves were then used as the input parameters in the likelihood ratio calculation described in next section.

The curve fitting is a procedure used for transforming a set of data points (the ones constituting the formant trajectories) into a small set of coefficients, thus performing data reduction. As explained in McDougall (2006: 102), the original data points prior to the curve fitting tend to assume a curvilinear relationship:

> The principles of regression[83] can be applied to a set of data points which appear to assume a curvilinear relationship, to determine a polynomial equation approximating the relationship. […]The regression procedure transforms the *x*-axis so that *y* is plotted against $a_0 + a_1x + a_2x^2$ and fits a straight line to the transformed data. […] This procedure can be further extended to a cubic or higher order polynomial or other curvilinear relationships such as exponential or logistic. (McDougall, 2006: 102)

a) *Curve fitting using polynomial functions*

This type of curve fitting approximates the formant-trajectory data points using polynomial functions of different degrees. The most basic polynomial function is the first-degree polynomial, which can be seen in Equation 2. This function includes an offset or constant value (α_0) and a slope coefficient (α_1) which corresponds to the linear function. The second type of function that we have considered is the second-degree polynomial function, described in Equation 3. This function includes a quadratic term with a α_2 coefficient. Again, for constructing the third-order polynomial functions, we add a cubic term with a α_3 coefficient, described in Equation 4.

[83] As defined by Barron (1997), "linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model" (Barron, 1997).

$$y(x) = \alpha_0 + \alpha_1 x \tag{2}$$

$$y(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 \tag{3}$$

$$y(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 \tag{4}$$

b) *Curve fitting using Discrete Cosine Transforms (DCT)*

The construction of a DCT function follows the same underlying idea that the polynomial curve fitting, but instead of using as basic elements the linear, quadratic and cubic functions, the DCT makes use of the sum of cosine functions with different amplitudes and frequencies as its building blocks or components. A cosine function is characterized by its amplitude (which will be multiplied by a coefficient α_k, being *k* the degree of the cosine function) and by its frequency (which will be higher as we increase the degree of the cosine function). It is by superposing the cosine curves of different degree that we obtain the fitting of the original curve.

In our study we consider cosine functions up to the third degree. The frequency of the first-degree component is set in a way that the curve fits half a period of a cosine (Equation 5). For the second and third-degree components we sum higher frequency cosines, which means that the curve will fit a full cosine period for the second-degree (Equation 6) and one period and a half for the third-degree component (Equation 7).

$$y(x) = \frac{\alpha_0}{\sqrt{N}} + \frac{2\alpha_1}{\sqrt{N}} C_1 \tag{5}$$

$$y(x) = \frac{\alpha_0}{\sqrt{N}} + \frac{2\alpha_1}{\sqrt{N}} C_1 + \frac{2\alpha_2}{\sqrt{N}} C_2 \tag{6}$$

$$y(x) = \frac{\alpha_0}{\sqrt{N}} + \frac{2\alpha_1}{\sqrt{N}} C_1 + \frac{2\alpha_2}{\sqrt{N}} C_2 + \frac{2\alpha_3}{\sqrt{N}} C_3 \tag{7}$$

where N is the number of points in the original curve and C_k is the k^{th}-degree DCT component:

$$C_1 = \cos\left(\frac{1}{2}\left(\frac{(2x+1)\pi}{N}\right)\right) \tag{8}$$

$$C_2 = \cos\left(\frac{(2x+1)\pi}{N}\right) \tag{9}$$

$$C_3 = \cos\left(\frac{3}{2}\left(\frac{(2x+1)\pi}{N}\right)\right) \tag{10}$$

Figure 14 shows two examples of curve fitting for the (F2) formant trajectory of VS [i̯e] corresponding to one token found in the second recording session of speaker ACL. The first graphical representation shows the polynomial fitting while the figure below shows the fitting using DCT functions.

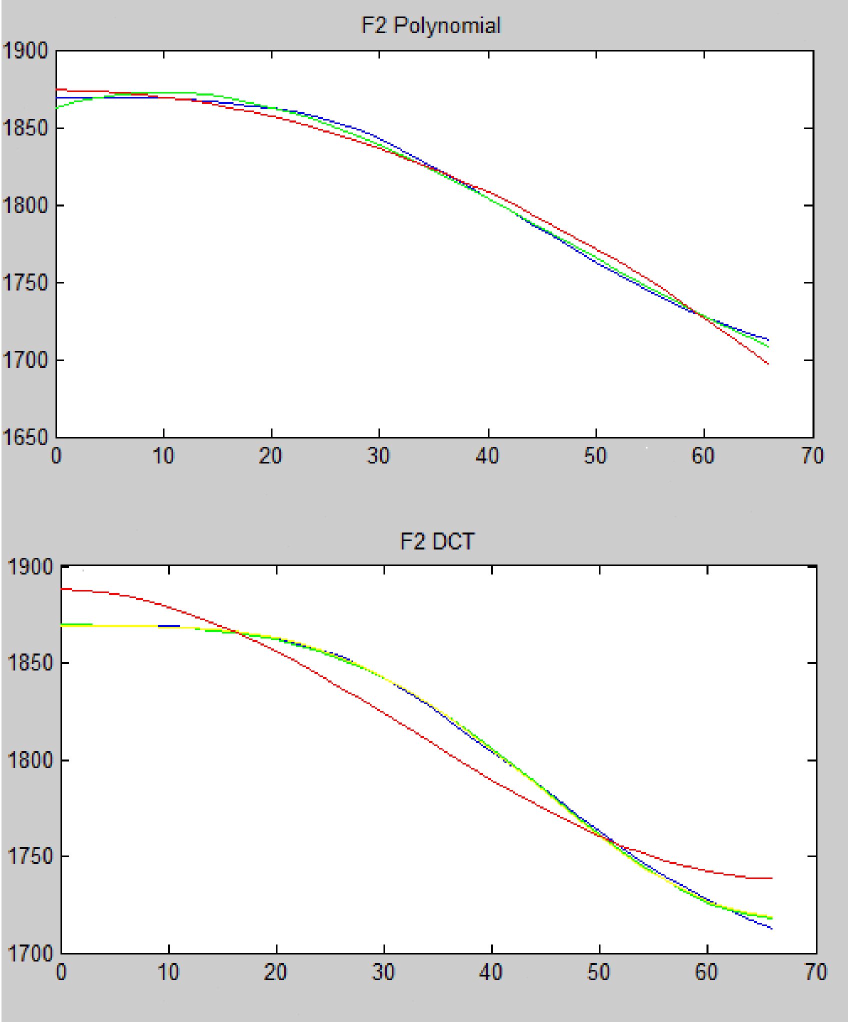


Figure 14. Example of curve fitting: the figure shows the approximation of the F2 of the VS [i̯e] for the speaker ACL, second session. The above figure represents the polynomial approximations while the figure below shows the formant-trajectory approximation using DCT functions. The original formant trajectories are depicted in blue, the first-degree approximations appears in red, the second-degree approximations in yellow, and the third-degree approximations in green.

Likelihood ratio calculation

We already mentioned in Chapter 3 (*cf. 3.6 The likelihood-ratio approach*) what a likelihood ratio (LR) is, what is the generic formula to calculate a LR and how its magnitude should be interpreted. For the LR calculation of the specific acoustic data of this study, we have used the *Multivariate Kernel Density (MVKD) formula* described in Aiken and Lucy (2004) and implemented by Morrison (2007).

With the formula described in Aitken and Lucy (2004) it is possible to obtain LRs from continuous multivariate data. It was originally envisaged for the evaluation of trace evidence in form of glass fragments, but afterwards it has also proven to be useful for the forensic comparison of voice and speech evidence (Enzinger, 2010; Kinoshita, Ishihara, & Rose, 2009; Morrison and Kinoshita, 2008; Rose, Kinoshita, & Alderman, 2006).

The MVKD formula allows an evaluation of a) the *similarity* of two speech samples with respect to the intra-speaker variation and b) the *typicality* of the speech samples with respect to an estimate of the probability density of a reference population[84]. In this formula, "the within-speaker variance is estimated via a normal distribution, and the between-speaker population probability density is estimated via a kernel model" (Morrison 2009c: 2390). Explaining the mathematical foundations of this formula lies beyond the scope of our study (cf. Aitken & Lucy, 2004 for these details). Yet, a more detailed description follows:

> In the MVKD the between-group distribution is modelled via the summation of a set of equally-weighted kernels with one kernel per group centred on the mean-vector of the measurements from that group (for application to forensic comparison of glass fragments, a group is a pane; for application to forensic voice comparison, a group is a speaker). Each kernel is a Gaussian whose covariance matrix is a scaled version of the pooled within-group covariance matrix. The scaling, and hence the degree of kernel smoothing, is determined by a function of the number of groups in the background database. (Morrison, 2011: 243)

The multivariate data used in our investigation are the coefficients obtained after approximating the formant trajectories of the VS by means of polynomial and DCT functions.

Besides, a *cross-validation procedure* has been adopted in this study for the calculation of each LR. By means of this procedure, the background database consisted of data from all speakers except for the speaker or speakers whose data were being compared.

More specifically, according to this procedure, each speaker's first session was compared:

a) With his own second session (which allow us to obtain non-contemporaneous *intra-speaker* comparisons).

b) With his brother's or speaking partner's second session. This allows us to obtain different-speaker comparisons of the following type: *intra-pair* MZ, DZ and B comparisons or just *inter-speaker* comparisons (for unrelated speakers, US).

[84] As indicated in Morrison (2009c: 2390), the estimate of the probability density of the population should be "based on the same sample of the population as is used to estimate the within-speaker variance".

c) With the first session of all the other speakers in the background database (which yield further *inter-speaker* comparisons).

Thus, cross-validated LRs were calculated separately for each VS, represented by the curve-fitting coefficients of each of their F1-F3 formant trajectories, as explained above.

Formants are combined directly in the MVKD formula while diphthongs need *a posteriori* fusion (see below). For our study we decided to combine F2-F3 (trajectory) information, leaving F1 aside for the reasons specified below. In other words, we aimed at characterizing each speaker by both his F2-F3 trajectories together. Other studies like Morrison (2009c) or Enzinger (2010) compared the performance of a system by fitting curves to the trajectories of F1, F2 and F3 with the performance of systems which only fitted curves to the trajectories of F2 and F3. Both of the above-mentioned studies concluded that the fused two-formant and three-formant systems yielded similar results, thus indicating that performance is not substantially deteriorated when F1 trajectories are not considered (Morrison, 2009c: 2395). Besides, it is well known (Künzel, 2001) that the first formant is usually compromised by the telephone network passband (0.3 – 3.4 kHz) which affect telephone transmissions in real forensic casework. For these reasons, we decided to fuse only F2 and F3 coefficients and not taking into account those of F1.

Goodness of fit (for the parametric curves) and fusion techniques

After having obtained the LRs for each VS using the different types of parametric curves, and having carried out the cross-validation procedure described above, the following step aimed at combining the results obtained per VS. This is done in order to improve the system performance, according to the state-of-the-art (e.g. González-Rodriguez et al., 2007; Morrison 2009c). There are several methods for combining (summing or fusing) the results yielded by different systems. In our investigation, we have 19 different forensic-comparison-systems (as many as VS have been studied) that yield different scores (the way to call pre-fused LRs) and we aim at fusing them all in a single LR for each speaker comparison. In next pages we will describe the different types of fusion procedures that have been tested. Before the fusion, we decided to choose only the results coming from the best fitting parametric curve. For that purpose, we tested the *goodness of fit* of both types of parametric functions (and their three degrees) by means of correlation. As shown in Section 4.4, both the cubic polynomial and the third-degree DCT outperform their second-degree counterparts. In turn, the latter outperform the first-degree functions. That is, a better approximation of the real formant trajectory is achieved using this kind of curves. For the rest of the investigation, we decided to work only with the results obtained from the third-degree polynomial and DCT functions. We selected the results yielded by these approximations for the subsequent fusion of VS.

As far as the *fusion techniques* are concerned, two basic distinctions can be made: The first procedure assumes statistical independence of the scores (systems) to be combined, while the second one does not, and therefore it needs some calibration. This second procedure is called logistic-regression fusion. Both types of combination procedures will be described below.

In a first step, we combined the scores obtained from the 19 systems (one per VS) by simply multiplying them together. This procedure is called *Naïve Bayes* (also Idiot's Bayes or Independence Bayes, cf. Rose, 2006: 171) and it assumes that the variables are independent, i.e. they are not correlated. Therefore, the value of the combined LR (LR_c) will be calculated as follows:

$$LR_c = Score_1 \times Score_2 \times Score_3 \times \cdots Score_{19} \qquad (11)$$

Yet, in order to avoid an overconfidence of the LR_c obtained, in a further step we proceeded to calculate the 19th root of the product, i.e. obtaining the geometric mean. Assuming statistical independence where there is actually correlation between variables, naïve Bayes fusion tends to yield overestimated LRs. Therefore, the calculation of the geometric mean of all the 19 scores instead of the simple product is recommended to compensate this overconfidence in the LRs (Daniel Ramos, personal communication), as follows:

$$LR_c = \sqrt[19]{Score_1 \times Score_2 \times Score_3 \times \cdots Score_{19}} \qquad (12)$$

In relation to the second type of score fusion, we used *logistic regression*, a well-known statistical classification model (see e.g. Hastie, Tibshirani and Friedman, 2009: 119-128). In its forensic application, the use of logistic regression implies not only *fusion* but also *calibration* (e.g. Brümmer et al., 2007; Brümmer and du Preez, 2006; Gónzalez et al., 2007; Morrison & Kinoshita, 2008; Pigeon, Druyts & Verlinde, 2000; Ramos-Castro, 2007; van Leeuwen & Brümmer, 2007; in Morrison, 2010a: 3061).

On the one hand, *calibration* is the process of designing and optimizing the transformation from the raw scores calculated by different systems into LRs in such a way that a cost function is minimized. On the other hand, *fusion* converts multiple sets of scores into LRs. In any case, what scores do is "quantifying the degree of similarity of pairs of samples while also taking account of their typicality" (Morrison, 2010a: 3061). These scores, or uncalibrated LRs, do not have an absolute meaning by themselves. However, it is the LR value after calibration what represents the weight of the evidence. As explained in Morrison (2013: 177), "a LR can be sequentially calculated for each data point, but what is needed is a LR that characterizes the strength of evidence with respect to the whole of the offender's speech on the offender recording, not with respect to multiple individual portions of the recording".

For the application of logistic regression, it is necessary to have some *training data*: scores from comparisons where it is known whether they are same-speaker comparisons or different-speaker comparisons. The appropriateness of this method for our acoustic data seems clear, as we have, per speaker and session, several tokens of different phonemes (the 19 different VS) from the same voice recordings. As specified in Morrison (2011: 245), "fusion requires multiple sets of scores of parallel comparisons made on the same data, e.g […] two or more acoustic-phonetic comparisons each run on tokens of a different phoneme from the same voice recordings". Calibration would be defined as "the application of an affine transformation to a set of scores, e.g., a linear shifting and scaling of the scores, so as to minimize a cost function" (Morrison, 2011: 245).

The logistic regression calibration objective (achieved in a first step called training) is to minimize the cost function C_{wlr} (Equation 13), which depends on a synthetic parameter P, the number of training scores known to be from the same origin N_{so} and the number of training scores known to be from different origin N_{do}. In typical forensic-phonetic scenarios, a value of P=0.5 is used, which means that the prior odds are 1 (cf. Brümmer, 2005). In this case, the C_{wlr} function becomes the cost function C_{llr}, described in next subsection.

$$C_{wlr} = \frac{P}{N_{so}} \sum_{j=1}^{N_{so}} \log\left(1 + e^{-f_j - logit(P)}\right) + \frac{1-P}{N_{do}} \sum_{j=1}^{N_{do}} \log\left(1 + e^{g_j + logit(P)}\right) \qquad (13)$$

where:

$$logit(P) = log\left(\frac{P}{1-P}\right) \qquad (14)$$

And where f_j represents the fused same-origin scores used in the training (s_{ij}) and g_j represents the fused different-origin scores used in the training (r_{ij}) with the weights to be used in the fusion stage α_i.

$$f_j = \alpha_0 + \sum_{i=1}^{19} \alpha_i s_{ij} \ ; \ g_j = \alpha_0 + \sum_{i=1}^{19} \alpha_i r_{ij} \ ; \qquad (15)$$

Once the appropriate weights (α_i) are obtained during the training phase after a series of one thousand iterations in order to achieve the minimal C_{wlr}, these weights will be used for performing the final step: the fusion of the scores in the so-called test phase. The formula for the score fusion is a simple weighted sum (linear fusion) of the scores, as follows (Equation 16):

$$LLRc = \alpha_0 + \sum_{i=1}^{19} \alpha_i Score_i = \alpha_0 + \alpha_1 Score_1 + \alpha_2 Score_2 + \cdots \alpha_{19} Score_{19} \qquad (16)$$

For carrying out calibration and fusion we used the logistic regression functions in the *FoCal Toolkit* (Brümmer, 2005), both for the training part and the fusion part. In relation to the training stage, and considering the size of our database (not as large as the ones normally used in statistical studies due to the inherent limitations of twin studies and due to budget and time constraints for speaker recruiting) we needed to ensure that the training population did not include any of the subjects under test. For that purpose we trained a different model (and thus obtained different weights) for each of the comparisons carried out by the 19 MVKD systems. The trained model did not include any of the two speakers being compared, in order to fulfill an honesty criterion, namely, that the scores used for training must be different from the scores to be fused.

Accuracy assessment: log-likelihood-ratio cost (C_{llr}) and Tippett plots

Assessing the output accuracy of a forensic-comparison system is a very relevant aspect in forensic sciences. Several measures and graphical ways have therefore been developed to evaluate such accuracy. For this study we have used the *log-likelihood-ratio cost* (C_{llr}), originally envisaged for its use in automatic speaker recognition (Brümmer & du Preez, 2006; van Leeuwen & Brümmer, 2007) but also applied in forensic-comparison studies based in traditional acoustic parameters (e.g. González-Rodríguez et al., 2007; Morrison & Kinoshita, 2008). Besides, *Tippett plots* have also been used as a graphical method to present the output of our forensic system based on VS and to assess its accuracy.

In relation to the C_{llr}, its most outstanding characteristics are, following Morrison (2010a): their continuous nature (worse results are more heavily penalized) and the fact that they are based on LRs. This measure is defined in Equation 17:

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{H_p}}\sum_{i=1}^{N_{H_p}} log_2\left(1+\frac{1}{LR_i}\right) + \frac{1}{N_{H_d}}\sum_{j=1}^{N_{H_d}} log_2\left(1+LR_j\right)\right) \qquad (17)$$

where N_{Hp} is the total amount of LRs for the *Hp* (hypothesis of the prosecution) and N_{Hd} is the total amount of LRs for the *Hd* (hypothesis of the defense). The LRs for the *Hp* are referred as LR_i and the LRs for the *Hd* are called LR_j.

In a typical forensic situation, *Hp* equals H_{ss}, i.e. "the offender and the suspect samples are from the same origin (same speaker)" while *Hd* equals H_{ds}, i.e. "the offender and the suspect are different speakers". The C_{llr} will depend on these hypotheses. However, for the speaker types that we are considering for our investigation, different hypotheses for the defense (*Hd*) can be established, while the *Hp* remains H_{ss}. Thus, when considering MZ twins, the *Hp* will be "the offender is not the suspect but his MZ twin" and the same with DZ twins and B siblings. The different C_{llr} values obtained according to the different hypotheses of the defense will be discussed in next Section 4.5.

According to the equation above, the lower the C_{llr}, the more accurate the performance of the system. This measure can be used to compare several systems which are based on the same set of data. For instance, we have compared for our study the performance of 19 systems, one per VS. (*cf. 4.4. Results*).

On the assumption that target comparisons (LR_i) should yield high LR values and non-target comparisons (LR_j) should yield low LR values for a forensic system to perform optimally, any deviation from this ideal situation is punished, with highly misleading LRs being charged heavier penalty (i.e. higher C_{llr} values) and vice versa (cf. González-Rodríguez et al., 2007: 2107). So, for every comparison system, large positive LLR (log-likelihood ratio) values which correctly support the same-speaker hypothesis (H_{ss} or H_p, ie. the hypothesis of the prosecution) are assigned very low C_{llr}. In contrast, negative LLRs which misleadingly support the different-speaker hypothesis (H_{ds} or H_d, ie. the hypothesis of the defense) are assigned high C_{llr} values. As specified in Morrison (2010a), this C_{llr} values get higher and higher as the LLRs become more negative and provide stronger contrary-to-fact support for the different-speaker hypothesis. Since LLRs close to zero do not provide a strong support for either H_{ss} or H_{ds} they are assigned moderate C_{llr} values.

Normally, not only the single measure C_{llr} is offered but also the so-called C_{llr}^{min} , which represents the C_{llr} obtained for a system without calibration errors. The difference between C_{llr} and C_{llr}^{min}, known as C_{llr}^{cal}, yields a numeric value which represents the calibration loss of the system.

Tippett plots represent another method for evaluating the performance of a forensic-comparison system but, as compared with the single measuring value of the C_{llr}, Tippett plots are graphical representations where more information can be found about the output of a LR-based comparison system. This type of representation was proposed by Evett and Buckelton (1996) in the field of DNA analysis and it owes its name to the work of Tippett et al. (1968) who first referred to the concepts of "within-source comparison" and "between-source comparison" (cf. Drygajlo, Meuwly, & Alexander, 2003). In this type of graph two curves are displayed, each one representing the probability for one of the competing hypothesis: H_p or H_d. Usually the hypothesis of the prosecution is that the offender and the suspect samples come from the same speaker, while the hypothesis of the defense is that they belong to different speakers. However, for the speaker types that we are testing (MZ, DZ, B or US), we will also draw Tippett plots based on a different H_d. (cf. Section 4.4.). Figure 15 provides an example of a Tippett plot based on the output of a hypothetical forensic-comparison system, where the line rising to the right represents the cumulative distribution of LLRs less than or equal to the value indicated in the x-axis, calculated for target (same-speaker) comparisons and the lines rising to the left show the cumulative distribution of LLRs greater than or equal to the value indicated in the x-axis, calculated for non-target (different-speaker) comparisons.

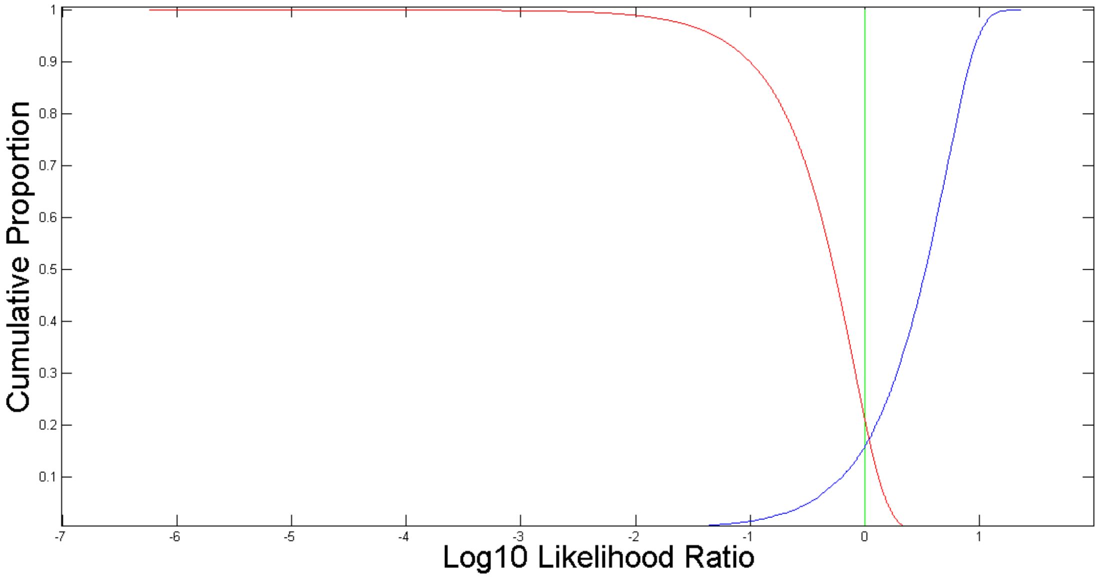


Figure 15. Hypothetical Tippett plot with fictitious data.

In Figure 16 we aim at providing a diagram which explains the different stages carried out for this first type of analysis (formant trajectories) and which have been described throughout this methodological subsection of Chapter 4.

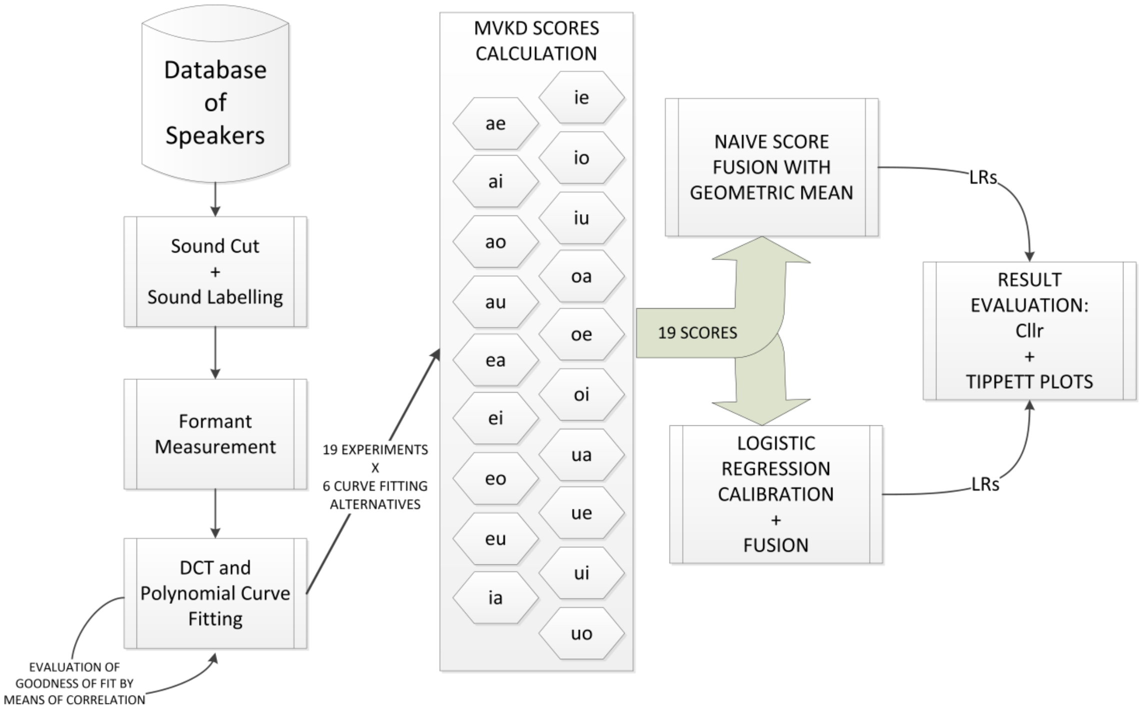

Database of Speakers
Sound Cut + Sound Labelling
Formant Measurement
DCT and Polynomial Curve Fitting
EVALUATION OF GOODNESS OF FIT BY MEANS OF CORRELATION
19 EXPERIMENTS X 6 CURVE FITTING ALTERNATIVES
MVKD SCORES CALCULATION
ae
ai
ao
au
ea
ei
eo
eu
ia
ie
io
iu
oa
oe
oi
ua
ue
ui
uo
19 SCORES
NAIVE SCORE FUSION WITH GEOMETRIC MEAN
LOGISTIC REGRESSION CALIBRATION + FUSION
LRs
LRs
RESULT EVALUATION: Cllr + TIPPETT PLOTS

Figure 16. Diagram showing the different stages carried out for the VS formant-trajectory analysis, from the speech material extraction to the accuracy evaluation of the forensic-comparison systems.

4.3. Parameters

We have explained above that the parameters used for the LR calculation are the coefficients extracted from the curve fitting of the formant trajectories in the vocalic sequences of interest. Here follows a brief description of the vocalic sequences.

Our study focuses on Spanish vocalic sequences. With this term we refer to the combination of vowel-vowel sequences as well as to the combination of glide-vowel sequences. In Spanish the first type of sequences are called *hiatuses* and the second type *diphthongs* (Aguilar, 1999), even though some terminological issues arise repeatedly regarding the phonological nature of diphthongs, the interpretation of glides or the syllabification process (see, for instance, Anderson, 1985; Navarro-Tomás, 1946; Alarcos, 1965; Hualde, 1991- In Aguilar, 1999: 58; and RAE, 2011: 335 and 342-343). In RAE (2011: 332), hiatuses are also called *heterosyllabic combinations* (i.e the elements making up the vocalic set belong to different syllables) while the label *tautosyllabic combinations* (i.e. belonging to the same syllable) is used to designate both diphthongs and triphthongs. Again, in the chapter devoted to vocalic sequences in RAE (2011), it is highlighted that the limits between tautosyllabic and heterosyllabic combinations are not always clear.

Yet, the differentiation between hiatuses and diphthongs is considered “a genuine feature in Spanish” (Aguilar, 1999: 59):

> The fact that a sequence can be pronounced as a hiatus – i.e. in two separate syllables – or must be pronounced as a diphthong – that is, in a single syllable – is a lexical property: the acquisition of a new word implies the knowledge about its syllabification. (Aguilar, 1999: 59)

We are also interested in the fact that both hiatus and diphthong pronunciations are sometimes allowed, for example in words like *cardíaco* / *cardiaco* (‘cardiac’)[85]. The language allows variation in some other cases, some of which we have described in Chapter 3, and we have included those words with variable pronunciation in our corpus design. On this basis, we think that considerable inter-speaker variation can be found, not only in specific words, but in general in the pronunciation of vocalic sequences in Spanish, which could be useful for forensic purposes.

Despite the fact that the “property of syllabicity is not phonetically defined in a precise way” (Aguilar, 1999: 58), some acoustic cues have been highlighted for the hiatus-diphthong distinction, such as the role of the *formant transition rate* (Borzone de Manrique, 1979; Quilis, 1981), the *onset duration*, *transition duration* and *offset duration* (Borzone de Manrique, 1979)[86]. In Aguilar (1999) a novel approach is taken to find which acoustic cues, in the temporal and frequency domain, distinguish hiatuses from diphthongs. For that purpose, second-order polynomial equations were fitted to the F1 and F2 trajectories of certain 24 combinations of Spanish vocalic sequences[87]. The results of the study carried out by Aguilar (1999) show that hiatuses and diphthongs differ in both the temporal and the frequential domain, with hiatuses having a longer duration and a greater degree of curvature in the F2 trajectory than diphthongs. Besides, Aguilar (1999) found that there were differences between the two categories (hiatus and diphthong) depending on the communicative situation and that they behaved differently as far as phonetic reduction is concerned: “there is […] an axis of reduction where a hiatus becomes a diphthong and a diphthong becomes a vowel” (Aguilar, 1999: 73)[88]. For the purposes of our thesis,

[85] In some derivational words, there is a double stress pattern which affects the derivational suffixes and may therefore imply a double hiatus-diphthong pronunciation of some vowel sequences. For example, in the case of the suffix *–íaco/-íaca* ~ *-iaco/-iaca* both the proparoxytone and the paroxytone form of the suffix are allowed (RAE, 2011: 398).

[86] In RAE (2011: 337), the main phonetic differences between diphthongs and hiatuses refer to three characteristics: sequence *duration,* formant *transitions* and *amplitude*: As regards duration, diphthongs would be shorter than their corresponding hiatuses. Although there is no agreement in this aspect, the transition between vowels would be slower in diphthongs than in hiatuses. In this respect, the curvature of the F2 transition between two vowels would be more prominent in the diphthong than in the hiatus. If we focus on amplitude, this parameter would be more similar between vowels in a diphthong than in a hiatus.

[87] In order to build up the corpus, the following variables were taken into account: phonetic category (hiatus and diphthong), stress and the vowel that follows the segment [i] or [u]: [a], [e] or [o] (Aguilar, 1999: 59).

[88] According to Aguilar (1999: 73), “these results will argue in favour of the existence of a phonological structure shared by all the speaking styles, but with different phonetic manifestations in function of extralinguistic factors, such as the speaker's attention to his speech”.

we are interested in testing whether similar kind of curve parameterization is useful for distinguishing speakers, rather than for a diphthong-hiatus differentiation.

Diphthongs have been traditionally classified in rising and falling diphthongs (Aguilar, 2010; Navarro Tomás, 1972; RAE, 2011). According to the description found in RAE (2011: 332), in *rising diphthongs*, the vowel marked with the feature [+high] appears in the first position of the vocalic sequence, while the vowel marked with the feature [-high] is in the second position. For the phonetic realization of these diphthongs (e.g. *miedo, justicia, tienda*), speech articulators move from a closure to an open position, making the second vowel in the sequence more salient. On the contrary, in *falling diphthongs* (e.g. *aula, boina, peine*), where the high vowel appears in second position, the speech articulators move from an open position to a closure. In these cases, the more salient vowel is the first one.

Diphthongs can also be made up by two different high vowels, like *ui*, as in *cui.das*. On numerous occasions, it has been stated that the Spanish language favors diphthongization[89], showing a clear tendency to avoid hiatuses (RAE, 2011:339). The RAE (2011: 333) adds that, furthermore, there is a preference for rising diphthongs in Spanish. This would be the reason why if two high vowels appear together, they would form a rising diphthong, as in *bui.tre, ciu.dad* or *viu.do*. Nevertheless, different factors may contribute to their realization as falling diphthongs, giving rise to pronunciation vacillations, as we will explain below.

Concerning the elements of a diphthong, this type of vocalic sequences have traditionally been said to be made up of a *semivowel* or *semiconsonant* and another vowel[90]:

> The *i* and *u* vowels are pronounced […] as semivowels when they appear at the end of the diphthong, and as semiconsonants when they appear at the beginning. […] In the groups *iu*, *ui* the predominant element of the diphthong is the second vowel, while the first one is reduced to a semiconsonant. (Navarro Tomás, 1972: 65; our translation)

According to the Spanish linguistic tradition, the semivowel *u* is transcribed as [u̯] and the semivowel *i* is transcribed as [i̯], while the semiconsonant *u* is transcribed as [w] and the semiconsonant *i* is transcribed as [j] (Navarro Tomás, 1972: 62-63). However, in RAE (2011), another convention is adopted: "According to the International Phonetic Alphabet, the transcription of these elements [*semivowels and semiconsonants*] is [i̯], [u̯], no matter whether they appear before or after the syllabic vowel" (RAE, 2011: 333). This is also the transcription

[89] The tendency to avoid hiatuses is especially prominent in fast speech (RAE, 2011: 349). This trend would explain many synaeresis and synalepha phenomena, being synaeresis the reduction to a single syllable of the vowels in a hiatus, taking place in a within-word context, while synalepha is the same phenomenon but occurring between words (RAE, 2011: 353).

[90] This vowel may receive different names: syllabic vowel or full vowel, for instance (RAE, 2011: 333).

convention adopted in Aguilar (2010: 25) and Gil (2007: 448), as well as the convention that we have used in this thesis. The main arguments for supporting the use of the non-syllabic diacritic [̯] instead of [j] or [w] are put forward by Aguilar (2010). On the one hand, adopting [j] and [w] –as in the Spanish linguistic tradition– implies using the same symbol for a vocalic segment and for a consonantal segment. On the other hand, it entails a non-unitary treatment of the vocalic sequences, as it only affects [i] and [u] (Aguilar, 2010: 24; see also the review of this book in San Segundo, 2010b). The cover terms 'vocal satélite' or 'vocal marginal' (*satellite vowel* or *marginal vowel*, in English) are proposed by RAE (2011) to refer to both semivowels and semiconsonants, although a more commonly used term (Aguilar, 2010; Gil, 2007) is 'paravocal' (*glide*, in English).

RAE (2011: 337) mentions two cases which present *pronunciation vacillations*:

- On the one hand, the combination of a vowel with the feature [+high] with a vowel with the feature [-high] may be submitted to pronunciation fluctuations. Examples of this kind are words like *anual*, *biombo*, *crueldad* or *diana*, which may be pronounced with hiatus or diphthong depending on several factors, not only geographic, sociolinguistic or stylistic, but also etymological or analogical.
- On the other hand, the combination of two high vowels (group *iu* or *ui*) also exhibits variation: while *buitre* or *cuita* are usually pronounced with diphthongs, the hiatus is preferred in words like *diurno* or *jesuita*.

Interestingly, since Navarro Tomás (1972: 149) the creation of rules to regulate such vacillations has been considered pointless, given the several factors conditioning the two possible pronunciations and hence the speakers' freedom towards these vocalic sequences (RAE, 2011: 337).

The vowel combinations for the different vocalic sequences making up the corpus in the second speaking task of this study are shown in Table 10. As can be seen, all 5 Spanish vowels are combined with each other with the exception of same-vowel sets[91]. The sequence *ou* was also discarded for the reasons established in Chapter 3.[92] Finally, we have to note that only intra-lexical vocalic sequences have been taken into account for this study. Vocalic sequences occurring in the limit of two words (e.g. -ai- en "un*a i*glesia") were not studied. The methodology to search the

[91] Same-vowel combinations are also described from the point of view of hiatus/vowel reduction differentiation (RAE, 2011: 339). However, for this study we have not considered their inclusion.

[92] In the search for words containing the original 20 vocalic sequences (resulting from combining all of them with each other, except same-vowel combinations) we found that there were almost no words with the combination –ou-; only the compound word *estadounidense*. The rest were foreign loanwords (e.g. *glamour, soul, country, boutique*) which have not been considered for this study.

words containing each of the selected vocalic sequences was described in Chapter 3 (cf. Section 3.3.2).

Table 10

Vowel combinations for the selected 19 vocalic sequences (VS)

	Second vowel of the VS				
First vowel of the VS↓	a	e	i	o	u
a	-	ae	ai̯	ao	au̯
e	ea	-	ei̯	eo	eu̯
i	i̯a	i̯e	-	i̯o	i̯u
o	oa	oe	oi̯	-	-
u	u̯a	u̯e	u̯i	u̯o	-

Note. Both rising and falling diphthongs are represented with [i̯] and [u̯].

4.4. Results

4.4.1. Curve fitting: best correlation values

As explained in previous sections, the formant trajectories of F1, F2 and F3 were fitted using two types of parametric curves: (linear, quadratic and cubic) polynomials and (first, second and third order) DCT functions. In a first step, we calculated the *goodness of fit* of each function by means of linear correlation (R values are shown in Table 11).

Of these functions, the quadratic and cubic polynomials, and the second and third order DCTs show the highest correlation values, while the first-order functions yielded much lower R values. Therefore, we have included in Table 11 only the correlation values corresponding to these coefficients. As explained in previous section, the F1 was discarded for fusion in the MVKD formula, so Table 11 shows the values exclusively for F2 and F3.

As can be seen in Table 11, the approximation of F2 trajectories is always better than the fitting of F3. This occurs for all VS and regardless of the type of parametric curve. We highlight in bold versus normal the values for F2 versus F3. Besides, the third-degree functions (both in polynomials and DCTs) outperform their second-degree counterparts. Again, this trend is observed in all VS and irrespective of the formant. In the table, the italics are used to highlight this.

Table 11

Correlation coefficients between the original formant (F2 and F3) trajectory and their fitted curves (polynomial and DCT)

		Type of curve fitting			
		Polynomial		DCT	
Vocalic sequence	Formant trajectory	Quadratic	Cubic	2nd-degree	3rd- degree
[ae]	F2	**0.9800**	***0.9923***	**0.9844**	***0.9915***
	F3	0.8486	0.9058	0.8588	0.9132
[ai]	F2	**0.9816**	***0.9941***	**0.9908**	***0.9952***
	F3	0.9112	0.9531	0.9250	0.9607
[ao]	F2	**0.9378**	***0.9851***	**0.9511**	***0.9809***
	F3	0.8401	0.9023	0.8360	0.8939
[au]	F2	**0.9273**	***0.9845***	**0.9451**	***0.9799***
	F3	0.8138	0.9156	0.8164	0.9075
[ea]	F2	**0.9673**	***0.9897***	**0.9790**	***0.9904***
	F3	0.8205	0.8907	0.8383	0.9050
[ei]	F2	**0.9658**	***0.9860***	**0.9619**	***0.9765***
	F3	0.8765	0.9380	0.8832	0.9380
[eo]	F2	**0.9600**	***0.9905***	**0.9790**	***0.9932***
	F3	0.8496	0.9194	0.8654	0.9302
[eu]	F2	**0.9350**	***0.9843***	**0.9638**	***0.9896***
	F3	0.8369	0.9023	0.8548	0.9201
[ia]	F2	**0.9743**	***0.9911***	**0.9844**	***0.9915***
	F3	0.8555	0.9295	0.8713	0.9327
[ie]	F2	**0.9710**	***0.9886***	**0.9740**	***0.9850***
	F3	0.9116	0.9599	0.9228	0.9598
[io]	F2	**0.9714**	***0.9919***	**0.9845**	***0.9941***
	F3	0.8849	0.9430	0.9029	0.9527
[iu]	F2	**0.9551**	***0.9894***	**0.9790**	***0.9944***
	F3	0.8837	0.9334	0.9028	0.9524
[oa]	F2	**0.9684**	***0.9885***	**0.9714**	***0.9842***
	F3	0.8494	0.9151	0.8447	0.9095
[oe]	F2	**0.9726**	***0.9940***	**0.9880**	***0.9959***
	F3	0.8284	0.9091	0.8422	0.9239
[oi]	F2	**0.9698**	***0.9900***	**0.9853**	***0.9937***
	F3	0.8539	0.9134	0.8704	0.9351
[ua]	F2	**0.9686**	***0.9898***	**0.9756**	***0.9873***
	F3	0.8375	0.9119	0.8396	0.9113
[ue]	F2	**0.9819**	***0.9940***	**0.9904**	***0.9951***
	F3	0.8186	0.9195	0.8202	0.9150
[ui]	F2	**0.9689**	***0.9892***	**0.9837**	***0.9928***
	F3	0.8907	0.9444	0.8989	0.9543
[uo]	F2	**0.9310**	***0.9707***	**0.9289**	***0.9591***
	F3	0.8008	0.8878	0.7897	0.8705

Note. Correlation coefficients between the original formant (F2 and F3) trajectory and their fitted curves (polynomial and DCT) show their goodness of fit. The values correspond to the average of all the speakers in this study. We highlight in bold the R values for F2, which are always larger than for F3, regardless of the VS or the curve fitting procedure, and we highlight in italics the R values for cubic polynomials and third-degree DCT, which are larger than the quadratic and second-order functions, across VS types and formants.

Some trends can be observed as regards which VS are best fitted. Having noted that the third-degree functions outperform the second-degree ones, we describe only the results for these functions. The following values are obtained for the polynomial:

- The R values for F2 range between a minimum of 0.9707 (corresponding to /uo/) and a maximum 0.9941 (/ai/), although /ue/ and /oe/ also yield very similarly high values (both are 0.9940).
- The R values for F3 range between 0.8878 (/uo/) and 0.9599 (/ie/), although /ai/ gets also high correlation values, with R = 0.9531

As far as the DCT is concerned, the following correlation values are obtained:

- The R values for F2 range between 0.9591 (/uo/) and 0.9959 (/oe/) but /ai/ and /ue/ get also high values (0.9952 and 0.9951, respectively).
- The R values for F3 range between 0.8705 (/uo/) and 0.9607 (/ai/) but /ie/ and /ui/ yield high correlation values as well (0.5998 and 0.9543, respectively).

All in all, /uo/ seems to be the VS where a comparatively worst fitting is achieved. This occurs for all types of curve fitting and degrees, and both for F2 and F3. In contrast, the VS with the maximum R value depend on the parametric function and formant considered, although there are slight differences between the maximum and the following highest R value. In general, we can observe that /ai/, /ie/, /ue/ and /oe/ tend to obtain high correlation values.

4.4.2. Combination/Fusion techniques: comparing MZ, DZ, B and US tests

We explained in the section devoted to the methodology (cf.Section 4.2.2) that after obtaining the results for each individual LR (before calibration, these are called *scores*), we would proceed to a fusion of these values in order to improve the system performance. Having tested that the third-order functions both for polynomial and DCT fit the real formant trajectories better than the other functions, we have considered the fusion of scores only for these (Poly3 and DCT3 from now on).

The first procedure that we carried out to combine the multiple scores resulting from the 19 different systems (one per VS) consisted in simply multiplying them together *à la naïve* Bayes. To the product of the multiplication, we further calculated the *n*th root with the aim of obtaining the geometric mean. The purpose of this was compensating the overconfidence expected in the

LRs obtained within the naïve-approach, i.e. without taking into account the correlation of variables. In Table 12 we can see the values obtained per speaker comparison.

Table 12

Results for the different speaker comparisons (geometrical mean)

		MZ (I)		MZ(O)	DZ(I)		DZ(O)	B(I)		B(O)	US(I)		US(O)
Cases →		01v01/02v02		01v02	13v13/14v14		13v14	21v21/22v22		21v22	25v25/26v26		25v26
Scores	**Poly3**	5.33	3.95	4.41	8.87	2.05	0.61	9.91	7.83	0.14	4.67	2.78	0.54
	DCT3	4.27	3.47	2.83	4.05	2.73	1.05	13.82	7.94	0.14	4.07	2.58	0.34
Cases →		03v03/04v04		03v04	15v15/16v16		15v16	23v23/24v24		23v24	27v27/28v28		27v28
Scores	**Poly3**	3.22	2.82	0.58	2.64	5.57	0.77	4.60	3.23	3.32	3.59	5.03	0.91
	DCT3	5.19	2.90	1.75	1.98	4.91	0.33	5.91	6.50	3.90	3.39	9.57	2.22
Cases →		05v05/06v06		05v06	17v17/18v18		17v18	47v47/48v48		47v48	29v29/30v30		29v30
Scores	**Poly3**	2.12	1.17	0.46	1.84	0.40	0.48	1.89	4.00	0.05	2.75	3.58	0.00
	DCT3	1.64	1.90	0.94	3.17	1.92	0.42	1.48	3.17	0.01	1.08	0.60	0.40
Cases →		07v07/08v08		07v08	19v19/20v20		19v20	49v49/50v50		49v50	31v31/32v32		31v32
Score	**Poly3**	7.90	5.13	0.34	6.84	9.24	0.40	2.97	1.62	0.12	3.94	3.20	0.01
	DCT3	5.70	3.21	0.16	5.10	7.65	0.76	2.73	1.49	0.16	6.99	2.75	0.02
Cases →		09v09/10v10		09v10	45v45/46v46		45v46				51v51/52v52		51v52
Score	**Poly3**	1.07	0.82	0.79	0.33	3.10	0.09				3.31	3.19	0.96
	DCT3	1.14	1.03	0.89	1.45	4.81	0.12				1.84	7.33	0.70
Cases →		11v11/12v12		11v12							53v53/54v54		53v54
Score	**Poly3**	1.80	1.07	0.21							3.57	4.24	0.65
	DCT3	2.99	2.36	1.37							1.61	2.82	0.17
Cases →		33v33/34v34		33v34									
Score	**Poly3**	4.52	5.64	1.87									
	DCT3	3.96	4.88	3.82									
Cases →		35v35/36v36		35v36									
Score	**Poly3**	1.05	2.36	2.69									
	DCT3	3.71	2.72	2.98									
Cases →		37v37/38v38		37v38									
Score	**Poly3**	0.40	4.01	2.09									
	DCT3	4.57	3.61	3.30									

Cases →		39v39/40v40		39v40
Score	**Poly3**	2.43	14.25	0.01
	DCT3	2.16	5.74	1.64
Cases →		41v41/42v42		41v42
Score	**Poly3**	2.70	2.22	1.47
	DCT3	4.26	2.42	1.39
Cases →		43v43/44v44		43v44
Score	**Poly3**	1.80	7.04	0.57
	DCT3	4.45	10.03	2.54

Note. Summary of the results for the different comparisons, after the sum of LRs (Geometrical Mean procedure). **The values shown are LRs and *not* LLRs.** MZ: Monozygotic twins; DZ: Dizygotic twins; B: Brothers; US: Unrelated Speakers; (I): intra-speaker tests; (O): inter-speaker tests. Divided columns are used for each pair member. Cases: xxvyy means speaker xx versus speaker yy. Blue is used for (I) and orange for (O). Shaded in grey are the values obtained for the B pair 23v24, strikingly high for a non-twin sibling pair (compare with the MZ intra-pair comparisons).

Table 12 shows the results of combining the scores of all the comparison systems under the first procedure (Naïve Bayes compensated after obtaining the Geometrical mean). This table classifies the results according to the type of comparison (intra-speaker or intra-pair) and according to the type of speaker (MZ, DZ, B or US) but the information for each specific speaker (e.g. 01v01) and each specific pair (e.g. 01v02) is also provided. In contrast, the results in Table 13 are pooled for all the speakers of the same type. This is done in order to better detect a possible trend in the direction of the LR values. That is to say, according to our research hypothesis, the following decreasing scale in the comparisons would be expected: IS (intra-speaker) > MZ > DZ > B > US. This would indicate that the acoustic parameters under study would be genetically influenced. The scaling in decreasing order from IS to US occurs always when considering both Poly3 and DCT3 except for the group of brothers, with higher values than DZ twins. It is worth noting that the standard deviation of this group is extremely high. Upon observation of Table 13, it is clear that the high mean value of this B group is basically due to a single pair: 23v24. Discarding this outlier, the values for the rest of non-twin siblings are lower than the values obtained by DZ twins, as expected. In next section, we provide some possible explanations for the unexpected values of the B pair 23v24.

The boxplot in Figure 17 shows the distribution of values according to the type of comparison (IS, MZ, DZ, B and US). Most values look normally distributed within each group with the notable exception of brothers, with an obvious skewed distribution. This is due, as explained in the previous paragraph, to a specific pair (23v24), with values located at the whisper of the bloxplot, which would overlap not with the average values of MZ comparisons but with the average values of IS comparisons, thus showing a striking similarity for this sibling pair. All in all, a scaling can be observed in the expected decreasing direction from IS to US: IS > MZ > DZ > B > US. Whereas considering the mean this trend was not observable, looking at the median (which is the value shown in boxplots), the expected trend appears more clearly.

Table 13

LR mean and standard deviation for the first type of combination (geometrical mean)

Type of comparison	Type of curve fitting	LR mean	LR std. dev.
IS (Intra-Speaker) comparisons	Poly3	3.81	2.69
	DCT3	3.96	2.54
MZ intra-pair comparisons	Poly 3	1.29	1.29
	DCT3	1.97	1.11
DZ intra-pair comparisons	Poly3	0.46	0.25
	DCT3	0.54	0.37
B intra-pair comparisons	Poly3	0.91	1.61
	DCT3	1.05	1.9
US inter-speaker comparisons	Poly3	0.44	0.54
	DCT3	0.42	0.42

Note. We show the mean and standard deviation values for the LRs for the first type of LR-combination (geometrical mean), according to the type of comparison (intra-speaker, intra-pair or inter-speaker) and speaker type: MZ, DZ, B or US. Results in red are against the expected scaling IS > MZ > DZ > B > US suggested in our research hypotheses. Red values (for B comparisons) are against-expectations higher than blue values (for DZ comparisons).

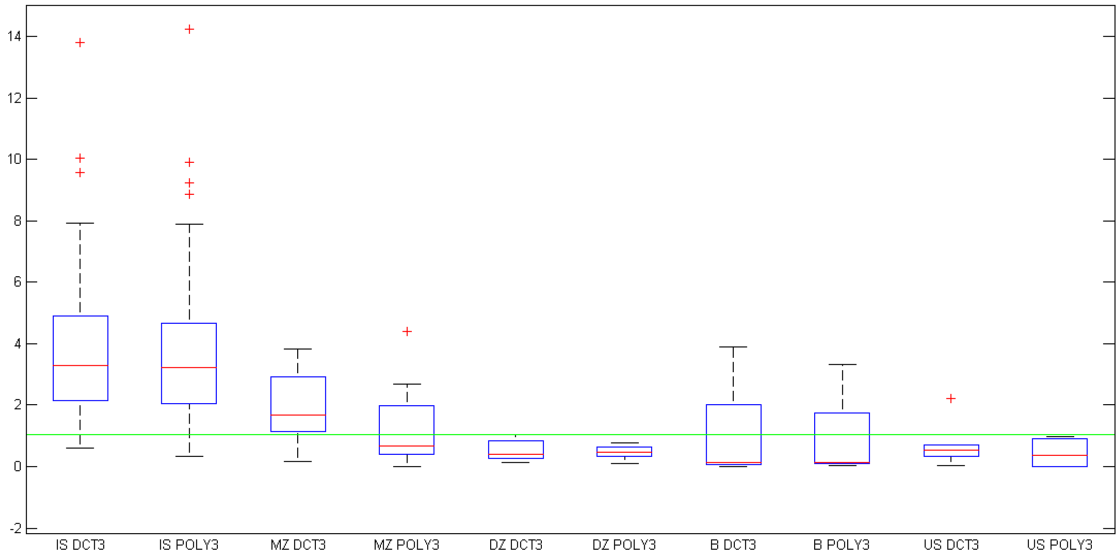


Figure 17. Boxplots showing the distribution of LR values (combined under the Geometrical mean procedure) per type of comparison: IS (intra-speaker comparisons), MZ (monozygotic intra-pair

comparisons), DZ (dyzigotic intra-pair comparisons), B (brother intra-pair comparisons) and US (unrelated-speaker intra-pair comparisons). The green line divides the graph in LRs > 1 and LRs < 1.

Table 14 shows the results of the score fusion using logistic regression. As in Table 12, the comparison results for each specific pair can be observed (orange). Besides, we include the results of every intra-speaker comparison (blue). Unlike Table 12, the results for this table are shown in logarithm (LLR). Therefore, it can be easily observed that the values are much higher when the fusion is conducted using logistic regression. For a comparison of the system performance using Geometrical Mean and Logistic Regression, see the C_{llr} plots in Figures 19 and 20.

Unlike what we did with the results of the LR-combination using Geometrical Mean, we do not show on this occasion the values pooled per speaker type, as doing so could lead to misleading results, as explained in the discussion section. We consider that the boxplots in Figure 18 depict more accurately how the values are distributed per speaker type. As it can be observed, the groups IS, MZ and DZ are normally distributed while the groups B and US show a markedly skewed distribution, indicating a strong heterogeneity for these speaker groups. In next section, we discuss these results in relation to the predominant influence of environmental over genetic factors affecting the parameters under study.

Table 14

Results for the different speaker comparisons (logistic regression)

		MZ (I)		**MZ(O)**	**DZ(I)**		**DZ(O)**	**B(I)**		**B(O)**	**US(I)**		**US(O)**
Cases →		01v01/02v02		01v02	13v13/14v14		13v14	21v21/22v22		21v22	25v25/26v26		25v26
Scores	**Poly3**	5.75	2.37	1.66	8.55	1.11	-6.28	9.34	9.34	-11.72	7.14	3.58	-4.35
	DCT3	6.58	-0.66	2.93	5.22	2.00	-1.21	12.57	6.60	-16.67	4.77	1.67	-5.57
Cases →		03v03/04v04		03v04	15v15/16v16		15v16	23v23/24v24		23v24	27v27/28v28		27v28
Scores	**Poly3**	5.97	2.44	-4.47	3.01	6.43	-0.95	6.01	3.21	5.85	4.34	6.07	**-0.34**
	DCT3	8.72	8.06	2.72	3.35	8.01	-4.79	6.99	8.14	7.25	6.52	9.80	**3.13**
Cases →		05v05/06v06		05v06	17v17/18v18		17v18	47v47/48v48		47v48	29v29/30v30		29v30
Scores	**Poly3**	4.76	0.66	-7.66	4.16	-8.14	-2.36	0.70	7.55	-14.32	4.28	5.35	-27.24
	DCT3	2.35	4.20	1.85	3.80	4.73	-8.66	-3.25	5.07	-20.59	-0.74	-5.13	-6.33

Cases →		07v07/08v08		07v08	19v19/20v20		19v20	49v49/50v50		49v50	31v31/32v32		31v32
Score	**Poly3**	9.25	5.77	-4.60	6.35	9.19	-6.50	1.40	3.87	-11.49	10.08	6.94	-24.74
	DCT3	8.90	4.43	-9.57	5.71	9.98	-0.71	4.19	0.36	-11.80	8.97	5.69	-13.92
Cases →		09v09/10v10		09v10	45v45/46v46		45v46				51v51/52v52		51v52
Score	**Poly3**	1.50	-1.07	-2.24	-7.05	3.25	-12.89				7.62	3.86	-2.54
	DCT3	1.37	-0.17	-1.25	0.11	6.43	-10.74				4.11	9.81	-0.26
Cases →		11v11/12v12		11v12							53v53/54v54		53v54
Score	**Poly3**	-0.70	1.70	-5.96							5.06	6.15	-7.11
	DCT3	4.56	3.67	0.33							1.40	3.44	-6.09
Cases →		33v33/34v34		33v34									
Score	**Poly3**	5.04	8.09	1.92									
	DCT3	5.83	7.49	8.91									
Cases →		35v35/36v36		35v36									
Score	**Poly3**	-3.62	6.07	4.08									
	DCT3	5.14	4.12	5.31									
Cases		37v37/38v38		37v38									
Score	**Poly3**	-13.69	3.86	4.99									
	DCT3	6.22	5.70	5.04									
Cases →		39v39/40v40		39v40									
Score	**Poly3**	1.83	13.82	-29.02									
	DCT3	-0.10	7.48	1.65									
Cases →		41v41/42v42		41v42									
Score	**Poly3**	3.15	2.26	1.18									
	DCT3	6.54	1.86	1.10									
Cases →		43v43/44v44		43v44									
Score	**Poly3**	2.38	7.72	0.46									
	DCT3	6.91	11.00	5.05									

Note. Summary of the results for the different comparisons, after the score fusion (logistic regression procedure). **In this case, the values shown are LLRs.** MZ: Monozygotic twins; DZ: Dizygotic twins; B: Brothers; US: Unrelated Speakers; (I): intra-speaker tests; (O): inter-speaker tests. Divided columns are used for each pair member. Cases: xxvyy means speaker xx versus speaker yy. Blue is used for (I) and orange for (O). Shaded in grey are the values obtained for the B pair 23v24, strikingly high for a non-twin sibling pair (compare with the MZ intra-pair comparisons).

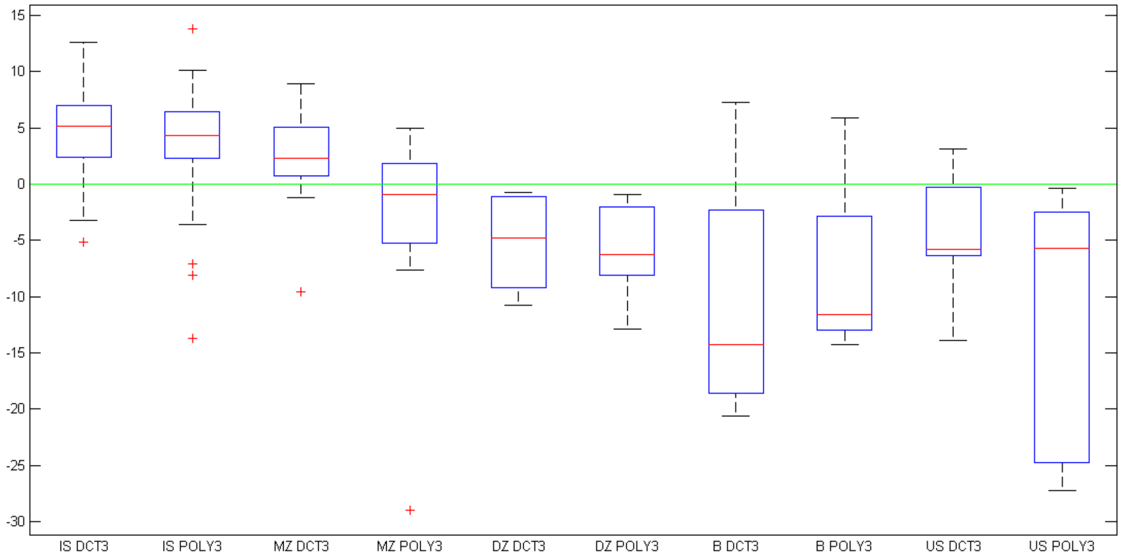


Figure 18. Boxplots showing the distribution of LLR values (combined under the Logistic Regression procedure) per type of comparison: IS (intra-speaker comparisons), MZ (monozygotic intra-pair comparisons), DZ (dyzigotic intra-pair comparisons), B (brother intra-pair comparisons) and US (unrelated-speaker intra-pair comparisons). The green line divides the graph in LLRs > 0 and LLRs < 0.

4.4.3. Accuracy assessment

We explained in Section 4.2.2 how both the cost function C_{llr} and Tippett plots allow us to evaluate the performance of our comparison system or systems. Both Figure 19 (DCT3) and Figure 20 (Poly3) show that the use of fusion techniques improves the performance of the forensic-comparison systems. While the C_{llr} values obtained when considering each VS independently may be as high as 0.78 after calibration ($C_{llr\ min}$), yielded by [ua] when considering DCT3, with either fusion procedure this value drops considerably: till 0.21 for the logistic-regression fusion or 0.15 for both the naïve combination and the geometric-mean combination. When considering the results for Poly3 (Figure 20), highest C_{llr} values are obtained when considering VS in isolation, being the highest /ua/ again with 0.82 in the $C_{llr\ min}$. The application of fusion techniques entails a drop in the function cost, with 0.30 obtained in the logistic-regression fusion, and 0.22 for both the naïve and the geometric-mean sum.

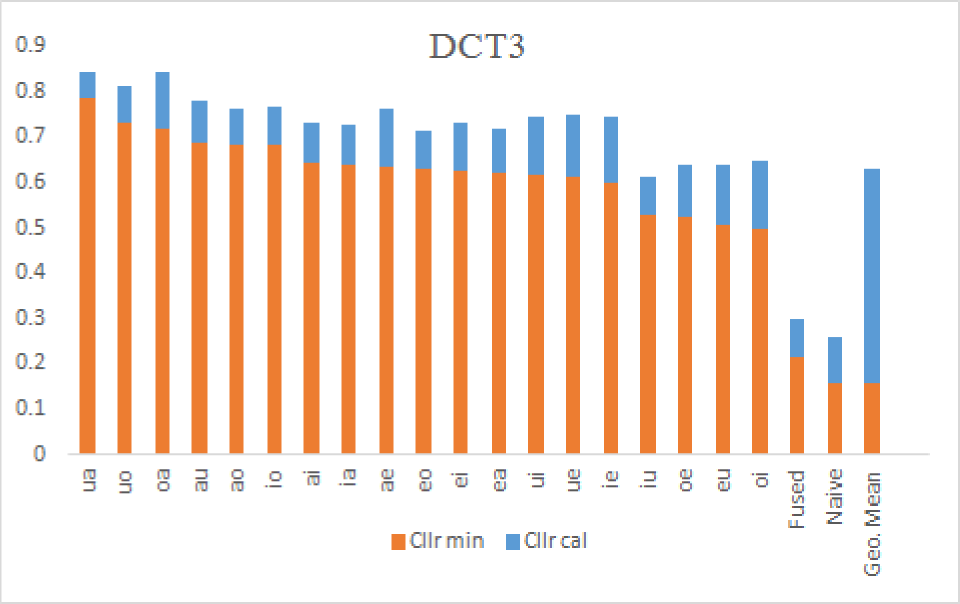


Figure 19. C_{llr} plot for the unfused and fused 19-diphthong discrimination (results for the curve fitting using third-order DCT function).

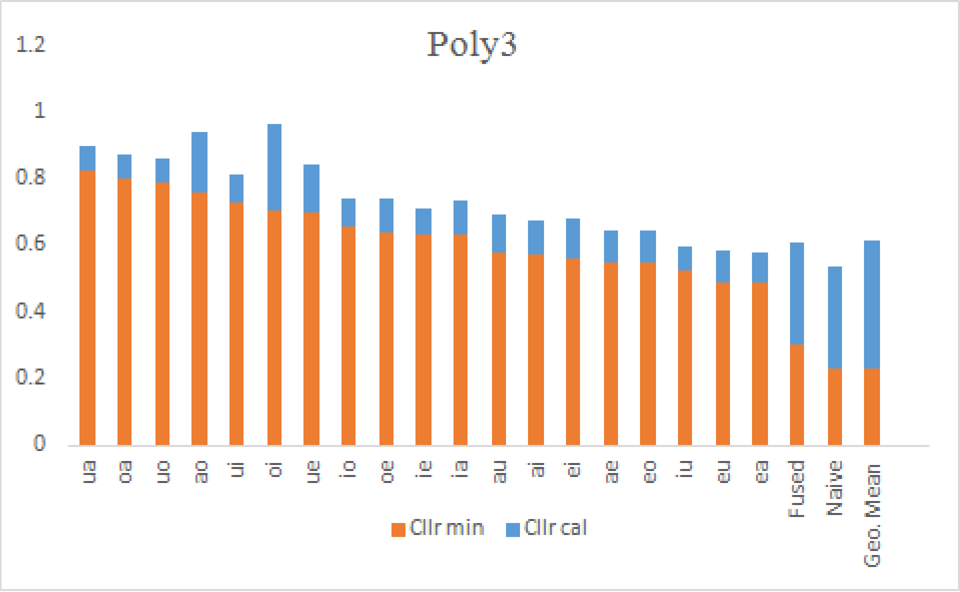


Figure 20.C_{llr} plot for the unfused and fused 19-diphthong discrimination (results for the curve fitting using third-order Poly3 function).

When evaluating the (logistic regression) fused scores of the system, we observe that they do not offer a better performance in terms of C_{llr} compared to the geometric mean of the naïve combination. Normally, this means that the training is not converging correctly. In order to confirm our suspicions, we tested a modified version of *FoCal*'s training function, which includes a scaling parameter called lambda that makes LRs lower, but it is associated with a more robust convergence of the training step. We will explain this aspect in the following section, as it is a further step that we have decided to take in relation to the discussion of results.

Finally, we present two Tippett plots in Figures 21 and 22, which show the LLR cumulative distribution according to the type of speaker comparison, for DCT3 and POLY3, correspondingly. As it was mentioned in the section devoted to methodology, Tippett plots are another way to measure the accuracy of a system. Unlike the C_{llr}, which is a single measure, Tippett plots allow us to see comparatively the system performance when considering US, B, DZ or MZ (as non-targets) and IS comparisons (as targets).

Figures 21 and 22 reveal that the system performs quite well when considering a standard forensic scenario with unrelated speakers as non-targets and same speakers as targets. Although different values are observed depending on the type of parametric curve, in general US comparisons (black line) presents almost no against-the-hypothesis cases (errors). In the case of targets (red line) more errors are observed in Figure 22, corresponding to POLY3.

If we compare the different performance of non-target comparisons, first we note the uneven jagged aspect of all the lines, as a clear indication of the small number of comparisons in each speaker category. This probably points to the inadequacy of this type of graph for comparing system performance depending on the type of speakers considered. The expected genetic effect, observable through the four different speaker types (MZ, DZ, B and US), was already discussed in Section 4.4.2.

The Tippett plot in Figure 23 presents the results of two types of comparisons: same-speaker comparision in the blue line, and different-speaker comparisons in the red line. On this ocassion, all the speakers have been considered together, i.e. not distinguishing per speaker type, as in the previous Tippett plots. As the different-speaker comparison includes more data now, the non-target curve is smoother. Besides, the system performance looks better, as small errors are observed. However, these results are deceptive. As it has been explained, we cannot consider that the against-the-hypothesis LRs do not exist but they are somehow hidden.

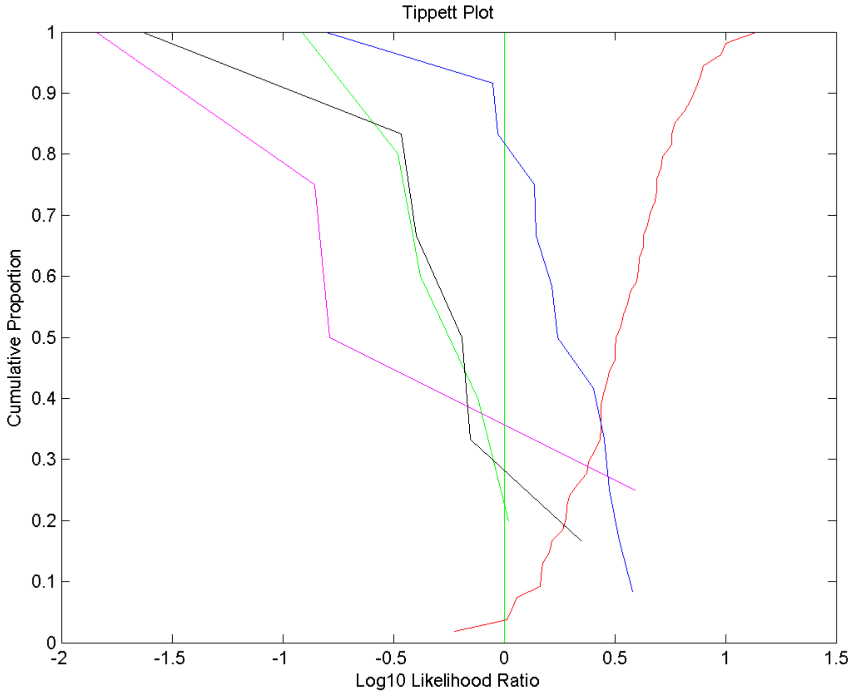


Figure 21. Tippett plot showing the cumulative distribution of LLRs using DCT3 and geometric mean fusion. Red is used for same-speaker comparisons and black for different-speaker (US) comparisons. The other three lines rising to the left represent one of the following IP comparisons: blue is for MZs, green for DZs and magenta for B.

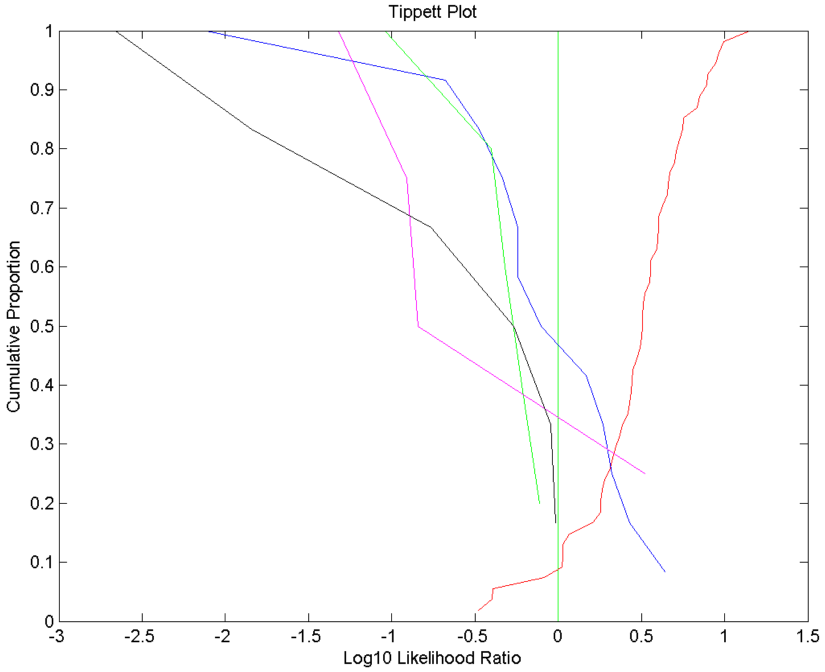


Figure 22. Tippett plot showing the cumulative distribution of LLRs using POLY3 and geometric mean fusion. Red is used for same-speaker comparisons and black for different-speaker (US) comparisons. The other three lines rising to the left represent one of the following IP comparisons: blue is for MZs, green for DZs and magenta for B.

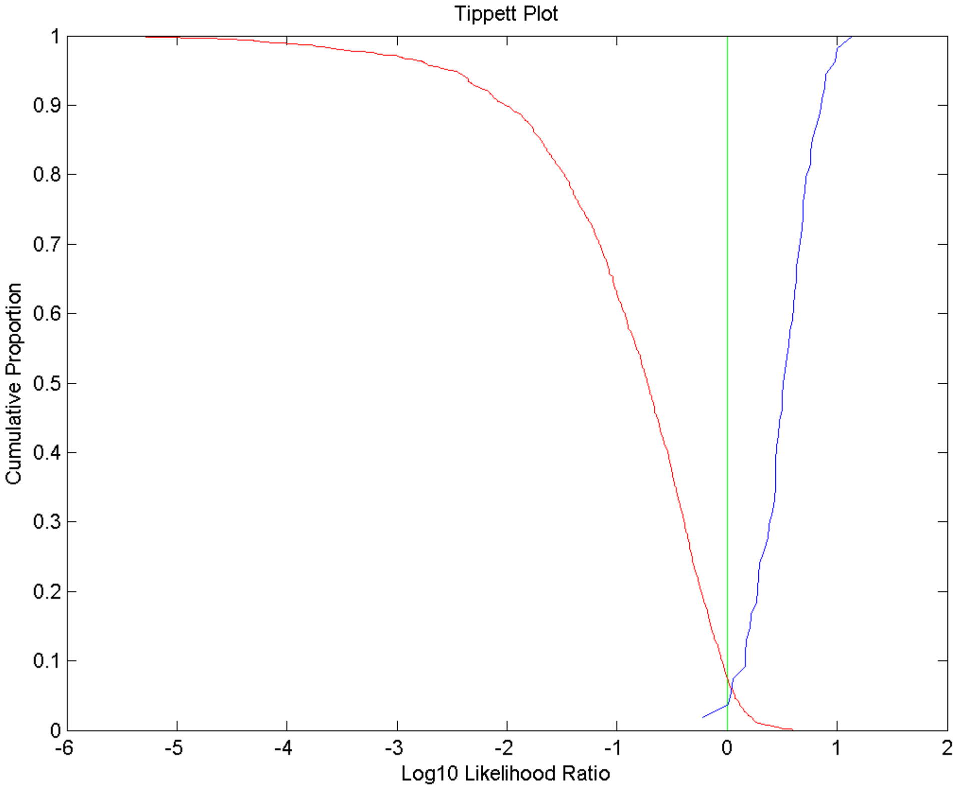


Figure 23. Tippett plot for the classical FSC scenario: the blue line represents same-speaker comparisons while the red line represents different-speaker comparisons (considering all the speaker types together).

4.5. Discussion

4.5.1. Curve fitting: best correlation values

From the results of the goodness-of-fit calculation, carried out by means of correlation, we draw the following conclusions. On the one hand, third-order functions fitted the trajectories better than the second-order functions. The fact that this trend is observed for all the VS and irrespective of the formant considered seems logical, as the presence of more coefficients for the curve fitting implies a more detailed or accurate approximation.

On the other hand, we found a better goodness of fit in the F2 values as compared with the F3, which could be due to the fact that F2 is more constrained by the linguistic system while the F3 is traditionally considered more speaker-specific (e.g. Battaner et al., 2003). This would imply that the variation found from one speaker to another makes the fitting of the trajectories for this formant more difficult than for F2. Nevertheless, the comparison scores yielded by both formants are later combined in the MVKD formula with the aim of making a forensic-comparison system more powerful.

In relation to the question of whether some VS were better fitted than others, we do not find clear trends. Results vary depending on the formant considered (F2 or F3) and on the parametric function (second- or third- degree DCT or polynomial). Since the third-degree functions were found to better correlate the formant trajectories than the second-degree functions, in the section devoted to the results we showed only the R values obtained when using third-degree functions. The most notable result was that the VS /uo/ always obtained the lowest R values, as compared with the other VS. This happened irrespective of the type of parametric function or degree. For example, in the parametric representation using third-degree DCT functions, the correlation value was 0.875. This could be due to the fact that this sequence is made up of two back vowels –known to be susceptible to formant detection errors, as F1 and F2 in these vowels are very close together, especially in /u/. Yet, further studies would be necessary in order to confirm that this is the cause of the low correlation values for this VS, in comparison with the others. In contrast, we could not spot a unique VS which was remarkably better correlated than the others. Rather, there were several VS with relatively high R values: /ai/, /ie/, /ue/ and /oe/ got values close to 1. The heterogeneity of this set of VS does not allow us to conclude that certain VS (e.g. rising diphthongs) are better correlated than others by means of the parametric functions used.

4.5.2. Combination/Fusion techniques: comparing MZ, DZ, B and US tests

Three combination techniques were proposed to fuse the scores (of each of the 19 systems, one per VS) obtained after using the MVKD formula: 1) naïve Bayes; 2) geometrical mean; and 3) logistic regression. Since the second one is just an amelioration of the first one, the results were shown per speaker comparison only for the second and the third technique. Some previous studies have compared both types of combination procedures and have not found that the procedure which implies calibration (logistic regression) yield much better results than the technique which assumes statistical independence (like the geometrical mean procedure). Indeed both of the studies which we are referring to (González-Rodríguez et al., 2007; Gil-Gil, 2009) combine the scores from systems based on different diphthongs, as in our study. In contrast, in their study of the formant trajectories of /o/, Morrison and Kinoshita (2008) concluded that substantial improvement was found in system performance when the output of MVKD was calibrated using logistic regression. By simply comparing Tables 12 and 14 and Figures 17 and 18, there does not seem to be large differences in the different-speaker comparisons using one fusion type or the other. Nevertheless, this can be better observed in the C_{llr} plots (Figures 19 and 20), where the accuracy of the forensic systems was tested. We will discuss these results in 4.5.3.

From the results shown in Tables 12 – 14, we can conclude that the expected decreasing scale *IS (intra-speaker) > MZ > DZ > B > US* in the comparison values exist. This happens irrespective of the type of score combination (geometric mean or logistic regression) and regardless of the type of function considered (Poly3 or DCT3), except for the group of brothers, with an average value higher than DZ twins. Due to the relatively small size of the brother group (four non-twin sibling pairs participated) it was possible to detect in Tables 12 and 14 the origin of the discordant mean value of this group. It turned out that the (intra-pair) comparison of brother pair 23v24 yielded very high LR values: 3.32 (Poly3) and 3.90 (DCT3) within the geometric-mean fusion approach, and also very high LLR values within the logistic-regression fusion approach: 5.85 (Poly 3) and 7.25 (DCT3). These values are very close to the most similar MZ pairs, and even close to typical intra-speaker comparison values. The fact that only a pair of non-twin siblings shows such high values, in comparison with the rest of brothers, makes the standard deviation of this group very large. This striking value makes the distribution of the brother group very skewed, as can be clearly seen in the boxplots (Figure 17 and 18). Although the first one shows the results in LR and the other in LLRs, it is worth mentioning how the dividing green line in the graph leaves the same groups above and below it.

Since a typical forensic scenario would uniquely contemplate IS comparisons and US comparison, we will begin describing the results for these two groups, and then explain what is observed for the groups MZ, DZ and B.

The most interesting finding is that either considering the geometric-mean combination (Figure 17) or the logistic-regression fusion (Figure 18), the comparison for the groups IS and US, which would correspond to targets (same-speaker comparisons) and to non-targets (different-speaker comparisons) yield consistent-with-fact (L)LR values, i.e. in agreement with reality. Therefore, most IS comparisons fall above the green line and most US comparisons fall below the green line. This can be observed both in Figure 17 and in Figure 18. Since in the former values are not logarithmic, LRs larger than 1 support the Hp (H_{ss}) while LRs below 1 indicate support for the competing hypothesis: Hd (H_{ds}). In the case of Figure 18, values are logarithmic, so LLRs above 0 point towards the Hp while the opposite happens with LLRs below 0, which would support the Hd. If we observe the two boxplots further to the left (IS comparisons: DCT3 and Poly3) and the two boxplots further to the right (US comparisons: DCT3 and Poly3), their distribution shows almost no points in the whiskers or outliers supporting contrary-to-fact hypotheses. This would indicate that in the most typical forensic situation, where comparisons are just made between same speakers (intra-speaker comparisons) and between unrelated speakers (inter-speaker comparisons), a forensic system based on the score combination/fusion of 19 Spanish VS would perform relatively well. Yet, in next section the output accuracy of the system will be better discussed in relation to the measure C_{llr} and the Tippett plots.

In order to see what happens when testing the performance of the system with related speakers, we have to look at the groups MZ, DZ and B. In general, it has to be said that values around 0 or 1 (depending on whether they are LRs or LLRs) do not indicate a strong support for neither of the competing hypotheses. For instance, in the case of MZ comparisons, the median values are close to the green line. Therefore, it would be difficult for the forensic-comparison system to decide whether the two speech samples come from the same or from different speakers. This closeness to the green line happens in both the DCT3 boxplot and the Poly3 boxplot, even though in one case (DCT3) the median falls above the line and in the other case (Poly3) it falls below it. The fact that the system cannot tip the balance in favor of one hypothesis or the other when comparing MZ twins is in agreement with the fact that these speakers are very similar. Depending on the specific twin pair considered, higher or lower (L)LRs are yielded by the system. This would indicate that the parameters considered are not uniquely and completely genetically related. On the contrary, non-genetic aspects (like learned habits) should be exerting a strong influence. In other words, the fact that some MZ twin pairs get high (L)LRs while other get lower values suggest that factors like the ones considered in the questionnaire should be taken into account to explain the variation. We refer to factors such as degree of relationship closeness, shared /non-shared leisure activities, shared / non-shared group of friends, time spent together, and so on.

Upon observation of the DZ distribution in boxplots (Figures 17 and 18), it seems that these speakers are very different from MZ twins. Even though this could be due to the small sample size of this group (5 DZ twin pairs vs. 12 MZ twin pairs), this difference points to a certain genetic influence of the parameters considered. Considering the Equal Environment Assumption [see Chapter 2], any excess of likeness found in MZ twins which is not present in DZ twins can only be attributed to genetic causes. In fact, it should not be forgotten that the VS formant trajectories are supposed to be related to the vocal tract anatomy (see Section 4.1.2), so this result seems reasonable.

Finally, when considering the B group, we have already marked how heterogeneous this group is. While the median clearly stays further below the median of DZ (or below any other group, indeed), which can be more clearly seen in Figure 18, there is one specific pair (23v24) with really high comparison values. This could only happen if the parameters studied are, to a high degree, environmentally influenced. With this we mean that the phonetic realization of the VS formant trajectories must be strongly subject to the specific and voluntary implementation of the acoustic target by the speaker, naturally within his anatomical constraints

The different behavior of the B pair 23v24 in comparison with the rest of speakers in his group can be explained by several factors. We have looked in detail at the responses given by these speakers in the questionnaire gathered at the first speaking session, and the following remarkable aspects are to be noted:

- Speakers 23 and 24 are the only ones among the B speakers who answer "Very often" to the question "How often do people confuse your voice with that of your brother?"
- They are the only ones in his group answering "Absolutely not. I think that we speak the same way" to the question "Do you consider that your voice/manner of speaking is very different from your brother's?"
- They are the only ones among the B speakers who share leisure activities.
- In comparison with the rest of non-twin brothers, speaker 23 and 24 see each other quite often (at least once a week) and they also talk to each other quite often (between twice and three times per week).
- In the open question "Mention a few aspects in which you think (or people have commented about how) your voice is different/similar from your brother's", they answer that many people have mentioned their same way of laughing, their similar intonation and their use of similar expressions. Besides, one of them mentions the anecdote of having been once recognized as brothers by certain person solely on the basis of their voice, without being both of them ever together before that person and without this person having beforehand knowledge of their family kinship.
- Finally, in the question related to their degree of closeness, from 1 to 5 (being 1 "not very close" and 5 "very close"), they gave a 4.5 points on average.

All of the above-mentioned responses given in their questionnaires could be indicative of the nurture factors outweighing the genetic ones for explaining the strikingly high LRs values obtained in their comparison. Furthermore, in a perceptual study in which these same speakers participated (San Segundo, 2013b), the laughter of these brothers was actually found to be very similar to each other. The question of whether this was due to a similar vocal tract or to imitated behavior was not tackled. All in all, we can conclude that this is a case of a non-twin sibling pair having a closee relationship than many of the MZ or DZ twins also participating in this study, which would have clearly exerted certain "intra-sibling mimetism" in the speech of these brothers. In Section 2.2 we already referred to the two opposite directions which sibling influences can adopt: towards accommodating or towards distancing their speech behavior; in this case, the accommodating direction may have been reinforced by the type of speaking task from which the parameters were extracted. As it was an information exchange between conversational partners, this may have specially triggered the convergence in their speech habits, possibly affecting their similar acoustic output for the VS. Indeed, there is a fast-growing research line investigating convergence and imitation patterns in speech occurring between speakers in the course of conversational interactions (see e.g. Pickering & Garrod, 2004; Pardo, 2006; Truong & Trouvain, 2012), although their methods may not be widely known and fully applied to Forensic Phonetics. Some studies focus specially on the convergence of phonetic features in close acquaintances (Kalmanovitch, 2012), or college roommates (Pardo et al., 2012; Coupland, 1984). The methodological approaches of this type of investigations are all indebted to the theory of accommodation (Giles, Coupland & Coupland, 1970), which had already emerged in the early 1970s. In light of its postulates, it would be extremely interesting to revisit the study of twins' vice similarities and differences (see Section 7.3).

4.5.3. Accuracy assessment

The cost function C_{llr} allowed us to evaluate the performance of our comparison system (fusing the scores of all the VS) or systems (considering the VS separately). Both Figure 19 (results for DCT3) and Figure 20 (results for Poly3) showed that the use of fusion techniques (either geometrical mean or logistic-regression fusion) improved the system performance. The worst individual (i.e. unfused) system was that of /ua/ with a C_{llr} of 0.78 after calibration (considering DCT3) while the best individual (unfused) system would be that of /oi/ (0.49). That is to say, a notable improvement was observed when fusing all the VS together. However, results did not vary considerably when using one procedure or another for score combination. This is in line with studies such as González-Rodríguez et al. (2007) or Gil-Gil (2009). This latter tested both the naïve Bayes procedure and logistic-regression fusion methods, and did not find that the one assuming statistical independence was much worse than the logistic-regression technique.

As introduced in previous section, a regularization factor (lambda) was used with the aim of evaluating the convergence of the logistic regression fusion. As it did not seem clear why the logistic-regression fusion yielded worse results than the other methods, we hypothesized that this could be due to a lack of training data (Daniel Ramos, personal communication), whose origin would be in the small database used The goal of lambda is therefore mitigating the effect of the lack of data for the training of the logistic regression, although at the cost of yielding LRs which would be more moderate (underconfident), i.e. less strong as they should, and therefore with a higher calibration error.

After evaluating the C_{llr} values obtained with the above-mentioned modified training function, we observed that the lower the lambda, the lower the C_{llr}^{min}. This helped us confirm that our logistic-regression model was diverging because our database was lacking a larger set of data for the training, and even if we did an “honest” training, it was not yielding accurate results, so the validity of this kind of approach (i.e. logistic regression) needs to be questioned in this kind of situations where a small database is being used (see conclusions in Section 4.6).

Furthermore, it should be noted that a modification of the training function provided by *FoCal* has been made with the aim of using a more exigent threshold. Despite this, the regression model did not converge either. In other words, the results presented in Figures 19 and 20 using the standard function in *FoCal* have been obtained using a convergence threshold of 10^{-12} instead of 10^{-5}, which is the default value in *FoCal*. The meaning of this threshold is the following. After finishing each of the iterations, a calculation was made of the difference between the new weights obtained in such iteration and the weights which had been obtained in the previous iteration. If the difference is smaller than the established threshold, the training phase is considered finished with the last obtained weights. This reduction in the convergence threshold implies that for our study we have been more exigent in relation to convergence. Despite this, convergence of the model has not been attained optimally, as it has been shown with the use of the correction factor lambda.

Finally, the magnitude of the LRs obtained deserves some discussion. All in all, we can conclude that the geometrical mean procedure for LR combination is the best of the three methods proposed, since the LRs within this approach are not as exaggeratedly large as in the case of the naïve and the logistic regression method. While LR values reaching 10^3 and 10^{-3} can be considered normal is FSC, it seems that numbers exceeding those values cannot be justified in this field, where discrimination is never perfect. Likewise, in DNA comparison, LRs do not usually exceed 10^{20}. It seems that nowadays the relation between the magnitude of LRs and their discrimination power is still an open research issue, as studies like Ramos-Castro and González-Rodríguez (2013) show.

4.6. Conclusions

The main objective of this analysis has been testing the forensic validity of formant trajectories extracted from Spanish vocalic sequences. For that purpose, the different analyses carried out in this chapter have focused on three hypotheses:

H1: Formant trajectories in the vocalic sequences under study will be somehow genetically influenced: higher similarity values will be found in MZ twins than in DZ twins, in siblings or in the reference population.

H2: A forensic-comparison system based on all the VS fused together will yield better performance than individual systems each based on a single VS.

H3: According to previous preliminary studies (San Segundo, 2010a), identification results will not be much better with one parameter curve fitting method as compared with the other.

To sum up, all the three hypotheses have been corroborated, with different degrees. The following comments are required in relation to the research questions formulated.

First hypothesis

The parameters studied in this first approach to the voice of twins and non-twin siblings have been the formant trajectories of 19 Spanish vocalic sequences. In order to investigate their forensic-phonetic relevance (i.e. how useful they are to discriminate between speakers in a forensic setting), we have established that the comparison of results found in 4 speaker types (MZ, DZ, B and US) would indicate whether these parameters are genetically influenced. The expected decreasing scaling to be found in the results of these speakers' intra-pair comparisons is: MZ > DZ > B > US. As the results are offered in LRs or LLRs, larger values are indicative of higher similarity between the speakers. Indeed, this is in agreement with the five general hypotheses established in Chapter 2 and repeated in the three chapters devoted to the analysis of data (4, 5 and 6). Therefore, we consider that a forensic-phonetic parameter will be more robust the more genetically influenced it is, since what is encoded in our genes we cannot change[93]; as compared with parameters more environmentally influenced (i.e. more subject to learned behavior or voluntary/contextual variation). This idea underlies all the investigation and also applies for the rest of parameters considered.

Our hypothesis that the parameters considered would be genetically related is only partially corroborated. We observe the hypothesized order MZ > DZ > B > US when considering the groups' median of the IP comparisons. Taking into account the mean impedes the corroboration of this hypothesis, due uniquely to a contrary-to-the-hypothesis sibling (B) pair. Several factors have been outlined in the results' discussion to explain why the B pair obtained so high values. From the results discussed not only for this B pair but also for the MZ, DZ, US as well as the IS comparisons, we conclude that in the analysis of the VS formant trajectories there is an important interplay between nature and nurture. In other words, if we were to classify the

[93] By any means this implies a deterministic view. Epigenetics has taught us that the alteration of the expression of specific genes is possible (see Section 2.2). However, this does not change the premise from which our investigation departs, i.e. that the more genetically influenced a parameter is, the more robust it will be for forensic purposes.

results obtained in one or another opposing directions (towards a genetic influence or towards an environmental/behavioral influence), we would obtain a table such as this:

Table 15

Interplay of nature (genetic) and nurture (environmental) influences for the results obtained

Results suggesting genetic influence	Results suggesting environmental influence
- The expected decreasing scale MZ>DZ>B>US is observed in the IP comparisons when considering the median of the speaker groups. According to the “classical twin method”, simply the scale MZ>DZ points to the greatest importance of genetic influence as compared with the environmental influence.	- The IP comparison of a non-twin sibling pair can reach as high LR values as those found for a MZ IP-comparison. This could not happen if the voice parameters considered were uniquely genetically influenced. The questionnaire responses of this pair point towards a ‘sibling mimetism’, synonym for (voluntary or unsconscious) convergence of the environmental sort.

Second hypothesis

Correlation was expected given that the measurements of the VS (per speaker) come from multiple sections in the same recording. For that reason, besides the Naïve Bayes combination, the systems corresponding to each individual VS were fused using a logistic-regression model. According to the results, it has been proved that a system based on all the VS fused together yields better performance than the systems based on a single VS separately. However, the question of whether the techniques assuming statistical independence (naïve Bayes and geometrical mean) perform better than the logistic regression procedure could not be answered. As mentioned in the discussion, our small database entailed a divergence of the logistic regression model. Thus, the results of this investigation do not allow confirming that logistic regression performs better that either of the other two techniques.

The following question arises from the above-described aspects: is the C_{llr} is a good measure of the accuracy for our specific comparison systems? Traditionally, the performance of a forensic comparison system has been assessed by means of the cost function C_{llr}. Although it has the advantage of showing how well a system is behaving, it has the drawback that it can hide LR calculation errors. In our case, we were obtaining extremely high scores using the MVKD formula for certain speaker comparisons and certain vocalic sequences, which affected the training step and finally yielded extraordinary large LRs for some comparisons. These kind of erroneous LRs were hidden under the broad C_{llr} figure, so a detailed overview of the scores is advised along with the presentation of the C_{llr} values. In any case, it seems that C_{llr} should not be the only measure of the performance of a system when dealing with a database of limited size.

All in all, we can conclude that in future studies, the use of a larger database would enable the inclusion of additional steps in the logistic regression model, namely the following three steps: training, tuning and test. The new step is the so-called tuning, where a separate population is used allowing to get feedback about the proper weights to use in the subsequent steps. This three-step procedure would imply a double cross-validation, which is computationally more demanding.

Simpler fusion procedures, like the ones assuming statistical independence, present the advantage of not needing a training phase. This links to a crucial issue in FSC. The need for large voice databases in order to be able to computationally fulfill certain research objectives cannot always be attained. Especially in twins' studies, it is remarkably difficult to gather a large number of twins' voices. For this reason, a fusion option which does not require training, like the geometric mean, has proved to be a better option. Besides, within this fusion approach the LRs are more conservative, compensating the overconfidence of the naïve approach.

Third hypothesis

We had hypothesized that the identification results would not be much better with one parameter curve fitting method as compared with the other. On the one hand, we can conclude that third-degree functions are the best way to fit the original formant-trajectory curves, irrespective of whether they are polynomial or DCT functions. This has been tested in 19 VS and the same trend has been observed for all of them. Likewise, the same result has been obtained irrespective of whether we considered F2 or F3, which are formants unlikely to be adversely affected by the telephone filtering characteristics.

In relation to the identification results, the hypothesis that there would not be strong differences in the system performance between using one type of parametric curve or the other was corroborated in view of Figures 19 and 20, which show that the C_{llr} values obtained using DCT3 or Poly3 fall in the same range. If we compare our results with previous studies, we can see how our findings agree with those of Morrison and Kinoshita (2008), in the sense that they found no differences the three and two-formant analyses –the best performance was achieved using third-degree polynomials–, although they equalized the duration of formant trajectories by means of linear-Hertz scaling. In the case of McDougall (2006), she found that the cubic-based approach did not always outperform the quadratic-based one.

5. GLOTTAL SOURCE ANALYSIS

5.1. Objectives and justification

We will set our research objectives and corresponding hypotheses for the glottal source analysis described in this chapter. Some studies related to this issue will be reviewed, which will serve as a state-of-the-art background and justification for the kind of analyses that will be carried out.

5.1.1. Objectives

The main objective of this analysis is testing the discriminatory power of a series of glottal features extracted from Spanish vowel fillers. This general objective can be split into the following specific or secondary objectives:

O1: Testing whether there is higher intra-pair similarity for this kind of parameters in MZ twins than in other speaker comparisons.

O2: Testing whether some glottal parameters yield better identification results than others.

For the above-mentioned objectives, we support the following hypotheses:

H1: Glottal parameters are genetically influenced: higher similarity values will be found in MZ twins than in DZ twins, in siblings or in the reference population. This is in agreement with the 5 basic hypotheses established for this thesis (Table 3; see Chapter 2).

H2: The biomechanical estimates of the glottal waveform will be especially speaker-specific, according to preliminary studies (San Segundo, 2012).

5.1.2. Justification

In this chapter we will describe the analysis of several parameters related to the glottal source that we have carried out in the voices recorded for this thesis. Before getting into the details of this kind of laryngeal examination, we will briefly review in this section the main studies on this topic. This will serve to show the main current trends followed in forensic studies focusing on this particular voice aspect. For a review of laryngeal studies specifically undertaken from a twin-related perspective, see Chapter 2.

Forensic investigations have traditionally relied on the information found in the vocal folds for speaker identification. Indeed, the chief forensic-phonetic overview works by most internationally relevant forensic phoneticians (e.g. Nolan, 1983; Künzel 1987; Baldwin & French, 1990; Hollien, 1990; Rose, 2002) pay attention to this aspect, dedicating several pages to the description of potential speaker-specific glottal parameters. They usually distinguish between supralaryngeal voice quality aspects and laryngeal[94] voice-quality aspects; logically it is in this last field where some relevant voice-source (i.e glottal) parameters are described. Leaving aside auditory-perceptual experiments like Köster and Köster (2004) and acoustic approaches to the study of specific phonation types like creaky, whisper or falsetto[95] (Moosmüller, 2001; Evans & Foulkes, 2009), we will focus our review in glottal-feature studies from a forensic point of view.

The speaker discriminatory potential of classical distortion parameters like jitter and shimmer[96] have been traditionally suggested (Künzel & Köster, 1992; Wagner, 1995). Later studies (Jessen, 1997) include further laryngeal parameters: H1*-H2*, H1*-A1 and B1[97]. In Table 16, taken from the study of Jessen, he refers to the underlying physiology for each acoustic parameter considered. For this literature review we are only interested in the laryngeal parameters, but it is worth noting that Jessen (1997) considers a three-folded division of voice quality in: supralaryngeal, laryngeal and sublaryngeal. For this distinction he follows Laver (1980) who distinguished voice quality in a narrow sense (only laryngeal settings) and in a broad sense (comprising also supralaryngeal settings). The further distinction in *sublaryngeal voice quality* made by Jessen (1997) is influenced by Fant (1968), Sluijter (1995) and Sluijter et al. (1995). With this term he refers to the contribution of pulmonic force to the voice source.[98] Jessen (1997) found that all of the investigated parameters carry speaker-specific information; among the laryngeal voice-quality features, the one which ranks higher (i.e. the most speaker-specific) being H1* - H2*.

Zheng (2005) offers a different perspective to the contribution of voice source features to speaker recognition. His thesis explores the usefulness of the combined use of vocal source and

[94] Jessen (1997) also considers sublaryngeal voice quality, as we will explain below.

[95] These studies are undoubtedly useful for forensic aims since they explore some important glottal features related to certain phonation types frequently used as voice disguise by criminals. However, the review of this kind of studies lies beyond our purposes for this thesis.

[96] The acoustic characteristics of these parameters will be described in Section 5.3.

[97] Jessen (1997:91) explains that while some of the parameters were taken directly as dependent variables, others were subjected to further calculations: “These calculations involved two steps. In the first step the events H1, H2 (first and second harmonics), A2, and A3 (second and third formants) were subjected to specific calculations that aimed to separate the laryngeal and sublaryngeal influence of interest from possible supralaryngeal influence” (Jessen, 1997:91). Jessen specifies that the calculations that he makes are proposed by “Stevens and students” following Fant (1960) and that the formulae are adopted from Sluijter (1995:108). For more information about this, see Jessen (1997:91). Regarding the second step of the calculations that he mentions, Jessen (1997:91) states: “In the second step, the calculated parameters H2*, A2*, and A3*, as well as A1, are substracted from the calculated parameter H1* to obtain the parameters H1* - H2*, H1* - A1, H1* - A2*, and H1* - A3*.

[98] “[…] not all characteristics of the voice source are necessarily due to larynx-internal mechanisms. Certain characteristics of the voice source that can be measured acoustically are the result of differences in subglottal pressure, which in turn in primarily due to pulmonic, rather than larynx-internal activity” (Jessen, 1997:85).

vocal tract information in order to improve a speaker recognition system, typically employing vocal-tract-related acoustic features, such as Mel-frequency cepstral coefficients (MFCC). The novel approach of this study lies in the representation of speaker-specific vocal source characteristics: “the linear predictive (LP) residual signal is adopted as a good representative of the vocal source excitation, in which the speaker-specific information resides on both time and frequency domains” (Zheng, 2005: iv). This study distinguishes two parts: one devoted to speaker identification and another one related to speaker verification. For speaker identification, the developed system achieves a relative improvement of 46.8%, compared with MFCC. Results also show that the combined use of source-tract information can improve the robustness of speaker recognition systems in mismatched conditions[99]: “[…] relative improvements of 15.3% […] have been achieved for speaker identification” (Zheng, 2005: v).

Table 16

Voice quality type and corresponding acoustic parameters and underlying physiology

Voice quality type	Acoustic parameter	Underlying physiology
Supralaryngeal	F1	Shape of vocal tract
	F2	
	F3	
Laryngeal	H1* - H2*	Open Quotient
	H1* - A1	Degree of glottal opening
	B1	Glottal leakage
Sublaryngeal	H1* - A2*,	Skewness of glottal pulse,
	H1* - A3*	Duration of closing portion

Note. Classification adapted from Jessen (1997). Retrieved from Table 2 of Jessen (1997: 92).

In Farrús and Ejarque (2007) and Farrús (2008) a method is proposed to improve a prosodic and voice spectral verification system by introducing new features based on jitter and shimmer measurements, extracted with *Praat* voice analysis software[100]: *jitter (absolute), jitter (relative), jitter (rap), jitter (ppq5), shimmer (dB), shimmer (relative), shimmer (apq3), shimmer (apq5) and shimmer (apq11)*. The results of these investigations show that both prosodic and spectral baselines, especially the prosodic one, are clearly improved when jitter and shimmer

[99] The mismatches are caused by (1) intra-speaker variation of speaking style, and by (2) acoustic environment variation (Zheng, 2005: 6).

[100] For a description of these jitter and shimmer measurements, see Farrús (2008: 69-70).

features are added. Besides, the absolute measurements of both features seem to be more discriminant than their relative measurements.

Finally, a series of publications have appeared since the early 2000s which have implied a great step forward to the examination of glottal-source parameters for forensic purposes, especially to test the speaker-discrimination relevance of novel features, such as the biomechanical estimates of the glottal waveform. Based on previous voice-pathology investigations, like Gómez-Vilda et al. (2007), other forensic-related publications by the same authors have appeared to show that their voice-analysis methodology, based in the decoupling of vocal tract from glottal source estimates, is useful for speaker identification. In Gómez-Vilda et al. (2008), they describe the advantages of splitting vocal from glottal information, as it opens the possibility of independently studying vocal and glottal components:

> It is well known that the vocal tract transfer function expressed by its resonances (formants) is of great interest for the biometrical characterization of the speaker […]. The glottal source descriptions in the time or frequency domain are well known for their capability of expressing speech pathology […]. But as both correlates, vocal and glottal, appear intermingled in the acoustic recording of speech, techniques relying on the analysis of the acoustic record of full voice resent from this juxtaposition and blurring, and become less efficient. A good approach will be to split voice into vocal tract and glottal source information for further analysis with current automatic pattern recognition engines. (Gómez-Vilda et al., 2008: 5-6)

In this study, they also describe the different parameters obtained from two neighbor glottal source cycles. Since most of them are used for this thesis dissertation, they will be explained in detail in Section 5.3. Regarding the speaker identification experiment carried out in Gómez-Vilda et al. (2008), the results showed that the proposed methodology improved the best method based on raw speech in a 50% over Equal Error Rates. The performance of the methodology was tested in a database of 240 speakers (Moreno et al., 1993), including a wide representation of their glottal characteristics.

Gómez-Vilda et al. (2009) describe in depth a methodology designed for effective pathology detection which could also be used for the biometric characterization of speakers (Gómez-Vilda et al., 2009: 765). As they indicate, “taking relations H1–H2, A1–A3, H1–A1 and H1–A3 as good correlates to pathology availed by other researchers’ results, a generalized signature is proposed on singularities detected on the Glottal Source spectral envelope (peaks and troughs)” (Gómez-Vilda et al., 2009: 760). They further elaborate on this:

> This generalization is based on the biomechanical dynamics of the vocal folds found on the Glottal Source spectral envelope (Gómez-Vilda et al., 2004), whose singularities may be shown to be strongly determined by the relations among parameters in well-known k-mass models (Story &

> Titze, 1995; Berry, 2001) once the influence of the vocal tract has been removed. (Gómez-Vilda et al., 2009: 760)

Later studies of the same research group (Gómez-Vilda et al., 2012) are of great interest since they propose a novel metric framework for the evaluation of forensic voice evidence. This metric is based on "distances among matrices of features obtained from questioned and suspect phonations of spontaneous fillers" (Gómez-Vilda et al., 2012: 1). Originally envisaged as an investigation to further develop the group's research line aimed at testing the forensic application of glottal parameters extracted for dysphonia-detection studies, Gómez-Vilda et al. (2012) spotted the "unfair behavior of the log-likelihood ratio when evaluated in terms of the squares of distances from the questioned to the suspect evidence". Therefore, a new metric was developed to avoid undesirable[101] situations in a forensic context. As a result of their experimental investigations with recordings from a set of 100 speakers, they found that the proposed metric framework "behave more fairly than classical likelihood ratios in supporting the hypothesis of the defense vs that of the prosecutor, thus offering a more reliable evaluation scoring" (Gómez-Vilda et al., 2012: 1).

Some pilot experiments have been carried out with a relatively small sample of (MZ and DZ) twin pairs, as well as siblings (San Segundo, 2012) with the same glottal parameters[102] used by the research group of Gómez-Vilda. The preliminary results of this study pointed to the discriminatory potential of this kind of features. The biomechanical estimates of the glottal waveform seemed especially speaker-specific (see Section 5.2. cf. *proof of concept*).

5.2. Speech material, analysis tools and method

5.2.1. Speech material

For the glottal source analysis, the complete speech material consisted in 853 tokens of the [eː] vowel (average tokens per speaker and session: 7.89) naturally sustained in hesitation speech, resulting in pause fillers[103]. These vowels were extracted from the fifth speaking task (interview with the researcher) since this task, together with the fourth one, was initially intended to elicit

[101] Gómez-Vilda et al. (2012: 10) refers to this kind of situations: "[…] situations where an innocent suspect could face erroneously a charge of evidence". Specifically, they found that using a classical LR-approach, "the questioned evidence produced a positive H_p (*Prosecutor's Hypothesis*) in base not to its similarity to the suspect, but in base to its dissimilarity to the line-up". (Gómez-Vilda et al., 2012:10)

[102] Other investigations carried out using the software developed by Gómez-Vilda and colleagues but tested with non-vocalic phonetic units will not be reviewed in this section, as they are of no interest for the purpose of this thesis dissertation.

[103] Stemming from a distinction between silent and filled pauses, Gil (2007) explains that in Spanish filled pauses usually consist in a nasal resonance [m:] or in the *hesitation vowel* [e:]. This is called so (cf. footnote) because its presence in speech is due to the speaker's hesitation about how his speech should carry on. This would be an unconscious technique to gain time to think.

this type of speech material. After having checked that in the fourth task not all of the speakers hesitated in their answers to the mathematical questions, we extracted the sustained [eː] from the fifth task. In this, as we described in Chapter 2 (cf. *Corpus elaboration*), the speakers are asked about what they have been talking before in the first task. Since there is a considerably long time gap between the first and the fifth task, the speakers do not remember clearly the whole conversation and they exhibit hesitating responses.

Speech material extraction

For the selection of the sustained [eː] vowels we made an auditory and spectrographic examination in *Praat* for every speaker and session's audio files recorded in the fifth task. We did not select those vowels where we perceived a marked creak realization, a high degree of nasalization, overlap with extraneous noise, laughter, etc. Besides, the duration criterion for the selection of [eː] samples was that they had to be longer than 160 milliseconds[104]. In average the vowels are around 200 miliseconds. These sustained [eː] could be found both while articulating (we refer to the instances when they lengthen the duration of [e] at the end of a word), as in the example "porque…" (Fig. 23) or between silent pauses, as in the example "pues… eh…." (Fig. 24).[105]

This is the kind of pause filler chosen in all speech fragments (instead of, for example, [i] as in the copulative conjunction "y….") because they had a sufficiently large number of tokens per speaker and per recording session. Using *Praat*, these phonetic units were manually located and the most stable part of them was marked and extracted, avoiding the beginning and the end of the vowel. Above all, we were interested in pause fillers or hesitation marks that most people use, as the name suggests, when they hesitate, or while they are thinking of what they are going to say next, when they are trying to remember something, etc. We find them useful because they are longer than vowels in connected speech (this is important in order to estimate the glottal parameters) and also because they are thought to be idiosyncratic.

[104] According to the user's manual (www.biometrosoft.com), speech samples should have a minimum duration of 50 ms for a typical female voice (f0 = 200 Hz) and 100 ms for a typical male voice (f0 = 100 Hz).

[105] Gil (2007) mentions this trend to lengthen the final vowel in her definition of the hesitation vowel in Spanish (cf. footnote, p. 299):

> *Llamada así porque su aparición se debe muchas veces a la vacilación del hablante acerca del modo en que debe continuar su discurso: en realidad, viene a ser un modo inconsciente de ganar tiempo. En ocasiones también se articulan elementos léxicos o cuasiléxicos; por ejemplo, en algunas variedades del español, la conjunción [iː] o el demostrativo [ˈe̹steː], y a veces, no llega a existir pausa real porque el hablante alarga un sonido dado mientras piensa cómo continuar: la entrada:::: imprevista del presidente…* (Gil, 2007: 299).

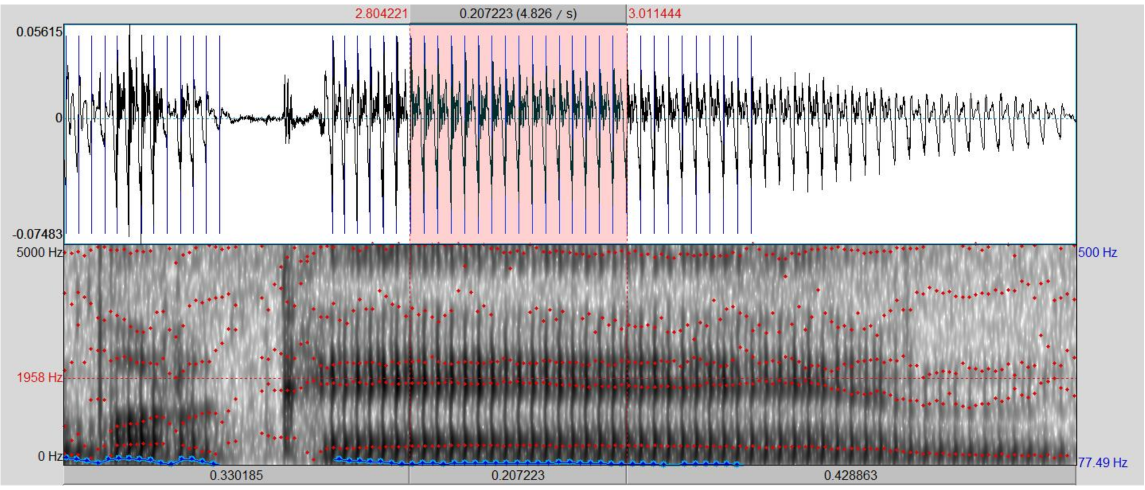


Figure 24. Extraction of the stationary part of the filler [eː] in the fragment “porque…”.

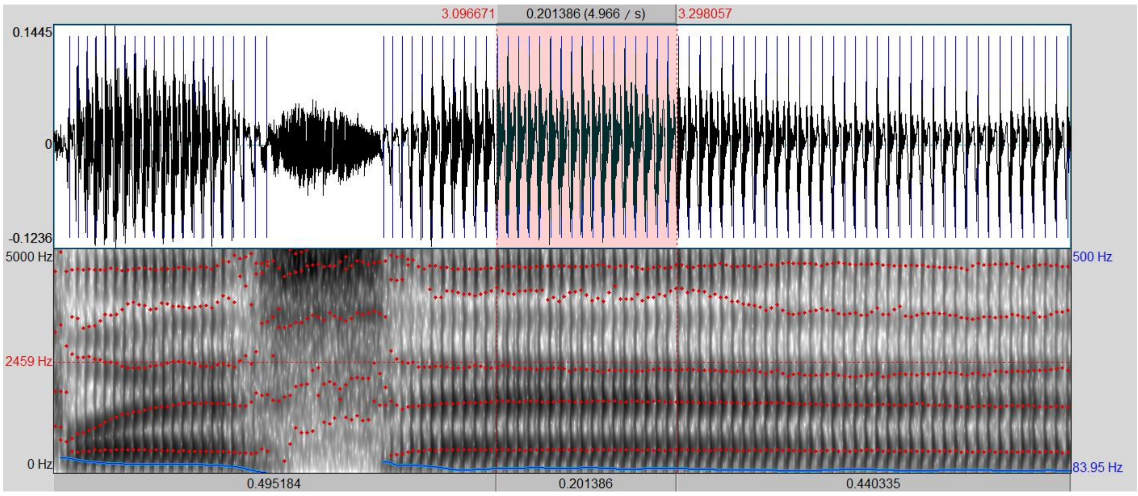


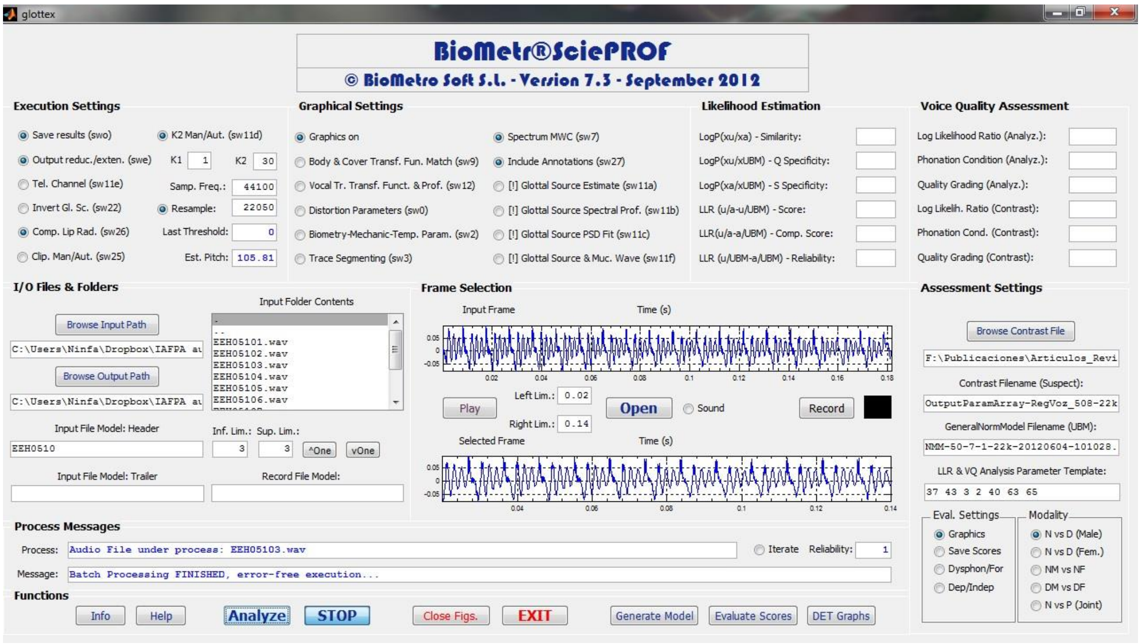
Figure 25. Extraction of the stationary part of the filler [eː] in the fragment “pues eh…”.

5.2.2. Analysis tools and method

Once the speech material was extracted, we proceeded to carry out several acoustic analyses using different tools and methodologies. More accurately, three types of software were used with varied purposes, which will be described in next pages: *BioMet®ScieProf*, *BioMet®PhonProf* and *BioMet®ForeProf*. All of them are developed by *BioMet®Soft*[106].

[106] “Founded in late 2010 as a ‘start-up’ company from Universidad Politécnica de Madrid, after winning the first prize in the VII Competition to Create New Companies with Technological Background among 260 other proposals, *BioMet®Soft* creates solutions for Security, Forensics and Medicine exploiting the pervasive character of voice and speech with a technology patented by Universidad Politécnica de Madrid.

BioMet®ScieProf[107] is designed, among other possible uses, to provide a thorough voice analysis based on the extraction of glottal source information, after the influence of the vocal tract has been eliminated thanks to an inverse filtering procedure. In other words, as specified by its developers "it has been designed for the evaluation of the voice quality and its biometrical properties by analyzing the glottal source obtained from the elimination of the vocal tract influence on voiced speech" (www.biometrosoft.com). We specifically used it for obtaining a whole set of voice parameters (68) from the recordings of the speakers participating in our study. The complete list of parameters can be found in Table 19 (see Section 5.3). In Figure 26 we can observe how the sounds were analyzed in batch-mode with a fixed time frame of 120 milliseconds. In Figure 27, an example of the kind of datasheet that this analysis yields. As will be detailed below (see Section 5.4), this software was useful in a first step to carry out a pilot experiment with a small database and a reduced parameter set. The parameters were entered into a MVKD formula, as described by Aitken and Lucy (2004) and implemented by Morrison (2007). The results of the forensic comparison carried out are explained in the section devoted to results.


glottex

BioMetr®SciePROF

© BioMetro Soft S.L. - Version 7.3 - September 2012

Execution Settings

- ● Save results (swo)
- ● Output reduc./exten. (swe)
- ○ Tel. Channel (sw11e)
- ○ Invert Gl. Sc. (sw22)
- ● Comp. Lip Rad. (sw26)
- ○ Clip. Man/Aut. (sw25)
- ● K2 Man/Aut. (sw11d)
- K1: 1 K2: 30
- Samp. Freq.: 44100
- ● Resample: 22050
- Last Threshold: 0
- Est. Pitch: 105.81

Graphical Settings

- ● Graphics on
- ○ Body & Cover Transf. Fun. Match (sw9)
- ○ Vocal Tr. Transf. Funct. & Prof. (sw12)
- ○ Distortion Parameters (sw0)
- ○ Biometry-Mechanic-Temp. Param. (sw2)
- ○ Trace Segmenting (sw3)
- ● Spectrum MWC (sw7)
- ● Include Annotations (sw27)
- ○ [!] Glottal Source Estimate (sw11a)
- ○ [!] Glottal Source Spectral Prof. (sw11b)
- ○ [!] Glottal Source PSD Fit (sw11c)
- ○ [!] Glottal Source & Muc. Wave (sw11f)

Likelihood Estimation

- LogP(xu/xa) - Similarity:
- LogP(xu/xUBM) - Q Specificity:
- LogP(xa/xUBM) - S Specificity:
- LLR (u/a-u/UBM) - Score:
- LLR(u/a-a/UBM) - Comp. Score:
- LLR (u/UBM-a/UBM) - Reliability:

Voice Quality Assessment

- Log Likelihood Ratio (Analyz.):
- Phonation Condition (Analyz.):
- Quality Grading (Analyz.):
- Log Likelih. Ratio (Contrast):
- Phonation Cond. (Contrast):
- Quality Grading (Contrast):

I/O Files & Folders

- C:\Users\Ninfa\Dropbox\IAFPA au
- C:\Users\Ninfa\Dropbox\IAFPA au
- Input Folder Contents: -, .., EEH05101.wav, EEH05102.wav, EEH05103.wav, EEH05104.wav, EEH05105.wav, EEH05106.wav
- Input File Model: Header: EEH0510
- Inf. Lim.: 3 Sup. Lim.: 3
- Input File Model: Trailer:
- Record File Model:

Frame Selection

- Left Lim.: 0.02
- Right Lim.: 0.14
- ○ Sound

Assessment Settings

- F:\Publicaciones\Articulos_Revi
- Contrast Filename (Suspect): OutputParamArray-RegVoz_508-22k
- GeneralNormModel Filename (UBM): NMM-50-7-1-22k-20120604-101028.
- LLR & VQ Analysis Parameter Template: 37 43 3 2 40 63 65
- Eval. Settings: ● Graphics ○ Save Scores ○ Dysphon/For ○ Dep/Indep
- Modality: ● N vs D (Male) ○ N vs D (Fem.) ○ NM vs NF ○ DM vs DF ○ N vs P (Joint)

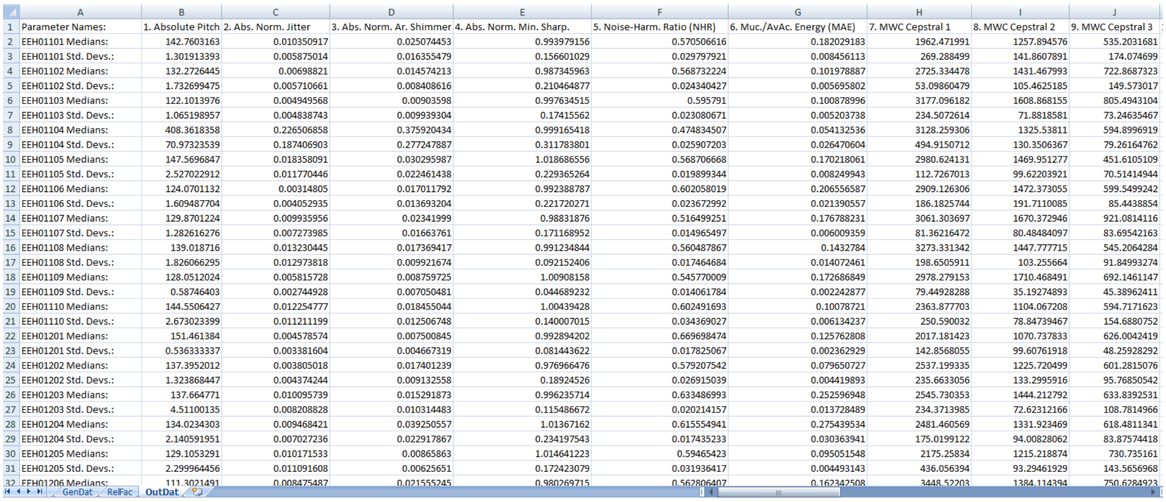
Process Messages

- Process: Audio File under process: EEH05103.wav
- ○ Iterate Reliability: 1
- Message: Batch Processing FINISHED, error-free execution...

Figure 26. BioMet®ScieProf configuration through the Graphical User Interface (GUI).

This technology is based on the extraction and characterization of the phonation profile of the speaker, and allows different levels of inference to be established." (http://www.glottex.com/en/about-us)

[107] We used version 7 – March 2012 (for the pilot experiment) and version 7.3 – Sept 2012 (for the final results).



	A	B	C	D	E	F	G	H	I	J
1	Parameter Names:	1. Absolute Pitch	2. Abs. Norm. Jitter	3. Abs. Norm. Ar. Shimmer	4. Abs. Norm. Min. Sharp.	5. Noise-Harm. Ratio (NHR)	6. Muc./AvAc. Energy (MAE)	7. MWC Cepstral 1	8. MWC Cepstral 2	9. MWC Cepstral 3
2	EEH01101 Medians:	142.7603163	0.010350917	0.025074453	0.993979156	0.570506616	0.182029183	1962.471991	1257.894576	535.2031681
3	EEH01101 Std. Devs.:	1.301913393	0.005875014	0.016355479	0.156601029	0.029797921	0.008456113	269.288499	141.8607891	174.074699
4	EEH01102 Medians:	132.2726445	0.00698821	0.014574213	0.987345963	0.568732224	0.101978887	2725.334478	1431.467993	722.8687323
5	EEH01102 Std. Devs.:	1.732699475	0.005710661	0.008408616	0.210464877	0.024340427	0.005695802	53.09860479	105.4625185	149.573017
6	EEH01103 Medians:	122.1013976	0.004949568	0.00903598	0.997634515	0.595791	0.100878996	3177.096182	1608.868155	805.4943104
7	EEH01103 Std. Devs.:	1.065198957	0.004838743	0.009939304	0.17415562	0.023080671	0.005203738	234.5072614	71.8818581	73.24635467
8	EEH01104 Medians:	408.3618358	0.226506858	0.375920434	0.999165418	0.474834507	0.054132536	3128.259306	1325.53811	594.8996919
9	EEH01104 Std. Devs.:	70.97323539	0.187406903	0.277247887	0.311783801	0.025907203	0.026470604	494.9150712	130.3506367	79.26164762
10	EEH01105 Medians:	147.5696847	0.018358091	0.030295987	1.018686556	0.568706668	0.170218061	2980.624131	1469.951277	451.6105109
11	EEH01105 Std. Devs.:	2.527022912	0.011770446	0.022461438	0.229365264	0.019899344	0.008249943	112.7267013	99.62203921	70.51414944
12	EEH01106 Medians:	124.0701132	0.00314805	0.017011792	0.992388787	0.602058019	0.206556587	2909.126306	1472.373055	599.5499242
13	EEH01106 Std. Devs.:	1.609487704	0.004052935	0.013693204	0.221720271	0.023672992	0.021390557	186.1825744	191.7110085	85.4438854
14	EEH01107 Medians:	129.8701224	0.009935956	0.02341999	0.98831876	0.516499251	0.176788231	3061.303697	1670.372946	921.0814116
15	EEH01107 Std. Devs.:	1.282616276	0.007273985	0.01663761	0.171168952	0.014965497	0.006009359	81.36216472	80.48484097	83.69542163
16	EEH01108 Medians:	139.018716	0.013230445	0.017369417	0.991234844	0.560487867	0.1432784	3273.331342	1447.777715	545.2064284
17	EEH01108 Std. Devs.:	1.826066295	0.012973818	0.009921674	0.092152406	0.017464684	0.014072461	198.6505911	103.255664	91.84993274
18	EEH01109 Medians:	128.0512024	0.005815728	0.008759725	1.00908158	0.545770009	0.172686849	2978.279153	1710.468491	692.1461147
19	EEH01109 Std. Devs.:	0.58746403	0.002744928	0.007050481	0.044689232	0.014061784	0.002242877	79.44928288	35.19274893	45.38962411
20	EEH01110 Medians:	144.5506427	0.012254777	0.018455044	1.00439428	0.602491693	0.10078721	2363.877703	1104.067208	594.7171623
21	EEH01110 Std. Devs.:	2.673023399	0.011211199	0.012506748	0.140007015	0.034369027	0.006134237	250.590032	78.84739467	154.6880752
22	EEH01201 Medians:	151.461384	0.004578574	0.007500845	0.992894202	0.669698474	0.125762808	2017.181423	1070.737833	626.0042419
23	EEH01201 Std. Devs.:	0.536333337	0.003381604	0.004667319	0.081443622	0.017825067	0.002362929	142.8568055	99.60761918	48.25928292
24	EEH01202 Medians:	137.3952012	0.003805018	0.017401239	0.976966476	0.579207542	0.079650727	2537.199335	1225.720499	601.2815076
25	EEH01202 Std. Devs.:	1.323868447	0.004374244	0.009132558	0.18924526	0.026915039	0.004419893	235.6633056	133.2995916	95.76850542
26	EEH01203 Medians:	137.664771	0.010095739	0.015291873	0.996235714	0.633486993	0.252596948	2545.730353	1444.212792	633.8392531
27	EEH01203 Std. Devs.:	4.51100135	0.008208828	0.010314483	0.115486672	0.020214157	0.013728489	234.3713985	72.62312166	108.7814966
28	EEH01204 Medians:	134.0234303	0.009468421	0.039250557	1.01367162	0.615554941	0.275439534	2481.460569	1331.923469	618.4811341
29	EEH01204 Std. Devs.:	2.140591951	0.007027236	0.022917867	0.234197543	0.017435233	0.030363941	175.0199122	94.00828062	83.87574418
30	EEH01205 Medians:	129.1053291	0.010171533	0.00865863	1.014641223	0.59465423	0.095051548	2175.25834	1215.218874	730.735161
31	EEH01205 Std. Devs.:	2.299964456	0.011091608	0.00625651	0.172423079	0.031936417	0.004493143	436.056394	93.29461929	143.5656968
32	EEH01206 Medians:	111.3021491	0.008475487	0.021555245	0.980269715	0.562806407	0.162342508	3448.52203	1384.114394	750.6284923

Figure 27. Example of datasheet (fragment) obtained after executing *BioMet®ScieProf* with information about the voice parameters values (median and standard deviation) per speaker/session/token. As can be seen in line 8 and 9 some pitch errors (second column: absolute pitch) occurred which in later stages were corrected.

Taking into account the results obtained in the above-mentioned proof of concept, fully described in the results section (see Section 5.4), and in view of some confusing values for some speakers, another tool *(BioMet®PhonProf)*[108] was used in order to detect possible pathologies in the voices of the speakers under study as well as with the aim of obtaining a more detailed evaluation of their voice quality. This software also allows the extraction of the same glottal parameters as *BioMet®ScieProf*. Its purpose and the application design are similar. Yet this kind of tool is mainly used by speech therapists in order to evaluate the voice of a patient. A useful feature of this software is that it shows the limits of normophony for the specified parameters. For instance, in Table 17, eight voice parameters are selected for speaker 09, first session, token two (EEH09102). The last column shows the value obtained for the parameter, while the second-third columns show the ranges of normality for a male voice. As can be observed, most parameters exceed the limits. Indeed only body mass and cover mass are within the normality range limits. As we will detail below, this does not necessarily mean that the voice is pathological. Other possible factors may merge to cause a specific parameter to be below or above the limits.

Table 17

Glottal Source Quality Analysis

Parameter number	Parameter name	Minimum value for normophony	Maximum value for normophony	Value obtained
35	Body Mass	0.018	0.025	0.020
37	Body Stiffness	9378.731	12594.897	12689.818

[108] We used the version 2.3 (7.3) July 2012.

38	Body Mass Unbalance	0.001	0.002	0.008
40	Body Stiffness Unbalance	0.003	0.014	0.034
41	Cover Mass	0.010	0.019	0.019
43	Cover Stiffness	4070.607	9616.993	28376.272
44	Cover Mass Unbalance	0.009	0.048	0.324
46	Cover Stiffness Unbalance	0.014	0.060	0.495

Note. Glottal Source Quality Analysis for the specific vowel EEH09102. The table shows eight parameters (first column) with the limits of normality established for each one: minimum (third column) and maximum (fourth column) and finally the value obtained for this specific speaker (09) in his first recording session (1) and second token (02). We mark in red the parameters exceeding the limits.

The tool *BioMet*®*PhonProf* is also of interest because it shows a figure of the glottal waveform per each vowel under analysis. This is important in order to complete the voice diagnosis of each speaker. As can be seen in Figure 28, this representation of the glottal waveform allows visual inspection to detect anomalies in the glottal cycle. For instance, on the left, Figure 28 shows the glottal waveform and flow of a vowel produced by the speaker 09 (specific token: EEH09102). On the right, all the cycles extracted for that vowel are represented, while for the figure on the left the mid cycle is used. The main characteristics that this speaker shows in his phonation are:

- Fast opening phase.
- Adduction defects, which are made clear through the irregularities in the waveform.
- Slow and long-lasting closing phase.
- Overall airflow contact escape (first half of the green line: until 0.3 ms).

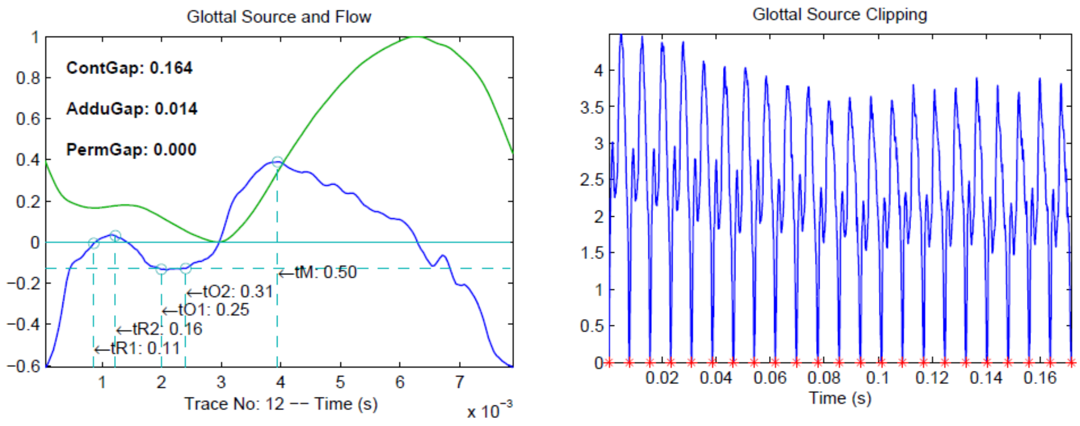


Figure 28. Glottal waveform and flow (left); and glottal cycles (right) of an [eː] sound by speaker 09.

If we compare speaker 09 above with his twin (speaker 10), we can appreciate how different the glottal waveform can be between speakers. Figure 29 shows on the left the glottal waveform and flow of a vowel produced by speaker 10 (second session, token one: EEH10201). On the right, this figure shows the glottal cycles extracted from the vowel under consideration. As an example of the voice characteristics which can be observed in this speaker (a more thorough description of these speakers can be found below in the section devoted to results), we can mention the following ones:

- Unclear opening and closing pattern (i.e. flat pattern of the waveform).
- Inappropriate opening and almost no closure phase.
- Airflow escape.
- Asymmetry of the glottal cycles (image on the right).

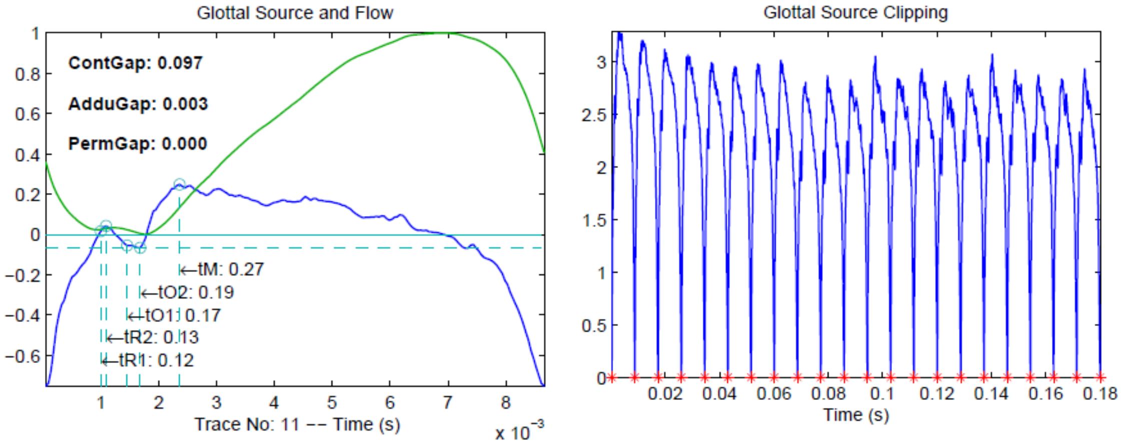


Figure 29. Glottal waveform and flow (left) and glottal cycles (right) of a [eː] sound by speaker 10.

If we carry on with the description of the methodological process undertaken for the glottal analysis, we have already explained that a first approach was taken with a small speaker subset, using the tool *BioMet*®*ScieProf* and a typical forensic comparison strategy, following the method explained in Morrison (2007). In a second step, we aimed to perform a visual inspection of the glottal waveform of some speakers, in view of certain unexpected values for the LRs obtained after the forensic comparison [see Section 5.4]. Some further methodological measures were taken in order to find possible scientific explanations for the results of the pilot experiment. Our goal was to find out whether the unexpected results could be due to the true idiosyncrasy of the speakers' voices or to either the presence of voice pathology or an evaluation (software) artifact. These measures consisted in:

a) First, in order to discard a possible pathology as cause of the unexpected results in the proof of concept, we examined the anamnesis of each speaker, with the information gathered in the two questionnaires that they had to fill in both recording sessions (*described in Chapter* 3). This evaluation will be detailed below.
b) Secondly, we checked the zygosity of a twin pair who showed high voice dissimilarities between them. The results of the DNA analysis carried out (*described in Chapter 3*) revealed that they were indeed MZ twin pairs, as they have stated in the questionnaire.
c) Thirdly, we analyzed again the recordings, this time with *BioMet®PhonProf*, as has been explained above, with the aim of visually inspecting the glottal waveform of the speakers and finding possible individual patterns which could explain the unexpected results in the pilot experiment.
d) Finally, we reexamined the values extracted from the first voice tool (*BioMet®ScieProf*) since the analysis was carried out in a batch-mode and this kind of processing may entail certain errors or artifacts (e.g. execution software artifacts, pitch artifacts or inversion artifacts) – [see Section 5.4].

So in view of both the anamnesis examination and the results of the zygosity test, and especially after the specific errors obtained with the previous software had been corrected, we proceeded to use a third voice evaluation tool: *BioMet®ForeProf*[109]. This allowed us to carry out a forensic comparison different from the one performed in the pilot test.

As observed in Figure 30, this new tool needs the user to input the path where he stored the datasheet (in .mat file format) with the values of the whole set of 65 voice parameters extracted from the previous software. This is needed for the *model*, the *control* and the *test*. With *control* and *test*, we just refer to the two speakers under comparison at each time, independently of the order. So, for instance, if we want to compare speaker 09 and speaker 10, one would be the test and the other one the control. The names test and control are given since the tool is designed to be used in forensic settings. They would represent what other researchers call the *suspect* and the *offender* (e.g. Morrison, 2009b)[110]. For research purposes, as is our case, the terminology is not so important. It just means "first element of comparison" and "second element of comparison". Note that for inter-speaker comparisons, the first element is one speaker and the second one, another one. However, for the intra-speaker comparisons, the first element would be the first

[109] We used the version 2.3 (7.3) Sept. 2012.
[110] Spanish: *dubitada* and *indubitada* (as used by the *Guardia Civil*).

session of a speaker while the second element would be his second session. By *model* we refer to the reference population, or background population. This was already described in Chapter 3 as the subset of speakers without kinship relationship who participated in our study carrying out the same speaking tasks as twin and non-twin siblings. Since the number of this type of speakers was not very high, we have increased it by adding all the other speakers except the one or two already used as *control* and *test*. This was made in the cross-validated fashion described by Morrison (2009c). For a complete picture of this method, we took into account the following aspects:

- Both the control and the test were not present in the model.
- Only the first session of each speaker was used so that the number of model tokens was not excessively large, in comparison with the data distribution of the control and the test.

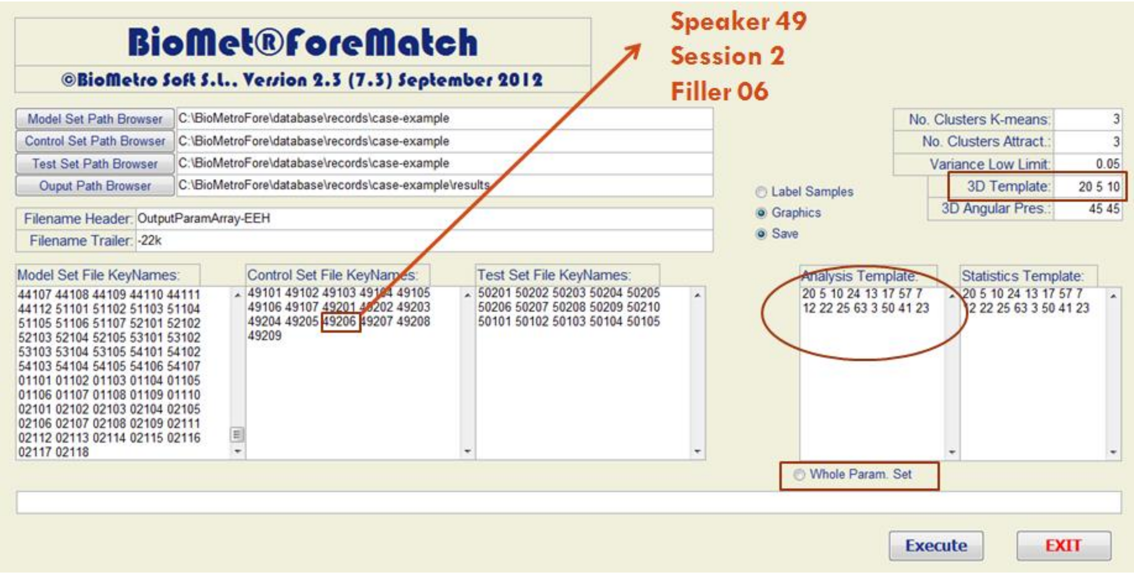
As can be seen in Figure 30, we had to specify the names of the model, the test and the control for each comparison. In this case, for the comparison of speaker 49 and 50, under the column "Model Set File KeyNames" we specified the names of the datasheet with the parameter values for the [eː] sounds (fillers) of all the speakers and sessions in our database, except those pertaining to speaker 49 and 50, which are specified under the column "Control Set File KeyNames" and "Test Set File KeyNames" respectively.

With the security that the datasheet with the values for the whole set of 68 parameters is now free of errors, after all the steps of the diagnosis, we proceeded to execute the program. These were the steps followed:

- Firstly, we run the program with the option "whole parameter set" selected. This gives us a first approximation to how similar or dissimilar are the elements of the comparison (test and control) when all the parameters are considered. The tool also lists the 68 parameters in order of relevance. This is possible thanks to a relevance analysis, carried out with LDA (Linear Discriminant Analysis).
- In a second step, we select the 16 more relevant parameters[111] (as they appear in the column "Statistics Template") and insert them in the column "Analysis Template". These features will be the ones considered in the next execution. With the button "Whole Parameter Set" unselected, we run again the program. The results of the execution are showed in Figure 31. In this second occasion, we also insert the three most relevant parameters in the box "3D Template" so that we can obtain graphs such as the one in Figure 32.

[111] This is done in order to carry out comparisons with a reduced set of parameters, namely the most characteristics of each speaker. Sixteen parameters are considered an adequate number to limit the acceptable margin of error (Gómez-Vilda, personal communication).

We have included Figure 31 as an example of the output of the tool *BioMet®ForeProf*. In it, the value of interest is circled. It is a match value which informs of the similarity between speakers taking also into account their respective distances towards the model (the mathematical model underpinning this comparison method is described in Gómez-Vilda et al., 2012). Interestingly, in the example proposed, the values of the inter-speaker[112] comparison (05v06) are very similar to the values obtained in the intra-speaker comparison (05v05). The values are LLRs (log-likelihood-ratios) and should be interpreted as indicated in Chapter 3.



BioMet®ForeMatch

©BioMetro Soft S.L., Version 2.3 (7.3) September 2012

Speaker 49
Session 2
Filler 06

Model Set Path Browser: C:\BioMetroFore\database\records\case-example
Control Set Path Browser: C:\BioMetroFore\database\records\case-example
Test Set Path Browser: C:\BioMetroFore\database\records\case-example
Ouput Path Browser: C:\BioMetroFore\database\records\case-example\results

Filename Header: OutputParamArray-EEH
Filename Trailer: -22k

No. Clusters K-means: 3
No. Clusters Attract.: 3
Variance Low Limit: 0.05
3D Template: 20 5 10
3D Angular Pres.: 45 45

○ Label Samples
● Graphics
● Save

Model Set File KeyNames:
44107 44108 44109 44110 44111
44112 51101 51102 51103 51104
51105 51106 51107 52101 52102
52103 52104 52105 53101 53102
53103 53104 53105 54101 54102
54103 54104 54105 54106 54107
01101 01102 01103 01104 01105
01106 01107 01108 01109 01110
02101 02102 02103 02104 02105
02106 02107 02108 02109 02111
02112 02113 02114 02115 02116
02117 02118

Control Set File KeyNames:
49101 49102 49103 49104 49105
49106 49107 49201 49202 49203
49204 49205 49206 49207 49208
49209

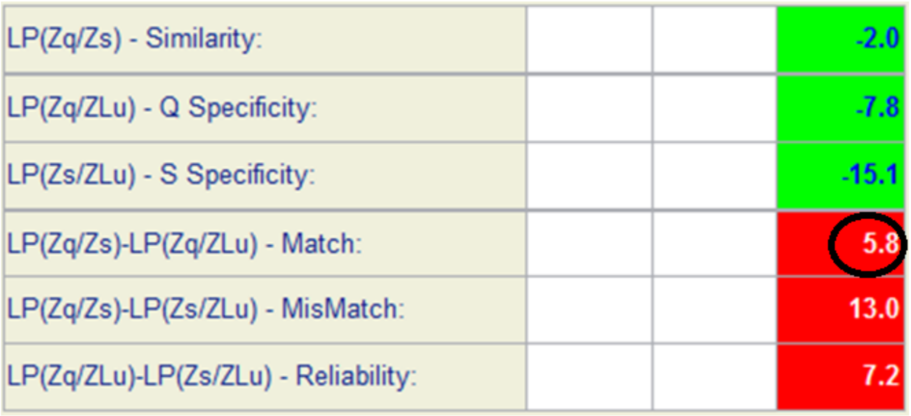
Test Set File KeyNames:
50201 50202 50203 50204 50205
50206 50207 50208 50209 50210
50101 50102 50103 50104 50105

Analysis Template:
20 5 10 24 13 17 57 7
12 22 25 63 3 50 41 23

Statistics Template:
20 5 10 24 13 17 57 7
12 22 25 63 3 50 41 23

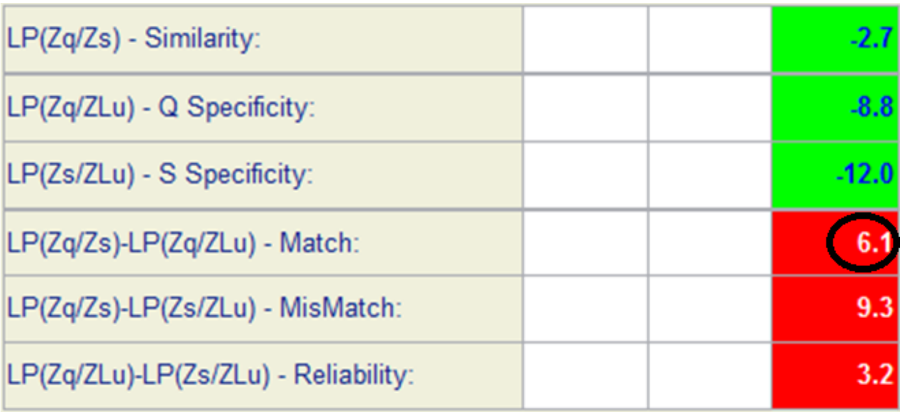
○ Whole Param. Set

Figure 30. *BioMet®ForeProf* configuration through the Graphical User Interface (GUI).



LP(Zq/Zs) - Similarity:			-2.0
LP(Zq/ZLu) - Q Specificity:			-7.8
LP(Zs/ZLu) - S Specificity:			-15.1
LP(Zq/Zs)-LP(Zq/ZLu) - Match:			5.8
LP(Zq/Zs)-LP(Zs/ZLu) - MisMatch:			13.0
LP(Zq/ZLu)-LP(Zs/ZLu) - Reliability:			7.2

[112] Note that this is an inter-speaker comparison but also an intra-pair comparison because the speakers are twins.



LP(Zq/Zs) - Similarity:			-2.7
LP(Zq/ZLu) - Q Specificity:			-8.8
LP(Zs/ZLu) - S Specificity:			-12.0
LP(Zq/Zs)-LP(Zq/ZLu) - Match:			6.1
LP(Zq/Zs)-LP(Zs/ZLu) - MisMatch:			9.3
LP(Zq/ZLu)-LP(Zs/ZLu) - Reliability:			3.2

Figure 31. Results obtained after the execution of *BioMet*®*ForeProf*. The table *above* shows the comparison of speaker 05 and speaker 06 (MZ twin pair). The table *below* shows the results of the comparison between the first session of speaker 05 and the second session of the same speaker. The values are LLRs (log-likelihood-ratios)

Figure 32 is an example of a 3D model obtained through the tool which we are describing. It shows the three most relevant parameters for each comparison. In this case, speaker 10 (session one) is compared with himself (session two). We can see how the combination of three parameters characterizes the speaker independently of the session (red color is session 1 and blue color session 2). In other words, those parameters place the speaker in a specific area of the whole data distribution. Especially parameter 42 and 21 allow a very good isolation of this speaker from the model or background population (shown in green). Other useful graphs like the series of box plots in Figure 33 evidence how idiosyncratic are the values of parameter 42 and 21 for speaker 10.

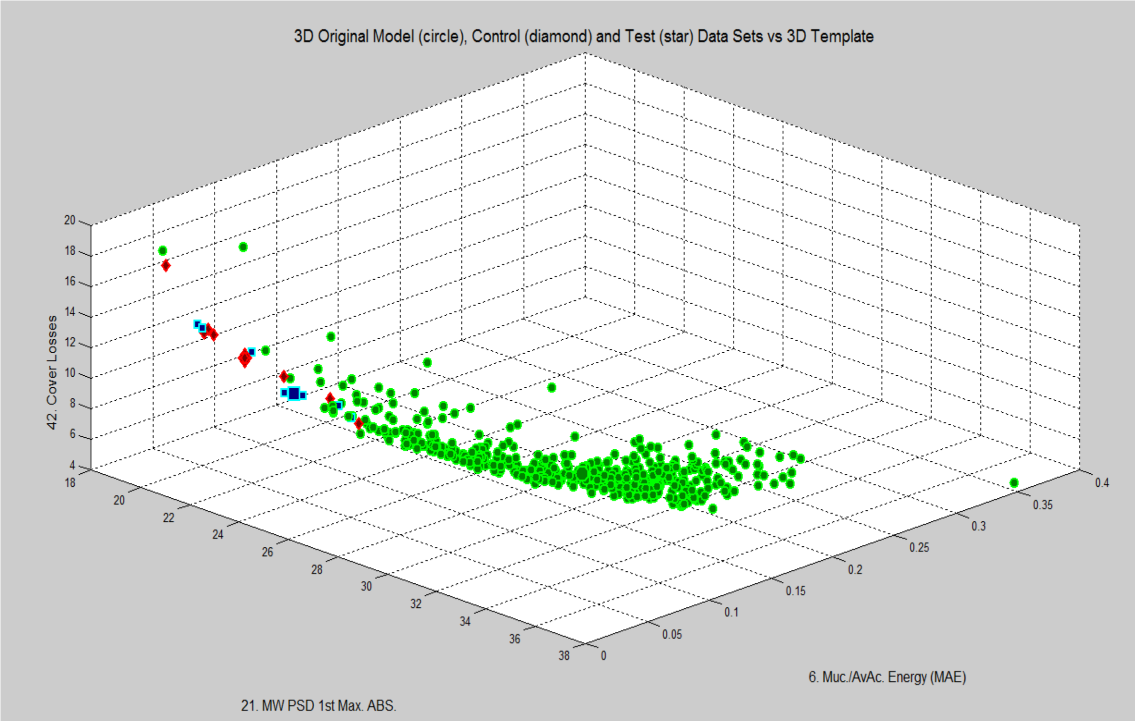


Figure 32. 3D Model for the comparison of speaker 10 with himself (one speaking session against the other). The parameters shown are the ones specified in the GUI of *BioMet*®*ForeProf*, in our case the three most relevant of the LDA.. The exponential relationship between parameter 42 and parameters 21 and 6 indicates a non-linear relationship.

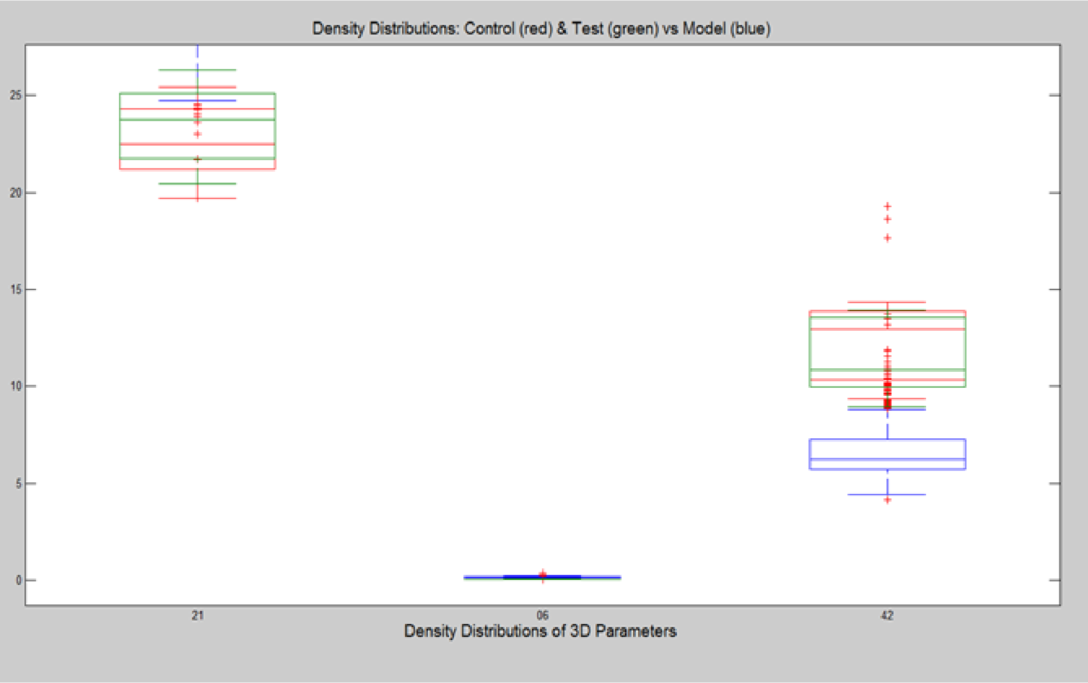


Figure 33. Boxplots for the parameters included in the 3D Model. The parameter values for the control and test (in this case, session one and two of speaker 10) are shown in red and green. The values for the model (background population) appear in blue.

So far we have described the three tools used for the voice analysis, including the forensic matching in the case of *BioMet*®*ForeProf*. The reasons for using one software or another at each specific research stage have also been detailed. For more information about the functioning of the tools develop by *BioMet*®*Soft*, see the specific bibliography in section 5.1.2.

5.3. Parameters

For a first approach to the parameters used for the glottal analysis, it seems useful to explain some basic aspects of the separation between the vocal tract biometry and the glottal waveform biometry (Gómez-Vilda, 2009):

Table 18

Vocal tract biometry and Glottal Waveform Biometry

Vocal tract biometry	Glottal waveform biometry
It is reflected in the vocal tract section and length.	It can be represented by (a) the biomechanics of the vocal folds and (b) by the spectral profile of the mucosal waveform.
It depends highly on the utterance message.	(a) In the *biomechanical characterization of the vocal folds*, body and cover can be distinguished.
It can be characterized by the reduction to MFCC of its Transfer Function or by the cepstral description of its parameterization by linear prediction (LPCC).	(b) The *spectral profile of the mucosal waveform* can be expressed by its singularity values or by its reduction to MFCC.

Note. Adapted from Gómez-Vilda (2009: 36).

Once we have mentioned the two different ways in which the glottal waveform biometry can be characterized (Table 18), the various parameters extracted from *BioMet*®*ScieProf,* and included in Table 19, can be classified in seven different subsets:

- Fundamental frequency (f_0) and distortion parameters (*p1-6*): jitter, shimmer and NHR
- Cepstral coefficients of the glottal source power spectral density (*p7-20*).
- Singularities of the glottal source power spectral density –profile- (*p21-34*).
- Biomechanical estimates of vocal fold mass, tension and losses (*p35-46*).
- Time-based Glottal Source coefficients (*p47-58*).
- Glottal gap (closure) coefficients (*p59-62*).
- Tremor (cyclic) coefficients (p63-68).

Table 19

Parameter set generated by BioMet®*ScieProf version 7.3 – Sept 2012*

Parameter	Description
1. Fundamental Frequency (f_0)	Inverse of each glottal cycle period, given in Hz
2. Abs. Norm. Jitter	Inverse of the difference between neighbor glottal cycle periods divided by their average
3. Abs. Norm. Ar. Shimmer	Difference between neighbor glottal cycle amplitudes divided by their average
4. Abs. Norm. Min. Sharp	Peak slenderness at the Maximum Flow Declination Rate: negative amplitude of the peak divided by its width
5. Noise-Harm. Ratio (NHR)	Ratio between the energy of the non-harmonic and the harmonic parts of the glottal source power spectral density
6. Muc./AvAc. Energy (MAE)	Ratio between the energy of the glottal source to average acoustic wave difference and the average acoustic wave
7. MWC Cepstral 1	First Cepstral Coefficient of the glottal wave correlate
8. MWC Cepstral 2	Second Cepstral Coefficient of the glottal wave correlate
9. MWC Cepstral 3	Third Cepstral Coefficient of the glottal wave correlate
10. MWC Cepstral 4	Fourth Cepstral Coefficient of the glottal wave correlate
11. MWC Cepstral 5	Fifth Cepstral Coefficient of the glottal wave correlate
12. MWC Cepstral 6	Sixth Cepstral Coefficient of the glottal wave correlate
13. MWC Cepstral 7	Seventh Cepstral Coefficient of the glottal wave correlate
14. MWC Cepstral 8	Eighth Cepstral Coefficient of the glottal wave correlate
15. MWC Cepstral 9	Ninth Cepstral Coefficient of the glottal wave correlate
16. MWC Cepstral 10	Tenth Cepstral Coefficient of the glottal wave correlate
17. MWC Cepstral 11	Eleventh Cepstral Coefficient of the glottal wave correlate
18. MWC Cepstral 12	Twelfth Cepstral Coefficient of the glottal wave correlate
19. MWC Cepstral 13	Thirteenth Cepstral Coefficient of the glottal wave correlate
20. MWC Cepstral 14	Fourteenth Cepstral Coefficient of the glottal wave correlate
21. MW PSD 1st Max. ABS.	First maximum of glottal source power spectral density
22. MW PSD 1st Min. rel.	First minimum of glottal source power spectral density
23. MW PSD 2nd Max. rel.	Second maximum of glottal source power spectral density
24. MW PSD 2nd Min. rel.	Second minimum of glottal source power spectral density
25. MW PSD 3rd Max. rel.	Third maximum of glottal source power spectral density
26. MW PSD End Val. rel.	Value of the glottal source power spectral density at half sampling frequency

27. MW PSD 1st Max. Pos. ABS.	Frequency of the first maximum of glottal source power spectral density
28. MW PSD 1st Min. Pos. rel.	Frequency of the first minimum of glottal source power spectral density relative to first maximum frequency
29. MW PSD 2nd Max. Pos. rel.	Frequency of the second maximum of glottal source power spectral density relative to first maximum frequency
30. MW PSD 2nd Min. Pos. rel.	Frequency of the second minimum of glottal source power spectral density relative to first maximum frequency
31. MW PSD 3th Max. Pos. rel.	Frequency of the third maximum of glottal source power spectral density relative to first maximum frequency
32. MW PSD End Val. Pos. rel.	Frequency of the glottal source power spectral density at half sampling frequency relative to first maximum frequency
33. MW PSD 1st Min NSF	Slenderness of the first “V groove” in the glottal source power spectral density: negative amplitude of the peak divided by its width
34. MW PSD 2nd Min NSF	Slenderness of the second “V groove” in the glottal source power spectral density: negative amplitude of the peak divided by its width
35. Body Mass	Equivalent dynamic mass of the vocal fold body for each glottal cycle
36. Body Losses	Equivalent resistive parameter of the vocal fold body for each glottal cycle
37. Body Stiffness	Equivalent lateral stiffness of the vocal fold body for each glottal cycle
38. Body Mass Unbalance	Difference between neighbor glottal cycle body masses divided by their average
39. Body Losses Unbalance	Difference between neighbor glottal cycle body losses divided by their average
40. Body Stiffness Unbalance	Difference between neighbor glottal cycle body stiffness divided by their average
41. Cover Mass	Equivalent dynamic mass of the vocal fold cover for each glottal cycle
42. Cover Losses	Equivalent resistive parameter of the vocal fold cover for each glottal cycle
43. Cover Stiffness	Equivalent lateral stiffness of the vocal fold cover for each glottal cycle
44. Cover Mass Unbalance	Difference between neighbor glottal cycle cover masses divided by their average
45. Cover Losses Unbalance	Difference between neighbor glottal cycle cover losses divided by their average
46. Cover Stiffness Unbalance	Difference between neighbor glottal cycle cover stiffness divided by their average
47. Rel. Recovery 1 Time	Ratio between the first recovery time and the total glottal cycle duration
48. Rel. Recovery 2 Time	Ratio between the second recovery time and the total glottal cycle duration
49. Rel. Open 1 Time	Ratio between the first opening time and the total glottal cycle duration
50. Rel. Open 2 Time	Ratio between the second opening time and the total glottal cycle duration
51. Rel. Maximum Amplit. Time	Ratio between the glottal source maximum amplitude instant and the total glottal cycle duration
52. Rel. Recov. 1 Amplitude	Ratio between the first recovery time amplitude and the peak-to-peak amplitude
53. Rel. Recov. 2 Amplitude	Ratio between the second recovery time amplitude and the peak-to-peak amplitude
54. Rel. Open 1 Amplitude	Ratio between the first opening time amplitude and the peak-to-peak amplitude
55. Rel. Open 2 Amplitude	Ratio between the second opening time amplitude and the peak-to-peak amplitude

56. Rel. Stop Flow Time	Ratio between the glottal flow minimum instant and the total glottal cycle duration
57. Rel. Start Flow Time	Ratio between the glottal flow start instant and the total glottal cycle duration
58. Rel. Closing Time	Ratio between the glottal flow maximum instant and the total glottal cycle duration
59. Val. Flow GAP	Ratio between the contact gap flow escape and the total glottal flow
60. Val. Contact GAP	Ratio between the escape flow and the total glottal flow during the contact phase
61. Val. Adduction GAP	Ratio between the diminished escape flow and the total glottal flow during the open phase
62. Val. Permanent GAP	Ratio between the escape flow and the total glottal flow during the recovery phase
63. 1st. Order Cyclic Coefficient	First PARCOR coefficient in the equivalent AR model of the unbiased vocal fold body stiffness
64. 2nd. Order Cyclic Coefficient	Second PARCOR coefficient in the equivalent AR model of the unbiased vocal fold body stiffness
65. 3rd. Order Cyclic Coefficient	Third PARCOR coefficient in the equivalent AR model of the unbiased vocal fold body stiffness
66. Tremor Frequency (Hz)	First harmonic of the unbiased vocal fold body stiffness
67. Tremor Est. Robustness	Proximity to the unity circle of the equivalent AR model first pole of the unbiased vocal fold stiffness
68. Tremor amplitude (rMSA)	Standard deviation of the unbiased vocal fold stiffness

Note. The description of each parameter has been adapted from *BioMet®PhonProf* User's Manual (March 2014).

Some parameters, such as jitter, shimmer or HNR have been used traditionally for voice pathology detection, which is well documented in several works (e.g. Boyanov & Hadjitodorov, 1997; Ritchings, McGillion, & Moore, 2002; as cited in Gómez-Vilda et al., 2009: 760). These parameters have also been found useful for speaker identification (see Section 5.1.2). However, some other parameters from the list above are less known in both disciplines, but especially in the forensic field. Therefore, in next pages, we carry out a brief description of the seven parameter subsets mentioned before.

1) *Fundamental frequency and classical perturbation estimates*

There are various algorithms for *fundamental frequency* (f_0) extraction in both the time and frequency domains. Defined as the inverse of the glottal cycle period, the f_0 can be estimated in the time domain with a glottal source trace technique which relies upon the quasi-periodicity of the voice (Murphy, 2008):

> Autocorrelation shows how well a voice signal correlates with itself over a range of delays, and any periodic signal will correlate with itself at very short delays as well as at those delays which correspond to multiples of the fundamental frequency. The fundamental frequency can be found

> by looking for peaks in the delay intervals which correspond to the normal frequency range of the voice signal given. (Murphy, 2008: 85)[113]

The frequency domain technique to extract the f_0 is based on the fact that equally spaced harmonics can give a measure of the f_0: "the f_0 may well be clear after FFT (Fast Fourier Transform) calculation, but it is also possible to subtract one harmonic from the next at any part of the transformation (harmonics being at equal spaces from each other and at multiples of the fundamental)" (Murphy, 2008: 85). A third way to extract the value of the f_0 is by calculating the cepstrum[114].

Frequency perturbation, or *period perturbation* – commonly called *jitter*- is defined as "the variability of the fundamental frequency from one cycle to the next. […] That is, jitter is a measurement of how much a given period differs from the period that immediately follows it, and not how much it differs from a cycle at the other end of the utterance. Jitter, then, is a measure of the frequency variability not accounted for by voluntary changes in f_0." (Baken & Orlikoff, 2000: 190). According to these authors, "the degree of frequency perturbation is intended to provide an index of the stability of the phonatory system" (Baken & Orlikoff, 2000: 190). Interestingly, Baken and Orlikoff (2000: 191) point to the idiosyncrasy of this parameter in the between-speaker variation: "The phonatory system is in no way a perfect machine and every speaker's vibratory cycles are erratic to some extent. But, on the face of it, one would guess that an abnormal larynx should produce a more erratic voice than a healthy one". Although increased vocal jitter is associated with voice disorder (being one of the physical correlates of perceived "hoarseness" or "harshness"), it should not be used as a sole diagnostic criterion. Besides, there are several possible sources of jitter. Their listing here is of interest for our study as far as the different potential causes of frequency perturbation may be behind the differences observed between speakers. According to Baken and Orlikoff (2000: 191-192), five are the potential factors causing jitter: *neurogenic, aerodynamic, mechanical, stylistics* and *chaotic oscillation*. For instance, regarding the mechanical factors, it is mentioned that "f_0 is also affected by the heartbeat, probably because of varying stiffness of the vascular bed of the vocal folds as a function of the cyclic change of blood pressure" (*inter alia*, Orlikoff & Baken, 1989; as cited in Baken & Orlikoff, 2000: 192). There are several jitter measures or numerical indices of frequency perturbation (cf. Baken & Orlikoff, 2000: 198-206). For our study, jitter is estimated as the ratio of the difference between neighbor periods with respect to its average value for the voice segment (Gómez-Vilda et al., 2009:769).

[113] Murphy (2008: 85) refers to www.phon.ucl.ac.uk, UCL Lecture 10: Speech Signal Analysis.

[114] "The cepstrum is the DFT (Discrete Fourier Transform) of the logarithmic amplitude spectrum of the original signal – a DFT of a DFT." (Murphy, 2008: 85).

As explained in Murphy (2008: 119), for the extraction of jitter "taking the measure of pitch at one cycle, we then subtract the value of pitch from the previous cycle before dividing by the average pitch for the whole window of analysis". Equation 18 is the calculation for jitter, where k is the cycle number and f_1 is the value of the fundamental frequency found at that cycle.

$$J_k = \frac{|f_{1k} - f_{1k-1}|}{\frac{1}{K}\sum_{k=1}^{K} f_{1k}} \qquad (18)$$

As it happened with jitter, Baken and Orlikoff (2000) describe different shimmer measures, like the *Directional Perturbation Factor,* originated by Hecker and Kreul (1971), the *Amplitude Variability Index (AVI)* of Deal and Emanuel (1978) or *Shimmer in dB* [115](cf. Baken & Orlikoff, 2000: 132-134). Orlikoff and Kahane (1991) found that shimmer (dB) increases significantly with age, while good physical condition (as measured by blood pressure, ventilator capacity, weight, and blood cholesterol levels) or participation in vigorous physical activity is associated with lower shimmer. For our thesis, the *absolute normalized area shimmer* used is estimated as the ratio of the difference between neighbor Glottal Source areas with respect to their average value for the voice segment (Gómez-Vilda et al., 2009:769).

The fourth parameter (*Abs. Norm. Min. Sharp*) is defined as the peak slenderness at the Maximum Flow Declination Rate: negative amplitude of the peak divided by its width. In order to understand how it is extracted, it seems useful to see first how Murphy (2008: 120) explains what the *normalized slope* is:

> […] a measure of the drift between two successive trough points of the glottal source […]. This slope measurement reveals the differences in the maximum pressures produced at the supraglottal lips of the vocal folds during the closure phase. The equation for the slope measurement is given as follows:
>
> $$Sl_k = \frac{\Delta y_k}{\Delta t_k} = \frac{y_{k2} - y_{k1}}{t_{k2} - t_{k1}} \qquad (19)$$
>
> Where t_{k1} and t_{k2} are points on the time axis, y_{k1} and y_{k2} represent the amplitude values per cycle and the blue line represents three separate cycles of the glottal source. We can define a triangle whose base is measured $t_{k2} - t_{k1}$, which we will call Δt, meaning the difference in time. We will further define the height of the triangle as $\Delta y = y_{k2} - y_{k1}$, or the difference in amplitude. Thus, the value for the slope is the differential given by Δy/ Δt. (Murphy, 2008: 120)

In Figure 34 we can see how the slope is evaluated between troughs in the glottal source:

[115] As described in Baken and Orlikoff (2000: 133), "given that the dB scale is based on a ratio of amplitudes it is an easy matter to use it for quantifying shimmer. The ratio need only be that of two contiguous cycles".

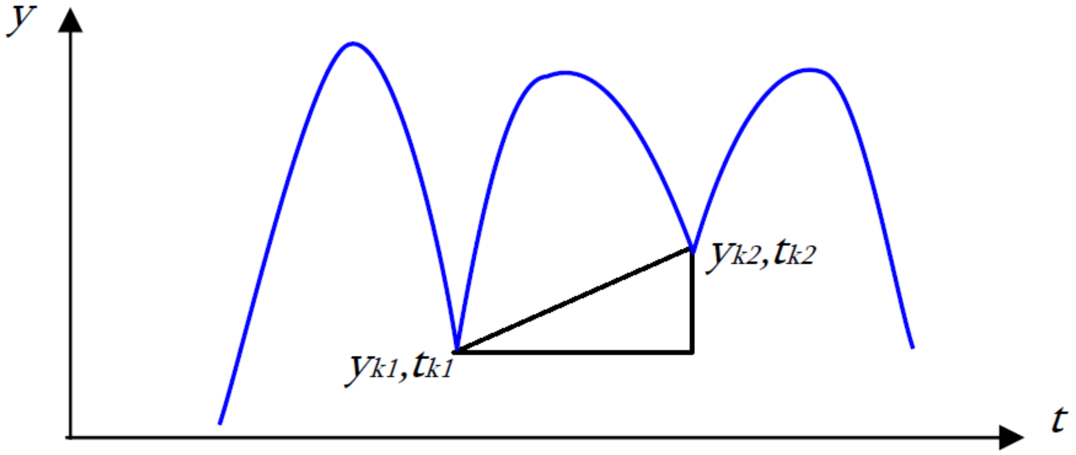


Figure 34. Evaluating the slope between troughs in the glottal source (Retrieved from Murphy, 2008: 120; *Figure 6.13*).

Sharpness, which is the parameter we are interested in (*p4 =Abs. Norm. Min. Sharp*), is, according to Murphy (2008: 120), "a calculation similar to slope and taken at the trough point between two cycles of vibration relating to the ratio of downward pressure to upward pressure at the glottal closure point." The equation for sharpness is as follows:

$$S_k = \frac{\left|y_{k0} - \frac{y_{k1} + y_{k2}}{2}\right|}{t_{k2} - t_{k1}} \qquad (20)$$

According to equation 20, "the difference in time, *Δt*, is given by $t_{k2} – t_{k1}$ and the value for *Δy* is given by the height of the upturned triangle. The value for sharpness is then given by the ratio between the two" (Murphy, 2008: 121). See Figure 35 for a graphical representation.

The fifth parameter (*Noise-Harm. Ratio, NHR*) is the ratio between the energy of the non-harmonic and the harmonic parts of the glottal source power spectral density. Parameter number six (*Muc./AvAc. Energy, MAE*) is the ratio between the energy of the glottal source to average acoustic wave difference and the average acoustic wave.

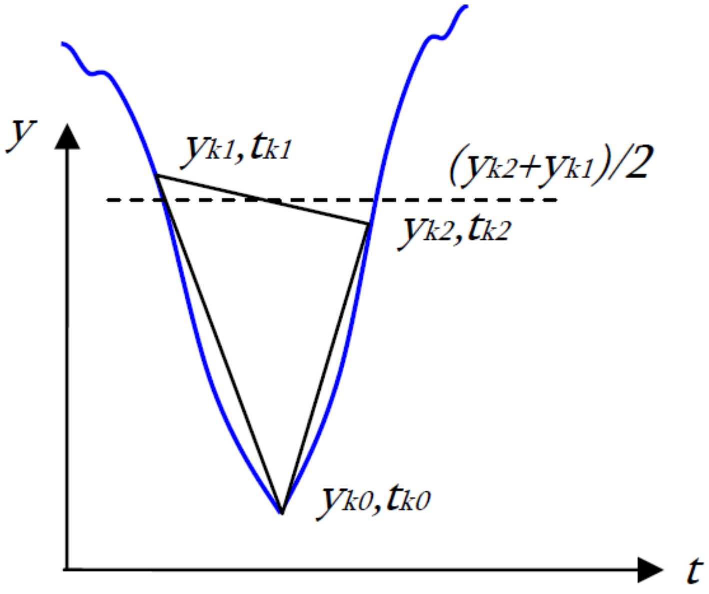


Figure 35. Evaluation of the trough sharpness (Retrieved from Murphy 2008: 121; *Figure 6.14*).

2) Cepstral coefficients of the glottal source power spectral density (p7-20).

This set of parameters aim at capturing the frequency-domain characteristics of the glottal source. According to Mazaira (2014: 111), the estimation of these parameters follows the same processing steps as in the extraction of mel-frequency cepstral coefficients (MFCCs) “except that in this case the input signal is no longer the speech signal, but the glottal residual obtained in the source-tract separation process”. MFCCs date back to the early 1980s in speech recognition, being afterwards adopted in speaker recognition (Davis & Mermelstein, 1980, as cited in Mazaira, 2014: 95). For a description of the steps involved in the MFCC computation, see Mazaira (2014: 95):

> The speech signal continuously changes due to articulatory movements, i.e. the temporal variation of the vocal tract shape during the utterance. This temporal variation, due to vocal tract characteristics, is relatively slow; therefore the speech signal is assumed to remain stationary in short periods of time. In other words, the speech signal can be regarded as having nearly constant characteristics in short periods such as those 20-40 ms in length (Furui, 1989). Once the signal is broken down in short frames, a spectral feature is extracted for each frame. (Mazaira, 2014: 95)

3) Singularities of the glottal source power spectral density profile (p21-34).

The Power Spectral Density (PSD) Profile of the glottal source refers to the envelope of the power spectral density of the glottal source, as explained by Mazaira (2014: 103):

If $S_g(n)$ represents the glottal source, then its DFT (Discrete Fourier Transform) will be defined as:

$$S_g(m) = \sum_{n=0}^{N-1} S_g(n) e^{jm\Omega n\tau} \qquad (21)$$

Where n represents the temporal index of the vector $S_g(n)$ inside a temporal window of N samples, $0 \leq n \leq N-1$, taken every τ seg. The frequency index is given by the integer variable m, which corresponds to an impulse given by mΩ, with frequency resolution Ω.

$$\Omega = \frac{f_s}{2N} \; ; \; f_s = \frac{1}{\tau} \; ; 0 \leq m \leq \frac{N}{2} - 1 \qquad (22)$$

Where f_s represents the sampling frequency and j denotes the imaginary unit. Under these assumptions the power spectral density of the glottal source will be represented by:

$$T_g(m) = \left\|S_g(m)\right\|^2 \qquad (23)$$

In Figure 36, retrieved partially from a figure included in the study of Mazaira (2014: 103), we can observe the PSD of the glottal source for a male voice, while Figure 37 shows the PSD of a male voice segment synchronously evaluated in a phonation cycle, which match the harmonic envelope or the PSD profile.

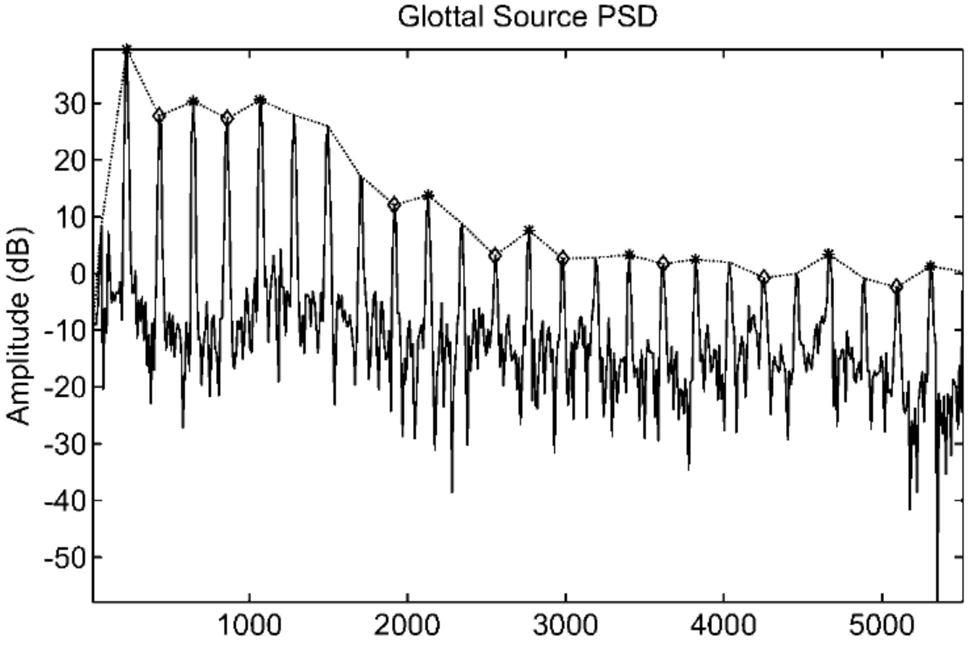


Figure 36. Power Spectral Density (PSD) of the glottal source evaluated over a temporal window which includes multiple glottal cycles. Retrieved from Mazaira (2014: 104; *Figure 2-37*): “The relative maxima of the distribution are marked by the harmonics present in the signal. The interconnection of these maxima is known as Harmonic Envelope or Power Spectral Density Profile” (Mazaira 2014: 104).

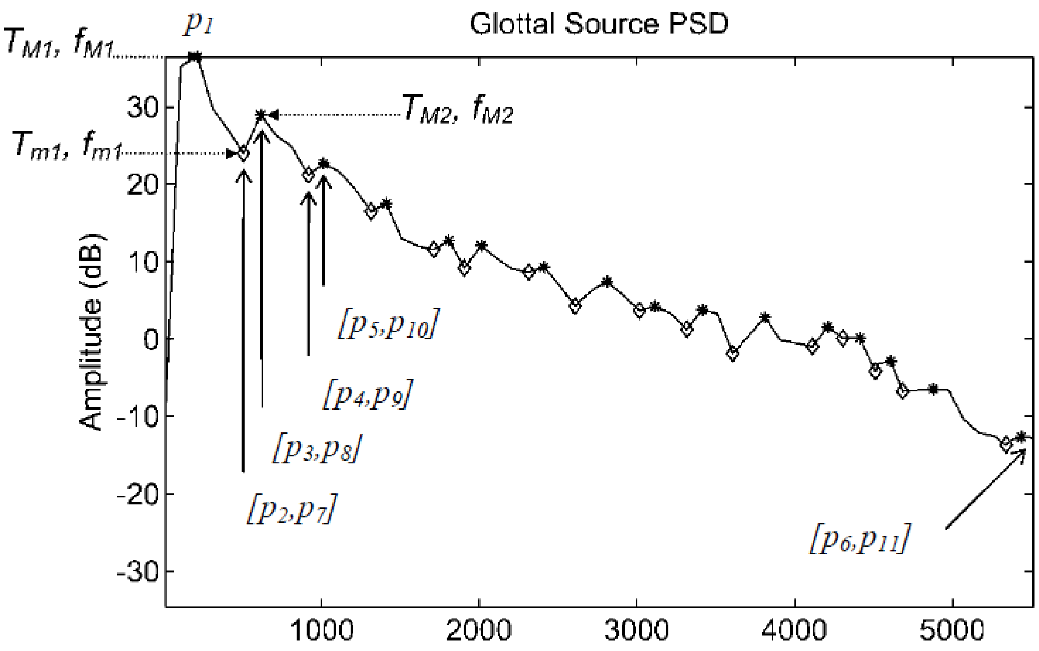


Figure 37. Power spectral density (PSD) of male voice segment synchronously evaluated in a phonation cycle, which match the harmonic envelope or the PSD profile. (Retrieved from Mazaira, 2014: 75; *Figure 2-11*).

In Figure 37 the following singular points have been highlighted:

- *P_1*: maximum PSD value in dB scale (*p21 in Table 19*)
- *P_2*: first minimum value related to the first maximum in dB scale (*p22 in Table 19*)
- *P_3*: second PSD maximum value related to the first maximum in dB scale (*p*23 *in Table 19*)
- *P_4*: second PSD minimum value related to the first maximum in dB scale *(p24 in Table 19*)
- *P_5*: third PSD maximum value related to the first maximum in dB scale (*p25 in Table 19*)
- *P_6*: PSD value at the maximum Nyquist value relative to the first maximum in dB scale (p26 *in Table 19*)
- *P_7*: relative position in frequency of the first minimum (*p28 in Table 19*)
- *P_8*: relative position in frequency of the second maximum (*p29 in Table 19*)
- *P_9*: relative position in frequency of the second minimum (*p30 in Table 19*)
- *P_{10}*: relative position in frequency of the third maximum (*p31 in Table 19*)
- *P_{11}*: relative position in frequency at the end for Nyquist frequency related to the first maximum (*p32 in Table 19*)

Besides the eleven parameters described above, two further parameters are included in this third set (*Singularities of the glottal source power spectral density profile)*. We refer to p33 (MW PSD 1st Min NSF) and p34 (MW PSD 2nd Min NSF), as described by Mazaira (2014: 105):

- *p33*: Slenderness factor of the first "V" profile, which is characterized by the first maximum, the first minimum and the second maximum. It can be defined as in equation 24:

$$\sigma_{m1} = \frac{f_{Mm}(2T_{m1} - T_{M2} - T_{M1})}{2(f_{M2} - f_{M1})} \qquad (24)$$

- *p34*: Slenderness factor of the second "V" profile, which is characterized by the third maximum, the second minimum and the fourth maximum. It can be defined as in equation 25:

$$\sigma_{m2} = \frac{f_{Mm}(2T_{m2} - T_{M3} - T_{M4})}{2(f_{M3} - f_{M4})} \qquad (25)$$

4) *Biomechanical parameters from the vocal fold body and cover dynamic correlates (p35-46)*

The biomechanical parameters from the vocal fold body and cover dynamic correlates (*Average Acoustic Waveform*[116] and *Mucosal Wave Correlate*) consists in estimations of the body dynamic mass, losses and tensions, assigned to *p35*, *p36* and *p37*, the cover equivalent parameters assigned to *p41*, *p42*, and *p43*, and their respective unbalances[117] evaluated cycle by cycle, assigned to *p38*, *p39*, and *p40* (body) and *p44*, *p45* and *p46* (cover).

5) *Time-based Glottal Source coefficients (p47-58).*

The following parameters can be found in Figure 38, representing an example of the glottal cycle temporal analysis of a typical male voice:

- Relative value of the True Recovery Instant (*p47*), and Relative value of the amplitude at the True Recovery Instant (*p52*), represented as *tR1*.

[116] "The Mucosal Wave Correlate (MWC) is a signal derived from the Glottal Source removing the Acoustic Average Wave (AAW) from it (Titze, 1994). The AAW […] can be seen as the Body Dynamic Component, because it may be associated to the one-mass/one-spring equivalent model of the vocal fold body. The residual left when removing the AAW from the Glottal Source signal is designed as the MWC (also the Cover Dynamic Component or CDC), as it can be associated to higher-order oscillation modes of the vocal folds related mainly with the dynamic behavior of the fold cover. Both signals can be considered correlates to the body and cover dynamics, and will be referred as such". (Gómez-Vilda et al., 2009: 761)

[117] Vocal folds are not completely symmetrical. Therefore, in the biomechanical *unbalance* estimates, considerable deviations are found between neighbor phonation cycles. This phenomenon occurs not only in disphonyic voices but also in normophonic ones.

- Relative value of the False Recovery Instant (*p48*), and Relative value of the amplitude at the False Recovery Instant (*p53*), represented as *tR*2.
- Relative value of the False Opening Instant (*p49*), and Relative value of the amplitude at the False Open Instant (*p54*), represented as *tO1*.
- Relative value of the True Opening Instant (*p50*), and Relative value of the amplitude at the True Open Instant (*p55*), represented as *tO2*.
- Relative value of the instant at the glottal source maximum (*p51*), represented as *tM*.

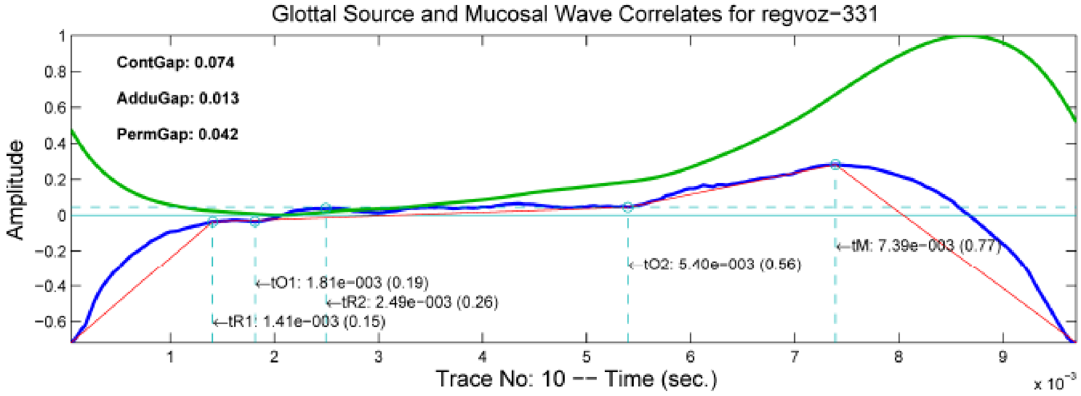


Figure 38. Example of the glottal cycle temporal analysis of a typical male voice. Green: Glottal Flow. Blue: Glottal Source [BioMet®Soft User's Manual 2011: 25-26].

Figure 38 shows one phonation cycle in detail (blue line). The kind of pattern depicted is known as the Liljencrants-Fant cycle and could be described as follows (BioMet®Phon User's Manual, 2014: 7):

- The cycle starts at the closing instant (0), where the pressure drastically drops below 0.

- Due to the elastic nature of the gas column in the vocal tract, a recovery is experienced reaching almost a stable value near 0 at tR1. This is known as the *recovery phase*.

- From the point at tR1 till tO2 (opening instant) the dynamic pressure stays close to 0 (*resting sub-phase*), as the vocal tract is closed by the vocal folds (except for slight escapes of flow represented by the green line).

- From tO2 till the end of the cycle, a burst of flow (green line) is expelled. This is called the *open phase*.

- The pressure rises during the first part of the burst injection till tM (*abduction sub-phase*). At this point the vocal folds begin an approximation (*adduction sub-phase*) to close the vocal tract again. This entails a steady and sharp decay in pressure till the end of the cycle (tC). [118]

Besides the nine coefficients described above, the following time-based parameters are also calculated:

- Rel. Stop Flow Time (*p56*): Ratio between the glottal flow minimum instant and the total glottal cycle duration.
- Rel. Start Flow Time (*p57*): Ratio between the glottal flow start instant and the total glottal cycle duration.
- Rel. Closing Time (*p58*): Ratio between the glottal flow maximum instant and the total glottal cycle duration.

6) *Glottal gap (closure) coefficients (p59-62).*
   - Val. Flow GAP (*p59*): Ratio between the contact gap flow escape and the total glottal flow.
   - Val. Contact GAP (*p60*): Ratio between the escape flow and the total glottal flow during the contact phase.
   - Val. Adduction GAP (*p61*): Ratio between the diminished escape flow and the total glottal flow during the open phase.
   - Val. Permanent GAP (*p62*): Ratio between the escape flow and the total glottal flow during the recovery phase.

7) *Tremor (cyclic) coefficients (p63-68).*
   - 1st Order Cyclic Coefficient (*p63*): First PARCOR (PARtial autoCORelation) coefficient in the equivalent AR model of the unbiased vocal fold body stiffness.
   - 2nd Order Cyclic Coefficient (*p64*): Second PARCOR coefficient in the equivalent AR model of the unbiased vocal fold body stiffness.
   - 3rd Order Cyclic Coefficient (*p65*): Third PARCOR coefficient in the equivalent AR model of the unbiased vocal fold body stiffness.

[118] In a nutshell, the LF cycle can be summarized as (BioMet®Phon User's Manual, 2014:7): closing + closed phase (0-tO2), divided in recovery sub-phase (0-tR1) and steady closure sub-phase (tR1-tO2) + open phase, divided in abduction sub-phase (tO2-tM) and adduction sub-phase (tM-tC).

- Tremor Frequency, in Hz (*p66*): First harmonic of the unbiased vocal fold body stiffness
- Tremor Est. Robustness (*p67*): Proximity to the unity circle of the equivalent AR model first pole of the unbiased vocal fold stiffness.
- Tremor amplitude, rMSA (*p68*): Standard deviation of the unbiased vocal fold stiffness

5.4. Results

5.4.1. Pilot experiment (proof of concept)

Overall results

In a first step, we carried out a pilot experiment with only 20 speakers (12 MZ twins and 8 DZ twins). Having entered 557 tokens of vowel fillers (8.5 per recording session per speaker with a mean duration of 12 milliseconds), we executed *BioMet®ScieProf* (version 7, March 2012) in batch-mode, with the configuration specified above (see Section 5.2.2). As a result of this processing, we obtained (per speaker, session and token) the median and standard deviation of 61 glottal parameters[119] (see Table 20). The features marked in grey were the subset of parameters used for the further calculation of cross-validated LRs. They can be classified in two main groups: (1) classical distortion parameters (i.e. jitter and shimmer measurements) and (2) biomechanical parameters. This second group can be further divided into *body* parameters and *cover* parameters. For both the body and the cover of the vocal folds, the software extracts the following features: dynamic mass, losses, stiffness, and their corresponding unbalances. The reason for selecting these parameters in this first proof of concept was that they are more semantic, in the sense that they are more self-explanatory, in comparison with other parameters like the cepstral coefficients, for example.

Table 20

Parameter set generated by BioMet®ScieProf version 7 – March 2012

Parameter	**Description**
1. Absolute Pitch	Value of the inverse of the glottal cycle in Hz.
2. Abs. Norm. Jitter	Ratio between next cycle duration difference and their mean
3. Abs. Norm. Cl. Shimmer	Ratio between next cycle amplitude difference and their mean
4. Abs. Norm. Sl. Shimmer	Ratio between next cycle sharpness difference and their mean
5. Abs. Norm. Ar. Shimmer	Ratio between next cycle area difference and their mean
6. GNE ratio	Glottal-to-Noise Energy ratio
7. MWC Cepstral 1	1st Cepstral Coefficient of the glottal wave correlate

[119] Note that for this pilot experiment, the version 7 (March 2012) of *BioMet®ScieProf* was used, which includes 61 parameters while in later versions the number of parameters rises to 68.

8. MWC Cepstral 2	2nd Cepstral Coefficient of the glottal wave correlate
9. MWC Cepstral 3	3rd Cepstral Coefficient of the glottal wave correlate
10. MWC Cepstral 4	4th Cepstral Coefficient of the glottal wave correlate
11. MWC Cepstral 5	5th Cepstral Coefficient of the glottal wave correlate
12. MWC Cepstral 6	6th Cepstral Coefficient of the glottal wave correlate
13. MWC Cepstral 7	7th Cepstral Coefficient of the glottal wave correlate
14. MWC Cepstral 8	8th Cepstral Coefficient of the glottal wave correlate
15. MWC Cepstral 9	9th Cepstral Coefficient of the glottal wave correlate
16. MWC Cepstral 10	10th Cepstral Coefficient of the glottal wave correlate
17. MWC Cepstral 11	11th Cepstral Coefficient of the glottal wave correlate
18. MWC Cepstral 12	12th Cepstral Coefficient of the glottal wave correlate
19. MWC Cepstral 13	13th Cepstral Coefficient of the glottal wave correlate
20. MWC Cepstral 14	14th Cepstral Coefficient of the glottal wave correlate
21. MW PSD 1st Max. ABS.	Value of the glottal wave correlate psd at the 1st maximum
22. MW PSD 1st Min. rel.	Id. at the 1st minimum
23. MW PSD 2nd Max. rel.	Id. at the 2nd maximum
24. MW PSD 2nd Min. rel.	Id. at the 2nd minimum
25. MW PSD 4th Max. rel.	Id. at the 3rd maximum
26. MW PSD End Val. rel.	Id. at the final spectral point
27. MW PSD 1st Max. Pos. ABS.	Position of the 1st maximum of the glottal wave correlate psd
28. MW PSD 1st Min. Pos. rel.	Id. at the 1st minimum
29. MW PSD 2nd Max. Pos. rel.	Id. at the 2nd maximum
30. MW PSD 2nd Min. Pos. rel.	Id. at the 2nd minimum
31. MW PSD 4th Max. Pos. rel.	Id. at the 3rd maximum
32. MW PSD End Val. Pos. rel.	Id. at the final spectral point
33. MW PSD 1st Min NSF	1st minimum slenderness
34. MW PSD 2nd Min NSF	2nd minimum slenderness
35. Body Mass	Mass of the vocal fold body
36. Body Losses	Loses of the vocal fold body
37. Body Stiffness	Stiffness of the vocal fold body
38. Body Mass Unbalance	Mass of the vocal fold body unbalance
39. Body Losses Unbalance	Loses of the vocal fold body unbalance
40. Body Stiffness Unbalance	Stiffness of the vocal fold body unbalance
41. Cover Mass	Mass of the vocal fold cover
42. Cover Losses	Loses of the vocal fold cover
43. Cover Stiffness	Stiffness of the vocal fold cover
44. Cover Mass Unbalance	Mass of the vocal fold cover unbalance
45. Cover Losses Unbalance	Loses of the vocal fold cover unbalance
46. Cover Stiffness Unbalance	Stiffness of the vocal fold cover unbalance
47. Rel. Recovery 1 Time	Relative value of the True Recovery Instant
48. Rel. Recovery 2 Time	Relative value of the False Recovery Instant
49. Rel. Open 1 Time	Relative value of the False Opening Instant
50. Rel. Open 2 Time	Relative value of the True Opening Instant
51. Rel. Maximum Amplit. Time	Relative value of the instant at the glottal source maximum
52. Rel. Recov. 1 Amplitude	Relative value of the amplitude at the True Recovery instant
53. Rel. Recov. 2 Amplitude	Relative value of the amplitude at the False Recovery instant
54. Rel. Open 1 Amplitude	Relative value of the amplitude at the False Open instant
55. Rel. Open 2 Amplitude	Relative value of the amplitude at the True Open instant
56. Val. Contact GAP	Residual glottal opening during the Contact Phase (mean)
57. Val. Adduction GAP	Residual glottal opening during the Closing Phase (mean)
58. Val. Permanent GAP	Residual glottal opening during the phonation cycle (mean)
59. 1^{st}. Order Cyclic Coefficient	First order AR coefficient from adaptive estimation
60. 2^{nd}. Order Cyclic Coefficient	Second order AR coefficient from adaptive estimation
61. 3^{rd}. Order Cyclic Coefficient	Third order AR coefficient from adaptive estimation

Note. Parameter set generated by *BioMet*®*ScieProf* (version 7.3 – Sept 2012). The features marked in grey were the subset of parameters used in the proof of concept.

In a further step, we took the median values of those parameters to calculate cross-validated LRs using the MVKD formula described in Aitken and Lucy (2004) and implemented in Morrison (2007). As specified in Figure 39, we made three types of comparisons: (1) same-speaker comparisons, (2) same twin-pair comparisons, and finally (3) different speaker comparisons. In each comparison, we consider three elements: *suspect* (first recording under comparison), *offender* (second recording under comparison), and *background population*. The first session of the suspect is always compared with the second session of the offender. To build the background population, we made it in a cross-validation fashion, taking into account the following aspects:

- Both the suspect and the offender are not present in the background population.
- In the case that either the suspect or the offender, or both of them, are part of a sibling-pair, also their brothers are discarded, in order not to bias the background population.
- Another prerequisite to build the background population with realistic (unbiased) characteristics is never to include the two speakers who make up a twin pair.

1st element of comparison	**-**	**2nd element of comparison**		
Speaker 1 (1st session)	-	Speaker 1 (2nd session)	→	**Same-speaker comparison**
Speaker 1 (1st session)	-	Speaker 2 (2nd session)	→	**Same-pair comparison**
Speaker 1 (1st session)	-	Speaker 3 (2nd session)	}	**Different-speaker comparisons**
Speaker 1 (1st session)	-	Speaker 4 (2nd session)		
Speaker 1 (1st session)	-	Speaker 5 (2nd session)		
…………..		…………		

Figure 39. Types of comparison carried out using the MVKD formula.

The results (LRs) of the speaker comparisons for the MZ pairs are shown in Table 21 and for the DZ pairs in Table 22. The results are divided in five groups, depending on the parameters which were gathered together in the MVKD formula[120]. These were the possible combinations: (1) jitter and shimmer, (2) jitter, shimmer and biomechanical parameters, (3) only-body biomechanical parameters, (4) only-cover biomechanical parameters, or (5) body and cover biomechanical parameters.

[120] The MKVD (Multivariate Kernel Density Formula) has been explained in Chapter 4.

Table 21

Summary of the LRs obtained in the pilot experiment for the MZ pairs

Speakers compared ↓	Jitter & Shimmer	Jitter Shimmer & Biomechanical	Only body (biomech.) parameters	Only cover (biomech.) parameters	Body & cover (biomech.) parameters
1 – 2	1.41	2.88	1.33	4.03	2.23
3 – 4	1.23	23.94	**4.72**	3.70	18.53
5 – 6	**1.47**	**99.53**	**4.68**	**11.41**	**68.73**
7 – 8	1.16	6.15	4.03	9.93	5.53
9 – 10	1.11	**80.89**	3.39	**36.87**	**88.63**
11 – 12	1.28	0.001	0.011	0.003	0.001

Note. In bold we show the highest values per column and grey-shaded appear the lowest values.

Table 22

Summary of the LRs obtained in the pilot experiment for the DZ pairs

Speakers compared ↓	Jitter & Shimmer	Jitter Shimmer & Biomechanical	Only body (biomech.) parameters	Only cover (biomech.) parameters	Body & cover (biomech.) parameters
13 – 14	0.001	4.59E-42	0.003	3.15E-06	8.69E-21
15 – 16	1.27	0.07	1.47	2.19	0.78
17 – 18	1.45	0.17	2.73	0.08	0.18
19 – 20	1.21	0.92	0.29	2.89	1.34

Note. Grey-shaded appear the lowest values.

The LR values greater than 1 indicate that the differences between the voice samples are more likely to occur under the same-speaker hypothesis while the LRs below 1 show that the differences are more likely to occur under the different-speaker hypothesis [see Chapter 3]. As far as the MZ twins are concerned, there are two pairs who obtain the largest LRs, almost independently of the parameters subset under consideration. These are pair 05-06 and pair 09-10. On the contrary, the LRs obtained by pair 11-12 are very low, in comparison with the other MZ twin pairs. In this case, for all the parameter sets considered, except for jitter and shimmer alone, there is more support for the different-origin hypothesis than for the same-origin hypothesis.

Regarding non-identical twins, for all the parameter sets the LR values are mainly around 1, so no decision can be taken on whether they are same speaker or different. The most surprising case is that of pair 13-14, with extremely low LRs for almost all the parameters: *4.59E-42* when considering jitter, shimmer and biomechanical parameters; *3.15E-06* when considering only cover parameters, and *8.69E-21* taking into account body and cover parameters. These results are striking as they imply a very strong dissimilarity of the speakers, which in the case of DZ twins is in disagreement with our hypothesis [see Chapter 3 and Section 5.1].

Except for the cases highlighted before (MZ pair 11-12 and DZ pairs 13-14), the overall trend seems to be that MZ pairs obtain higher LRs than DZ twins (in agreement with our hypothesis) and this occur in all the parameter sets, with the exception of jitter and shimmer alone, for which both type of twins have the same LR values. For an analysis of the causes behind the relatively low values of MZ pair 11-12 as well as for the extremely low values of DZ pair 13-14, we carried out the diagnosis steps described in next subsection.

Diagnosis of specific cases (unexpected results)

The diagnosis methodology aimed at establishing the causes of the voice dissimilarities found in the speakers 11-12 and 13-14 in the pilot experiment included the following steps:

1. *Anamnesis review*: This was carried out in order to check whether some relevant information had been specified by the twins in the questionnaires filled at the first and the second recording session.
2. *Zygosity test:* In view of the answers in the questionnaire of MZ pair 11-12 (see below) and their apparent physical dissimilarity, a DNA test was carried out with the aim of confirming that they were MZ twins. The results of the DNA analysis (*described in Chapter 3*) revealed that they were indeed MZ twin pairs.
3. *Thorough voice examination:* This in-depth voice analysis was performed with a specific tool (*BioMet®PhonProf*), used by speech therapists. The objective was carrying out a visual inspection of the glottal waveform of the speakers, to detect some possible anomalies or idiosyncrasies of their voices.
4. *Error-correction process*: Since the pilot-experiment voice analysis was carried out in a batch-mode and this kind of processing may entail certain errors, like execution (software) errors, pitch errors or inversion errors, a correction-phase was included as the final step of this diagnosis methodology.

We have to note that, besides speaker pair 11-12 and 13-14, we enlarged the study to also include speaker pair 09-10 for two reasons. First, because in the pilot study they obtained high LR values, indicating a strong similarity. This would be just the opposite of what happened to pairs 11-12 and 13-14. For that reason, we considered that comparing the glottal behavior of the similar pair 09-10 with the dissimilar pairs 11-13 and 13-14 could be of interest. Furthermore, speakers 09-10 showed perceptually a harsh voice which could also yield interesting results in the in-depth voice study carried out with *BioMet®PhonProf.*

As far as the anamnesis review of the speakers is concerned, the main information concerning health habits, current (at the date of the recording) health state, age and other voice-related data are summarized in Table 23.

According to the information in Table 23, the voice differences found in the pilot experiment for speakers 11-12 are in agreement with the data filled by them in the questionnaire. On the one hand, there seems to be different smoking habits in one speaker (11) in comparison with his twin (12). The existence of nodules and occasional sore throat in speaker 11, as compared with speaker 12 could also explain the different results yielded in the proof of concept. On the other hand, further interesting information about this MZ twin pair is found in the section 5 and 6 of the questionnaire, related to "preferences and personal traits" and "similarity and confusion between twins", respectively. We include in Table 24 their answers to certain relevant questions.

Table 23

Health-related questionnaire answers for speakers 09-10, 11-12 and 13-14

Speaker	Twin Type	Age	Smoking habits	Health state	Other data
09	MZ	20	Both smoke since 16, more than 6 cigarettes per day	Recovering from flu	Feeling usual throat pain when speaking
10	MZ	20			
11	MZ	33	He smokes more than a packet/day for more than 15 years	Good	Nodules and occasional sore throat
12	MZ	33	He smokes for 6 years, only occasionally		None
13	DZ	36	None of them smoke	Good	Feeling usual throat pain when speaking. He speaks a lot because of his profession. Medical intervention in thyroid and adenoids. Deviated nasal bridge. Hormonal imbalances. Gastric reflux
14	DZ	36			

Table 24

Similarity-related questionnaire answers for speakers 11-12

Question	Speaker 11	Speaker 12
In general, do you like being twins?	Yes	Indifferent
How close is your relationship with your twin? (1-5)	4	3
Do you think you and your twin are very different?	Yes, we are different especially in the physical aspect.	Yes, we are different both in the physical aspect as well as in personality.
How often do people confuse your voice with your twin's?	Very seldom	Never
Note some aspects in which your voice/speech is different or similar from your twin's	Some people have observed that we use the same expressions	We have similar vocabulary

We considered the answers of speakers 11-12 in Table 24 atypical in the sense that they diverge from the average responses given by other MZ pairs. More accurately, the most frequent answer to the first question in Table 24 is a shared yes-reply, with a 4.8[121] mean answer in the second question. The typical response to the third question is "We are very similar in general" and, accordingly, MZ twins usually answer that their voices are confused with high frequency. The last question is of interest because, even though they have noted that they are not very often confused because of their voice, they agree in the fact that they may have similar expressions and vocabulary. Despite the fact that the speakers are not voice experts and their answers should not be considered specially relevant or detailed, it is frequent in other MZ pairs' responses to this question that they also note a similarity in *tone, timbre*, *intensity* or *speech rate*. The shared characteristics pointed out by speakers 11-12 are not influenced by their glottal configurations. Their shared vocabulary and common expressions reflects speaking manner similarities rather than voice likeness. Therefore, the twins' lack of awareness of other voice similarities is in agreement with their dissimilarity results in the pilot experiment.

Apart from the somehow atypical answers of speakers 11-12, mentioned above, the rest of the questionnaire responses fit in with the expected shared environmental conditions for a MZ twin pair: they meet relatively often (once a week) and talk to each other quite often (two-three time per week); they went together to school, being in the same classroom (for 18 years) and have lived in the same home until 30. Their unexpected answers to the above-mentioned questions, especially their lack of physical similarity (also noted by them) led us to check their zygosity, as

[121] This question, as well as the first one (*Do you like being twins?*) does not necessarily imply greater voice similarity, as previous studies have shown [see Chapter 2] but it may have an influence on the specific twin pair that we are considering here.

they were not totally sure of whether they were MZ or DZ twins. They were confirmed as MZ twins (by means of a DNA test as explained in Section 5.2.1), and hence further diagnosis steps followed, with these as well as with the rest of speakers, as we will describe below.

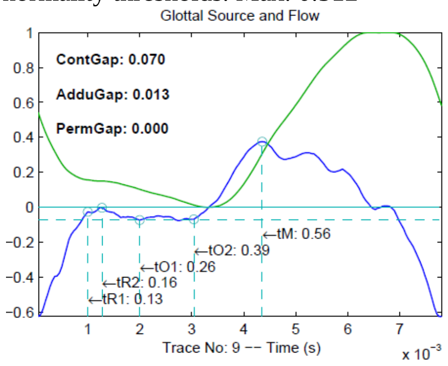
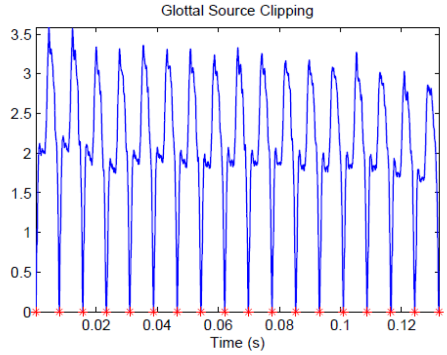
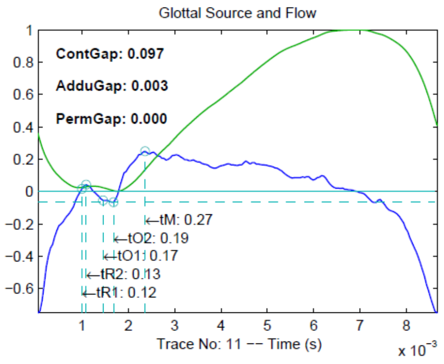
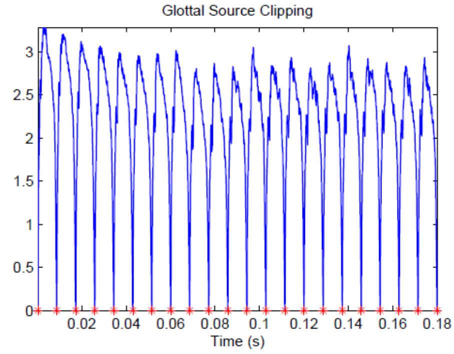
As far as the DZ pair 13-14 is concerned, their results in the proof of concept were especially striking for their dissimilarity values. Their health-related answers in the questionnaire reveal that one of them has undergone or still present special voice-related difficulties. Speaker 14 has been operated from the thyroid and has needed rehabilitation from a speech therapist. His hormonal imbalances and his habitual gastric reflux also affect badly the behavior of his vocal folds. All this could explain the extreme difference found in LRs, as compared with other DZ pairs. For a more accurate voice evaluation, we undertook the examination explained below.

We have also included a "control" MZ pair in this diagnosis process. The selected speakers (09-10) were among the two pairs who ranked highest in the pilot experiment (i.e. with high LRs and therefore, more alike). Their answers to the questionnaire evidence a close similarity in the health habits and conditions of both of them at the time of the recording. The next pages contain a description of a thorough voice examination with the tool *BioMet$^{®}$PhonProf* of all the three speaker pairs being diagnosed in this section.

Table 25

Glottal Source Quality Analysis for Speakers 09-10 (MZ)

	Speaker 09	Speaker 10

Glottal description	- Fast vocal folds' opening, reflected in the steep slope of the glottal cycle at the opening/recovery phase. See Relative value of the True Recovery Instant, tR1 (EEH09102 & EEH09203) -Slow and long-lasting closure (EEH09204). -No clear opening-closing pattern (EEH09103). -Characteristic (recurrent) opening defect, as can be seen in the little lump in the recovery slope (EEH09105) - Air escape (see green line). - Overall hypertension.	- Fast vocal folds' opening, reflected in the steep slope of the glottal cycle at the opening/recovery phase. See Relative value of the True Recovery Instant, tR1 (EEH10102). - Very flat glottal pattern, with almost no complete closure (EEH10108). - Dentate profile, due to opening or closing defects (EEH10108 & EEH10105). - Vibration asymmetry, reflected in the fact that each glottal cycle is different (EEH10203 & EEH10204). - Overall hypertension and air escape.
Parameters below or above normality thresholds[122]	- *Cover Stiffness:* high or very values, reaching over 28000 an over 34000 in two phonations (EEH09102 & EEH09208, respectively).[123] - *Body Stiffness Unbalance:* above normality thresholds in several phonations. Max. 0.034 - *Cover Stiffness Unbalance*: exceeding normality thresholds. Max. 0.512	- *Cover Stiffness:* values exceeding the limits of normality but not as high as his cotwin. Max. 20018.277 (EEH10205). - *Body Stiffness Unbalance:* above normality thresholds in several phonations. Max. 0.078 - *Cover Stiffness Unbalance*: exceeding normality thresholds. Max. 0.297
Typical glottal waveform	  	 

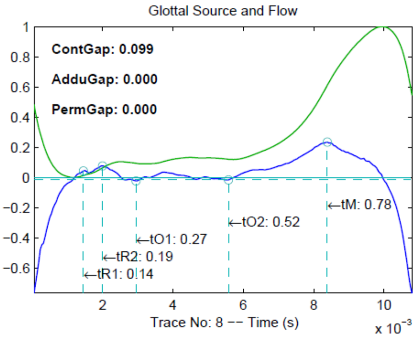
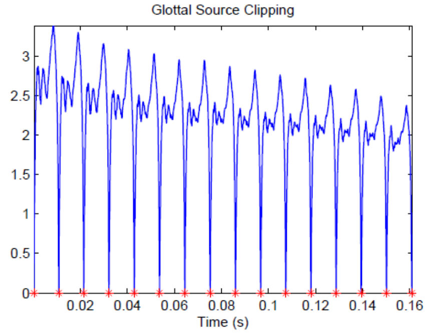
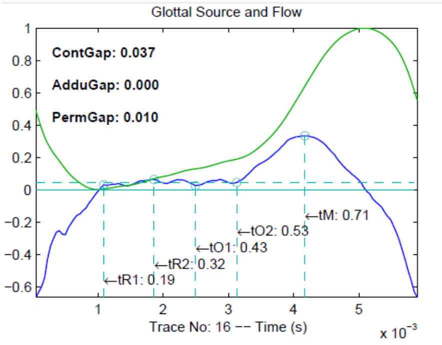
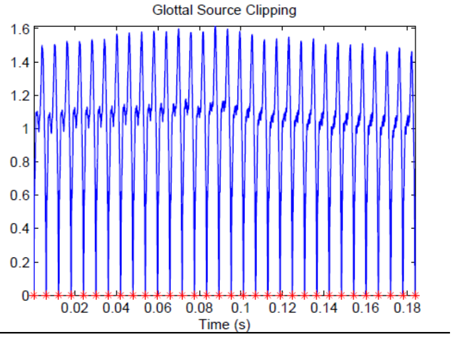
Note. The figures belong to token EEH09206 and EEH10201, respectively.

[122] We have selected for inclusion in this table the (up-to-three) most striking values, either because they are under the minimum value set in the normality threshold or above that threshold. It does not mean that other parameters may be also, for some phonations, below or above.

[123] The threshold of normophony is set in 9378.731 (min) and 12594.897 (max).

Table 26

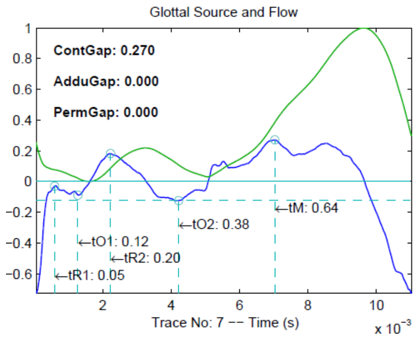
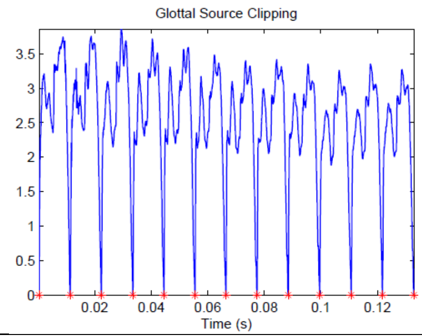
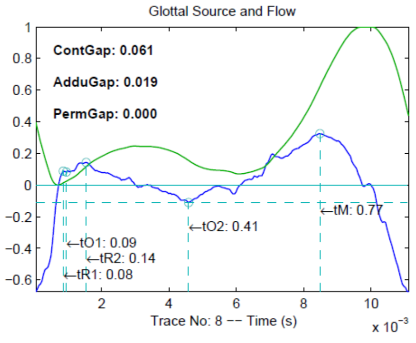
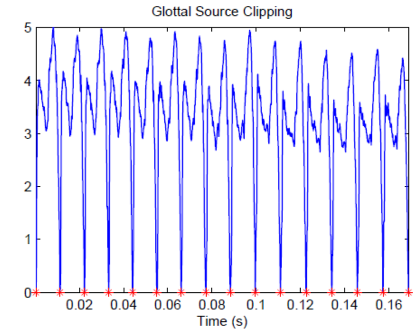
Glottal Source Quality Analysis for Speakers 11-12 (MZ)

	Speaker 11	**Speaker 12**
Glottal description	- Fast vocal folds' opening, reflected in the steep slope of the glottal cycle at the opening/recovery phase. See Relative value of the True Recovery Instant, tR1 (EEH11202 & EEH11104) - Very flat glottal pattern, with almost no complete closure. - No clear opening-closing pattern (EEH11106). - Characteristic (recurrent) defects in the closing phase (EEH11206). - Air escape (see green line). - Overall hypertension.	- Slow recovery phase, only sometimes presenting defects (EEH12105, EEH12111, EEH12201). - Slow opening phase. -Presence of many cycles per phonation.
Parameters below or above normality thresholds	- *Body Mass:* slightly above the threshold: value 0.026 often repeated (EEH11102, EEH11104, EEH11205, EEH11206). - *Body Stiffness Unbalance:* above normality thresholds in several phonations. Max. 0.026	- *Body Stiffness:* values mostly exceeding the limits of normality (around 15000) but occasionally also below (EEH12105).
Typical glottal waveform	 	 

Note. The figures belong to token EEH11206 and EEH12201, respectively. The difference in f_0 between these speakers is especially striking.

Table 27

Glottal Source Quality Analysis for Speakers 13-14 (DZ)

	Speaker 13	Speaker 14
Glottal description	- Very fast recovery phase - Very irregular glottal pattern, with almost no complete closure. - No clear opening-closing pattern. - Very serrate profile, due to opening or closing defects. - Air escape (see green line). - Overall hypertension.	- Very fast recovery phase. - Very irregular glottal pattern, with almost no complete closure. - No clear opening-closing pattern. - Serrate profile, due to the presence of numerous defects. - Air escape (see green line). - Overall hypertension.
Parameters below or above normality thresholds	- *Body Mass:* above the threshold. Max. 0.032 - *Cover Stiffness Unbalance:* greatly above the threshold in several phonations. Max. 0.189 - *Body Stiffness:* below the threshold. Min. 7867 - *Cover Stiffness:* below the threshold. Min. 2621	- *Body Mass:* above the threshold. Max. 0.030 - *Cover Stiffness Unbalance:* greatly above the threshold in several phonations. Max. 0.102 - *Body Mass Unbalance:* greatly above the threshold. Max. 0.045 - *Body Stiffness:* below the threshold. Min. 8400
Typical glottal waveform	 	 

Note. The figures belong to token EEH13107 and EEH14205, respectively.

In view of previous tables (25, 26 and 27), we can draw the following conclusions. As far as the MZ pair 09-10 is concerned, both speakers present hypertension of their vocal folds, which is noticeable mainly through the high values of cover stiffness. This aspect, together with the unbalances existing both for the cover and the body of the vocal folds, point to the possible existence of a voice pathologic behavior. All in all, the results of the glottal examination for these speakers are in line with their medical record, as gathered in our questionnaire. Their similarity in the anamnesis agrees with their similar values in the voice analysis and with their LRs obtained in the pilot experiment.

For the MZ pair 11-12, we can conclude that the phonation of speaker 11 is worse than his cotwin's. There is more hypertension, air escape and adduction defects in his glottal cycles while speaker 12 presents more cycles per phonation and no imbalances in his biomechanical parameters. Upon glottal examination, it is not always easy to distinguish between what may constitute voice pathology and what could be simply idiosyncratic marks of a speaker's voice (i.e. its biometric signature). We have considered the presence of hypertension (reflected in cover stiffness) together with cover and body unbalances as symptoms of pathology (Gómez-Vilda, personal communication). On the contrary, the existence of high values for certain parameters does not necessarily imply that the voice under consideration is pathological. For instance, it is agreed that an excess in the values for body mass can entail a decrease in the body stiffness values, this phenomenon being normophonic. Therefore, for speaker 11 a high body mass is compensated with relatively low body stiffness. In speaker 12 the opposite trend is observed: high body stiffness compensates low body mass. This could be a possible explanation of why these MZ twins were so different in the pilot experiment. This last voice examination has allowed us to find more detailed information about their phonation characteristics than could be observed in the proof of concept. Furthermore, their different glottal behavior agrees with their anamnesis.

As far as the DZ pair 13-14 is concerned, they show very similar glottal patterns as well as similar values for the biomechanical parameters under consideration. This is not in accordance with their different medical records and with the dissimilarity values in the pilot experiment. According to his medical precedents, the results of the voice examination were only expected for speaker 14. However, the unbalances, irregular glottal pattern and hypertension of both speakers point to the potential existence of voice pathology in both speakers 13 and 14.

As was specified before, a last step follows in our methodology aimed at correcting errors in the system for analysis. For this step, we first run the program *BioMet*$^{®}$*ScieProf* now with all the 54 speakers making up our corpus. We did it in batch-processing mode and detected three types of errors or artifacts, which were then corrected following the instructions in the manual (*BioMet®Soft* User's Manual, 2010).

1) *Pitch errors:* It could be the case that the pitch was either too low or too high. This requires adjusting or changing certain execution settings in the GUI of the program.

2) *Inversion errors:* In these cases, the glottal source shows the glottal closure spikes (saliences) downside-up. This requires the activation of the button “Invert Glottal Source (sw22)”. [124]
3) *Software artifacts:* These kinds of errors are “execution errors” and they appear in a text file saved together with the mat files and the rest of the output files. They will be listed first in Table 28. The rest of errors have to be detected one by one by examining the figure with representations of the glottal source and mucosal wave correlated and the figure with the glottal source clipping. The recordings containing these errors will be listed afterwards.

After all the detected errors had been corrected, we proceeded to conduct a forensic comparison using the tool *BioMet*®*ForeProf* and following the guidelines described in next section.

Table 28

Classification of errors according to one of the following types: inverted signal, low or high pitch and software errors.

Type of error	**Recording**
Software	05109
Software	13106
Software	21107
Software	25102
Software	28108
Software	40201
Low pitch	01212
High pitch	02110
High pitch	03104
Low pitch	04104
Inversion	05101
Inversion	05109
Inversion	11201
High pitch	13106
Low pitch	13108
Inversion	14108
Low pitch	16102
Inversion	16202
Low pitch	19103
Inversion	20105

[124] As stated in the manual (*BioMet®Soft* User’s Manual, 2010: 26) “this situation is infrequent in normophonic voicing given the accuracy of the glottal source reconstruction algorithms embedded in Glottex®. Nevertheless the visual inspection of the glottal source is recommended to detect and correct these rare cases” It follows that “This is a relatively uncommon error. Although the clipping function has a robust sign detector to apply the proper polarity to the glottal wave, in certain degenerate cases, or when batch processing fragments come from very different speakers in age or gender, it is possible that the trace may appear inverted. One of the symptoms is that the acute spikes (saliences) of the glottal source appear upside.” (*BioMet®Soft* User’s Manual, 2010: 27).

Inversion	21105
Low pitch	21107
Low pitch	21108
High pitch	21204
Low pitch	23204
Low pitch	24106
Low pitch	24111
High pitch	27205
Low pitch	29102
Low pitch	29103
Low pitch	29105
Inversion	39103
Inversion	39107
Inversion	40201
High pitch	41204
High pitch	42107
High pitch	42110
Inversion	43106
Inversion	43110
Inversion	44102
High pitch	45202
Low pitch	46102
Low pitch	49203
Inversion	51101
Inversion	51104
Inversion	51106
Low pitch	52207
Low pitch	52206

5.4.2. Final voice analysis and forensic comparison

After the pilot experiment described above and after having carried out a diagnosis consisting of several stages in order to find out possible explanations for the unexpected results in two twin pairs, we analyzed the whole speaker population and obtained a set of 68 parameters per recording (vowel filler), session and speaker.

With the aim of conducting a forensic comparison, we created a feature vector of 68 parameters given as x_{sij}, where *s* refers to the subject, *i* is for the session, and *j* for the filler. Pair-wise parameter matching experiments were carried out by likelihood ratio contrasts used in forensic voice matching (Ariyaeeinia et al,. 2008). The test is based on two-hypotheses contrasts: that the conditional probability between voice samples $Z_a=\{x_{aij}\}$ and $Z_b=\{x_{bij}\}$ (from the two subjects under test, a and b) is larger than the conditional probability of each subject relative to a Reference Speaker's Model Γ_R in terms of logarithmic likelihood

$$\lambda_{ab} = \log\left(\frac{p(Z_b|\Gamma_a)}{\sqrt{p(Z_a|\Gamma_R)p(Z_b|\Gamma_R)}}\right) \tag{26}$$

where conditional probabilities have been evaluated using Gaussian Mixture Models (Γ_a, Γ_b, Γ_R) as:

$$p(Z_b|\Gamma_a) = \Gamma_a(Z_b) \tag{27}$$

$$p(Z_a|\Gamma_R) = \Gamma_R(Z_a) \tag{28}$$

$$p(Z_b|\Gamma_R) = \Gamma_R(Z_b) \tag{29}$$

Following this background, the forensic voice evaluation framework is a two-step process:

- Step 1. *Model Generation*. A model representative of the normative population set considered (male subjects between 18-52 years-old) was created on recordings Z_R={x_{Rjk}}, as a Gaussian Mixture Model Γ_R={w_R, μ_R, C_R}, w_R, μ_R and C_R being the set of weights, averages and covariance matrices associated to each Gaussian Probability Distribution in the set.
- Step 2. *Score Evaluation*. The material under evaluation will be composed of different parameterized voice samples in matrix form Z_a={x_{aj}}, where $1 \leq j \leq J_a$ is the sample index, each sample being a vector x_{aj}={$x_{aj1}\ldots x_{ajM}$} from vowel-like segments conveniently parameterized. Similarly, the set of the correspondent speaker to be matched will be given as Z_b={x_{bj}}, where $1 \leq j \leq J_b$ will be the sample index, each sample being a vector x_{bj}={$x_{bj1}\ldots x_{bjM}$}.

The conditioned probability of a sample from speaker *a* x_{aj} matching speaker *b* will be estimated as

$$P(x_{bj}|\Gamma_a) = \frac{1}{(2\pi)^{M/2}|C_a|^Q} \cdot e^{-1/2(x_{bj}-\mu_a)^T C_S^{-1}(x_{bj}-\mu_a)} \tag{30}$$

Similarly the conditioned probability of a sample from speaker *a* matching the Reference Model will be

$$P(x_{aj}|\Gamma_R) = \frac{1}{(2\pi)^{M/2}|C_R|^Q} \cdot e^{-1/2(x_{aj}-\mu_R)^T C_S^{-1}(x_{aj}-\mu_R)} \tag{31}$$

Finally the conditioned probability of a sample from speaker b matching the Reference Model will be

$$P\left(x_{bj}|\Gamma_R\right) = \frac{1}{(2\pi)^{M/2}|C_R|^Q} \cdot e^{-1/2\left(x_{bj}-\mu_R\right)^T C_S^{-1}\left(x_{bj}-\mu_R\right)} \tag{32}$$

A full description of this methodology is given in Gómez-Vilda et al. (2012)

The participants were distributed in: 24 subjects were MZ siblings in 12 pairs (numbered as 01-02, 03-04, 05-06, 07-08, 09-10, 11-12, 33-34, 35-36, 37-38, 39-40, 41-42 and 43-44), 10 subjects were DZ siblings in 5 pairs (corresponding to speakers numbered as 13-14, 15-16, 17-18, 19-29 and 45-46), 8 subjects were non-twin brothers (B) in 4 pairs (numbered as 21-22, 23-24, 47-48 and 49-50) and 12 subjects were not known to have any familiar relationship (unrelated speakers: US), grouped also as 6 pairs (25-26, 27-28, 29-30, 31-32, 51-52 and 53-54).

Speakers were matched in: a) different-session intra-speaker tests (I: intra-speakers), b) inter-speaker tests (O: inter-speakers). A priori expectations assume that MZ should show the largest LLRs, followed by DZ, then by non-twin siblings; non-related speakers are expected to show the lowest LLRs. The baseline is defined by a reference background set composed of 20 speakers (set B). Scores are qualified as Strong Likeness if above 1, Weak Likeness if between 1 and -1 and Unlikeness if below -1 (see Figure 40). The hypotheses tested were the following:

H_1 Intra-speaker tests should show large LLRs.
H_2 MZ inter-speaker tests should show large LLRs.
H_3 DZ inter-speaker tests should show also large LLRs although not as large as H_1 or H_2.
H_4 B inter-speaker tests should show LLRs at least over the background baseline (fixed at the LLR value λ = -10).
H_5 US inter-speaker tests should show LLR's aligned with the background baseline.

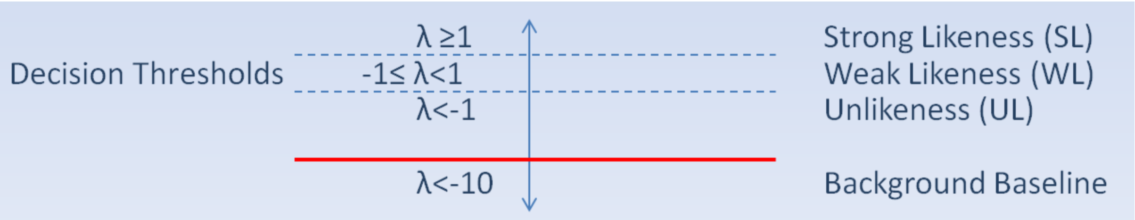


Figure 40. Decision Thresholds: Scores are qualified as Strong Likeness if above 1, Weak Likeness if between 1 and -1 and Unlikeness if below -1. The background baseline is fixed at λ = -10

According to the decision thresholds described above, our five hypotheses could also be represented as:

H_1: MZ(I), DZ(I), B(I), US(I) → $\lambda > -1$

H_2: MZ(O) → $\lambda > -1$

H_3: DZ(O) → $\lambda > -10$

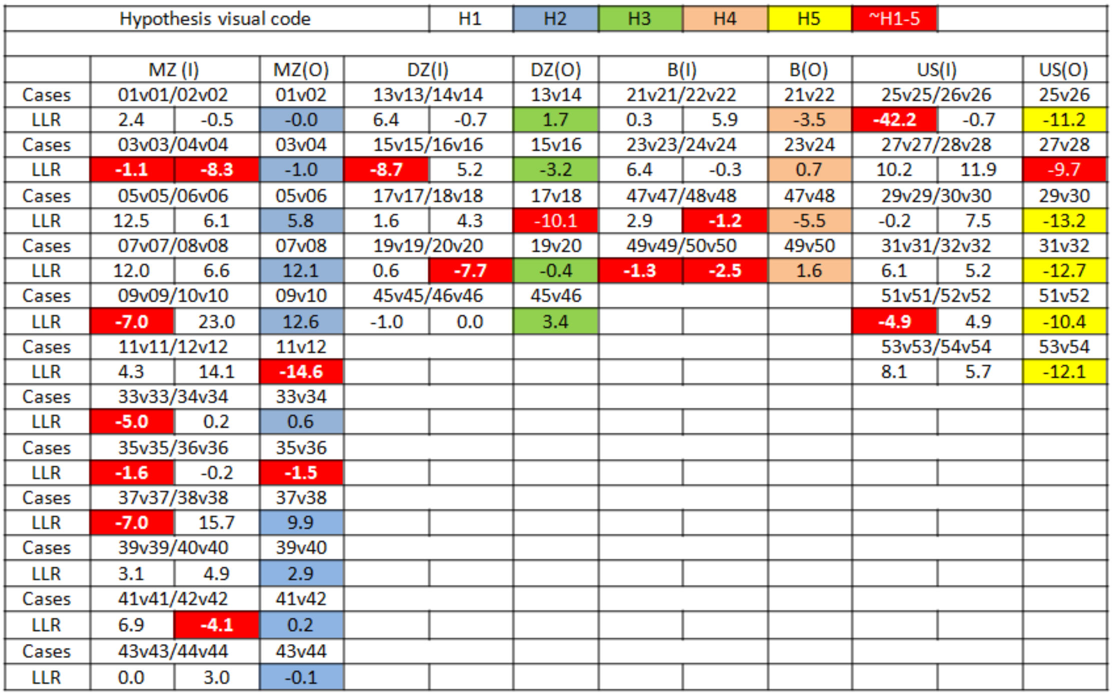
H_4: B(O) → $\lambda > -10$

H_5: US(O) → $\lambda < -10$

The results of the matching tests appear in Table 29. The results contradicting the strongest hypotheses (H_1 and H_2) are marked in bold. Five speakers out of the total of 54 appear to be in the limit of H_1 (03, 35, 48, 49 and 50), eight speakers show strong intra-speaker dissimilarity (04, 09, 15, 20, 33, 37, 42 and 51), and one shows very strong self-dissimilarity (25), therefore 14 out of 54 do not fulfill H_1. The rest of the speakers show weak or strong self-similarity in inter-session tests, fulfilling H_1. Regarding H_2 we find two out of 12 pairs not fulfilling it (11 vs 12, 35 vs 36). Hypothesis 3 is not fulfilled in one out of five pairs (17 vs 18). H_4 is fulfilled in all four cases. Only one pair of unrelated subjects is slightly over the baseline (27 vs 28) out of 5 cases fulfilling H_5.

Table 29

Summary of the results for the different tests



Hypothesis visual code					H1	H2	H3	H4	H5	~H1-5			
	MZ (I)		MZ(O)	DZ(I)		DZ(O)	B(I)		B(O)	US(I)		US(O)
Cases	01v01/02v02		01v02	13v13/14v14		13v14	21v21/22v22		21v22	25v25/26v26		25v26
LLR	2.4	-0.5	-0.0	6.4	-0.7	1.7	0.3	5.9	-3.5	**-42.2**	-0.7	-11.2
Cases	03v03/04v04		03v04	15v15/16v16		15v16	23v23/24v24		23v24	27v27/28v28		27v28
LLR	**-1.1**	**-8.3**	-1.0	**-8.7**	5.2	-3.2	6.4	-0.3	0.7	10.2	11.9	-9.7
Cases	05v05/06v06		05v06	17v17/18v18		17v18	47v47/48v48		47v48	29v29/30v30		29v30
LLR	12.5	6.1	5.8	1.6	4.3	-10.1	2.9	**-1.2**	-5.5	-0.2	7.5	-13.2
Cases	07v07/08v08		07v08	19v19/20v20		19v20	49v49/50v50		49v50	31v31/32v32		31v32
LLR	12.0	6.6	12.1	0.6	**-7.7**	-0.4	**-1.3**	**-2.5**	1.6	6.1	5.2	-12.7
Cases	09v09/10v10		09v10	45v45/46v46		45v46				51v51/52v52		51v52
LLR	**-7.0**	23.0	12.6	-1.0	0.0	3.4				**-4.9**	4.9	-10.4
Cases	11v11/12v12		11v12							53v53/54v54		53v54
LLR	4.3	14.1	**-14.6**							8.1	5.7	-12.1
Cases	33v33/34v34		33v34									
LLR	**-5.0**	0.2	0.6									
Cases	35v35/36v36		35v36									
LLR	**-1.6**	-0.2	**-1.5**									
Cases	37v37/38v38		37v38									
LLR	**-7.0**	15.7	9.9									
Cases	39v39/40v40		39v40									
LLR	3.1	4.9	2.9									
Cases	41v41/42v42		41v42									
LLR	6.9	**-4.1**	0.2									
Cases	43v43/44v44		43v44									
LLR	0.0	3.0	-0.1									

Note. Summary of the results for the different tests. MZ: Monozygotic twins; DZ: Dizygotic twins; B: Brothers; US: Unrelated Speakers; (I): intra-speaker tests; (O): inter-speaker tests. Divided columns are used for each pair member. Cases: xxvyy means speaker xx versus speaker yy. Matches: Strong Likeness (SL): $\lambda \geq 1$; Weak Likeness (WL): $-1 \leq \lambda < 1$; Unlikeness (UL): $\lambda < -1$. In bold: results contrary to hypotheses H_1 and H_2 (MZ should be SL or WL, Intra-speaker's should be SL or WL). H_1: MZ(I), DZ(I), B(I), US(I) → $\lambda > -1$; H_2: MZ(O) → $\lambda > -1$; H_3: DZ(O) → $\lambda > -10$; H_4: B(O) → $\lambda > -10$; H_5: US(O) → $\lambda < -10$.

In view of the results, the degree of hypotheses corroboration can therefore be summarized as:

H_1: 40/54, but relaxing the threshold[125]*, it could be 45/54=5/6=83.3%*

H_2: 10/12=5/6=83.3%

H_3: 4/5=80%

H_4: 4/4=100%

H_5: 5/6=83.3%

Parameter Discrimination Capability

With the aim of finding which parameters show the best discrimination capability, we have selected the three most relevant parameters in each of the 81 comparisons carried out with *BioMet®ForeProf*, resulting from 54 speakers compared with themselves –intra-speaker comparison– and with their pairs[126]. That is, we selected the parameters ranking highest in the *relevance listing* made by *BioMet®ForeProf*, using the Fisher's Ratio[127]. These same three features are the ones which allow the representation of the *3D Original Model, Control and Test Data Sets vs 3D Template* (see Figure 32) and also the *Density Distribution of 3D Parameters* (see Figure 33).

The table with all the data (i.e. the three most relevant parameters per comparison) can be found in appendix E. This kind of information allows us to create Figure 41, where the most relevant parameters outstand above the others. As it can be observed, the five parameters with a higher number of hits (occurrences) are:

1. p6 (*Muc./AvAc. Energy, MAE*), with 18 hits.
2. p32 (*MW PSD End Val. Pos. rel.*), with 14 hits.
3. p8 (*MWC Cepstral 2*), with 12 hits.
4. p21 (*MW PSD 1st Max. ABS.*) and p59 (*Val. Flow GAP*); both with 11 hits.

[125] As we said before, 14 out of 54 do not fulfill H_1. However, five speakers out of the total of 54 appear to be in the limit of H_1 (03, 35, 48, 49 and 50). That is why we say that a "relaxing threshold" would be 45/54.
[126] Their pairs were their siblings in the case of MZ, DZ and brothers, and their speaking partners in the case of the reference population.
[127] Fisher's ratio is a measure for (linear) discriminating power of some variable.

Of this list, two parameters belong to the “singularities of the glottal source PSD” (subset 3: *p21* and *p32*), one to the “fundamental frequency and distortion parameters” (subset 1: *p6*), another one is a cepstral coefficient of the glottal source PSD (subset 2: *p8*) and a further one belongs to the glottal gap coefficients (subset 6: *p59*). For a full description of these parameters, see Section 5.3 and Table 19.

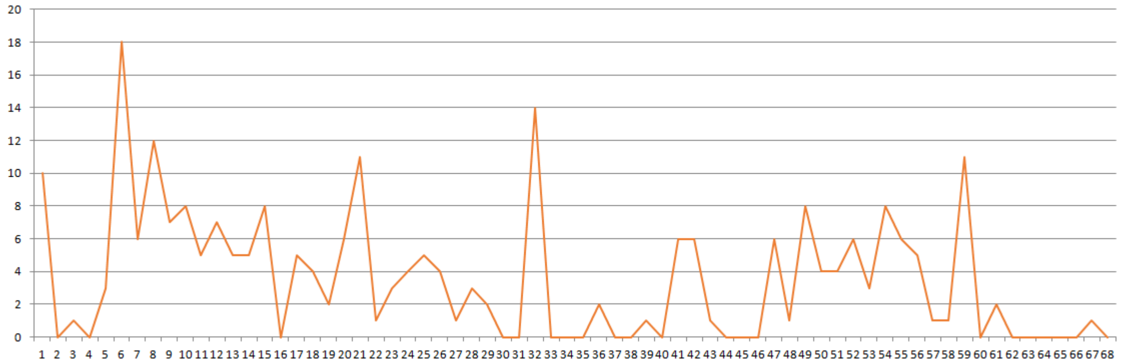


Figure 41. Line graph showing in the x axis the 68 parameters analyzed and in the y axis the total number of hits, regardless of the type of comparison (intra- or inter-speaker) and the type of speaker (MZ, DZ, B or US).

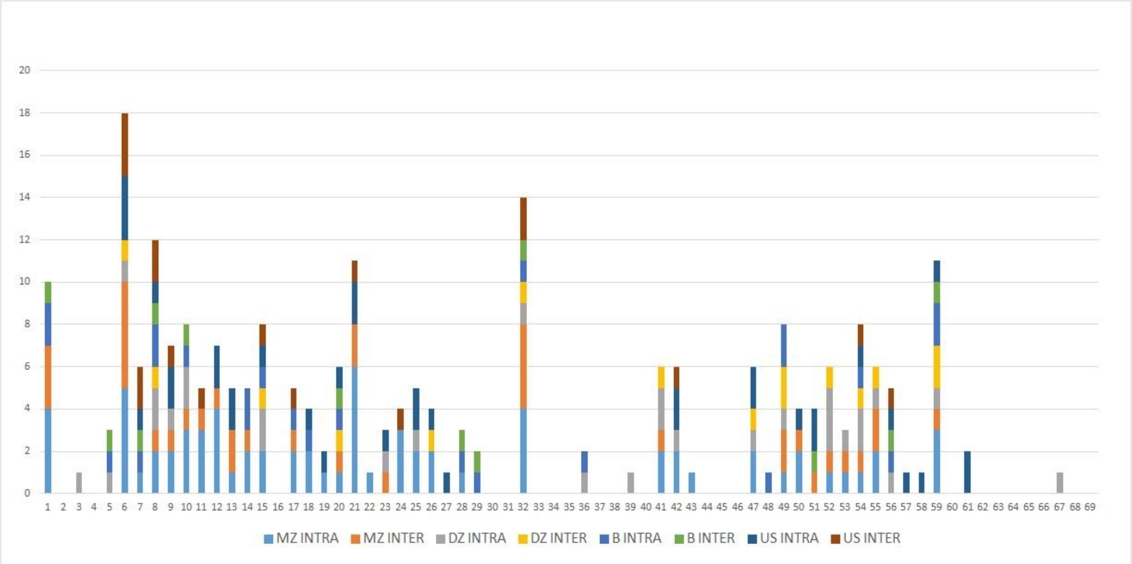


Figure 42. Bar chart showing in the x axis the 68 parameters analyzed and in the y axis the total number of hits, per type of speaker (MZ, DZ, B and US) and per type of comparison (intra- and inter-speaker comparison).

Figure 42 shows that the parameters with most discriminatory potential appear in all types of comparisons, regardless of the type of speaker or type of comparison. Even though the parameter may look equally distributed, it should be taken into account that the total number of speakers per type is uneven.

In order to investigate which parameters were the most relevant in their corresponding subset, Tables 30-36 were created. In these, the total number of hits per parameter and the specific comparisons where these parameters are relevant are indicated. Finally in the last column, a percentage[128] has been calculated to indicate how many of the comparisons are intra-speaker or inter-speaker. The most relevant parameters per subset are marked in grey.

According to this, for the first subset (Table 30), p6 (*Muc./AvAc. Energy, MAE*) is the most relevant parameter, followed by p1 (*fundamental frequency*). For the parameter subset 2 (Table 31), p8 (*MWC Cepstral 2*) slightly outstands among the other cepstral parameters, but p9, p10, p11 and p15 (*MWC Cepstral 3, 4, 6* and *9*, correspondingly) seem to be also relevant. For the parameter subset 3 (Table 32), the most relevant parameters are p21 (*MW PSD 1st Max. ABS.*) and p32 (*MW PSD End Val. Pos. rel*). Interestingly, these two parameters are among the most relevant parameters in the top 6-parameter list that we mentioned above. For the parameter subset 4, the most outstanding parameters are p41 (*cover mass*) and p42 (*cover losses*), although with a relatively small number of hits: 6 each parameter. For the parameter subset 5, the most relevant parameters seem to be p49 (*Rel. Open 1 Time*) and p54 (*Rel. Open 1 Amplitude*). For the parameter subset 6, the most relevant parameter is p59 (*Val. Flow GAP*), which turns out to be one of the most relevant parameters in general. Finally, for the last parameter subset, none of the parameters are outstanding. Only p67 (*Tremor Est. Robustness*) appears once, for the intra-speaker comparison of speaker 20.

Table 30

Parameter Subset 1: Absolute pitch and distortion parameters (p1-6)

Parameter	**Hits**	**Cases**	**Percentage** **%Intra/%Inter**
Fundamental frequency (p1)	10	01v02; 09v09; 11v11; 11v12; 01v01; 23v23; 43v44; 21v21; 21v22; 43v43	60%-40%
Abs. Norm. Jitter (p2)	0		
Abs. Norm. Ar. Shimmer (p3)	1	20v20	
Abs. Norm. Min. Sharp (p4)	0		
Noise-Harm. Ratio (p5)	3	18v18; 49v50; 49v49	
Muc./AvAc. Energy (MAE) (p6)	18	09v10; 29v30; 33v34; 38v38; 41v42; 28v28; 05v05; 05v06; 06v06; 10v10; 17v17; 17v18; 25v26; 37v37; 37v38; 52v52; 53v53; 53v54	50%-50%

[128] The percentage has been only calculated for the parameters with a high number of occurrences.

Note. For the parameter subset 1, information about the total number of hits per parameter, specific comparisons where these parameters are relevant (column *cases*) and inter-/intra- speaker percentage. The most relevant parameters per subset are marked in grey.

Table 31

Parameter Subset 2: Cepstral coefficients of the glottal source power spectral density (p7-20)

Parameter	**Hits**	**Cases**	**Percentage %Intra/%Inter**
MWC Cepstral 1 (p7)	6	07v07; 21v21; 21v22; 27v27; 27v28; 31v32	50%-50%
MWC Cepstral 2 (p8)	12	13v13; 13v14; 23v23; 23v24; 35v36; 51v52; 14v14; 27v28; 34v34; 40v40; 48v48; 51v51;	58% - 42%
MWC Cepstral 3 (p9)	7	14v14; 32v32; 37v38; 40v40; 42v42; 51v51; 51v52;	72% - 28%
MWC Cepstral 4 (p10)	8	01v01; 04v04; 16v16; 49v49; 01v02; 08v08; 19v19; 49v50;	75% - 25%
MWC Cepstral 5 (p11)	5	25v26; 34v34; 41v41; 41v42; 42v42;	60% - 40%
MWC Cepstral 6 (p12)	7	27v27; 39v39; 54v54; 03v03; 03v04; 04v04; 07v07	85.7% - 14.3%
MWC Cepstral 7 (p13)	5	27v27; 39v39; 39v40; 53v53; 03v04	60% - 40%
MWC Cepstral 8 (p14)	5	07v08; 08v08; 22v22; 44v44; 50v50	80% - 20%
MWC Cepstral 9 (p15)	8	15v15; 15v16; 16v16; 26v26; 33v33; 34v34; 50v50; 53v54	75% - 25%
MWC Cepstral 10 (p16)	0		
MWC Cepstral 11 (p17)	5	22v22; 25v26; 35v35; 35v36; 43v43	60% - 40%
MWC Cepstral 12 (p18)	4	11v11; 22v22; 54v54; 06v06	100% - 0%
MWC Cepstral 13 (p19)	2	35v35; 54v54	100% - 0%
MWC Cepstral 14 (p20)	6	19v20; 26v26; 35v35; 35v36; 49v49; 49v50	50%-50%

Note. For the parameter subset 2, information about the total number of hits per parameter, specific comparisons where these parameters are relevant (column *cases*) and inter-/intra- speaker percentage. The most relevant parameters per subset are marked in grey.

Table 32

Parameter Subset 3: Singularities of the glottal source power spectral density -profile (p21-34)

Parameter	Hits	Cases	Percentage %Intra/%Inter
MW PSD 1st Max. ABS. (p21)	11	09v09; 09v10; 10v10; 31v32; 37v37; 38v38; 52v52; 07v07; 31v31; 37v38; 44v44	73% - 27%
MW PSD 1st Min. rel. (p22)	1	42v42	
MW PSD 2nd Min. rel. (p23)	3	03v04; 15v15; 29v29	
MW PSD 2nd Max. rel. (p24)	4	29v30; 33v33; 02v02; 03v03	
MW PSD 4th Max. rel. (p25)	5	03v03; 14v14; 30v30; 32v32; 02v02	
MW PSD End Val. rel. (p26)	4	19v20; 30v30; 37v37; 41v41	
MW PSD 1st Max. Pos. ABS. (p27)	1	31v31	
MW PSD 1st Min. Pos. rel. (p28)	3	36v36; 47v47; 47v48	
MW PSD 2nd Max. Pos. rel. (p29)	2	47v48; 47v47	
MW PSD 2nd Min. Pos. rel. (p30)	0		
MW PSD 4th Max. Pos. rel. (p31)	0		
MW PSD End Val. Pos. rel. (p32)	14	43v43; 43v44; 44v44; 07v08; 13v13; 13v14; 21v21; 21v22; 01v01; 01v02; 11v12; 12v12; 27v28; 51v52	43% - 57%
MW PSD 1st Min. NSF (p33)	0		
MW PSD 2nd Min. NSF (p34)	0		

Note. For the parameter subset 3, information about the total number of hits per parameter, specific comparisons where these parameters are relevant (column *cases*) and inter-/intra- speaker percentage. The most relevant parameters per subset are marked in grey.

Table 33

Parameter Subset 4: Biomechanical estimates of vocal fold mass, tension and losses (p35-46)

Parameter	**Hits**	**Cases**	**Percentage %Intra/%Inter**
Body Mass (p35)	0		
Body Losses (p36)	2	19v19; 24v24	
Body Stiffness (p37)	0		
Body Mass Unbalance (p38)	0		
Body Losses Unbalance (p39)	1	20v20	
Body Stiffness Unbalance (p40)	0		
Cover Mass (p41)	6	07v08; 11v11; 17v17; 17v18; 08v08; 16v16;	67% - 33%
Cover Losses (p42)	6	38v38, 53v53; 53v54; 10v10; 17v17; 28v28	83% - 17%
Cover Stiffness (p43)	1	09v09	
Cover Mass Unbalance (p44)	0		
Cover Losses Unbalance (p45)	0		
Cover Stiffness Unbalance (p46)	0		

Note. For the parameter subset 4, information about the total number of hits per parameter, specific comparisons where these parameters are relevant (column *cases*) and inter-/intra- speaker percentage. The most relevant parameters per subset are marked in grey.

Table 34

Parameter Subset 5: Time-based Glottal Source coefficients (p47-58)

Parameter	**Hits**	**Cases**	**Percentage %Intra/%Inter**
Rel. Recovery 1 Time (p47)	6	15v15; 15v16; 26v26; 32v32; 33v33; 02v02	83.3% - 16.7%
Rel. Recovery 2 Time (p48)	1	24v24	
Rel. Open 1 Time (p49)	8	19v20; 48v48; 50v50; 11v12; 12v12; 15v16; 19v19; 43v44;	50%-50%
Rel. Open 2 Time (p50)	4	30v30; 41v41; 41v42; 06v06	
Rel. Maximum Amplit. Time (p51)	4	09v10; 25v25; 47v48; 52v52	
Rel. Recov. 1 Amplitude (p52)	6	05v05; 18v18; 33v34; 45v45; 45v46; 46v46	66.7% - 33.3%
Rel. Recov. 2 Amplitude (p53)	3	05v05; 05v06; 18v18	
Rel. Open 1 Amplitude (p54)	8	31v31; 36v36; 45v46; 05v06; 31v32; 45v45; 46v46; 47v47;	63%-37%
Rel. Open 2 Amplitude (p55)	6	36v36; 45v45; 45v46; 33v34; 11v11; 39v40	50% - 50%
Rel. Stop Flow Time (p56)	5	23v24; 24v24; 29v29; 29v30; 46v46	60% - 40%
Rel. Start Flow Time (p57)	1	28v28	
Rel. Closing Time (p58)	1	25v25	

Note. For the parameter subset 5, information about the total number of hits per parameter, specific comparisons where these parameters are relevant (column *cases*) and inter-/intra- speaker percentage. The most relevant parameters per subset are marked in grey.

Table 35

Parameter Subset 6: Glottal gap (closure) coefficients (p59-62)

Parameter	Hits	Cases	Percentage %Intra/%Inter
Val. Flow GAP (p59)	11	40v40; 39v40; 48v48; 04v04; 13v13; 13v13; 17v18; 23v23; 23v24; 29v29; 39v39	73% - 27%
Val. Contact GAP (p60)	0		
Val. Adduction GAP (p61)	2	51v51; 25v25	
Val. Permanent GAP (p62)	0		

Note. For the parameter subset 6, information about the total number of hits per parameter, specific comparisons where these parameters are relevant (column *cases*) and inter-/intra- speaker percentage. The most relevant parameters per subset are marked in grey.

Table 36

Parameter Subset 7: Tremor (cyclic) coefficients (p63-68)

Parameter	Hits	Cases	Percentage %Intra/%Inter
1st. Order Cyclic Coefficient (p63)	0		
2nd. Order Cyclic Coefficient (p64)	0		
3rd. Order Cyclic Coefficient (p65)	0		
Tremor Frequency (Hz) (p66)	0		
Tremor Est. Robustness (p67)	1	20v20	
Tremor amplitude (rMSA) (p68)	0		

Note. For the parameter subset 7, information about the total number of hits per parameter, specific comparisons where these parameters are relevant (column *cases*) and inter-/intra- speaker percentage. The most relevant parameters per subset are marked in grey.

5.5. Discussion

5.5.1. Discussion for the first hypothesis

Our first general hypothesis was that *glottal parameters are genetically influenced, i.e. higher similarity values will be found in MZ twins than in DZ twins, in siblings or in the reference population*. According to this, we specifically established the following hypotheses to be tested:

H_1 Intra-speaker tests should show large LLRs.
H_2 MZ inter-speaker tests should show large LLRs.
H_3 DZ inter-speaker tests should show also large LLRs although not as large as H_1 or H_2.
H_4 B inter-speaker tests should show LLRs over the background baseline (fixed at $\lambda = -10$).
H_5 US inter-speaker tests should show LLR's aligned with the background baseline.

According to the decision thresholds established in Figure 40, our five hypotheses were represented as:

H_1: MZ(I), DZ(I), B(I), US(I) → $\lambda > -1$

H_2: MZ(O) → $\lambda > -1$

H_3: DZ(O) → $\lambda > -10$

H_4: B(O) → $\lambda > -10$

H_5: US(O) → $\lambda < -10$

In Section 5.4.2. we showed the results of the different matching tests carried out (see Table 29), highlighting the cases contradicting our hypotheses. In view of those results, the degree of hypothesis corroboration was very high: three of our hypotheses were corroborated in 83.3% of the cases (H_1, H_2 and H_5), another one was corroborated in 80% of the cases (H_3) and another one was corroborated in all the cases under study (H_4):

H_1: 40/54, but relaxing the threshold[129], it could be 45/54=5/6=83.3%

H_2: 10/12=5/6=83.3%

H_3: 4/5=80%

H_4: 4/4=100%

H_5: 5/6=83.3%

[129] As we said before, 14 out of 54 do not fulfill H_1. However, five speakers out of the total of 54 appear to be in the limit of H_1 (03, 35, 48, 49 and 50). That is why we say that a "relaxing threshold" would be 45/54.

In the following sections we will discuss the results of our analysis, distinguishing between the intra-speaker comparisons and the inter-speaker comparisons.

Intra-speaker results

As regards the confirmation of H_1, it is unclear why 14 out of 54 speakers do show self-unlikeness in a larger or smaller extent when their phonation samples belonging to one recording session are tested against the other session's tokens. Several reasons could be suggested, such as changes in phonation due to emotional stress or even temporary pathological conditions. Excluding weak self-unlikeness[130] the number of cases not fulfilling the hypothesis would be 9 out of 40, which is still a considerably large figure (16.6%). As suggested in San Segundo and Gómez-Vilda (2013: 255), "the immediate reflection is if these could be labeled as 'goats' in Doddington's Zoo (Doddington et al., 1998)." It is a truism in speaker recognition that there are "striking performance inhomogeneities among speakers within a population" (Doddington et al., 1998: 1). In other words, we can establish a difference between speakers depending on how well they behave in automatic recognition systems. According to this, "goat" is the name used by Doddington et al. (1998: 1) to refer to "those speakers who are particularly difficult to recognize":

> Goats tend to adversely affect the performance of systems by accounting for a disproportionate share of the missed detections. The goat population can be an especially important problem for entry control systems, where it is important that all users be reliably accepted. (Doddington et al., 1998: 1)

The "goats", according to this definition, would be opposed to the "sheep", which (in Doddington's zoo) comprise the default speaker type. Fortunately, sheep are predominant in a population and "systems perform nominally well for them" (Doddington et al., 1998: 1).

In our study, the kind of speaker showing remarkable differences from one recording session to another is not limited to one out of the four speaker types we have considered (MZ, DZ, B or US). On the contrary, we can find examples of these speakers in all groups, as it can be easily observable in Table 29. A thorough examination of the questionnaires filled at both recording session by each speaker would be necessary in order to gain complete understanding of the causes behind the large negative LLR values. Independently of the fact that this type of speakers ("goats") is acknowledged since long in automatic speaker recognition, the question of what may make a speaker so different from one occasion to another is a key question in forensic

[130] With the expression *weak self-unlikeness* we refer to the following cases: speaker 03 (LLR = -1.1), speaker 35 (LLR = -1.6), speaker 48 (-1.2), speaker 49 (-1.3) and speaker 50 (-2.5). The LLR values for their intra-speaker comparisons are much lower than for the rest of the intra-speaker comparisons which exhibit self-unlikeness.

phonetics and one issue which is still open for further research. For our study, it is possible that some normalization techniques used in the selection of the speaker's most characteristic phonation patterns could help in reducing the 16.6% of speakers exhibiting large intra-speaker variation.

Most intra-speaker comparisons yield LLRs around -8 , -7 or lower but there is one striking case of LLR = -42.2. This is a clear exception which should be in-depth analyzed. Looking at the anamnesis of this speaker it is revealed that he suffers from hypothyroidism. Often called underactive thyroid, this is a common endocrine disorder in which the thyroid gland does not produce enough thyroid hormone. Interestingly, one of the symptoms of this hormonal problem is hoarse voice (Longo & Fauci, 2011). This hormonal problem could explain the strikingly large intra-speaker variation for this speaker.

Inter-speaker results

Regarding the inter-speaker results, we have to separately consider H_2, H_3, H_4 and H_5. As far as H_2 is concerned, the number of non-fulfillments of the hypothesis is quite small: only 2 out of 12 pairs do not obtain LLRs above -1. One case is that of the MZ pair 11 and 12 (LLR = -14.6) and the other case is that of the MZ pair 35 and 36 (LLR= -1.5). It is evident that their cases are not comparable. If we look at pair 11-12, this is the MZ pair who was deeply examined after their awkward results in the pilot experiment. In this proof of concept they were found to be very different from each other (see Table 21). Besides, some unexpected answers that they gave in the questionnaire, together with their lack of physical similarity led us to check their zygosity. Although they were confirmed as MZ twins, their unlikeness was made clear in the voice diagnosis, where apart from an evident f_0 difference, other dissimilarities were observed in their respective glottal description. The most plausible reasons for their striking differences are: on the one hand, a physical explanation: the existence of smoking habits in one of them, which made his f_0 much lower than that of his cotwin[131]; and on the other hand, psychological or behavioral factors: according to their questionnaires, their attitude towards being twins made them clearly separate trying to be independent and different since they were children. In this case, the learned speech habits aimed at attaining divergence patterns may have outweighed their anatomical similarities. The other MZ pair with negative LLRs is the pair 35-36. Their LLR value (-1.5) is not as striking as in the other pair. In this case, checking their responses to the questionnaire, a plausible explanation for their results could be an involuntary divergence in their speech patterns due to their lack of contact: not living together and not having frequent communication, or maybe a voluntary desire to sound different. Actually, the learned or voluntary factors influencing speech

[131] In the case of the smoking twin, he also suffered from frequent sore throat and occasional nodules.

behavior in MZ twins remain greatly unexplored, as we have explained in the literature review [see Chapter 2].

As far as the H_3 is concerned, only in one DZ pair out of five this hypothesis is not corroborated. In our hypothesis, we established that DZ twins should show large LLRs but not as large as MZ twins, being the decision threshold $\lambda > -10$. The only case where this is not fulfilled is that of DZ pair 17-18, who obtained a LLR = -10.1. This exception is then almost irrelevant. In all the other cases, the LLRs values are as expected, relatively large LLRs but not as large as in the MZ pairs. Of course, it depends on the specific pairs we are considered. It could turn out that a DZ pair (45-46) obtains a LLR = 3.4, which is larger than the LLR = 2.9 observed for MZ pair (39-40). Even though we have to take into account that the number of MZ pairs doubles the number of DZ pairs, and it is then hard to make comparisons and establish a trend, the third hypothesis seems to be corroborated for the majority of cases under study.

Considering H_4, all the sibling cases corroborate our hypothesis. LLR values above -10 are obtained in 100% of the pairs analyzed. Since full siblings (B) and DZ twins both share the same genetic load, H_3 and H_4 were established at the same level: $\lambda > -10$. Indeed, the average LLR similarity between the two groups of speakers (DZ and B) is evident. This could indicate that the glottal parameters under study are actually genetically influenced. This is made clearer if we look at the results obtained by US, which are discussed below in relation to H5.

The fifth hypothesis established that US would obtain LLRs aligned with the background baseline, fixed at $\lambda < -10$. This is fulfilled in almost 100% of the cases. The only exception is found in speakers 27-28 with LLR = -9.7. Strictly applying our decision threshold, this pair of speakers would not fulfill the hypothesis but it seems clear that the difference between -9.7 and -10 is not very significant when we are expressing the results in logarithmic figures. The degree of H5 corroboration is then very satisfactory. The confirmation of this hypothesis is especially important, as it indicates that in a typical forensic situation (when unrelated speakers are compared) our system performs very well, with none of the speakers being misidentified (*false alarms*). And again, more evidence is gained in favor of our hypothesis that the glottal parameters are genetically influenced, as none of the unrelated speakers show any similarity, in comparison with the somehow genetically related DZ and B, with larger LLR values and the much genetically related MZ, with still larger (in average) LLR values.

5.5.2. Discussion for the second hypothesis

Our second hypothesis referred to the discrimination capability of the glottal parameters under study. According to preliminary studies (San Segundo, 2012), which are part of the pilot experiment described above, we suggested that the biomechanical estimates of the glottal waveform would be especially speaker-specific, i.e. showing a great discrimination potential. In San Segundo (2012), these biomechanical parameters (*parameter subset 4: biomechanical estimates of vocal fold mass, tension and losses*), clearly outweighed the other features analyzed, namely jitter and shimmer estimates. However, in the more complete analysis carried out for this thesis, we have considered the whole parameter set in *BioMet®ScieProf*, comprising 68 parameters. Therefore, upon consideration of more parameters than in the pilot experiment, different results are obtained as far as their discrimination capability is concerned.

Taking into account the total number of occurrences per parameter (*hits*)[132], we listed the five most frequent (i.e. with higher occurrences). The first one (*p6: Muc./AvAc. Energy, MAE*) belongs to the "fundamental frequency and distortion parameters (subset 1); the second (*p32*: *MW PSD End Val. Pos. rel*) and the fourth one (*p21: MW PSD 1st Max. ABS*) belong to the "singularities of the glottal source PSD" (subset 3); the third one (*p8: MWC Cepstral 2*) is a cepstral coefficient of the glottal source PSD (subset 2) and the fifth one[133] (*p59: Val. Flow GAP*) belongs to the glottal gap coefficients (subset 6).

Out of the 68 parameters analyzed, we established that the five mentioned above are the most discriminant since they are listed in the top-three positions of the relevance list in *BioMet®ScieProf* in more occasions than the others. We also tried to reveal which parameter was more outstanding among the other features in their corresponding subset. For that purpose we marked the parameters with an occurrence relatively higher than the other features in each subset. For the subset 1, not only p6 (which was relevant in the general classification) but also p1 (f_0) stood out among the others. For the subset 2, most cepstral parameters are homogenously distributed in number of occurrences but p8, p9, p10, p12 and p15 could be highlighted as slightly more discriminant. For the subset 3, interestingly p21 and p32 clearly outstand among the other singularities of the glottal source PSD. This seems logical, as p21 is the maximum PSD value in dB scale while p32 is the relative position in frequency at the end for Nyquist frequency related to the first maximum. Some correlation between these two parameters is possible. All the parameters in the subset 4 (biomechanical estimates) rank quite low in number of occurrences, as compared with the other subsets. Only p41 (cover mass) and p42 (cover losses), both with 6 occurrences, obtain a good number of hits, while the others (e.g. body losses, body losses unbalance) occur in one isolated case or even in none. From this, it can be concluded that the parameters which are discriminant for one pair may not be so for another pair. So their relevance

[132] Only the three most relevant parameters per comparison (i.e. ranking highest in the relevance list made by *BioMet®Fore*) were taken into account.

[133] Note however that both p21 and p59 appear in 11 occasions (i.e. 11 hits), so they are both at the fourth position of most relevant parameters.

should be studied on a case-by-case basis. For the subset 5, the distribution is again quite uneven: p49 and p54 rank highest, followed by p47, p52 or p55 but others occur only occasionally in one speaker. It also seems logical that p49 and p54 rank similarly high as the former is the "ratio between the first opening time and the total glottal cycle duration" (p49) and the latter is the "ratio between the first opening time amplitude and the peak-to-peak amplitude" (p54). For subset 6, the difference between one parameter (p59) and the others is very large. While p59 (Val. Flow GAP) obtains 11 hits, being among the top-five of the general list, the others either never occur or appear on few occasions (e.g. p.60 with two hits). Finally, subset 7 is the group of parameters with fewer occurrences. Only p67 occur once for a single intra-speaker comparison (20-20).

5.6. Conclusions

As regards the first hypothesis (*glottal parameters are genetically influenced, i.e. higher similarity values will be found in MZ twins than in DZ twins, in siblings or in the reference population*), we can conclude that the glottal parameters analyzed, as a whole, could be genetically influenced. With few exceptions, DZ and non-twin brothers' behavior is similar ($\lambda >$ -10) while MZ twins obtain larger LLRs and US are homogenously around the baseline ($\lambda <$ -10). This is in agreement with our hypotheses, as we predicted that the LLRs values of the forensic comparison would be distributed in a line going from the largest positive LLRs for the MZ twins, at one end of the line, and the largest negative LLRs for the US, at the other end of the line. The former share 100% of their genes while the latter share 0%. In between, there are the DZ twins and the B, sharing on average 50% of their genetic information.

Looking at the LLR results obtained by DZ and B, no large differences are found, which could correlate well with their lack of genetic differences. Only an awkward result of strong dissimilarity is found in the DZ pair 17-18 (LLR= -10.1). It is striking that no such result is found in B pairs, which share less environmental factors than DZ, due to the age gap. We could argue that this is a case of speech divergence, according to their answers to the questionnaire[134], either voluntary (due to a desire to sound different and form different personalities) or involuntary (due to different external factors and unshared environmental influences). All in all, the degree of

[134] Although these DZ twins are only 18 years old and still live together at the family house, their answers to the questionnaire indicate a strong divergence in lifestyle and independence regarding friends and other environmental influences. For instance, they go to the same school but have always been in different classes, to the question "do you like having a twin", they answer "indifferent", to the question "how close is your relationship with your twin", they give "2" and "3" points in a 1-5 scale, they highlight that they are very different both in physical aspects as in personality, they have a different group of friend and different leisure activities.

hypothesis corroboration for H_4 is 100% and for H_3 80% with only the single exception that we have mentioned above, out of five DZ pairs.

As far as the MZ twins are concerned, the two only cases against our hypothesis are explained in previous section: the case of pair 11-12, with the most striking against-the-hypothesis value (LLR = -14.6) is not only explainable by physical causes but also in view of their anamnesis. For the other exception (pair 35-36), only their answers to the questionnaire could suggest a desire to sound different causing a voluntary divergence in speech habits. These two exceptions make the degree of corroboration of H_2 be of 83.3%. The same percentage is obtained for the fulfillment of H_5. Out of the 6 pairs considered, only one does not reach the $\lambda < -10$ baseline threshold established. Yet, the LLR value of this exception is really close to the threshold (-9.7). In a real case, this would imply that two different speech samples (i.e. coming from different speakers) would be deemed, with a strong support, to come from a different speaker. Therefore, even though we have to consider as non-corroboration of our hypothesis, it is not really an important exception. The threshold established at $\lambda < -10$ was actually arbitrary.

In conclusion, really large LLR values, such as 12.1, 12.6 or 9.9 are only obtained by (some) MZ pairs (such as 07-08, 09-10 or 37-38, correspondingly). According to the twin methodology used in other disciplines like Psychology or Medicine [see Chapter 2], when comparing MZ and DZ twins all the excess of similarity that is found in the former which does not occur in the latter is only explainable by genes, since both groups share the same environmental characteristics. Nevertheless, a more thorough study of the specific glottal parameters which have been relevant in the comparison of those very similar MZ pairs would probably shed more light on this issue.

Besides, our results are in agreement with previous studies about twins insofar as different results are found for different twin pairs, indicating a lack of homogeneity in twin pairs. The idiosyncrasies of each relationship could be only studied on a case-by-case basis to find the causes for speech convergence or divergence, which probably indicates that the weight of external factors (like psychological aspects, educational and environmental influences) is more important that it could be thought a priori in this type of voice studies.

Finally, we have also taken advantage to study intra-speaker variation in all the four type of speakers participating in this study. We have found that in more cases than desirable in a forensic context, the system performance is not completely good when two speaker sessions are tested one against the other. Some intra-speaker comparisons yield LLRs around -8 , -7 or lower but there is one striking case of LLR = -42.2. This is a clear exception which should be in-depth analyzed. A possible explanation for this large intra-speaker variation could be found in a hormonal disease suffered by this speaker. Yet the other cases suppose still large figures which are not desirable in a forensic system. They are called *missed hits*, and they are as important as *false alarms*. Since intra-speaker variation is an important issue in FSC, from this investigation it can be clearly concluded that more studies are necessary to investigate which factors influence the high intra-speaker variation found in a relatively large number of speakers.

As explained in Chapter 4, Tippett plots are a standard graphical method to represent the LR results of a forensic comparison system. Figure 43 shows the Tippet plot for our glottal source analysis. As it can be seen, the blue line (intra-speaker comparisons) extend largely on the right, which implies good performance of the system but there are still some LLRs which support the contrary-to-fact hypothesis, represented in the blue line from 0 to the left[135]. If we look at the red lines, for the US the system performance is optimal as there are only LLRs supporting the consistent-with-fact hypothesis. The results for MZ, DZ and B are different. The strongest support for the contrary-to-fact hypothesis can be observed for MZ twins (dashed line from 0 to the right) while for the DZ and B, with a similar performance, most cases fall within the consistent-with-fact hypothesis and we find only some cases supporting the contrary-to-fact hypothesis. This could be explained by the genetical influence of the glottal parameters analyzed. As MZ, DZ and B exhibit a similarity that is not present in the US, the support for the contrary-to-fact hypothesis in these speakers simply indicates that the system sometimes fails to support that these speaker pairs are different.

As regards the second hypothesis, we have tried to investigate which parameter subsets allow better speaker discrimination. Firstly, not all of them behave in exactly the same way. For instance, it is clear that the subset 7 (tremor –cyclic- coefficients) have almost no occurrences in the speaker comparisons; at least not ranking among the three most relevant parameters (see Section 5.4.2). The same happens with the subset 6 (biomechanical estimates). Although these parameters might seem very promising for biometric purposes, as they are very semantic, as compared with the others, almost only cover mass (p41) and cover losses (p42) have a sufficiently large number of occurrences. This is explainable due to the fact that both subset 7 and 4 are made up of parameters which are usually associated to pathological phonation. In our study only speakers with no voice pathologies have participated.

[135] Note that the value -42.2 corresponding to speaker 25 has been excluded from the intra-speaker comparisons used to create this Tippett plot. As it has just been explained, this value was considered an outlier and the possible reasons explaining this contrary-to-the-fact value were suggested in Section 5.5.1.

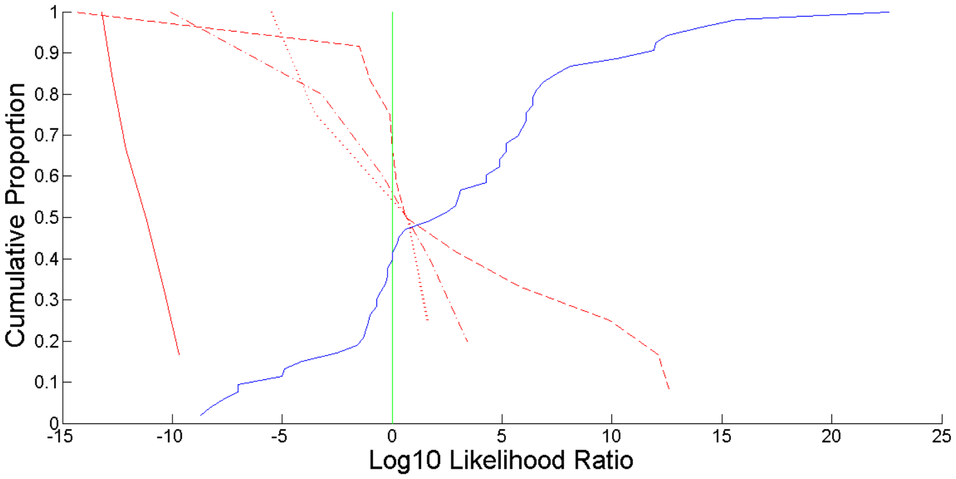


Figure 43. Tippett plot showing in the blue solid line the intra-speaker comparisons (for all the speaker types), and in the red lines the inter-speaker comparisons, being the solid line for US, the dashed line for MZ, the dot-dashed line for DZ and the dot line for B.

On the other end, we find the subset 2 (cepstral coefficients of the glottal PSD), with almost all of its parameters having high occurrences in a relative homogeneous distribution. If we were to detect the most relevant parameters independently of how discriminant the other parameters of its subset are, p16 (*Muc./AvAc. Energy, MAE*) is clearly the most discriminant with 18 hits, followed by the singularity of the glottal source p32 (*MW PSD End Val. Pos. rel*), the cepstral p8 that we have already mentioned, and another parameter included in the subset 3 (singularities of the glottal source PSD), namely p21 (*MW PSD 1st Max. ABS*). *Val. Flow GAP* (p59) is also quite relevant, even though other glottal gap coefficients of its group are not. Finally, p49 and p54 of the subset 5 (time-based glottal source coefficients) are also of relative importance. Both are related to the closed phase of the L-F cycle (see Figure 38: example of the glottal cycle temporal analysis of a typical male voice).

6. AUTOMATIC ANALYSIS

6.1. Objectives and justification

In this section we will identify our research objectives and hypotheses for the automatic analysis described in the rest of the chapter. Some relevant studies related to this approach will be reviewed, although the main studies which have investigated twins' voices from an automatic perspective, and specifically using the same recognition system (namely, Kim, 2009; and Künzel, 2010) were already described in Chapter 2.

6.1.1. Objectives

The main objective of this analysis is testing the performance of the automatic speaker recognition system *Batvox* (version 4) for discriminating MZ, DZ and non-twin siblings. Following the same forensic comparison procedure as in previous analyses, *inter-speaker comparisons* and *intra-speaker comparisons* were carried out. This general objective can be formulated as:

O1: Testing whether there is higher intra-pair similarity for this kind of parameters in MZ twins than in other speaker comparisons.

For the above-mentioned objective, we support the following hypothesis:

H1: Since the cepstral features in which the automatic system is based depends largely on anatomical-physiological foundations, we propose that they must be somehow genetically related. Therefore, higher similarity values will be found in MZ twins than in DZ twins, in siblings or in the reference population. This is in agreement with the 5 basic hypotheses established for this thesis (Table 3; see Chapter 2).

6.1.2. Justification

As far as we know, only two studies have used the automatic system *Batvox* for the analysis of twins' voices so far. We refer to Kim (2009) and Künzel (2010), who studied Korean and German twins, respectively. These studies were already reviewed in Chapter 2, where a detailed account of voice-related studies on twins was carried out. Our focus in this section will be then on other relevant aspects which concern automatic systems in general and specifically the one used for this analysis. Describing the state-of-the-art research related to automatic systems lies completely out of our scope in this thesis. We will just give an overview of the automatic speaker recognition technology and challenges, which could serve as a justification for its use in our investigation. Furthermore, we will review two recent studies which have specifically tested the performance of *Batvox* under typical forensic situations.

Kinnunen and Li (2009) provide a relatively recent overview of both the classical and state-of-the-art methods in (text-independent) automatic speaker recognition. According to these authors, session variability (i.e. any mismatch between the training and testing conditions) remains to be the most challenging problem in this field, as this decreases the accuracy of speaker recognition, sometimes dramatically. Therefore, the main focus of speaker recognition research nowadays lies in tackling this mismatch basically through normalization and adaptation methods. For more details about score normalization, see Kinnunen and Li (2009: 14).

The first stages of any automatic system could be very briefly summarized as:

- *Parameter extraction*: transformation of the raw signal into feature vectors in which speaker-specific properties are emphasized and statistical redundancies suppressed (Kinnunen & Li, 2009: 2-3).

- *Speaker modeling*: using feature vectors extracted from a given speaker's training utterance(s), a speaker model is trained and stored in the system database (Kinnunen & Li, 2009: 4) According to Kinnunen and Li (2009), "classical speaker models can be divided into *template models* and *stochastic models* (Campbell, 1997) , also known as *nonparametric* and *parametric* models, respectively. […] The Gaussian mixture model (GMM) (Reynolds & Rose, 1995; Reynolds, Quatieri & Dunn, 2000), is the most popular model for text-independent recognition" (Kinnunen & Li, 2009: 4).

Since Section 6.3 (cf. *Parameters*) is aimed at describing the parameters for the current investigation, a more detailed description of short-term spectral features can be found in that section. In the following pages we review two recent studies which have specifically used a widely known automatic recognition system (*Batvox*) with the purpose of tackling relevant issues and challenges in Automatic Speaker Recognition (ASR).

Künzel (2013) addresses the issue of the so-called "language gap" in automatic systems for speaker recognition. This refers to the reduction in performance of such systems due to language mismatch between the voice samples under comparison and/or the reference population[136]. Although several types of mismatch are possible, Künzel (2013) investigates the

[136] It is well known that "virtually all state-of-the-art speaker recognition systems use a set of background speakers or cohort speakers in one form or another to enhance the robustness and computational efficiency of the recognizer" (Kinnunen & Li, 2009: 3).

most typical forensic situation, where a suspect's speech sample is in language A and the test sample in language B; thus assuming that a reference population matching the language of the speaker model, A, is available (Künzel, 2013: 25).[137] The results of this study showed that "the overall performance of the automatic system for cross-language voice comparisons was equal to or at times slightly better than for same-language comparisons". Apart from other factors which may have influenced these good results (e.g. homogeneity of speech material, identical recording conditions and transmission channels, etc; cf. Künzel, 2013: 34-35), the most important reason seems to lie in the double normalization procedure of this specific automatic system. As specified in Künzel (2011: 28), the option to use the so-called 'case impostors'[138] is a special feature of the normalization procedure of *Batvox* which would be particularly useful in cross-language speaker recognition. In other automatic systems, it is acknowledged that some normalization procedure may be as well the solution to the cross-language problem but "due to their individual architecture, systems may require different types of normalization for the effects of language and transmission channel" (Künzel, 2013: 37).

Comparing the magnitude of the language mismatch with the effect of other sources of mismatch (specifically, three types of transmission channels[139]) on the same voice data, this investigation shows that the impact of the language mismatch effect on the system performance was not as large as the impact of the channel transmission characteristics. While equal error rates (EERs) for same-language and cross-language comparisons were approximately the same (ranging from zero to 5.6%), the different types of transmission channels caused the EERs to rise by less than 1% on average. These results should be interpreted taking into account the specific conditions of the experiment, e.g. the specific languages considered (German, Spanish, Russian, Polish, English and Chinese) and the fact that only female voices were studied.

In a more recent study, Künzel and Alexander (2014) assessed the effect of several types of signal degradations on the performance of the automatic speaker recognition system *Batvox* and tested diverse enhancement algorithms which could compensate those degradations. They

[137] As stated in Künzel (2013: 25-26), matching reference population and test sample for language is also possible but its consequences are described as: "if the test samples are made more 'similar' to the reference population than to the speaker model, the amount of false rejections will be reduced but at the same time the number of false acceptances will increase, which is unacceptable from a forensic point of view".

[138] According to Künzel (2013: 28), "the term 'case impostors' denotes a set of speakers who are definitely not identical with the speaker under test but exhibit some similarities, primarily in terms of channel characteristics, and in this case language. Thus the system may recognize certain acoustic resemblances as irrelevant and reduce *a priori* the probability for false acceptance errors. Technically, the impostors are used as a Z-norm cohort in what may be called a second normalization process, after application of the T-norm, and serve to reduce the misalignment in the event that the available T-norm cohort is less-than-perfect. Since the number of impostors is usually small (between 3 and 10) the second normalization is based only on the mean of the cohort scores but not on its variance".

[139] Transmission of the speech data was carried out via landline telephone, GSM and, for part of the corpus, VoIP (using Skype).

found that, out of the seven types of degradations of the acoustic signal considered[140], the performance of the system was most affected by pop music in both single-channel and 2-channel recordings, and also by noise inside a fast moving car, while road traffic and restaurant noise did not affect the system's performance significantly. Several types of enhancement processes were tested which could reduce the harmful effects of signal degradations and thus improve the performance of the automatic system. According to the results of their study, the authors suggest that "speech enhancement cannot be generally rejected as a tool for the pre-processing of speech samples that have to be used for forensic speaker recognition" (Künzel & Alexander, 2014: 251).

Although the two aspects referred above (i.e. language mismatch and signal distortion) are not characteristic of the recordings used for our investigation, they are still relevant for this thesis in order to show the performance adequacy of *Batvox* in forensically realistic and challenging conditions. It is then to expect that this tool also yields good identification results in a challenging situation of extreme similarity between speakers, as it is the case of the twins. For instance, it would be of interest to compare the magnitude in the performance decrease, (measured in EERs) when considering twins as model and test speakers or when considering signal distortion or language mismatch conditions.

6.2. Speech material, analysis tools and method

6.2.1. Speech material

For the automatic analysis, we extracted a speech fragment of around 120 seconds per speaker and per recording session. According to Künzel (2010), this is the recommended duration of a voice sample to be analyzed using the automatic speaker recognition system *Batvox*. Also following the recommendation of this author, we selected the fifth speaking task of our corpus for the extraction of the speech material.

Speech material extraction

For the selection of the speech fragments (120 seconds of duration in average), the audio files belonging to the first and the second recording session of each speaker (average duration of 5

[140] These seven types of acoustic degradations are considered by Künzel and Alexander (2014: 245) the most typically found in forensic speaker recognition tasks: 1) amplitude clipping due to recording overload caused by wrong setting of the recording device and/or very loud speech; 2) background music, particularly pop and folklore; 3) noise inside a restaurant; 4) road traffic; 5) noise inside a moving vehicle; 6) reverberation caused by the local environment, e.g., sparsely-furnished rooms, prison cells, hallways; 7) background music and speech, two-channel recording.

minutes) were entered in *Praat.* The speech material chosen for further analyses was selected from approximately the middle of the audio file, in order to avoid the beginning of the conversation, where the speaker has not already settled to his ordinary speaking style. Prior to the labeling and extraction, the audio files were first perceptually examined in order to remove extraneous noise, laughter, clicks, cough, etc, according to the recommendations suggested in Künzel (2010: 256).

6.2.2. Analysis tools and method

For the automatic analysis, we have used *Batvox* (version 4), which is based on parameters related to the resonances of the vocal tract, basically cepstral coefficients (see Section 6.3). A good summary of how this type of automatic speaker identification systems work can be found in Künzel (2010: 253-4) where he cites relevant bibliographic references in this field (Gónzalez-Rodríguez et al., 2003; Drygajlo, 2007; Ramos-Castro, 2007, Przybocki et al., 2007). One of the main assets of automatic systems is that between-sample differences in the speech content are not relevant. As Künzel (2010: 254) states: "Automatic systems exploit the 'sound' of the voice and disregard the speech content almost completely". For more details about the advantages and disadvantages of automatic systems versus the traditional (acoustic-phonetic-linguistic) method, see Chapter 1.

As explained in Künzel and Alexander (2014: 247), the main characteristics of Batvox are "a 38-dimensional feature vector consisting of 19 MFCCs plus their deltas, GMM-Channel-Factor analysis for the compensation of speaker models (Kenny et al., 2005) and nuisance attribute projection (Campbell et al., 2006) for the test files". The working principle of *Batvox*[141] will follow with a comparison between the statistical model for the reference speaker and the results for the target speaker's model. The similarity score obtained after this procedure is then weighed using a reference population, which, as indicated by Künzel (2010: 257), "can be composed in terms of number of speakers, type of speech material (spontaneous, read, interview etc.) transmission channel characteristics (microphone, landline telephone, GSM, VoIP, analogue radio, TV) and other variables, according to the conditions of the case".

For our study, the system was set to "Identification Mode", where results are indicated as *normalized scores* that can be used to calculate False Alarms (FA) and False Rejections (FR)

[141] This description refers to the *LR (likelihood-ratio) mode of operation*, which, according to Künzel (2010: 256), "corresponds to the typical forensic paradigm of identifying an individual in an open set of individuals, *where* the system matches the voice sample of one or more known (reference) speakers with a sample of unknown (target) speakers".

rates, and finally, equal-error rates (EERs)[142], an accepted measure of the performance of an identification system (see below). This "Identification Mode" of operation was deemed the most appropriate for the purpose of this investigation (see *Batvox 4.1 Basic User Manual*, in Agnitio Voice Biometrics, 2013). Furthermore, it is the mode which was also used in the above-mentioned Künzel (2013) and Künzel and Alexander (2014) (cf. 6.1.2. *Justification*). As reference population, a cohort of 31 Spanish male speakers was used (spontaneous conversation and high quality recordings), coming from *Batvox* databases.

The following tests were carried out:

1. Each speaker's session one was compared with the same speaker's session two. In ASR this is called a *match* (target trial). According to the terminology that we have been using in the rest of the analyses, this would be an *intra-speaker comparison*.
2. Each speaker's session one was compared with all other speakers' session two. This would be called *no-matches* (impostor trials) in ASR. If we want to be consistent with the terms used for the rest of analyses, these comparisons can be described as *inter-speaker comparisons*.

The results of these two types of comparisons can give us an idea of the general performance of the system without taking into account the fact that some speakers are MZ, DZ or siblings. Yet, for our research question, the most interested aspect to investigate deals with the magnitude of the "sibling effect" in comparison with the general performance of the system. This is the reason why the following further test was carried out:

3. Each speaker's first session was compared with the first session of his sibling (or conversation partner in the case of the reference population). This type of tests yields *intra-pair comparisons*.

6.3. Parameters

It is well known that the vocal tract is made up of the oral, nasal and pharyngeal cavities. Each of these cavities has a resonance profile which is supposed to be somehow typical and idiosyncratic for each speaker, at least similarly to what happens with other physical aspects of the human being, which are more or less individual (Künzel, 2011: 40). Automatic methods in general (as explained above), and *Batvox* specifically, extract a set of features representing the resonance profile of the vocal cavities of a speaker (i.e. the Mel FCC coefficients) and creates a multidimensional vector. These would be the kind of parameters used in this third type of analysis,

[142] EERs were calculated using the *Biometrics 1.2* software (Biometrics 1.2, 2012).

as compared with high-level features, such as the ones described in previous approaches to the study of twins' voices [see Chapters 4 and 5].

The general description of *short-term spectral features* by Kinnunen and Li (2009) states that "as the name suggests, they are computed from short frames of about 20-30 milliseconds in duration […] *being* usually descriptors of the short-term spectral envelope, which is an acoustic correlate of timbre, i.e. the 'color' of sound, as well as the resonance properties of the supralaryngeal vocal tract" (Kinnunen & Li, 2009: 3). The explanation for the breakdown in short frames is that, as the speech signal continuously changes due to articulatory movements, the signal must be decomposed in short intervals where the signal is assumed to remain stationary (Kinnunen & Li, 2009: 4). After this breakdown in short frames, a spectral vector can be extracted from each frame. In the case of *Batvox*, Künzel (2010: 256) specifies that the 38-dimensional feature vector is calculated every 10 ms. Usually the frame is pre-emphasized and multiplied by a smooth window function prior to further steps (Kinnunen & Li, 2009: 4).[143]

From the description of the *parameter extraction* above, it is clear that no separation of linguistic or phonetic units is made under the automatic approach. This is why Jessen (2009: 699) classifies this type of automatic methods as *holistic*: "The distribution of the MFCCs over the entire course of the recording of a speaker is determined. […] no segmentation of the speech stream into different linguistic categories, such as consonants, vowels or syllables is performed" (Jessen, 2009: 699).

For a summary of the three sequential stages of the automatic speaker identification process (*parameter extraction*, *parameter modeling* and *calculation of distances*), see Jessen (2009: 698- 703). González-Rodríguez et al. (2006) or Müller (2007) can also be read for further information about automatic speaker identification.

6.4. Results

6.4.1. Overall system performance

As specified above (see Section 6.2.2), we carried out three types of tests, which yielded results for intra-speaker, inter-speaker and intra-pair comparisons. If we first look at the results for intra-

[143] As explained in Jessen (2009: 699), "as a means of smoothing the spectral shape and of making the outcome more realistic psycho-acoustically, the spectrum is then passed through a filterbank based on the non-linear Mel scale. The logarithms of the filter coefficients are transferred to the cepstrum by application of the Discrete Cosine Transform. The resulting vectors are now called cepstral coefficients (Bimbot et al., 2004 for more details and further literature)" (Jessen, 2009: 699).

speaker and inter-speaker comparisons alone, we obtained similarly high coefficients of recognition for all the pooled four speaker types. This can be observed in Figure 44, which shows a 0% EER. The input values for the creation of this figure were of two types:

- *Matches* (shown in the blue line): the values were obtained from the comparison of each speaker's session one with his own session two.
- *No-matches* (shown in the red line): the values were obtained from the comparison of each speaker's session one and all other speakers' session two. [144]

The equal-error rate (EER) is an accepted measure of the performance of an identification system. It is the point of intersection of the distributions for Matches and No-Matches. An EER = 0% indicates that there is no overlap of matches and no-matches, so neither False Alarms (FA) nor False Rejections (FR) occur. This shows that the overall system performance with high-quality recordings and without taking into account the "sibling effect" (intra-pair comparison) is perfect.

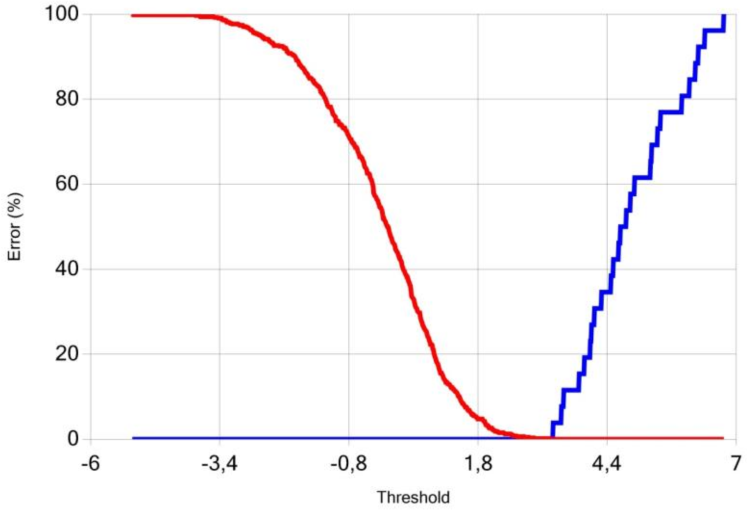


Figure 44. Cumulative distribution of scores for same-speaker comparisons or matches (blue line) and different-speaker comparisons or no-matches (red line). The EER obtained is 0%, indicating that there is no overlap between matches and no-matches (i.e. neither False Acceptances nor False Rejections exist).

6.4.2. Sibling effect

When considering also the intra-pair comparisons, the recognition coefficients were expected to be much lower, as the comparison is not between the same individuals. However, different

[144] To avoid comparing a speaker with his sibling or conversational partner, at least in this first analysis which does not take into account the "sibling effect", only the even members of each speaker pair were selected, both for the matches and no-matches. That is, only speakers 02, 04, 06 and so on were used in the analysis. Following Künzel (2011)'s methodology, in order to facilitate this task, one member of the speaker set was labeled "red" (the odd numbers) and the other member was labeled "blue" (the even numbers). Figure 1 shows the EER (0%) using the blue speakers. The same test was repeated using only the red speakers and a similar EER was obtained (0.07%), which is a very insignificant difference.

patterns are observed depending on the type of speaker (MZ, DZ, S or US). This can be seen in Table 37, where the values obtained are classified per speaker (i.e. his intra-speaker coefficients) and per speaker pair (i.e. their intra-pair coefficients), depending on whether they are MZ, DZ, S or US. As it can be observed in this table, all intra-speaker comparisons yield similarly high coefficients of recognition (blue color values). In relation to the inter-speaker (e.g. intra-pair) comparisons (orange color values), Table 37 is useful to observe the different values obtained by different speaker pairs. This table has been done analogous to the tables created for the other two types of analyses carried out for this thesis [see Chapters 4 and 5] so that the performance of the system can be analyzed per speaker or per speaker pair. The fact that the speakers in this investigation are not very numerous is an advantage in order to carry out this kind of detailed examination. For instance, if we look at within-group differences, the values of MZ pair 39-40 (0.64) are very different from the other pairs' coefficients (much higher in average).

If we are interested in the behavior of the groups in general, and not specifically in each pair, Table 38 is more insightful and probably more appropriate to assess the system performance depending on the speaker type. According to the information in this table, MZ intra-pair comparisons yield the highest values (i.e. the dissimilarity is the lowest). From the average values obtained by the MZ pairs to the coefficient values yielded for US, we observe a gradation from largest to lowest, all through the average values of the DZ intra-pair comparisons and the B intra-pair comparisons. This trend is thus in agreement with our hypothesis, where we predicted the following scale (from more to less similar): MZ > DZ > S > US. In other words, the coefficient grading goes in the same direction as the "magnitude" of kinship relationship[145].

We have added to the table the average coefficients obtained in MZ intra-speaker comparisons. As expected, the same-speaker comparisons yield the highest coefficients. The inclusion of these matches[146] in the table is intended to serve as a baseline to which the rest of (intra-speaker) coefficients can be compared, under the assumption that nobody could be more similar to anyone than to himself, although some exceptions may occur in the case of MZ twins, as we describe below (*cf. Special case study: MZ twins*).

[145] See Chapter 2 to see how we understand the interplay between genetic load and shared environmental factors to explain the hypothesized scale of similarity MZ > DZ > B > US.

[146] It should be noted that in Table 38 –and its corresponding Figure 45– we could alternatively (or also) have included the average coefficients for the same-speaker comparisons of subjects other than MZ. As indicated above (cf. *overall system performance*) and as it can be observed in Table 37, the values for the intra-speaker comparisons are homogenously distributed, regardless of the speaker type (i.e. the coefficients are very high, in general). Indeed, the average coefficient for intra-speaker comparisons (all the speaker types pooled together) is 4.86, a very similar value to 4.83, which is the average coefficient for MZ intra-speaker comparisons.

Table 37

Summary of the results for the different tests

	MZ (I)		MZ(O)	DZ(I)		DZ(O)	B(I)		B(O)	US(I)		US(O)
Cases	01v01/02v02		01v02	13v13/14v14		13v14	21v21/22v22		21v22	25v25/26v26		25v26
Score	4.22	3.48	3.79	5.25	6.17	3.77	4.51	6.24	0.64	4.93	4.47	0.39
Cases	03v03/04v04		03v04	15v15/16v16		15v16	23v23/24v24		23v24	27v27/28v28		27v28
Score	4.82	4.79	2.65	4.27	4.87	2.53	7.76	5.27	3.31	3.99	4.29	0.64
Cases	05v05/06v06		05v06	17v17/18v18		17v18	47v47/48v48		47v48	29v29/30v30		29v30
Score	4.29	4.95	3.45	5.13	6.35	0.18	5.53	4.63	0.79	5.29	5.42	-0.66
Cases	07v07/08v08		07v08	19v19/20v20		19v20	49v49/50v50		49v50	31v31/32v32		31v32
Score	4.23	4.14	2.31	3.51	5.46	2.17	2.78	3.31	0.36	2.92	4.67	0.25
Cases	09v09/10v10		09v10	45v45/46v46		45v46				51v51/52v52		51v52
Score	3.64	4.06	2.66	3.44	3.83	0.40				3.80	3.52	0.71
Cases	11v11/12v12		11v12							53v53/54v54		53v54
Score	3.24	5.29	1.34							4.03	5.22	0.22
Cases	33v33/34v34		33v34									
Score	4.55	6.06	3.20									
Cases	35v35/36v36		35v36									
Score	6.44	3.94	4.93									
Cases	37v37/38v38		37v38									
Score	5.41	4.52	3.54									
Cases	39v39/40v40		39v40									
Score	6.05	6.74	0.64									
Cases	41v41/42v42		41v42									
Score	4.68	5.9	3.53									
Cases	43v43/44v44		43v44									
Score	4.43	4.08	2.59									

Note. MZ: Monozygotic twins; DZ: Dizygotic twins; B: Brothers; US: Unrelated Speakers; (I): intra-speaker tests; (O): inter-speaker tests. Divided columns are used for each pair member. Cases: xxvyy means speaker xx versus speaker yy. Blue is used for (I) and orange for (O).

Table 38

Average coefficients per speaker type and test type

Speaker Type	Test type	Average coefficient
Unrelated Speakers (US)	Intra-pair	0.26
Non-twin brothers (B)	Intra-pair	1.28
Dyzigotic twins (DZ)	Intra-pair	1.81
Monozygotic twins (MZ)	Intra-pair	2.89
Monozygotic twins (MZ)	Intra-speaker (matchs)	4.83

Note. We show all the intra-pair values per speaker type but also the intra-speaker values for MZ twins (last row), in order to highlight the grading in values (from lowest to largest), where the lowest means more dissimilar and the largest more similar.

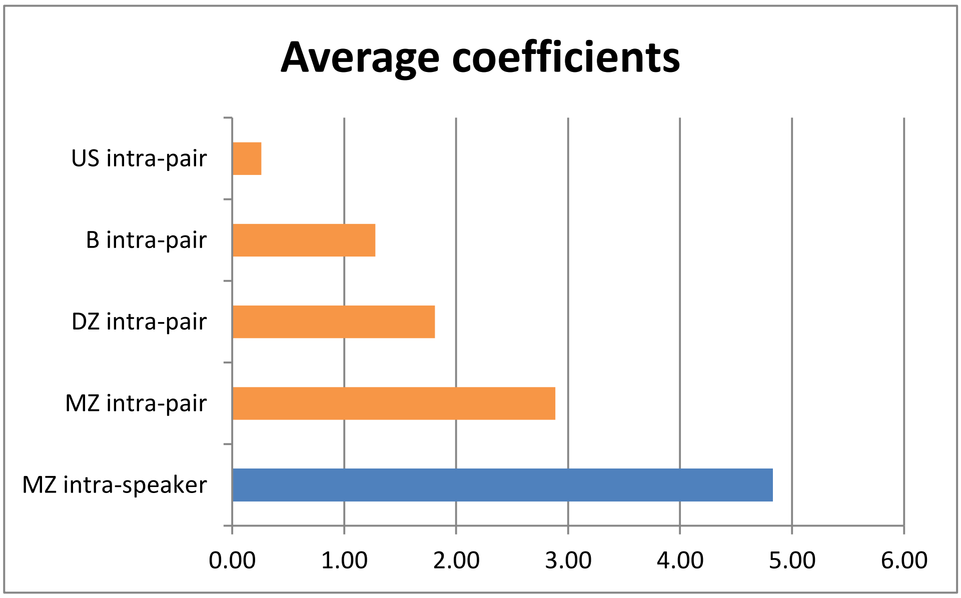


Figure 45. Grading of average coefficients from US to MZ intra-speaker comparisons (US < B < DZ < MZ). MZ intra-speaker (matches or same-speaker) comparisons yield the highest coefficients. The larger the value, the more similarity between test and model (i.e. the two elements for comparison). Orange is used for intra-pair comparisons while blue is used for intra-speaker comparisons.

6.4.3. Special case study: MZ twins

The MZ intra-pair comparisons deserve special consideration. As they represent the cases of highest similarity in human beings, they have been more often studied than the other types of kinship relationships considered in this investigation [see Chapter 2 for a literature review]. In the case of FSC carried out using automatic recognition methods, the existence of previous studies which have also used *Batvox* for the voice comparison of MZ twins gives us the opportunity to compare our results with previous findings (cf. 6.5. *Discussion*).

For the MZ twins participating in our study, we have considered useful to compare the coefficients obtained by each speaker in the intra-speaker comparisons (IS) with the coefficients obtained by these same speakers in the intra-pair comparisons (IP). Table 39 contains this information, extracted from the general results shown in Table 37:

Table 39

Measuring MZ discrimination capability: the IS-IP value

MZ pair	**Intra-speaker (IS) comparison coefficient**	**Intra-pair (IP) comparison coefficient**	**IS-IP Difference**
01v02	3.48	3.79	-0.31
03v04	4.79	2.65	2.14
05v06	4.95	3.45	1.50
07v08	4.14	2.31	1.83
09v10	4.06	2.66	1.40
11v12	5.29	1.34	3.95
33v34	6.06	3.20	2.86
35v36	3.94	4.93	-0.99
37v38	4.52	3.54	0.98
39v40	6.74	0.64	6.10
41v42	5.9	3.53	2.37
43v44	4.08	2.59	1.49

Note. For each of the MZ twin pairs, we show the *IS-IP value*, calculated as the difference between the intra-speaker (IS) comparison coefficient and the intra-pair (IP) comparison coefficient. Only 2/12 cases (highlighted in red) show negative values, indicating that in those cases the automatic system would not be able to discriminate the twin members of the MZ pair. (Note that only the IS coefficient for one twin member are chosen)

We have calculated an *IS-IP value* to measure the difference between the intra-speaker (IS) comparison coefficient and the intra-pair (IP) comparison coefficient. This has been done per speaker and speaker pair. Note however that for the IS coefficients, we have only taken into account the values obtained by one member of the pair: the twin member with the even number in his pair (i.e. 02, 04, 06, 08, etc.). The selection of the IS coefficients of the odd pairs did not yield any negative value. That is the reason why we show the results of the even numbers; as explained above, the interest of this calculation lies in finding any possible speaker pair subject to discrimination errors by the system under test.

As shown in Table 39, only two cases out of twelve MZ pairs show a negative value in their IS-IP value, meaning that the IP coefficient is larger than the IS coefficient. This implies that in these two cases the automatic system *Batvox* would not be able to discriminate between one twin and the other. In positive values, we can say that in 83.3% of the total MZ cases, the system identifies an identical twin without falsely accepting his cotwin. In Figure 46 we draw the IP and IS coefficient values per MZ twin pair, in IS-decreasing order to show how the trend “large IS – small IP” is followed in all cases except in the last two, corresponding to the MZ pairs 01v02 and 35v36, as we could also observe in Table 39. These two pairs account for the 16.7% not confirming the hypothesis that IS comparisons are always larger than MZ IP comparisons. However, as discussed in Section 6.5, this small percentage is in agreement with previous studies.

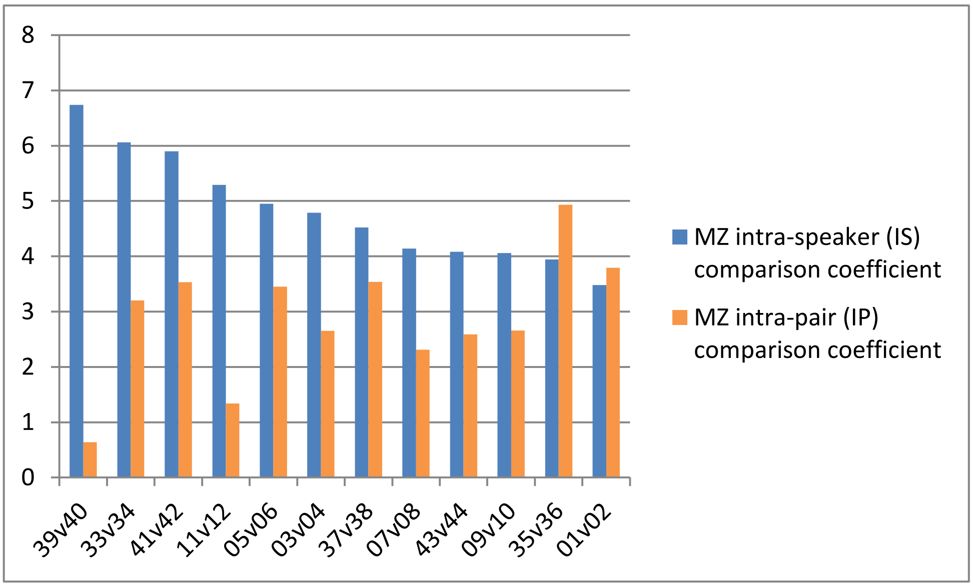


Figure 46. IS-IP difference per speaker pair. We show in the x axis the 12 MZ pairs and in the y axis the coefficient values for IS comparisons (blue) and IP comparisons (orange). Only the two last twin pairs would not be discriminated by the system.

The two specific cases of MZ twins which would not be recognized by the system explain the 9.9% EER obtained in Figure 47, where the line for matches (blue) is used in this case for intra-pair comparisons (only MZ) and the line for no-matches (red) represents the inter-speaker comparisons. In Figure 48 we have plotted the lines already shown in Figure 44, which show the overall system performance. The black line is for IS (intra-speaker) comparisons of all the speakers in the corpus, and the blue line represents the IP (intra-pair) comparisons, only for MZ.

In this new figure, one can distinguish a *left-shift* from the general IS-curve to the MZ IP-curve, which indicates the performance deterioration from a situation where the system has to recognize same speakers to a situation where identical-twin recognition takes place. The lines for the no-matches in both cases (compare yellow and red lines) are practically identical. In both cases, they represent different-speaker comparisons, while in one case (yellow line, i.e. no-matches in Figure 44) these tests compared the first session of each speaker with the first session of all the other speakers in our corpus; and in the other case (red line; i.e no-matches in Figure 47), the different-speaker tests were obtained from comparing each speaker's first session with all the other speakers' second session.

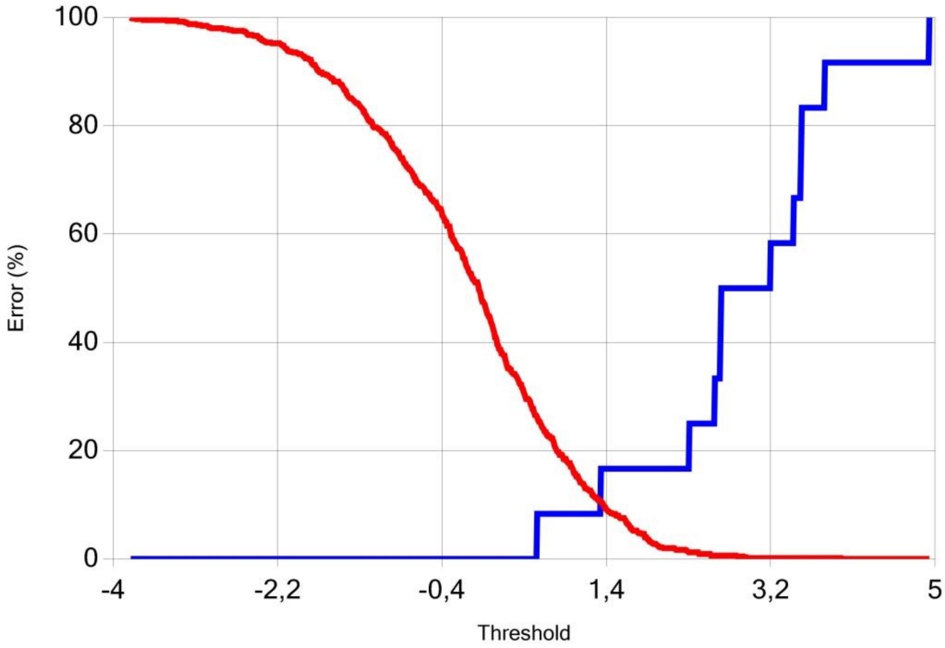


Figure 47. Cumulative distribution of scores for MZ intra-pair comparisons or matches (blue line) and inter-speaker comparisons or no-matches (red line). The EER obtained is 9.9%, indicating that some overlap between matches and no-matches exist.

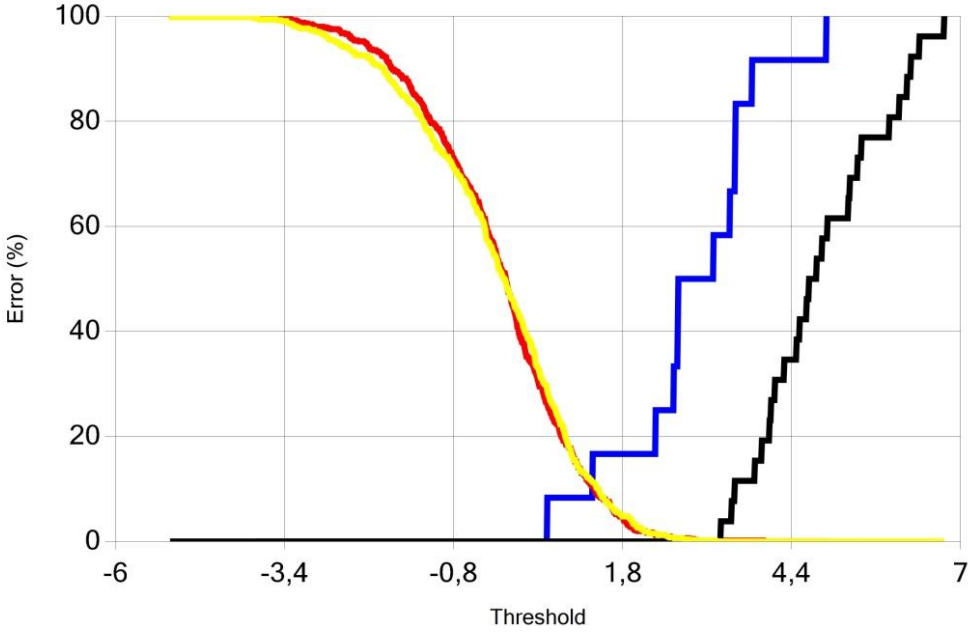


Figure 48. Cumulative distributions of scores for intra-pair (IP) comparisons (only MZ) in the blue line and intra-speaker (IS) comparisons for all the speakers in the black line. Lines yellow and red are for inter-speaker (different speakers) comparisons or no-matches. The only difference between both is that one (yellow) compared first session of every speaker with first session of all other speakers, while the other (red) compared first session of every speaker with second session of all other speakers.

6.5. Discussion

Several aspects can be discussed about the results obtained with the automatic system *Batvox*. On the one hand, we have tested the overall system performance with our speakers as *tests* and *models*, i.e. without taking into account the fact that part of these speakers are twins or siblings. This test has yielded intra- and inter-speaker comparisons. In other words: *matches* (for same-speaker comparisons) and *no-matches* (for different speaker comparisons). The 0% EER obtained for this first test shows that there were no FA or FR, which indicates a perfect performance of the system.

On a second test, we introduced the concept of "intra-pair (IP) comparison" while taking into account the fact that out of the 54 speakers considered, 24 are MZ twins, 10 are DZ twins, 8 are non-twin siblings and 12 are unrelated speakers. The results of comparing each speaker with his pair corroborated the hypothesis that higher similarity values would be found in MZ twins than in DZ twins, in siblings or in the reference population. On average, higher coefficients were obtained by MZ IP-comparisons, followed by DZ twins, brothers and unrelated speakers, in that order. This is the scale that we expected taking into account the degree of shared genes and shared environmental factors by pairs in these four speaker types [see Chapter 2; Section 2.2].

Finally, when the intra-pair comparison values only for the MZ twins are compared with the no-matches, we obtain a 9.9% EER, so a left-shift was observed in Figure 48 from the general IS-curve to the MZ IP-curve. This represents the deterioration in the system performance from a situation where the recognition is between same speakers to a situation where identical-twin recognition takes place. These results could be compared with the 11% EER obtained by Künzel (2010) who also studied MZ twins. Although he studied both male and female twins, and two speaking styles (read speech and spontaneous speech) we have considered here only the results for male twins and spontaneous speech. The male participants in Künzel's study were 9 MZ pairs while in our investigation there are 12 pairs. Yet the EER percentages are very similar, indicating that the rate of false acceptance of other twin by this system is around 10%. Having a closer look at the data for the individual twin pairs (i.e. comparing the IP and the IS values), Künzel found that some speakers were more easily identified than others. Our study also points in this direction, as the coefficients in the IS and IP comparisons differ between pairs, sometimes considerably (see Table 39 and Figure 46). In fact, as it follows from the literature review carried out in Chapter 2, this heterogeneity appears as a common factor in most studies on twins' voices. In the previous analyses for this investigation we have observed the same phenomenon: that different twin pairs exhibit different results when an intra-pair comparison is carried out, regardless of the type of phonetic-acoustic examination (be it formant trajectories or glottal characteristics). Indeed, this need not be a characteristic exclusively linked to twins but common in speaker recognition. As Doddington et al. (1998) explain, different speaker typologies could be established on the basis on how easily recognized/imitated they are. This implies that, in terms of FA and FR, "a considerable amount of the errors in an experiment, may be linked to only a few speakers" (Künzel, 2010: 264).

The other study which has analyzed twins' voices using *Batvox* (version 3.0) focused only in female voices (Kim, 2009), so the results in that study are not comparable with ours. From the investigation of Künzel (2010) we know that there is an important sex-related difference in the performance of the automatic system, this being superior for male as compared to female voices [see Chapter 2, Section 2.4.4; and Künzel, 2010, cf. *Are women's voice more problematic?*]. Yet, it is worth-mentioning that Kim (2009) also found that in 9 out of 22 cases, twins could be misidentified. She specifically refers to a situation where intra-twin LRs in the same speaking style condition were higher than intra-speaker LRs in different speaking style condition.

6.6. Conclusions

The most important conclusion which can be drawn from this analysis is that the similarity coefficients yielded by the automatic system decrease exactly as the kinship relationship of the speaker pairs decreases. In Chapter 2 we explained our reasons for sustaining the hypothesis that higher similarity values (hence worse recognition) would be found in MZ intra-pair (IP) comparisons than in DZ IP-comparisons. In turn, these speakers would be more similar than non-twin brothers (B) and the latter more similar than unrelated speakers (US). The justification for this lies in the fact that MZ twins share 100% of their genetic information and in general they also share educational and environmental background, while DZ twins share 50% of their genes but usually the same external influences as MZ twins. Sharing the same genetic information as DZ twins, B are supposed to share less environmental characteristics due to the age gap; and finally US share neither their genes nor their environmental background. This reasoning gives rise to the scale: MZ > DZ > B > US, where '>' means 'more similar than'; for the aim of our investigation, at least in voice terms.

After conducting the other voice analyses suggested in this thesis (namely, glottal analysis and formant-trajectory approaches), tackling the issue of voice similarity also from an automatic perspective seemed appropriate and necessary for a more thorough understanding of how voice characteristics can differ or concur in very similar speakers. Since the cepstral features in which automatic systems are usually based depend largely on anatomical-physiological foundations, we suggested as general hypothesis for this analysis that these parameters would be genetically related and, therefore, that higher similarity values would be found in MZ twins than in DZ twins, in siblings or in the reference population. To our knowledge, this is the first time that this hypothesis is tested for an automatic system using the four types of speakers mentioned (MZ, DZ, B and US). The underlying idea behind this hypothesis is not foreign to phonetic studies, however. For instance, Künzel (2010: 251) sustains that "the more similar the geometry of two vocal tracts is, the more similar will be the respective similarity coefficients, or LRs" and that "this problem is particularly relevant to related speakers, most extremely for identical (MZ) twins" (Künzel, 2010: 251). As a matter of fact, the issue of how the comparison of very similar speakers can affect the recognition performance of an automatic system has been investigated before [see Chapter 2], albeit almost exclusively using MZ twins as participants.

In our study, the results indicate a very good performance of the system when only inter-speaker comparisons (as no-matches) and intra-speaker (as matches) are taken into account. With an EER of 0%, we can say that the system performance is perfect, as no FA or FR occur. Yet, when intra-pair comparisons for MZ twins are considered, the performance deterioration is not as high as one could expect for cases of extremely similar speakers. The results indicate a relatively good performance of the automatic system, as only 2 out of 12 MZ pairs would not be recognized by the system. This represents a 16.7% of the total MZ cases, where the values in the intra-speaker comparisons are smaller than in their intra-pair comparison.

When comparing our results with previous findings by other authors who have tested the same automatic system with twins, we have been able to corroborate the widely-reported finding in ASR[147] that some speakers are simply more easily identified than others. The EER 9.9% in our study (see crossover point in Figure 47), corresponding to 2 out of 12 MZ twins who would be misidentified, is comparable to the EER 11% in Künzel (2010), indicating that confusion or non-distinction between twins occurred. The issue of the "striking performance inhomogeneities among speakers within a population" was already raised by Doddington et al. (1998) and we already referred to it in Chapter 5 (glottal analysis), where some cases (16.6%) were found of speakers exhibiting large self-unlikeness (i.e. they were very dissimilar when comparing their first and second recording session).

To sum up, testing the performance of an automatic speaker recognition system using identical twins implies a strong reduction of inter-speaker variation and, as explained by Künzel (2010), this is a most challenging task since "the *a priori* chances for a target voice to be *very* similar to the reference voice is much larger than within a set of unrelated speakers" (Künzel, 2010: 269). We agree with him in considering that "a system that identifies an identical twin without falsely accepting the other twin is probably fit for use in the forensic environment" (Künzel, 2010: 274). The explanation for this seems logical: the system works even when it is being tested in a disadvantageous situation, which could be compared with a situation where there is distortion or cross-language samples to compare. All these are challenging situations. However, a real case where twins' voices should be compared is not the most frequent situation in a forensic setting, basically because of the low incidence of twin births [see Chapter 2].

Yet, the importance of investigating twins' voices goes beyond this pragmatic view, i.e. it is relevant *per se*, regardless of how many real cases involve the comparison of twins. First, the comparative study of MZ and DZ twins can reveal the genetic influence of the parameters under study [see Chapter 2; Section 2.2]. Hence the importance of carrying out studies with both types of twins, not only MZ twins. The finding that certain voice parameter or parameters is/are genetically marked entails a good performance of any system which would be based on such parameters because the typical speakers for comparison would be usually genetically unrelated, which means that the system would be good at separating them. Second, the consideration of further types of kinship relationships, apart from MZ and DZ twins, such as non-twin siblings can help clarify certain unresearched issues, such as the interplay between genetical and environmental influences in voice. If environmental factors had no effect on the parameters under

[147] In Künzel (2010: 264): "At first glance the large differences between twin pairs corroborate a finding reported in nearly all studies on speaker identification, be it by man or machine, that some speakers are identified more easily than some others and that a considerable amount of errors in an experiment may be linked to only a few speakers (cf. Doddington et al., 1998).

study, and these were exclusively based on genes, there would not have been differences in the coefficients obtained by DZ and B. However, some differences were observed, despite the small sample of both DZ and B speakers.

From our investigation, it seems clear that the cepstral parameters in which the automatic system *Batvox* is based are genetically influenced. It is well-known that these features relate to the geometry of the vocal tract, so some physical similarity between twins is expected to be encoded in DNA. Yet, the different use and configuration of the vocal apparatus could be exploited by twins in different ways, which could leave a generous margin for intra-pair variation (Nolan & Oh, 1996; Loakes, 2006a). These different usages would be more related to learned aspects than to inborn characteristics.

Besides, it has not been mentioned so far that neither the group of MZ twins nor the DZ twin group are homogenous as far as their genes are concerned. As explained in Chapter 2, MZ twins can be monochorionic or dichorionic, depending on whether they shared the same placenta or have two different placentas instead; they can also be monoamniotic or diamnotic, depending on whether they share the same amniotic sac or not. How this can affect the differences found between one twin pair and another, as well as the influence of *epigenetics* in twin differences, will be addressed in next chapter [see Chapter 7; cf. *Conclusions*], as they do not only affect the automatic analysis described in this chapter but the whole of the investigation.

7. CONCLUSIONS

7.1. Summary of the research approach and main conclusions

The main objective of this investigation has been analyzing the speech and voice characteristics of a set of speakers with the aim of shedding light on a relevant but under-researched topic: the interplay of genetic and non-genetic forces in speakers' voice similarity. This issue is of special interest to Forensic Phonetics (and more specifically to Forensic Speaker Comparison), a discipline at the crossroads between Applied Linguistics, the Law and Biometric Sciences. Nevertheless, the conclusions drawn from this study could be relevant for General Phonetics or for other disciplines, as explained in Section 7.4.

For our research purpose, four types of speakers were recorded: 24 monozygotic (identical) twins, 10 dizygotic (non-identical) twins, 8 non-twin siblings and 12 unrelated speakers. The characteristics of these speakers (i.e. gender, age, dialect) and their selection criteria are explained in detail in Chapter 3. The most important aspect to highlight here is the reason for having chosen these specific types of speakers. In line with what is known about the twin method, the *equal environment assumption* and previous studies related to twins, five working hypotheses were put forward. Firstly, we assumed that a speaker would be similar to himself (in his voice characteristics), i.e. from one recording session to another. For this reason, intra-speaker variation was measured. This assumption (H_1) was made for all speaker types. Secondly, accepting that MZ twin pairs are the most similar speakers that can exist (because of their shared genes and shared environmental influences), we hypothesized (H_2) that MZ intra-pair comparisons would yield matching scores similar to those obtained in intra-speaker comparisons. The third hypothesis (H_3) implied that DZ intra-speaker comparisons would yield relatively large matching scores but not as large as in the case of MZ twins (on the one hand, because they share the same environmental characteristics as MZ cotwins; on the other hand, because the genetic load shared by DZ cotwins is half that of MZ twins). In the fourth hypothesis (H_4), we stated that the intra-pair comparisons in the case of brothers would yield matching scores over certain *background baseline* (i.e. the values obtained by the background population, namely the unrelated speakers). That means that brothers should be more similar than unrelated speakers because they share 50% of their genes, exactly the same proportion as DZ twins, and they usually have environmental influences in common, although to a lesser extent than DZ twins. Finally, we hypothesized (H_5) that a background baseline should exist for the matching scores obtained by the unrelated speakers.

All in all, what has been suggested is that the confirmation of the above-described hypotheses would point to the genetic influence of the parameters under study, and thus to their robustness for their use in a typical FSC situation where the speech samples of a suspect and an offender are available for comparison. As these two speakers would be in principle genetically unrelated, making a comparison based on genetically based features could allow distinguishing them.

The general hypotheses described above were applicable to the three analytical approaches considered: 1) analysis of the formant trajectories vocalic sequences; 2) glottal-source analysis; and 3) automatic analysis using the software *Batvox*. However, taking into account the differences existing between them, both in the acoustic characteristic of the parameters and in the analysis methodology, specific hypotheses were formulated for each type of approach. In the following paragraphs we list the main conclusions drawn from this investigation, depending on the parameters analyzed and the perspectives adopted.

1. *Analysis of the VS formant trajectories:*

- The first and more important hypothesis to test within this approach was the one entailing more difficulties for corroboration, as different results (depending on the type of speaker or type of comparison) pointed to two opposite directions to explain those results: either the stronger influence of genetic endowment or the prevalence of non-genetic (i.e. external, learned, environmental) factors. All in all, we can conclude that the genetic influence on these parameters is large, as the hypothesized decreasing scale MZ > DZ > B > US in LR-based results occurs in all cases except for the B pair 23-24, with strikingly high LRs. Although several reasons could be found to explain this discordant result, we should further conclude that learned habits must also play an important role in the formant trajectories of VS. While genetic factors are undoubtedly influencing the acoustic parameters studied, their impact might not be as clear as for the other parameters studied from the other two analysis perspectives.

- Our investigation has clearly shown that a forensic-comparison based on all the 19 Spanish VS fused together yields better performance than individual systems, each based on a single VS.

- Finally, out of the two parametric procedures used for the curve fitting of the VS formant trajectories (polynomial functions or DCT functions) we cannot conclude that one outperforms the other when comparing system accuracy. Yet, cubic polynomials and third-degree DCT functions were found to better correlate with the original formant trajectories.

2. *Glottal-source analysis*

    - The glottal parameters examined within this second approach were found to be genetically influenced (H1), as higher LLRs were yielded by the intra-pari (IP) comparisons of MZ twins than in DZ twins, non-twin brothers or unrelated speakers (i.e. MZ > DZ > B > US). The only cases where the comparison results contradicted our established hypotheses could be explained upon examination of their anamnesis/questionnaires. The most worth-mentioning case is that of the strikingly dissimilar MZ pair 11-12, where their different smoking habits and certain attitudinal factors favoring voice divergence towards their cotwin must have outweighed their expected similarity due to anatomical resemblance.

    - In relation to the question of whether some glottal parameters yield better identification results than others, the hypothesis that the biomechanical estimates of the glottal waveform would be especially speaker-specific (H2), established according to preliminary studies (San Segundo, 2012), was not corroborated. Upon consideration of the whole set of parameters available in the voice-analysis software used, other parameters outranked those already mentioned. The five most relevant (considering all the comparisons for this study, regardless of the type of speaker and the type of comparison) were: *p6* (ratio between the energy of the glottal source to average acoustic difference and the average acoustic wave), *p32* (frequency of the glottal source power spectral density at half sampling frequency relative to first maximum frequency), *p21* (first maximum of glottal source power spectral density), *p8* (second cepstral coefficient of the glottal wave correlate) and *p59* (ratio between the contact gap flow escape and the total glottal flow).

3. *Automatic analysis*

    - The only research objective that we aimed at investigating within this last approach was whether there is more intra-pair similarity in MZ than in other speaker comparisons for the type of cepstral parameters in which *Batvox* is based. Since such parameters depend largely on anatomical-physiological foundations, we suggested that they should be somehow genetically influenced. This hypothesis was corroborated, as the similarity coefficients yielded by the automatic system decreased exactly as the kinship relationship of the speaker pairs decreases. In other words, the score sorting from largest to smallest resulted in the following scale of values: MZ > DZ > B > US.

Irrespective of the type of analysis carried out, an important conclusion can be highlighted, which is related to the large intra-speaker variation found for some speakers. We have also referred to this as “high self unlikeness” and it is a phenomenon that stretches over all the parameters and the analyses considered. It basically means that the comparison of the first recording session of a speaker with his own second recording session yields remarkably low values (either scores or LRs), suggesting that these speakers would be difficult to be recognized by a forensic-comparison system as the same person. In FSC terms, this would result in missed hits (false rejections). This lack of homogeneity in the intra-speaker comparisons would deserve special attention in future studies. Closely linked to this heterogeneity, another relevant conclusion that can be drawn from this study is that different results are found depending on the twin pair under comparison. Regardless of the voice-analysis perspective considered, some twins are found to be more similar than others. This is in agreement with previous studies about twins, where the comparison results are never found to be the same for all the pairs. In our case, this is notably marked for MZ twins. Maybe because it is the largest group, with 12 different pairs, it never behaves as a homogenous group. Precisely derived from what has just been said, the third and final general conclusion that we can draw is that the more voice-analysis angles from which we tackle speaker comparisons, the more opportunities for avoiding missed hits and false acceptances. It has been made sufficiently clear that certain parameters and analyses can show that speakers are (misleadingly) very different to themselves while, fortunately, other approaches can yield consistent-with-fact high similarity values.

7.2. Original contributions to the research field

In this section we will highlight the main contributions of this investigation to the field of Phonetics, and especially to the discipline of Forensic Phonetics and to the twin-voice-studies realm. Apart from the results of the different analyses carried out, which represent novel, original contributions to the field, the following aspects are also completely original contributions of this thesis:

- *Research topic*: First of all, it should be stressed that this study represents, to our knowledge, the first investigation into the voice characteristic of Spanish twins and non-twin siblings. Previous studies for this language are pilot experiments undertaken by this author and coauthors (San Segundo, 2010a; San Segundo, 2010c; San Segundo, 2012; San Segundo, 2013a; and San Segundo & Gómez-Vilda, 2013).

- *Type of speakers*: Furthermore, the consideration of the four types of speakers already mentioned (MZ, DZ, B and US) seems also novel and original of this thesis. Most previous investigations in other languages have studied either MZ and DZ twins or MZ twins alone. Only a minority of the studies involves either only non-twin siblings or MZ and non-twin siblings together.

- *Literature review*: Around thirty voice-related twin studies have been reviewed, distinguishing between perceptual, acoustic, articulatory and automatic approaches. In Appendix F, these studies have been classified following a chronological order in what is, to our knowledge, the first attempt to summarize, for all the voice-related studies on twins which could be identified in the literature, the following aspects: the speaker sample (n), differentiating whether the twins were MZ or DZ, the gender of the twins and the data collection method. Due to the scarcity of twin registries, this could be useful for future researchers undertaking a study of the voice characteristics of twins.

- *Database collection*: A relevant aspect in which this study contributes to the advances in the field is the creation of a database with the voices of twins and non-twin siblings. Other studies (see Appendix E) have carried out different types of acoustic analyses on the basis of a previously gathered database or corpus. For our investigation, the corpus of twins' voices was designed and collected 'ad hoc', as no previous databases of Spanish twins and non-twin siblings existed. It is also the first time, as far as we know, that a Spanish corpus is collected with part of its speaking tasks having been telephone-filtered with the method described in Chapter 3. Other Spanish corpora which contain telephone-filtered recordings are described in San Segundo, Alves and Fernández (2013), although the filtering procedure differs in this case. For the originality of the speaking tasks designed in this corpus, see the next point, as it specifically refers to the methodology.

- *Methodological approach*: In the first chapter of this thesis the main current methodologies in FSC have been described. As a result of this review, it was concluded that adopting a hybrid perspective, which combines traditional and automatic analyses, each with its strengths and weaknesses, could be considered the most comprehensive approach to speaker comparison. For that reason, our study draws on a three-folded analysis that combines (1) traditional phonetic-acoustic parameters with (2) not only features but also techniques which are rather characteristic of automatic methods. The joint consideration of traditional and automatic perspectives is not novel in FSC, but it *is* the study of twins' voices from these three angles: VS formant trajectories, glottal-source features and vocal-tract cepstral parameters.

Within this methodological subsection, we would also like to highlight the following aspects related to the corpus design which represent original contributions. Firstly, we could point out the different speaking tasks in which the corpus is divided. This implies that different speaking styles are available for comparison, even though this type of comparison was not among the goals of our research. In addition to the most commonly found speaking styles in twins' research, (i.e. reading vs. spontaneous conversation), we have distinguished between "informal interview with the researcher" and "spontaneous conversation between the siblings" (or conversational partners in the case of unrelated speakers). Considering the importance of the "intra-sibling mimetism", separating between a communicative situation where the speakers are familiar to each other and another situation where they speak with someone more unfamiliar is very relevant from a sociolinguistic perspective. As far as we know, the creation of these two different conversational situations in twins' interactions is original of this thesis. Furthermore, the design of the second speaking task (fax exchange to elicit specific VS) is also novel, albeit indebted to Morrison, Rose and Zhang (2012). The originality of our design lies in the following phases: 1) methodology for search of words containing the VS of interest, and 2) procedure for the creation of the fax sheets with a semantic context.

7.3. Implications

The main implications of this thesis are related to real forensic casework. Since all the parameters tested for this investigation have proved to be genetically conditioned, to a greater or lesser extent, they would be useful for comparing speech samples of known and unknown origin, as found in legal cases. Moreover, as different parameters have been tested depending on the type of analysis conducted, we could indicate separately which features were found more useful in the formant-trajectory analysis, on the one hand, and in the glottal-source study, on the other hand. For the automatic approach, based on cepstral parameters, there was no attempt to distinguish the speaker discriminatory potential between features.

In relation to the analysis of formant trajectories, although there is not a VS which outranks among the others when considering the 19 forensic-comparison systems (one per VS) separately, we have found that the /uo/ sequence is especially difficult to fit by means of a parametric curve with a high correlation degree. Considering the time limitations which usually characterize phonetic forensic casework, if all the VS could not be labeled and extracted for further analysis, /uo/ would be a good candidate to discard, at least until a better-correlated curve technique is found. Anyway, the best results are yielded by a combination of as many VS as possible. Another implication of this study for forensic applications is related to the fact that third-degree parametric functions (either DCT or polynomial) fit better the formant trajectories of VS than second- or first- order functions. This happens both for F2 and F3, so in future forensic casework analyzing these parameters, it would not be necessary to test smaller degrees than the third ones to fit the trajectories.

For the glottal-source analysis, seven parameter subgroups were distinguished. We listed the three most relevant parameters in each of the 81 comparisons carried out (54 intrapair + 27 intraspeaker) and the five parameters with a higher number of occurrences turned out to belong to four parameter subsets: “singularities of the glottal source Power Spectral Density” (p21 and p32); “fundamental frequency and distortion parameters” (p6), “cepstral coefficients of the glottal source PSD” (p8) and glottal gap coefficients (p59). In Section 7.1 we have listed those specific features, found to be relevant in more comparisons than the others. The implication of this is that in case of time constraints, the forensic investigation should focus on these specific parameters, which seem to be the most important to differentiate between same speakers and different speakers. However, the software used is intended to provide the results of the offender-sample comparison based on the total 68 parameters and not considering them separately. Since this is a rather automatic approach, the consideration of the whole set of parameters is not more time-consuming than the consideration of isolated features. It is also important to note that the parameters which seem relevant for distinguishing a pair of speakers may not be relevant for another pair. Moreover, the use of the whole set of parameters might be recommended, as certain features can be more explanatory or illustrative, exhibiting less discriminatory potential, while others can be more speaker specific but less difficult to relate to physical characteristics. In other words, while some parameters like the ratio between the contact gap flow escape and the total glottal flow (p59) may be more easily explainable to and understood by the court, others like the cepstral coefficients of the glottal PSD (with almost all of its parameters having high occurrences in a relative homogeneous distribution) can be more powerful for speaker discrimination. This is in line with the trade-off between automatic and traditional features described by Rose (2006) and mentioned in Chapter 1.

7.4. Limitations of the study and directions for future research

This study has some limitations as regards, for instance, the number of recorded speakers or the absence of female voices, both of which suggest further directions for research. That is, apart from considering the enlargement of this study to include also female twins and non-twins, future studies should also contemplate the collection of a larger database. Despite not being an easy task (at least for the case of Spanish adult twins), the search for more twins is undoubtedly a key issue for conducting better statistical analyses, and probably of different sort. Yet, not having considered many speakers in this study has facilitated the detailed examination of their questionnaires, which allowed us to explain certain discordant results.

The most important limitation of this study, however, is related to the strict application of the twin method. From the *equal environment assumption* usually linked to this method, we have learned that the excess of similarity (for an investigated parameter) exhibited by MZ twins which is not present in DZ pairs must be due to genetic causes. Although we have taken advantage of the principles and assumptions of the twin method for our forensic-related study, it would be sensible to use a strict application of this twin methodology before assuring that the results of our study undoubtedly point to the genetic influence of the phonetic features studied. We refer to the use of heritability estimates or concordance rates, in which the expected elevated similarity in MZs over DZs is often reported, depending on whether it is a continuous or a dichotomous trait (cf. Tomblin & Buckwalter, 1998; and also Section 2.2).

Other minor limitations of the study refer to the fact that the recordings used for the three types of analyses were high-quality recordings, i.e. not presenting any kind of channel distortion, like the ones affecting telephone transmission. We say that this is a minor limitation since this corpus was designed to specifically take into account forensically realistic conditions. For that reason, in the recording set-up described in Chapter 3 speakers held a telephone conversation in several of the speaking tasks designed. Besides, part of the corpus has been telephone-filtered, so speech samples are available for future studies analyzing speech features in telephone-degraded conditions. Yet, for a first approach to Spanish twins' voices, it seemed prudent to undertake these analyses in the best possible recording conditions.

According to previous studies, further possible methodological innovations that could be added to our study of twins' voices are mentioned below. On the one hand, for the formant-trajectory analysis, the application of time-normalization techniques to the trajectories has been considered in studies like Enzinger (2010) and Morrison (2008, 2009c). In both cases, the computation of time-equalized trajectories entailed better performance than the use of raw trajectories. On the other hand, our formant-trajectory analysis could also benefit from another procedure for the calculation of LRs different from the MVKD used on this occasion. We refer to the Gaussian mixture model-universal background model (GMM-UBM), widely used in ASR (Alexander & Drygajlo, 2004; Becker, Jessen & Grigoras, 2008) and also applied specifically to the analysis of VS formant trajectories (Morrison, 2009c). In this study, the comparison of MVKD and GMM-UBM resulted in the outperformance of the latter. Finally, it could be of interest to analyze separately each VS distinguishing between stress conditions (stressed or unstressed) in order to check whether this entails better system performance. Although for this investigation this distinction has not been made, the corpus design encompasses various examples per stress condition that are available for future research. Likewise, the corpus includes some extra words (e.g. *hueso*, *hielo*) containing VS in certain phonetic contexts which were deemed to give rise to varying pronunciations. Therefore, we suggest the examination of these units in next-coming investigations.

In relation to the glottal-source analysis, the most interesting aspect to investigate in future studies would be the analysis of possible correlations between the parameters listed as the most relevant. Since many of them do not appear to be completely independent from each other, investigating their relationship and taking into account their potential correlation would possibly lead to more robust results.

In addition, it seems that future studies could benefit from the fusion of the three different systems tested for this investigation. The independence of glottal features from vocal-tract characteristics makes them specially promising for an improvement of an overall forensic system performance.

This study has also allowed us to glimpse other disciplines or research fields which could help not only elucidate the origins of twin's voices similarities and differences, but also suggest new methodologies for their study. In Chapter 6 we concisely raised the topic of speech convergence in relation to the sibling mimetism apparently found in some pairs. In Section 4.5.2 we discussed the possible reasons behind the great similarity found for a non-twin sibling pair in relation to an important research line investigating convergence and imitation patterns in speech which occurs between speakers in the course of conversational interactions (e.g. Pickering & Garrod, 2004; Pardo, 2006; Truong & Trouvain, 2012), with some studies focusing specially on the convergence of phonetic features in close acquaintances (Kalmanovitch, 2012), or college roommates (Pardo et al., 2012; Coupland, 1984). Since the methodological approaches of these recent investigations –indebted to the theory of accommodation (Giles, Coupland & Coupland, 1970) – may have not been fully applied to Forensic Phonetics, it would be extremely interesting to revisit the study of twins' voice similarities and differences in light of the postulates of this research approach.

Finally, future research focusing on twins' voices should pay more attention to the concept of epigenetics, which we briefly described in Chapter 2. We have continuously referred throughout this thesis to two basic forces which would intermingle to explain the (dis)similarities in twins and non-twins' voices, namely, genetic and environmental factors. The often-neglected third factor, i.e, epigenetics, might explain how the alteration in the expression of specific genes (caused by mechanisms other than changes in the underlying DNA sequence) is behind the striking dissimilarities found for some twin pairs. It was succinctly mentioned in Chapter 6 that, despite being frequently assumed to be so, a set of twins do not constitute a homogenous group as far as their genetic endowment is concerned. For instance, MZ twins can be monochorionic or dichorionic, depending on whether they share the same placenta or not. The fact that spontaneous mutations tend to occur more often in dichorionic MZ twins makes them more likely to differ genetically than monochorionic MZ twins (Stromswold, 2006; cf. Chapter 2). Whether the existence of different types of MZ twins affects their voice similarity or not is an open research question, which, in any case, would require specific DNA testing to obtain detailed information about the zygosity of the twin pairs.

All in all, the most important aspect to highlight in relation to the limitations of this study, strongly linked to the suggestions for future studies, is the fact that this is the first investigation into the voice characteristics and speech patterns of Spanish twins and non-twin siblings. Therefore, a strong effort has been done in order to obtain a considerable number of these speakers, as no previous databases existed. In order to fill this gap, the occasion seemed ripe to design a corpus comprising many speaking styles and other tasks like the vocal control techniques explained in Chapter 3. In other words, many more analyses (and probably also more computationally demanding) could have been conducted using the speakers' voice samples recorded 'ad hoc' for this thesis. However, this has been limited by the fact that an important contribution of this investigation has already been the design and collection process of the corpus itself. This has been created with special technical care in order to attain a database made up of high-quality recordings and multiple speaking styles, in addition to a complete questionnaire about the speakers. Moreover, a strong emphasis has been placed in achieving forensically realistic conditions (e.g. two recording sessions, spontaneous conversations, etc.) without neglecting important phonetic issues, such as the control of variables (phonetic context of the studied segments), the use of phonetically balanced texts or the use of eliciting techniques. All of these aspects open, not only this investigation but also the collected corpus, for more detailed studies on twin-related topics as well as on non-twin Spanish male voices in general.

REFERENCES

Abril, A., Ambrosio, E., de Blas, M., Caminero, A., García, C., & de Pablo, J. (2009). *Fundamentos de psicobiología*. Madrid: Sanz y Torres.

Agnitio Voice Biometrics (2013). Batvox 4.1 Basic User Manual [Computer software].

Aguilar, L. (1999). Hiatus and diphthong: Acoustic cues and speech situation differences. *Speech Communication*, *28*(1), 57--74.

Aguilar, L. (2010). *Vocales en grupo*. Madrid: Arco/Libros.

Aitken, C., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 53*(1), 109--122.

Alarcos Llorach, E. (1965). *Fonología española* (4th ed.). Madrid: Editorial Gredos.

Alexander, A., & Drygajlo, A. (2004). *Scoring and direct methods for the interpretation of evidence in forensic speaker recognition*. Paper presented at the 8th International Conference on Spoken Language Processing (ICSLP), Jeju, Korea.

Alpert, M., Kurtzberg, R., Pilot, M., & Friedhoff, A. (1963). Spectral characteristics of the voices of twins. *Acta Geneticae Medicae et Gemellologiae*, *12*, 335--341.

Alves, H., Gil, J., Pérez, C., & San Segundo, E. (2014). La cualidad individual de la voz y la identificación del locutor: el proyecto CIVIL. In *Fonética Experimental, Educación Superior e Investigación* (pp.591--611). Madrid: Arco/Libros.

Alves, H., Rico, J., & Roca, I. (2010). *BuFón: Buscador de patrones fonológicos*. Retrieved from: http://www.estudiosfonicos.cchs.csic.es/fonetica/bufon?p=presentacion.

Anderson, A. H., Badger, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., … & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351--366.

Anderson, S. (1985). *Phonology in the twentieth century*. Chicago: University of Chicago Press.

Ariyaeeinia, A., Morrison, C., Malegaonkar, A., & Black, S. (2008). A test of the effectiveness of speaker verification for differentiating between identical twins. *Science & Justice*, *48*(4), 182--186.

Aronson, A., & Bless, D. (2009). *Clinical voice disorders* (4th ed.). New York: Thieme.

Baken, R., & Orlikoff, R. (2000). *Clinical measurement of speech and voice* (2nd ed.). San Diego:

Singular Thomson Learning.

Baldwin, J., & French, P. (1990). *Forensic phonetics*. London: Pinter Publishers.

Barron, A. (1997). Lecture on *Introduction to Statistics. Personal Collection of A. Barron*. Lecture, Department of Statistics, Yale University. Retrieved from http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

Battaner, E., Gil, J., Marrero, V., Llisterri, J., Carbó, C., & Machuca, M., … & Ríos, A. (2003). VILE: Estudio acústico de la variación inter e intralocutor en español. In *SEAF 2003: Actas del II Congreso de la Sociedad Española de Acústica Forense* (pp. 59--70).

Becker, T., Jessen, M., & Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian mixture models. In *Proceedings of Interspeech 2008* (pp. 1505--1508).

Benson, M., & Deal, J. (1995). Bridging the individual and the family. *Journal of Marriage and The Family*, 561--566.

Berger, C., Robertson, B., & Vignaux, G. (2010). Interpreting scientific evidence. In I. Freckelton & H. Selby, *Expert Evidence*. Sydney: Thomson Reuters.

Bernard, J. (1967). *Some measurements of some sounds of Australian English* (Doctoral dissertation). Sydney University.

Berry, D. (2001). Mechanisms of modal and nonmodal phonation. *Journal of Phonetics*, *29*, 431--450.

Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., & Meignier, S., … & Reynolds, D. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, *2004*, 430--451.

BioMet®Phon. (2014). *User's Manual*. Universidad Politécnica de Madrid. [Computer software]

BioMet®Soft. (2010). Universidad Politécnica de Madrid. [Computer software] Retrieved from www.biometrosoft.com.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer [Computer software] (Version 5.3.79). Retrieved from http://www.praat.org/.

Boomsma, D. (1998). Twin registers in Europe: An overview. *Twin Research*, *1*(01), 34--51.

Boone, D. (1977). *The voice and voice therapy*. Englewood Cliffs, N.J.: Prentice-Hall.

Borzone de Manrique, A. M. (1979). Acoustic analysis of the Spanish diphthongs. *Phonetica, 36*(3), 194--206.

Boyanov, B., & Hadjitodorov, S. (1997). Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. *IEEE Engineering in Medicine and Biology Magazine*, *16*(4), 74--82.

Brümmer, N. (2005). *FoCal: Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers*. Retrieved from: https://sites.google.com/site/nikobrummer/focal. [Computer software]

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2), 230--275.

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., & Karafiat, M., … & Strasheim, A. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2072--2084.

Bruyninckx, M., Harmegnies, B., Llisterri, J., & Poch, D. (1994). Language-induced voice quality variability in bilinguals. *Journal of Phonetics*, *22,* 19-31.

Burlingham, D. (1952). *Twins; a study of three pairs of identical twins*. New York: International Universities Press.

Cambier-Langeveld, T. (2007). Current methods in forensic speaker identification: Results of a collaborative exercise. *International Journal of Speech, Language and the Law*, *14*(2), 223--243.

Campbell, J. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE 85, 9* (pp. 1437--1462).

Campbell, W., Campbell, J., Reynolds, D., Singer, E., & Torres-Carrasquillo, P. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language, 20*(2), 210--229.

Champod, C., & Evett, I. (2007). Commentary on APA Broeders (1999) ‘Some observations on the use of probability scales in forensic identification’, Forensic Linguistics 6 (2): 228--41. *International Journal of Speech Language and the Law*, *7*(2), 239--243.

Charlet, D., & Peral, V. (2007). Voice Biometrics within the Family: Trust, Privacy and Personalisation. *E-business and Telecommunication Networks* (*Second International*

Conference, ICETE 2005, Reading, UK, October 3-7, 2005. Selected Papers), 3, 93--100.

Cicres, J. (2011). Transcripció i autenticació de gravacions en contextos judicials. *Llengua, Societat i Comunicació. Revista de Sociolingüística de la Universitat de Barcelona, 9*, 23-62.

Cielo, C., Agustini, R., & Finger, L. (2012). Vocal features of monozygotic twins. *Revista CEFAC*, *14*(6), 1234--1241.

Cornut, G. (1971). Génèse de la voix de l'enfant. *Journal Français d'Otorhinolaryngolie, Audiophonologie et Chirurgie Maxillofaciale.*, *20*(2), 411--416.

Coulthard, M., & Johnson, A. (2007). *An introduction to forensic linguistics*. Abingdon, Oxon: Routledge.

Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language*, *46*, 49--70.

Dallapiccola, B., Stomeo, C., Ferranti, G., Di Lecce, A., & Purpura, M. (1985). Discordant sex in one of three monozygotic triplets. *Journal of Medical Genetics*, *22*(1), 6--11.

Daly, L. (1983). *Family communication: A sociolinguistic perspective* (Doctoral dissertation). Georgetown University, Washington.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *28*(4), 357--366.

Deal, R., & Emanuel, F. (1978). Some waveform and spectral features of vowel roughness. *Journal of Speech and Hearing Research*, *21*(2), 250--264.

Debruyne, F., Decoster, W., Van Gijsel, A., & Vercammen, J. (2002). Speaking fundamental frequency in monozygotic and dizygotic twins. *Journal of Voice*, *16*(4), 466--471.

Decoster, W., Van Gysel, A., Vercammen, J., & Debruyne, F. (2000). Voice similarity in identical twins. *Acta Oto-Rhino-Laryngologica Belgica*, *55*(1), 49--55.

Delgado, C. (2001). *La identificación de locutores en el ámbito forense* (Doctoral dissertation). Universidad Complutense de Madrid.

Delgado, C., Márquez, M., Olivas, M., & Barrios, L. (2009). Identificación forense de locutores (IFL): Categorización de parámetros acústicos y fono-articulatorios del español. *Revista Española de Lingüística, 39*(1), 33--60.

Dellwo, V., Kolly, M., & Leemann, A. (2012). *Speaker identification based on speech temporal information: A forensic phonetic study of speech rhythm in the Zurich variety of Swiss German*. Paper presented at the 21st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Santander, Spain.

Dodd, B., & McEvoy, S. (1994). Twin language or phonological disorder? *Journal of Child Language, 21*, 273--273.

Doddington, G. (2001). Speaker recognition based on idiolectal differences between speakers. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (pp. 2521--2524).

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proceedings of the 5th International Conference on Spoken Language Processing,* 1--5.

Drygajlo, A. (2007). Forensic automatic speaker recognition. *IEEE Signal Processing Magazine, 24*, 132--135.

Drygajlo, A., Meuwly, D., & Alexander, A. (2003). Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 689--692.

Dudenhausen, J. (2003). Die Mehrlingsschwangerschaft. In H. Bender, K. Diedrich & W. Künzel, *Handbuch der Frauenheilkunde und Geburtshilfe* (pp. 301--309). Munich: Urban & Fischer.

Eckel, F., & Boone, D. (1981). The s/z ratio as an indicator of laryngeal pathology. *Journal of Speech and Hearing Disorders*, *46*(2), 147--149.

Eckert, P., & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology, 21,* 461--490.

Enzinger, E. (2010). *Characterising Formant Tracks in Viennese Diphthongs for Forensic Speaker Comparison*. Proceedings of the 39th International AES Conference: Audio Forensics, Practices and Challenges, 47--52.

Eriksson, E., & Sullivan, K. (2008). An investigation of the effectiveness of a Swedish glide+ vowel segment for speaker discrimination. *International Journal of Speech, Language and* *the Law*, *15*(1), 51--66.

Evans, I., & Foulkes, P. (2009). *Speaker identification in whisper*. Paper presented at the 18th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Cambridge, UK.

Evett, I. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, *38*(3), 198--202.

Evett, I., & Buckleton, J. (1996). Statistical analysis of STR data. In A. Carraredo, B. Brinkmann & W. Bär, *Advances in Forensic Haemogenetics* (pp. 79--86). Heidelberg: Springer-Verlag.

Explorable.com. (2009). Snowball Sampling - Chain Referral Sampling. Retrieved 4 June 2014, from https://explorable.com/snowball-sampling

Fant, G. (1970). *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*. The Hague: Mouton.

Farrús, M. (2008). *Fusing prosodic and acoustic information for speaker recognition* (Doctoral dissertation). Universitat Politècnica de Catalunya.

Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. In *Proceedings of Interspeech* (pp. 778-781).

Feiser, H. (2009). *Acoustic similarities and differences in the voices of same-sex siblings*. Paper presented at the 18th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Cambridge, UK.

Feiser, H., & Kleber, F. (2012). *Voice similarity among brothers: evidence from a perception experiment*. Paper presented at the 21st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Santander, Spain.

Fierro, P. (2013). Dizygotic Twins - Twins, Triplets, and More. *Netplaces.com*. Retrieved 30 October 2013, from http://www.netplaces.com/twins-triplets-multiples/an-introduction-to-zygosity/dizygotic-twins.htm

Flach, M., Schwickardi, H., & Steinert, R. (1968). Zur Frage des Einflusses erblicher Faktoren auf den Stimmklang (Zwillingsuntersuchungen). *Folia Phoniatrica et Logopaedica*, *20*(6), 369--378.

Forrai, G., & Gordos, G. (1982). A new acoustic method for the discrimination of monozygotic and dizygotic twins. *Acta Paediatrica Hungarica*, *24*(4), 315--322.

French, P. (1994). An overview of forensic phonetics with particular reference to speaker identification. *International Journal of Speech, Language and the Law*, *1*(2), 169--181.

French, P., & Harrison, P. (2007). Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by Peter French & Philip Harrison. *International Journal of Speech Language and the Law, 14*(1), 137--144.

French, P., Nolan, F., Foulkes, P., Harrison, P., & McDougall, K. (2010). The UK position statement on forensic speaker comparison; a rejoinder to Rose and Morrison. *International Journal of Speech Language and the Law, 17*(1), 143--152.

French, P., & Stevens, L. (2013). Forensic speech science. In M. Jones & R. Knight, *The Bloomsbury companion to phonetics* (pp. 183-197). London: Bloomsbury.

Fuchs, M., Oeken, J., Hotopp, T., Täschner, R., Hentschel, B., & Behrendt, W. (2000). Die Ähnlichkeit monozygoter Zwillinge hinsichtlich Stimmleistungen und akustischer Merkmale und ihre mögliche klinische Bedeutung. *HNO*, *48*(6), 462--469.

Furui, S. (1989). *Digital speech processing, synthesis, and recognition*. New York: Marcel Dekker.

Galton, F. (1875). The history of twins, as a criterion of the relative powers of nature and nurture. *Journal of the Anthropological Institute of Great Britain and Ireland*, *5,* 391--406.

Gedda, L., Fiori-Ratti, L., & Bruno, G. (1960). La voix chez les jumeaux monozygotiques. *Folia Phoniatrica et Logopaedica*, *12*(2), 81--94.

Geschwind, N. (1983). Genetics: fate, chance, and environmental control. In C. Ludlow & J. Cooper, *Genetics aspects of speech and language disorders* (pp. 21--33). New York: Academic Press.

Gigerenzer, G. (2003). *Reckoning with risk*. London: Penguin.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684.

Gil, J. (2007). *Fonética para profesores de español*. Madrid: Arco/Libros.

Gil, J., & San Segundo, E. (2014). La cualidad de voz en fonética judicial. In E. Garayzábal, M. Jiménez & M. Reigosa, *Lingüística Forense: La Lingüística en el ámbito legal y policial.* (2nd ed., pp. 154--187). Madrid: Euphonia Ediciones.

Gil-Gil, J. (2009). *Identificación forense de locutor mediante el empleo de relaciones de* *verosimilitud sobre secuencias vocálicas como función discriminante y uso de la entropía cruzada empírica como medida de precisión de los resultados* (Master's thesis). Universidad Autónoma de Madrid.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in Applied Sociolinguistics*. Cambridge: Cambridge University Press.

Gold, E., & French, P. (2011). An international investigation of forensic speaker comparison practices. In *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China* (pp. 1254-1257).

Goldstein, U. (1976). Speaker-identifying features based on formant tracks. *The Journal of the Acoustical Society of America*, *59*(1), 176--182.

Goldstein, A., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society*, *17*(5), 217--220.

Gómez-Vilda, P. (2009). Lecture on *Análisis de la Señal Acústica y Procesado Digital de la Voz.* Personal Collection of P. Gómez-Vilda. Lecture, Máster en Fonética y Fonología, CSIC-UIMP.

Gómez-Vilda, P., Álvarez-Marquin, A., Mazaira-Fernández, L.M., Fernández-Baillo, R., Nieto-Lluis, V., & Martínez-Olalla, R., … & Rodellar-Biarge, M.V. (2008). Decoupling vocal tract from glottal source estimates in speaker's identification. *Language Design*, (Special Issue), 111--118.

Gómez-Vilda, P., Álvarez-Marquina, A., Mazaira-Fernández, L.M., Fernández-Baillo, R., Rodellar-Biarge, M.V., & Nieto-Lluis, V. (2010). *Glottal Biometric Features: Are Pathological Voice Studies appliable to Voice Biometry?* Paper presented at the I Workshop de Tecnologías Multibiométricas para la Identificación de Personas, Las Palmas de Gran Canaria.

Gómez-Vilda, P., Fernández-Baillo, R., Nieto, A., Díaz, F., Fernández-Camacho, F.J., Rodellar-Biarge, M.V., … & Martínez, R. (2007). Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters. *Journal of Voice*, *21*(4), 450--476.

Gómez-Vilda, P., Fernández-Baillo, R., Rodellar-Biarge, M.V., Nieto-Lluis, V., Álvarez-Marquina, A., & Mazaira-Fernández, L.M., … & Godino-Llorente, J.I. (2009). Glottal source biometrical signature for voice pathology detection. *Speech Communication*, *51*(9), 759--781.

Gómez-Vilda, P., Mazaira-Fernández, L.M., Martínez-Olalla, R., Álvarez-Marquina, A., Hierro, J., & Nieto, R. (2012). *Distance Metric in Forensic Voice Evidence Evaluation using Dysphonia-relevant Features*. Paper presented at the VI Jornadas de Reconocimiento Biométrico de Personas (JRBP), Las Palmas de Gran Canaria, Spain.

González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., & Ortega-García, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language*, *20*(2), 331--355.

González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D., & Ortega-García, J. (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(7), 2104--2115.

Greisbach, R., Esser, O., & Weinstock, C. (1995). Speaker identification by formant contours. In A. Braun & J. Köster, *Studies in Forensic Phonetics*. Trier: Wissenschaftlicher Verlag.

Grigorenko, E. (2009). Speaking genes or genes for speaking? Deciphering the genetics of speech and language. *Journal of Child Psychology and Psychiatry*, *50*(1-2), 116--125.

Harris, J. (1998). *The nurture assumption*. New York: Free Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer.

Haworth, C., Asbury, K., Dale, P., & Plomin, R. (2011). Added value measures in education show genetic as well as environmental influence. *Plos One*, *6*(2), 16006.

Hayiou-Thomas, M. (2008). Genetic and environmental influences on early speech, language and literacy development. *Journal of Communication Disorders*, *41*(5), 397.

Hazen, K. (2002). The family. In J. Chambers, P. Trudgill & N. Schilling-Estes, *The Handbook of Language Variation and Change* (pp. 500--525). Malden, MA: Blackwell.

Hecker, M., & Kreul, E. (1971). Descriptions of the speech of patients with cancer of the vocal folds. II. Judgments of age and voice quality. *The Journal of the Acoustical Society of America*, *49* (4), 1275--1282.

Hellín, E. (2010). *Peritaje 2.0: usos de la telefonía móvil*. Paper presented at Jornadas de Lingüística Forense, Universidad Autónoma de Madrid, Madrid, Spain.

Hirano, M. (1981). *Clinical Examination of the Voice*. New York: Springer-Verlag.

Hollien, H. (1990). *The acoustics of crime*. New York: Plenum Press.

Holmes, J. (1999). Preface. *Language in Society*, *28* (2), 171--173.

Holmes, J., & Meyerhoff, M. (1999). The community of practice: Theories and methodologies in language and gender research. *Language in Society*, *28*(2), 173--183.

Homayounpour, M., & Chollet, G. (1995). Discrimination of voices of twins and siblings for speaker verification. In *Proceedings of Eurospeech* (pp. 345--348).

Hualde, J. I. (1991). On Spanish syllabification. In H. Campos & F. Martínez Gil, *Current Studies in Spanish Linguistics* (pp. 475--493). Washington: Georgetown University Press.

Hualde, J. I. (2005). The sounds of Spanish. Cambridge: Cambridge University Press (pp. 19--21).

Ingram, J., Prandolini, R., & Ong, S. (1996). Formant trajectories as indices of phonetic variation for speaker identification. *International Journal of Speech Language and the Law*, *3*(1), 129--145.

Jain, A., Duin, R., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4--37.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *International Journal of Speech Language and the Law*, *4*(1), 84--103.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, *2*(4), 671--711.

Johnson, K., & Azara, M. (2000). The perception of personal identity in speech: Evidence from the perception of twins' speech. *Unpublished Manuscript*.

Kalmanovitch, Y. (2012). *Interpersonal long-term phonetic accommodation-patterns in close acquaintances*. Paper presented at the International Symposium on Imitation and Convergence in Speech, Aix-en-Provence, France.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2005). Factor analysis simplified. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *1*, 637--640.

Kim, K. (2009). *Automatic Speaker Identification of Korean Male Twins*. Paper presented at the 18th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Cambridge, UK.

Kinga, P. (2007). Hereditary phonetic parameters of the human voice. *Magyar Nyelvor (Hungarian Language Guardian)*, *131*(3), 306--315.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, *52*(1), 12--40.

Kinoshita, Y., & Osanai, T. (2006). Within speaker variation in diphthongal dynamics: what can we compare? In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 112--117).

Knörnschild, M., Von Helversen, O., & Mayer, F. (2007). Twin siblings sound alike: isolation call variation in the noctule bat, Nyctalus noctula. *Animal Behaviour*, *74*(4), 1055--1063.

Koeppen-Schomerus, G., Spinath, F., & Plomin, R. (2003). Twins and non-twin siblings: Different estimates of shared environmental influence in early childhood. *Twin Research*, *6*(02), 97--105.

Köster, O., & Köster, J. (2004). The auditory-perceptual evaluation of voice quality in forensic speaker recognition. *The Phonetician*, *89,* 9--37.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies*. Malden, MA: Wiley-Blackwell.

Künzel, H. J. (1987). *Sprechererkennung*. Heidelberg: Kriminalistik-Verlarg.

Künzel, H. J. (1994). Current Approaches to Forensic Speaker Recognition. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (pp. 135-141).

Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmissions on the measurement of formant frequencies, *Forensic Linguistics, 8 (1),* 80--99.

Künzel, H. J. (2010). Automatic speaker recognition of identical twins. *International Journal of Speech, Language and the Law*, *17*(2), 251--277.

Künzel, H. J. (2011). La prueba de voz en la investigación criminalística. *Ciencia Forense, INACIPE-Academia Iberoamericana de Criminalística y Estudios Forenses*, *1*(1), 37--50.

Künzel, H. J. (2013). Automatic speaker recognition with cross-language speech material. *International Journal of Speech, Language and the Law, 20*(1), 21--44.

Künzel, H. J., & Alexander, P. (2014). Forensic automatic speaker recognition with degraded and enhanced speech. *Journal of the Audio Engineering Society*, *62*(4), 244--253.

Künzel, H.J., & González-Rodríguez, J. (2003). Combining automatic and phonetic-acoustic speaker recognition techniques for forensic applications. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1619--1622).

Künzel, H. J., & Köster, J. (1992). Measuring Vocal Jitter in Forensic Speaker Recognition. In *Proceedings of the 44th Annual Meeting, American Academy of Forensic Sciences*. (pp. 113-114).

Labov, W. (1972). The transformation of experience in narrative syntax. In W. Labov, *Language in the Inner City* (pp. 354--396). Philadelphia: University of Philadelphia Press.

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, *30*(1), 9--26.

Lejeune, J. (1963). Autosomal disorders. *Pediatrics*, *32*(3), 326--337.

Leemann, A., Dellwo, V. & Kolly, M-J. (2012). *Exploring speech temporal features in twins: the case of %V*. Paper presented at Phonetik und Phonologie 8, Jena, 12-13 October 2012.

Lewontin, R., Rose, S., & Kamin, L. (1984). *Not in our genes*. New York: Pantheon Books.

Loakes, D. (2006a). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins* (Doctoral dissertation). University of Melbourne.

Loakes, D. (2006b). Variation in long-term fundamental frequency: measurements from vocalic segments in twins' speech. In *Proceedings of the 11th Australian International Conference on Speech Science & Technology,* (pp. 205--210).

Loakes, D., & McDougall, K. (2010). Individual variation in the frication of voiceless plosives in Australian English: a study of twins' speech. *Australian Journal of Linguistics, 30(2),* 155--181.

Locke, J., & Mather, P. (1989). Genetic factors in the ontogeny of spoken language: Evidence from monozygotic and dizygotic twins. *Journal of Child Language*, *16*(3), 553--559.

Loehlin, J., & Nichols, R. (1976). *Heredity, environment, & personality: A study of 850 sets of twins*. Austin: University of Texas Press.

Longo, D., & Fauci, A. (2011). Disorders of the thyroid gland. In D. Longo, A. Fauci, D. Kasper, S. Hauser, J. Jameson & J. Loscalzo, *Harrison's Principles of Internal Medicine* (18th ed., pp. 2224-2247). New York: McGraw-Hill.

López-Escobedo, F. (2010). *El análisis de las características dinámicas de la señal de habla como posible marca para la comparación e identificación forense de voz: Un estudio para el español de México* (Doctoral dissertation). Universitat Pompeu Fabra.

Luchsinger, R., & Arnold, G. (1965). *Voice, speech, language*. Belmont, Calif.: Wadsworth.

Lundström, A. (1948). *Tooth size and occlusion in twins*. Basel: Karger.

Ma, K. (2011, August). Top 10 Weirdest Twin-Crime Stories. *Time*. Retrieved from http://content.time.com/time/specials/packages/article/0,28804,2090549_2090540_2090539,00.html

Markel, J., & Gray, A. (1976). *Linear prediction of speech*. Berlin: Springer-Verlag.

Matheny, A., & Bruggemann, C. (1973). Children's speech: heredity components and sex differences. *Folia Phoniatrica*, *25* (6), 442--449.

Mazaira, L. (2014). *Nueva metodología para la integración de rasgos biométricos en sistemas de identificación de locutor en entornos de seguridad* (Doctoral dissertation). Universidad Politécnica de Madrid.

McDougall, K. (2004). Speaker-specific formant dynamics: an experiment on Australian English /ai/. *International Journal of Speech Language and the Law*, *11*(1), 103--130.

McDougall, K. (2005). *The role of formant dynamics in determining speaker identity* (Doctoral Dissertation). University of Cambridge.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies. *International Journal of Speech Language and the Law*, *13*(1), 89--126.

McDougall, K., & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. In *Proceedings of the 16th International Congress of Phonetic Sciences*, (pp. 1825--1828).

Merriman, C. (1924). The intellectual resemblance of twins. *Psychological Monographs: General and Applied*, *33*(5), 1--57.

Meuwly, D. (2006). Forensic individualisation from biometric data. *Science & Justice*, *46*(4), 205--213.

Miller, P. (2012, January). A thing or two about twins. *National Geographic*. Retrieved from http://ngm.nationalgeographic.com/2012/01/twins/miller-text

Ministerio del Interior, Gobierno de España. (2011). *Anuario Estadístico del Ministerio del Interior*. Retrieved from: http://www.interior.gob.es/documents/642317/1204756/Anuario+estad%C3%ADstico+de+2011.pdf/1d35a1c8-f2e1-4417-bc5a-ca4e17bb7e66.

Moosmüller, S. (2007). The influence of creaky voice on formant frequency changes. *International Journal of Speech Language and the Law*, *8*(1), 100--112.

Mora, M. (2013, February 10). La policía francesa detiene a dos gemelos para aclarar una ola de ataques sexuales. *El País*. Retrieved from http://internacional.elpais.com/internacional/2013/02/10/actualidad/1360530132_840599.html

Moreno, A., Poch, D., Bonaforte, A., Lleida, E., Llisterri, J., Mariño, J., & Nadeu, C. (1993). Albayzín speech database: design of the phonetic corpus. In *Proceedings of Eurospeech* (pp. 175-178).

Moreno, F. (2011). La entrevista sociolingüística: esquemas de perspectivas. *Linred: lingüística en la Red, 9*: 1--16.

Morrison, G. S. (2007). *Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation*. Available from http://geoff-morrison.net/#MVKD.

Morrison, G. S. (2008). Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /ai/. *International Journal of Speech, Language and the Law, 15*(2), 249--266.

Morrison, G. S. (2009a). Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework. *Australian Journal of Forensic Sciences*, *41*(2), 155--161.

Morrison, G. S. (2009b). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*(4), 298--308.

Morrison, G. S. (2009c). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, *125*(4), 2387--2397.

Morrison, G. S. (2010a). Forensic Voice Comparison. In I. Freckelton & H. Selby, *Expert Evidence*. Sydney: Thomson Reuters.

Morrison, G. S. (2010b). *Sound File Cutter Upper*. [Computer software] Retrieved from:

http://geoff-morrison.net/#CutUp.

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic--phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model--universal background model (GMM--UBM). *Speech Communication*, *53*(2), 242--256.

Morrison, G. S. (2012). *SoundLabeller: Ergonomically designed software for marking and labelling sections of sound files.* [Computer software] Retrieved from: http://geoff-morrison.net/#SndLbl.

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, *45*(2), 173--197.

Morrison, G.S., & Kinoshita, Y. (2008). Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. In *Proceedings of Interspeech* (pp. 1501--1504).

Morrison, G.S., & Nearey, T. (2011). *FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories*. [Computer software] Retrieved from: http://geoff-morrison.net/#FrmMes.

Morrison, G.S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, *44*(2), 155--167.

Mowrer, E. (1954). Some factors in the affectional adjustment of twins. *American Sociological Review, 19* (4), 468--471.

Müller, C. (2007). *Speaker classification*. Berlin: Springer.

Murphy, K. (2008). *Digital signal processing techniques for application in the analysis of pathological voice and normophonic singing voice* (Doctoral dissertation). Universidad Politécnica de Madrid.

National Organization of Mothers of Twins Clubs, Inc. (n.d.). *National Organization of Mothers of Twins Clubs, Inc*. Retrieved February 26, 2013, from http://www.nomotc.org/index.php?option=com_content&task=view&id=67&Itemid=55

Navarro Tomás, T. (1946). *Estudios de fonología española*. Syracuse, N.Y.: Syracuse University Press.

Navarro Tomás, T. (1972). *Manual de pronunciación española* (17th ed.). Madrid: Consejo Superior de Investigaciones Científicas.

Nearey, T., Assmann, P., & Hillenbrand, J. (2002). Evaluation of a strategy for automatic formant tracking. *The Journal of the Acoustical Society of America*, *112*(5), 2323--2323.

Newman, H., Freeman, F., & Holzinger, K. (1937). *Twins, a study of heredity and environment*. Chicago: University of Chicago Press.

Nichols, R., & Bilbro Jr, W. (1966). The diagnosis of twin zygosity. *Human Heredity*, *16*(3), 265--275.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Nolan, F. (1997). Speaker recognition and forensic phonetics. In W. Hardcastle & J. Laver, *The handbook of phonetic sciences* (pp. 744--767). Oxford: Blackwell.

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations and roles. In *Proceedings of the conference "Law and language: Prospect and retrospect"* (pp. 1--19.). Levi, Finnish Lapland.

Nolan, F. (2003). A recent voice parade. *International Journal of Speech, Language and the Law*, *10*(2), 277—291.

Nolan, F., & Oh, T. (1996). Identical twins, different voices. *International Journal of Speech Language and the Law, 3*(1), 39--49.

Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language and the Law*, *16*(1), 31--57.

Núñez, F., & Suárez, C. (1998). *Manual de evaluación y diagnóstico de la voz*. Oviedo: Servicio de publicaciones de la Universidad de Oviedo.

Ochs, E., & Schieffelin, B. (1983). *Acquiring conversational competence*. London: Routledge & Kegan Paul.

Orlikoff, R., & Baken, R. (1989). Fundamental frequency modulation of the human voice by the heartbeat: preliminary results and possible mechanisms. *The Journal of the Acoustical Society of America*, *85*(2), 888--893.

Orlikoff, R., & Kahane, J. (1991). Influence of mean sound pressure level on jitter and shimmer measures. *Journal of Voice*, *5*(2), 113--119.

Ortega-García, J., González-Rodríguez, J., & Marrero-Aguiar, V. (2000). AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. *Speech Communication*, *31*(2), 255--264.

Pakstis, A., Scarr-Salapatek, S., Elston, R., & Siervogel, R. (1972). Genetics contributions to morphological and behavioral similarities among sibs and dizygotic twins: Linkages and allelic differences. *Social Biology*, *19*, 185--192.

Paluszny, M., Selzer, M., Vinokur, A., & Lewandowski, L. (1977). Twin relationships and depression. *The American Journal of Psychiatry*, *134*, 988--990.

Pardo, J. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382--2393.

Pardo, J., Gibbons, R., Suppes, A., & Krauss, R. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190--197.

Parker, N. (1964). Twins: A psychiatric study of a neurotic group. *The Medical Journal of Australia*, *2*, 735--742.

Peeters, H., Van Gestel, S., Vlietinck, R., Derom, C., & Derom, R. (1998). Validation of a telephone zygosity questionnaire in twins of known zygosity. *Behavior Genetics*, *28*(3), 159--163.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169--190.

Pigeon, S., Druyts, P., & Verlinde, P. (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing*, *10*(1), 237--248.

Pinker, S. (1997). *How the mind works*. New York: Norton.

Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from one another? *Behavioral and Brain Sciences*, *10*(01), 1--16.

Plomin, R., DeFries, J., McClearn, G., & McGuffin, P. (2008). *Behavioral Genetics* (5th ed.). New York: Worth.

Plomin, R., & Kosslyn, S. (2001). Genes, brain and cognition. *Nature Neuroscience*, *4*, 1153--1154.

Posthuma, D., & Boomsma, D. (2000). A note on the statistical power in extended twin designs. *Behavior Genetics*, *30*(2), 147--158.

PRESEEA. (2003). *Metodología del "Proyecto para el Estudio Sociolingüístico del Español de España y de América"*. Retrieved from: http://www.linguas.net/LinkClick.aspx?fileticket=%2FthWeHX0AyY%3D&tabid=474&mid=928.

Przybocki, M., Martin, A., & Le, A. (2007). NIST speaker recognition evaluations utilizing the Mixer corpora—2004, 2005, 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 1951--1959.

Przybyla, B., Horii, Y., & Crawford, M. (1992). Vocal fundamental frequency in a twin sample: looking for a genetic effect. *Journal of Voice*, *6*(3), 261--266.

Quilis, A. (1981). *Fonética acústica de la lengua española*. Madrid: Gredos.

Ramos-Castro, D. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems* (Doctoral dissertation). Universidad Autónoma de Madrid.

Ramos-Castro, D., & Gonzalez-Rodríguez, J. (2013). Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, *230*(1), 156--169.

Real Academia Española y Asociación de Academias de la Lengua Española (RAE). (2011). *Nueva gramática de la lengua española. Fonética y Fonología.* (pp. 332--354). Madrid: Espasa.

Rende, R., Plomin, R., & Vandenberg, S. (1990). Who discovered the twin method? *Behavior Genetics*, *20*(2), 277--285.

Reynolds, D., & Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*(1), 72--83.

Reynolds, D., Quatieri, T., & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*(1), 19--41.

Ritchings, R., McGillion, M., & Moore, C. (2002). Pathological voice quality assessment using artificial neural networks. *Medical Engineering & Physics, 24*(7), 561--564.

Roberts, H. (2004). Statistical evaluation in forensic DNA typing. In I. Freckelton & H. Selby, *Expert evidence*. Sydney: Thomson Reuters.

Romaine, S. (1984). *The acquisition of communicative competence*. New York: Blackwell.

Rose, P. (2002). *Forensic speaker identification*. London: Taylor & Francis.

Rose, P. (2006a). Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language, 20*(2), 159--191.

Rose, P. (2006b). The intrinsic forensic discriminatory power of diphthongs. In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 64--69).

Rose, P. (n.d.). Catching criminals by their voice - Combining automatic and traditional methods for optimum performance in forensic speaker identification. (Project report).

Rose, P., Kinoshita, Y., & Alderman, T. (2006). Realistic extrinsic forensic speaker discrimination with the diphthong /ai/. In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 329--334).

Rose, P., & Morrison, G. S. (2009). A response to the UK position statement on forensic speaker comparison. *The International Journal of Speech, Language and the Law*, *16*(1), 139--163.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association, 95*(451), 760--768.

Ryalls, J., Shaw, H., & Simon, M. (2004). Voice onset time production in older and younger female monozygotic twins. *Folia Phoniatrica et Logopaedica*, *56*(3), 165--169.

Saks, M., & Koehler, J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892--895.

Sambur, M. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech and Signal Processing, 23*(2), 176--182.

San Segundo, E. (2010a). Parametric representations of the formant trajectories of Spanish vocalic sequences for likelihood-ratio-based forensic voice comparison. *The Journal of the Acoustical Society of America*, *128*(4), 2394.

San Segundo, E. (2010b). Review of the book *Vocales en grupo*, by L. Aguilar. [This is a description of form and content, not a title] *Español Actual, 93*, 196--204.

San Segundo, E. (2010c). Variación inter e intralocutor: parámetros acústicos segmentales que caracterizan fonéticamente a tres hermanos. *Interlingüística*, *21*, 352--363.

San Segundo, E. (2012). *Glottal source parameters for forensic voice comparison: an approach to voice quality in twins' voices*. Paper presented at the 21st Annual Conference of the

International Association for Forensic Phonetics and Acoustics (IAFPA), Santander, Spain.

San Segundo, E. (2013a). A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application. *Procedia-Social and Behavioral Sciences*, *95*, 59--67.

San Segundo, E. (2013b). *Guess who is laughing: A perceptual experiment on twin and non-twin siblings' identification*. Paper presented at the 31st International Conference AESLA (Asociación Española de Lingüística Aplicada), Universidad de La Laguna, San Cristóbal de La Laguna.

San Segundo, E. (2014). El entrenamiento musical y otros factores que pueden influir en el reconocimiento perceptivo de hablantes. In *Fonética Experimental, Educación Superior e Investigación* (pp.571--588). Madrid: Arco/Libros.

San Segundo, E., Alves, H. & Fernández, M. (2013). CIVIL Corpus: Voice Quality for Speaker Forensic Comparison. *Procedia-Social and Behavioral Sciences, 95*, 587--593.

San Segundo, E., & Gómez-Vilda, P. (2013). Voice biometrical match of twin and non-twin siblings. In *Proceedings of the 8th International Workshop Models and analysis of vocal emissions for biomedical applications, Firenze, Italy* (pp. 253--256).

Scarr, S., & Carter-Saltzman, L. (1979). Twin method: Defense of a critical assumption. *Behavior Genetics*, *9*(6), 527--542.

Scheffer, N., Bonastre, J., Ghio, A., & Teston, B. (2004). Gémellité et reconnaissance automatique du locuteur. *Actes des Journées d'Étude sur la Parole (JEP)*, 445--448.

Seeman, M. (1937). Die Bedeutung der Zwillingspathologie für die Erforschung von Sprachleiden. *Arch Sprach-Stimmheilk.*, *1*, 88--98.

Segal, N. (1984). Cooperation, competition, and altruism within twin sets: A reappraisal. *Ethology and Sociobiology*, *5*(3), 163--177.

Segal, N. (1990). The importance of twin studies for individual differences research. *Journal of Counseling & Development*, *68*(6), 612--622.

Segal, N. (1993). Implications of twin research for legal issues involving young twins. *Law and Human Behavior*, *17*(1), 43.

Shields, J. (1962). *Monozygotic twins brought up apart and brought up together*. London: Oxford Univ. Press.

Siemens, H. (1924). *Die Zwillingspathologie*. Berlin: Springer.

Sluijter, A. (1995). *Phonetic correlates of stress and accent*. The Hague: Holland Academic Graphics.

Sluijter, A., Shattuck-Hufnagel, S., Stevens, K., & Heuven, V. (1995). Supralaryngeal resonance and glottal pulse shape as correlates of stress and accent in English. In *Proceedings of the International Congress of Phonetic Sciences* (pp. 630-633).

Smith, J., Renshaw, D., & Renshaw, R. (1968). Twins who want to be identified as twins. *Diseases of the Nervous System*, *29*(9), 615--618.

Spielman, A., Brand, J., Buischi, Y., & Bretz, W. (2011). Resemblance of tongue anatomy in twins. *Twin Research and Human Genetics, 14*(03), 277--282.

Spuhler, J. (1977). Biology, speech, and language. *Annual Review of Anthropology*, *6*(1), 509--561.

Srihari, S., Huang, C., & Srinivasan, H. (2008). On the discriminability of the handwriting of twins. *Journal of Forensic Sciences*, *53*(2), 430--446.

Stevens, K. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In P. Denes & E. David Jr., *Human Communication: A Unified View* (pp. 51--66). New York: McGraw Hill.

Story, B., & Titze, I. (1995). Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America*, *97*(2), 1249--1260.

Stromswold, K. (2006). Why aren’t identical twins linguistically identical? Genetic, prenatal and postnatal factors. *Cognition*, *101*(2), 333--384.

Sutcliffe, A., & Derom, C. (2006). Follow-up of twins: health, behaviour, speech, language outcomes and implications for parents. *Early Human Development*, *82*(6), 379--386.

Thaitechawat, S., & Foulkes, P. (2011). Discrimination of speakers using tone and formant dynamics in Thai. In *Proceedings of the 17th International Congress of Phonetic Sciences,* (pp. 1978-1981).

Thompson, P., Cannon, T., Narr, K., Van Erp, T., Poutanen, V., Huttunen, M.,… & Toga, A. (2001). Genetic influences on brain structure. *Nature Neuroscience*, *4*(12), 1253--1258.

Tippett, C., Emerson, V., Fereday, M., Lawton, F., Richardson, A., Jones, L., & Lampert, M. (1968). The evidential value of the comparison of paint flakes from sources other than vehicles. *Journal of the Forensic Science Society, 8*(2), 61--65.

Tomblin, J., & Buckwalter, P. (1998). Heritability of poor language achievement among twins. *Journal of Speech, Language, and Hearing Research*, *41*(1), 188.

Trouvain, J., & Truong, K. (2012). *Convergence of laughter in conversational speech: effects of quantity, temporal alignment and imitation*. Paper presented at the International Symposium on Imitation and Convergence in Speech, Aix-en-Provence, France.

van Dommelen, W. (2001). Identification of twins by spoken syllables. *Perceptual and Motor Skills*, *92*(1), 8--10.

van Leeuwen, D., & Brümmer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In C. Müller, *Speaker Classification I: Fundamentals, Features, and Methods.* (pp. 330-353). Heidelberg: Springer-Verlag.

van Lierde, K., Vinck, B., De Ley, S., Clement, G., & Van Cauwenberge, P. (2005). Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach. *Journal of Voice*, *19*(4), 511--518.

Vandenberg, S. (1966). Contributions of twin research to psychology. In M. Manosevitz, G. Lindzey & D. Thiessen, *Behavioral genetics: Method and research* (pp. 145--164). New York: Appleton-Century-Crofts.

Vandenberg, S., & Wilson, K. (1979). Failure of the twin situation to influence twin differences in cognition. *Behavior Genetics*, *9*(1), 55--60.

von Bracken, H. (1934). Mutual intimacy in twins. *Character and Personality*, *2*(4), 293--309.

Wagner, I. (1995). A new jitter-algorithm to quantify hoarseness: an exploratory study. *International Journal of Speech Language and the Law*, *2*(1), 18--27.

Weirich, M. (2010). Articulatory and acoustic inter-speaker variability in the production of German vowels. *ZAS Papers in Linguistics*, *52*, 19--42.

Weirich, M. (2011). *The influence of NATURE and NURTURE on speaker-specific parameters in twins' speech: Articulation, acoustics and perception.* (Doctoral dissertation). Humboldt-Universität zu Berlin.

Weirich, M., & Lancia, L. (2011). Perceived auditory similarity and its acoustic correlates in twins and unrelated speakers. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2118--2121).

Whiteside, S., & Rixon, E. (2000). Identification of twins from pure (single speaker) and hybrid (fused) syllables: An acoustic and perceptual case study. *Perceptual and Motor Skills*, *91*(3), 933--947.

Whiteside, S., & Rixon, E. (2001). Speech patterns of monozygotic twins: an acoustic case study of monosyllabic words. *The Phonetician*, *82*(2), 9-22.

Whiteside, S. P., & Rixon, E. (2004). Speech characteristics of monozygotic twins and a same-sex sibling: an acoustic case study of coarticulation patterns in read speech. *Phonetica, 60(4)*, 273-297.

Wolf, J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, *51*(6B), 2044--2056.

Yarmey, A., Yarmey, A., Yarmey, M., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology*, *15*(3), 283--299.

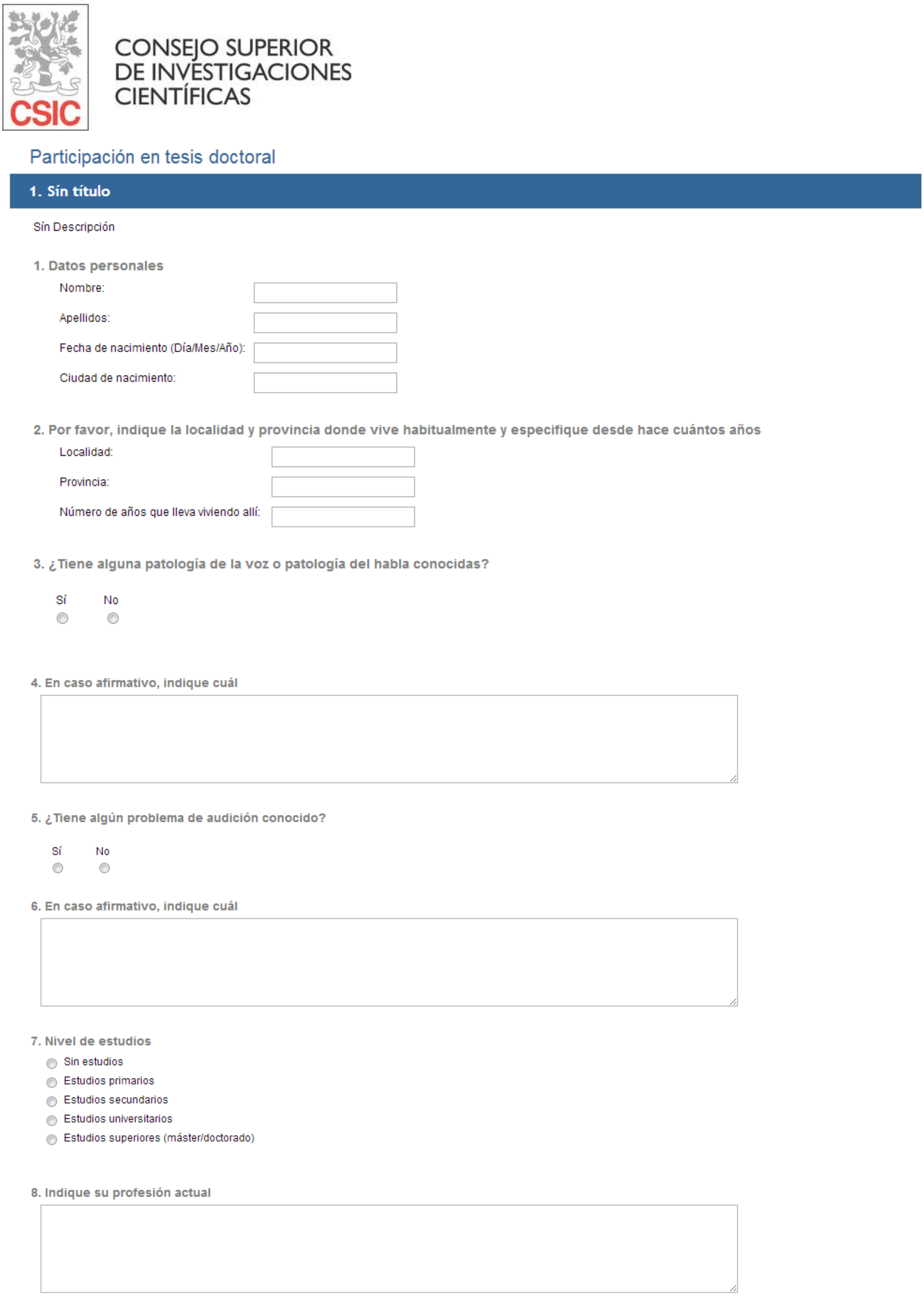
Zazzo, R. (1978). Genesis and peculiarities of the personality of twins. In W. Nance, G. Allen & P. Parisi, *Twin research (Part A): Psychology and method* (pp. 1--11). New York: Alan R. Liss.

Zhang, C., Morrison, G. S., & Thiruvaran, T. (2011). Forensic voice comparison using Chinese/iau/. *Proceedings 17th International Congress of Phonetic Sciences*, 2280--2283.

Zheng, N. (2005). *Speaker Recognition Using Complementary Information from Vocal Source and Vocal Tract* (Doctoral dissertation). The Chinese University of Hong Kong.

Appendix A: Questionnaires

A1: Online questionnaire



CONSEJO SUPERIOR
DE INVESTIGACIONES
CIENTÍFICAS

CSIC

Participación en tesis doctoral

1. Sín título

Sín Descripción

1. Datos personales

Nombre:

Apellidos:

Fecha de nacimiento (Día/Mes/Año):

Ciudad de nacimiento:

2. Por favor, indique la localidad y provincia donde vive habitualmente y especifique desde hace cuántos años

Localidad:

Provincia:

Número de años que lleva viviendo allí:

3. ¿Tiene alguna patología de la voz o patología del habla conocidas?

Sí ○ No ○

4. En caso afirmativo, indique cuál

5. ¿Tiene algún problema de audición conocido?

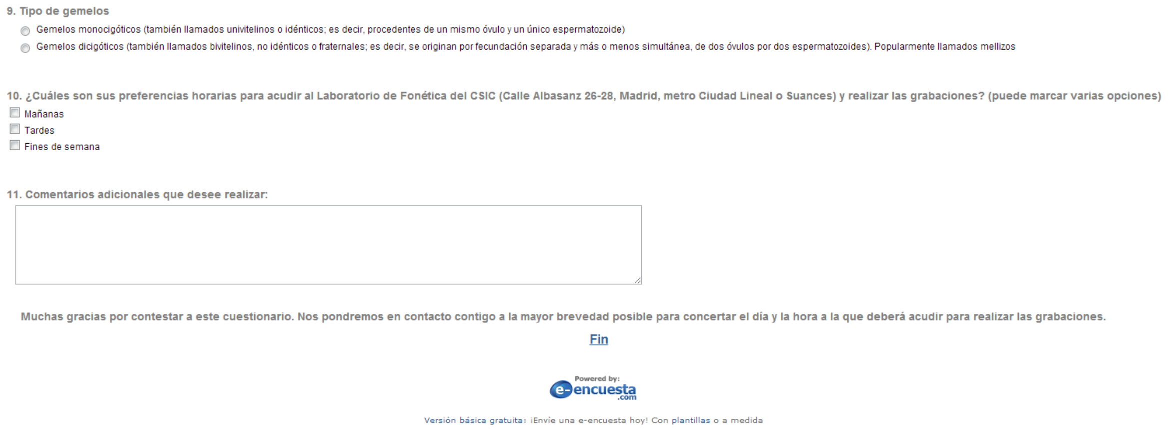
Sí ○ No ○

6. En caso afirmativo, indique cuál

7. Nivel de estudios

○ Sin estudios
○ Estudios primarios
○ Estudios secundarios
○ Estudios universitarios
○ Estudios superiores (máster/doctorado)

8. Indique su profesión actual



9. Tipo de gemelos

○ Gemelos monocigóticos (también llamados univitelinos o idénticos; es decir, procedentes de un mismo óvulo y un único espermatozoide)

○ Gemelos dicigóticos (también llamados bivitelinos, no idénticos o fraternales; es decir, se originan por fecundación separada y más o menos simultánea, de dos óvulos por dos espermatozoides). Popularmente llamados mellizos

10. ¿Cuáles son sus preferencias horarias para acudir al Laboratorio de Fonética del CSIC (Calle Albasanz 26-28, Madrid, metro Ciudad Lineal o Suances) y realizar las grabaciones? (puede marcar varias opciones)

□ Mañanas

□ Tardes

□ Fines de semana

11. Comentarios adicionales que desee realizar:

Muchas gracias por contestar a este cuestionario. Nos pondremos en contacto contigo a la mayor brevedad posible para concertar el día y la hora a la que deberá acudir para realizar las grabaciones.

A2: Participant questionnaire for the first recording session

Note. We include a sample of the type of questionnaire filled by twins and non-twin siblings. This comprises a part A (questions about the speaker) and a part B (questions about the speaker and his twin). The questionnaire filled by unrelated speakers is made of part A only.

CUESTIONARIO

Recogida de datos para tesis doctoral

PARTE A – Preguntas sobre usted

1. Datos personales

Por favor, rellene los espacios en blanco con la información que se le solicita.

Nombre:..

Apellidos:

....

NIF:

....

Fecha de nacimiento (día, mes y año):

....

Dirección de correo electrónico:

....

Número de teléfono:

....

2. Perfil lingüístico

Por favor, indique la localidad y provincia donde nació, donde vive habitualmente y donde ha residido con anterioridad:

- Incluya únicamente aquellas ciudades en la que haya residido durante **tres meses** seguidos o más

	Ciudad, provincia, país	**¿Durante cuánto tiempo vivió allí?**	**¿Qué edad tenía cuando vivió allí?**
Lugar de nacimiento			
Lugar de residencia actual			
Otros lugares de residencia			
Otros lugares de residencia			
Otros lugares de residencia			
Otros lugares de residencia			

Por favor, marque con una cruz donde corresponda, o bien rellene los espacios en blanco:

Lengua materna:..

¿Habla otros idiomas, además del español?

□ **Sí**

□ **No**

En caso afirmativo, indique en la siguiente tabla los idiomas que habla o que ha estudiado, y el nivel aproximado de conocimiento que estima que tiene en cada idioma.

→ Marque también, a partir de la tercera columna, la opción u opciones que más se adecuen a la forma en que ha aprendido cada idioma (puede marcar **varias** opciones):

Idioma	**Nivel estimado (básico, intermedio o avanzado)**	**¿Lo estudió en el colegio/instituto?**	**¿Lo estudió en una escuela de idiomas / academia /profesor particular?**	**¿Lo aprendió en una estancia en el extranjero?**	**Otras formas de aprendizaje (escriba cuál)**
……………….	…………………….	□	□	□	………………………
……………….	…………………….	□	□	□	………………………
……………….	…………………….	□	□	□	………………………
……………….	…………………….	□	□	□	………………………
……………….	…………………….	□	□	□	………………………
……………….	…………………….	□	□	□	………………………

Por favor, rellene los espacios en blanco con la siguiente información sobre las personas de su entorno familiar:

Persona cercana	**Lugar de nacimiento**	**Lugar donde ha residido más tiempo**	**Lengua en la que se comunica con ellos**
Padre	…………………….	…………………….	…………………….
Madre	…………………….	…………………….	…………………….
Pareja	…………………….	…………………….	…………………….

Por favor, rellene los espacios en blanco con la siguiente información sobre su perfil lingüístico con relación a sus amigos y ámbito laboral:

Mayoritariamente, ¿qué lengua utiliza para comunicarse con sus amigos más cercanos?

...

En el trabajo, ¿qué lengua utiliza mayoritariamente para comunicarse con sus compañeros de trabajo, jefes, colaboradores, etc.?

……………………………………………………………...

3. Salud

a. ¿Tiene algún problema de voz o de habla conocidos?

- □ **Sí**
- □ **No**

En caso afirmativo, indique cuál: ……………………………………………

b. ¿Tiene algún problema de audición conocido?

- □ **Sí**
- □ **No**

En caso afirmativo, indique cuál: ……………………………………………

c. ¿Padece asma?

- ○ **No.**
- ○ **Sí, hoy tengo síntomas de asma.**
- ○ **Sí, pero únicamente cuando hago deporte o algún ejercicio físico extenuante.**
- ○ **Otras respuestas:..**

d-1. ¿Fuma?

- □ **Sí**
- □ **No**

En caso afirmativo, indique cuánto tiempo hace que fuma: ……………………………………

d-2. ¿Ha fumado alguna vez de forma habitual? (Responda solo en caso de que usted haya fumado anteriormente, PERO NO FUME AHORA)

□ **Sí**

□ **No**

En caso afirmativo, indique cuánto tiempo hace (años o meses) que dejó de fumar:

……

En caso afirmativo, indique durante cuántos años estuvo fumando:

……

d-3. En caso de que haya respondido afirmativamente a la pregunta d-1. o d-2, indique con qué frecuencia fuma (si fuma actualmente) o fumaba (si ha fumado en el pasado PERO NO FUMA AHORA):

○ **Muy de vez en cuando (bodas, celebraciones, etc.)**

○ **De vez en cuando (cada fin de semana, cuando salgo con los amigos)**

○ **A diario, menos de 6 cigarrillos**

○ **A diario, más de 6 cigarrillos**

○ **Más de una cajetilla al día**

○ **Otras respuestas:...**

e. Marque con una cruz los casos que le correspondan.
→ En la columna de la derecha, escriba una descripción o explicación si es necesario, o bien marque con una cruz una de las opciones.

○	**Tengo desviado el tabique nasal**	
○	**Me han extraído alguna muela del juicio**	**→ Indique cuántas: 1** ○ **2** ○ **3** ○ **4** ○
○	**Me falta alguna pieza dental**	**→ Indique cuántas: ....**
○	**Llevo aparato corrector de dientes**	**→ Tipo (ortodoncia fija / ortodoncia removible / otros):** **...** **→¿Desde hace cuántos años?: ....**
○	**He llevado aparato corrector de dientes alguna vez**	**→ Tipo (ortodoncia fija / ortodoncia removible / otros):** **...** **→ ¿Durante cuántos años?: ....**
○	**Tengo reflujo gástrico**	
○	**Tengo algún problema hormonal**	**→ Especifique cuál: ....**
○	**Tengo las adenoides inflamadas o hipertróficas (vegetaciones)**	
○	**Me han operado de vegetaciones**	

○	**Me han extraído las amígdalas**	
	Me han realizado una intervención quirúrgica en alguna de las siguientes partes: ○ **nariz** ○ **garganta** ○ **mandíbula** ○ **dientes** ○ **cuerdas vocales**	→ **Especifique qué tipo de operación:** ....

f. <u>En este momento</u>, ¿padece alguna molestia o infección de: garganta y/o nariz, o bien sufre algún dolor en los dientes, la boca, los oídos o la mandíbula? Conteste en la siguiente tabla:

Dolor/infección/molestia actual de:	**Sí**	**No**	**Describa brevemente la molestia o el dolor**
Garganta			
Nariz			
Dientes			
Oídos			
Mandíbula			

g. Marque con una cruz donde corresponda.

Hoy (en este momento), ¿presenta usted síntomas de….?

○ **Resfriado o catarro**

○ **Rinitis alérgica**

○ **Otro tipo de obstrucción nasal (sinusitis, etc.):**

○ **Asma**

○ **Dolor de garganta**

○ **Reflujo gástrico**

4. Otros datos de interés

a. Nivel de estudios:

○ **Sin estudios**

○ **Estudios primarios**

○ **Estudios secundarios**

○ **Estudios universitarios *Indique cuáles:***
……………………………………………………………………………………..

○ **Estudios superiores (máster/doctorado) *Indique cuáles:***
……………………………………………………………...

b. Profesión / Ocupación

- **Indique su profesión actual:**
……………………………………………………………………………………………...

- **¿Cuántos años hace que la ejerce?**
……………………………………………………………………………….

- **Debido a su profesión, ¿pasa usted mucho tiempo hablando?**

○ **Sí, bastante**

○ **Ni mucho ni poco**

○ **Más bien poco**

- **¿Nota molestias debidas a un abuso vocal (ya sea por pasar excesivo tiempo hablando o por hacerlo en condiciones perjudiciales para su salud, como gritar, o hablar en lugares ruidosos, etc.)?**

○ **Con mucha frecuencia**

○ **Con frecuencia / A veces**

○ **Muy de vez en cuando, cuando hago algún esfuerzo puntual**

○ **Nunca**

c. Actividades de ocio y tiempo libre

- **Por favor, indique las actividades que realiza más frecuentemente durante su tiempo de ocio:**

………
………

- **Independientemente de su actividad profesional, ¿pasa usted mucho rato hablando, cantando o realizando cualquier otra actividad que implique el uso de la voz, tanto en su tiempo libre (conversaciones con amigos, por ejemplo), como debido a ciertas actividades de ocio que realice (teatro, cantar en un coro, etc.)?**

○ **Sí, bastante**

○ **Ni mucho ni poco**

○ **Más bien poco**

- **¿Nota molestias debidas a un abuso vocal (ya sea por pasar excesivo tiempo hablando o por hacerlo en condiciones perjudiciales para su salud, como gritar, o hablar en lugares ruidosos, etc.)?**

○ Con mucha frecuencia

○ Con frecuencia / A veces

○ Muy de vez en cuando, cuando hago algún esfuerzo puntual

○ Nunca

Ha llegado al final de la primera parte del cuestionario.
Gracias por su colaboración.

PARTE B – Preguntas sobre usted y sobre su hermano

1. Tipo de gemelos

Por favor, marque con una cruz donde corresponda.

□ **Gemelos monocigóticos (también llamados univitelinos o idénticos; es decir, procedentes de un mismo óvulo y un único espermatozoide)**

□ **Gemelos dicigóticos (también llamados bivitelinos, no idénticos o fraternales; es decir, se originan por fecundación separada y más o menos simultánea de dos óvulos por dos espermatozoides). Popularmente llamados mellizos.**

2. Relaciones familiares y personales

Por favor, marque con una cruz donde corresponda, o bien, rellene los espacios en blanco.

- **¿Existen más hermanos en su familia, aparte de su gemelo y de usted?**

  □ **Sí**

  □ **No**

En caso afirmativo, indique cuántos y el sexo y la edad de cada uno:
(Ejemplo) Hermano 1: Mujer, 35 años; Hermano 2: Hombre, 21 años

……

……

- **¿Tiene actividades de ocio compartidas con su gemelo? Es decir, ¿realizan alguna actividad juntos en su tiempo libre (practicar el mismo deporte, ir a una escuela de idiomas, etc.)**

  □ **Sí**

  □ **No**

En caso afirmativo, indique cuáles:

……

- **¿Tiene amigos en común con su gemelo?**

  □ **Sí**

  □ **No**

En caso afirmativo, indique con cuánta frecuencia suelen quedar todos juntos:

○ Con mucha frecuencia, casi a diario

○ A veces

○ Muy de vez en cuando

○ Casi nunca

- ¿Tiene un círculo de amigos propios, fuera de los que comparte con su gemelo?

□ Sí, algunos

□ Sí, bastantes

□ Sí, pero la mayoría de mis amigos son también amigos de mi gemelo

□ No

3. Usted y su gemelo: *comunicación*.

- ¿Con cuánta frecuencia ve a su hermano gemelo?

○ A diario, ya que vivimos juntos

○ A diario, aunque no vivimos juntos

○ De dos a tres veces a la semana

○ Al menos una vez a la semana

○ Muy de vez en cuando, cada 15 días o cada mes

○ Casi nunca, solo en ocasiones especiales (Navidad, celebraciones familiares, etc.)

○ Otras respuestas: ……………………………………………………………………...……….

- ¿Con cuánta frecuencia habla (por teléfono, por Skype, etc.) con su gemelo?

○ A diario, varias veces al día

○ Al menos una vez al día

○ De dos a tres veces a la semana

○ Al menos una vez a la semana

○ Muy de vez en cuando, cada 15 días o cada mes

○ Casi nunca

○ Otras respuestas: ……………………………………………………………………..

4. Usted y su gemelo: *convivencia*.

- ¿Vive con su hermano?

  □ Sí

  □ No

- En caso de respuesta negativa a la pregunta anterior, ¿cuánto tiempo (especifique el número de años o de meses) llevan viviendo separados (es decir, desde hace cuánto tiempo no viven en el mismo hogar familiar)?

……

- Aparte de la convivencia en el hogar familiar, indique si volvieron a vivir juntos en algún momento, por algún motivo (ejemplo: realizaron el servicio militar juntos, convivieron en la misma residencia universitaria durante sus estudios, etc.) e indique durante cuánto tiempo.

→ Motivo: ………

→ Tiempo de convivencia: ……………………………………………………………………………………

- ¿Usted y su hermano acudieron juntos al colegio de educación primaria?

  □ Sí

  □ No

- En caso afirmativo, ¿estaban en la misma clase?

  □ Sí

  □ No

- En caso afirmativo, indique hasta qué edad acudieron juntos al colegio: ………………………… y hasta qué edad estuvieron en la misma clase: ……………………………………………………………

5. Usted y su gemelo: *preferencias y rasgos personales*.

- En general, ¿le gusta tener un hermano gemelo?

  □ Sí

  □ No

  □ Indiferente

- **Indique brevemente porqué le gusta o porqué no le gusta:**

……

- **¿Considera que está muy unido a su hermano? Indique en una escala del 1 al 5 en qué grado considera que está unido a su hermano gemelo (1 significa muy poco y 5 significa muchísimo):**

1○ 2○ 3○ 4○ 5○

- **En general, ¿cree que su hermano gemelo y usted son muy distintos?**

□ **Sí, somos distintos tanto en el aspecto físico como en la personalidad**

□ **Sí, somos distintos principalmente en la personalidad**

□ **Sí, somos distintos principalmente en el aspecto físico**

□ **No, la verdad es que somos muy parecidos, en general**

□ **Otras respuestas: ……………………………………………………………………...**

- **¿Quién cree que es más seguro de sí mismo?**

□ **Yo**

□ **Mi hermano**

□ **Ninguno de los dos es más seguro de sí mismo que el otro**

- **Indique brevemente cómo definiría a su hermano:**

……………………………………………………………………………………………………….

……………………………………………………………………………………………………….

- **¿Qué rasgos o características propias considera que le diferencian más de su hermano?**

……………………………………………………………………………………………………….

……………………………………………………………………………………………………….

6. Usted y su gemelo: *parecido y confusión*.

- **¿Con qué frecuencia la gente confunde su voz con la de su hermano (por teléfono, en el portero automático de una casa o a través de una puerta cerrada, por citar algunos ejemplos)?**

□ **Con mucha frecuencia**

□ **Alguna vez**

□ **Muy de vez en cuando**

□ **Nunca**

- **¿Considera que su forma de hablar es diferente a la de su hermano?**

  □ **Sí, es bastante diferente**

  □ **Sí, es un poco diferente**

  □ **En absoluto. Creo que hablamos igual**

  □ **No sé, nunca me he parado a pensarlo**

- **Si ha respondido afirmativamente a la pregunta anterior, indique por qué considera que su forma de hablar es diferente a la de su hermano. Si alguna vez otras personas han comentado que usted y su hermano hablan de forma diferente, explique cuáles son esas diferencias que otras personas han observado:**

……

……

…………………………………………………………………………………

- **Si ha respondido "no" a la pregunta anterior, indique por qué considera que su forma de hablar es la misma que la de su hermano. Si alguna vez otras personas han comentado que usted y su hermano hablan de forma muy parecida, explique cuáles son esas semejanzas que otras personas han observado:**

……

……

…………………………………………………………………………………

Fin de la segunda parte del cuestionario.
Gracias por su colaboración.

A3: Participant questionnaire for the second recording session

CUESTIONARIO

Recogida de datos para tesis doctoral

SEGUNDA SESIÓN DE GRABACIÓN

1. Datos personales

Por favor, rellene los espacios en blanco con la información que se le solicita.

Nombre:...

Apellidos:...

NIF:...

2. Posibles cambios entre sesiones de grabación

Por favor, marque con una cruz donde corresponda, o bien rellene los espacios en blanco:

a. **Desde la anterior sesión de grabación, ¿le han detectado alguna patología de la voz o patología del habla?**

□ **Sí**

□ **No**

En caso afirmativo, indique cuál: ……………………………………………

b. **Desde la anterior sesión de grabación, ¿le han detectado algún problema de audición?**

□ **Sí**

□ **No**

En caso afirmativo, indique cuál: ……………………………………………

c. **En este momento, ¿padece alguna molestia o infección de: garganta y/o nariz, o bien sufre algún dolor en los dientes, la boca, los oídos o la mandíbula? Conteste en la siguiente tabla:**

Dolor/infección/molestia actual de:	Sí	No	Haga una breve descripción de la molestia o del dolor
Garganta	□	□	
Nariz	□	□	
Dientes	□	□	
Oídos	□	□	
Mandíbula	□	□	

d. **Desde la anterior sesión de grabación, ¿han cambiado sus hábitos relacionados con el tabaquismo?**

□ **Sí**

□ **No**

e. **En caso afirmativo, marque con una cruz donde corresponda:**

○ **He dejado de fumar (*Indique cuándo:* …………………………………………………………………)**

○ **He empezado a fumar (*Indique cuándo:* ……………………………………………………………)**

○ **La frecuencia con la que ahora fumo es mayor/menor (*Indique la frecuencia aproximada en cigarros por día*: ………………………………………………………………………………………..)**

f. **Desde la anterior sesión de grabación, indique si se ha producido alguno de los siguientes cambios:**

○	**Tengo desviado el tabique nasal**	
○	**Me han extraído alguna muela del juicio**	**→ Indique cuántas: 1 ○ 2 ○ 3 ○ 4 ○**
○	**Me falta alguna pieza dental**	**→ Indique cuántas: ....**
○	**Llevo aparato corrector de dientes**	**→ Tipo: .... ¿Desde hace cuánto?:..................**
○	**Tengo reflujo gástrico**	
○	**Tengo algún problema hormonal**	**→ Especificar cuál: ....**
○	**Tengo las adenoides inflamadas o hipertróficas (vegetaciones)**	
○	**Me han operado de vegetaciones**	
○	**Me han extraído las amígdalas**	
	Me han operado o realizado alguna intervención quirúrgica en alguna de las siguientes partes: ○ **nariz** ○ **garganta** ○ **mandíbula** ○ **dientes** ○ **cuerdas vocales**	**→ Especificar qué tipo de operación:** ....

g. **Marque con una cruz donde corresponda.**
Hoy (en este momento), ¿presenta usted síntomas de….?

○ **Resfriado o catarro**

○ **Rinitis alérgica**

○ **Otro tipo de obstrucción nasal (sinusitis, etc.): ....**

○ **Asma**

○ Dolor de garganta

○ Reflujo gastroesofágico

h. ¿Ha cambiado de profesión u ocupación laboral?

□ Sí

□ No

i. Únicamente en caso afirmativo, responda a las siguientes preguntas:

- Indique su profesión actual: ………………………………………………………………………………………………………

- Debido a su profesión, ¿pasa usted mucho tiempo hablando?

  ○ Sí, bastante

  ○ Ni mucho ni poco

  ○ Más bien poco

- ¿Nota molestias debido a un abuso vocal (ya sea por pasar excesivo tiempo hablando o por hacerlo en condiciones perjudiciales para su salud, como gritar, o hablar en lugares ruidosos, etc.)?

  ○ Con mucha frecuencia

  ○ Con frecuencia / A veces

  ○ Muy de vez en cuando, cuando hago algún esfuerzo puntual

  ○ Nunca

j. Actividades de ocio y tiempo libre

- Independientemente de su actividad profesional, ¿pasa usted mucho rato hablando, cantando o realizando cualquier otra actividad que implique la voz, tanto en su tiempo libre (conversaciones con amigos, por ejemplo), como debido a ciertas actividades de ocio que realice (teatro, cantar en un coro, etc.)?

  ○ Sí, bastante

  ○ Ni mucho ni poco

- ○ Más bien poco

- ¿Nota molestias debido a un abuso vocal?

  - ○ Con mucha frecuencia
  - ○ Con frecuencia / A veces
  - ○ Muy de vez en cuando, cuando hago algún esfuerzo puntual
  - ○ Nunca

Fin del cuestionario.
Gracias por su colaboración.

Appendix B: Corpus tasks and instructions

Note. We include here the instructions for the twin and non-twin siblings. The unrelated speakers received the same instructions but logically the addressee differ (e.g. "call your brother" is substituted by "call your friend").

- Instructions for the first task: spontaneous conversation

Instructions for speaker A (original, in Spanish):

Llama por teléfono a tu hermano al 2838. A partir del texto que acabas de leer sobre dos gemelos, mantén una conversación con tu hermano sobre alguna anécdota que hayáis vivido vosotros como gemelos/mellizos. Te puede servir de inspiración la anécdota que has leído antes sobre dos gemelos y su profesora de inglés, que les confundía. También podéis hablar sobre vuestra relación como gemelos, a partir de lo que habéis leído sobre los dos gemelos del texto, cuando uno se cayó al suelo y su hermano estaba muy angustiado.

NOTA 1: Podéis hablar de ambas cosas (anécdota o relación entre hermanos) o solamente de una de ellas.

NOTA 2: Podéis hablar durante el tiempo que queráis. Yo iré a llamaros al despacho, para empezar la siguiente tarea.

Instructions for speaker A (original, in Spanish):

Espera a que tu hermano te llame por teléfono. A partir del texto que acabas de leer sobre dos gemelos, mantén una conversación con tu hermano sobre alguna anécdota que hayáis vivido vosotros como gemelos/mellizos. Te puede servir de inspiración la anécdota que has leído antes sobre dos gemelos y su profesora de inglés, que les confundía. También podéis hablar sobre vuestra relación como gemelos, a partir de lo que habéis leído sobre los dos gemelos del texto, cuando uno se cayó al suelo y su hermano estaba muy angustiado.

NOTA 1: Podéis hablar de ambas cosas (anécdota o relación entre hermanos) o solamente de una de ellas.

NOTA 2: Podéis hablar durante el tiempo que queráis. Yo iré a llamaros al despacho, para empezar la siguiente tarea.

-Instructions for the second task: fax exchange

Note. The fax samples can be found in appendix D.

Instructions for speaker A (original, in Spanish):

Hay una serie de faxes sobre la mesa. La calidad de los mismos no es muy buena y parte de la información es difícil de leer.

Tu hermano también ha recibido estos faxes. Quizá los suyos tienen mejor calidad que los tuyos. Llámale por teléfono al número 2838 y pídele la información que te resulta difícil leer en tu fax.

1) Anota esta información en tu fax y dila en voz alta según la escribes.

2) Cuando hayas terminado de preguntarle a tu hermano por la información que te falta, comprueba que la has entendido bien.

Tu hermano deberá hacer lo mismo con la información que le falte. Por tanto, a sus faxes les faltará información que en los tuyos sí que aparece. Ofrécele esa información cuando te la pida.

Instructions for speaker B (original, in Spanish):

Hay una serie de faxes sobre la mesa. La calidad de los mismos no es muy buena y parte de la información es difícil de leer.

Tu hermano también ha recibido estos faxes. Quizá los suyos tienen mejor calidad que los tuyos. Espera a que te llame por teléfono y pídele la información que te resulta difícil leer en tus faxes.

1) Anota esta información en tu fax y dila en voz alta según escribes.

2) Cuando hayas terminado de preguntarle a tu hermano por la información que te falta, comprueba que la has entendido bien.

Tu hermano deberá hacer lo mismo con la información que le falte. Por tanto, a sus faxes les faltará información que en los tuyos sí que aparece. Ofrécele esa información cuando te la pida.

-Instructions for the third task: reading of two phonetically-balanced texts

Instructions for speaker A and B (original, in Spanish):

Por favor, lee los siguientes textos. Hazlo con su velocidad y entonación habituales. Si te equivocas o te trabas, no te preocupes y vuelve a leer solo desde el principio de la frase en la que se ha trabado.

TEXT 1:

Hay algo ahí, en el aire, que cambia el sentido de las cosas. Ese viento suave vuela, te toca la cara, mientras cuentas las hojas de los árboles. El agua corre buscando los campos. Al abrir las puertas de mi casa pienso: este país, una mañana más.

A mi edad, comienzan a faltarme las fuerzas, ya casi no soy joven, y la muerte de mi mujer en la guerra me pesa mucho. Cuando el cuerpo llega a esa hora, la ciencia de los doctores no logra detener el paso del tiempo.

De niño, allá en mi tierra, solía pasarme los días revolviendo de un lado a otro. Poco a poco, los coches de la ciudad fueron llamando mi atención. Mi madre decía que tuviera cuidado, pero yo me creía muy mayor, así que no tenía ni interés ni tiempo para mi propio signo.

Pero sigo, es cierto, cuántas cosas buenas encontré entre su gente. Si cuento los queridos veranos de entonces, no son siete, ni nueve, ni veinte. Debe ser que soy niño de nuevo en este cuerpo triste.

Source: Ortega, González, and Marrero (2000).

TEXTO 2:

El joyero Federico Vanero ha sido condenado por la Audiencia de Santander a ocho meses de arresto mayor y cincuenta mil pesetas de multa por un delito de compra de objetos robados. La vista oral se celebró el miércoles pasado y, durante ella, uno de los fiscales, Carlos Valcárcel, pidió para el joyero tres años de prisión menor y una multa de cincuenta mil pesetas. Gracias a las revelaciones de Vanero de hace dos años y medio se llegó a descubrir la existencia de una sospechosa mafia policial en España, parte de la cual se vio envuelta en el llamado "caso El Nani".

Source: Bruynickx, Harmegnies, Llisterri, and Poch (1994).

-Instructions for the fourth task: reading of two phonetically-balanced texts

Instructions for speaker A and B (original, in Spanish):

Tienes delante una serie de operaciones matemáticas que tu hermano debe calcular. Llámale al 2838 y pregúntale por la solución de cada operación. Dile que tú le vas a calcular el tiempo de respuesta. Cuando conteste, le darás la solución.

Cuando tu hermano haya terminado sus cálculos matemáticos, tú deberás hacer lo mismo con las operaciones que él te pida.

-Instructions for the fifth task: informal interview with the researcher

This task does not require specific and written instructions for the participants. The researcher just phones each participant separately and informs him that a telephone interview is going to take place. In the meantime, the other participant, who is in different separate room, has to fill the questions of the questionnaire which belongs to the specific recording session.

Appendix C: Search for words containing the vocalic sequences of interest

-Words containing two vowels where none of them is “i” or “u” (-aeo- group)

Unstressed –aeo-					
a + V		e + V		o + V	
ae	ao	ea	eo	oa	oe
israelí	baobab*	aérea	núcleo	almohadilla	poesía
maestría	ahogado	aleatorio	acordeonista	Joaquín	coherencia
aerodinámico	ahorrador	apeadero	aéreo	toallero	roedor
aeróbic	maorí	arbórea	aleonado	coatí	cohesión
paellada		argéntea	leonés		poemario
saetero		bronceador	arbóreo		incoherencia
		beatíficación	arqueológico		Noemí
			beodez		Bengoechea

			mediterráneo		héroe
			cutáneo		coeficiente
			crustáceo		
			cetáceo		
			geográfico		
			teoría		
			vídeo		
			espontáneo		

Stressed –aeo-											
a + V				e + V				o + V			
ae		ao		ea		eo		oa		oe	
áe	aé	áo	aó	éa	eá	éo	eó	óa	oá	óe	oé
Herráez	maestro	bacalao*	cantaor	aldea	alboreada	apogeo	león	barbacoa	almohada	aloe	bohemio
Narváez	paella	Callao	cantaora	jalea	aldeano	ateo	apoteosis	anchoa	cloaca		cohete
Arbeláez	saeta	caos	caoba	asamblea	alineado	ateneo	beodo	Balboa	croata		proeza
Sáez	Israel	colacao	karaoke	azotea	teatro	bloqueo	gaseoso	canoa	guipuzcoano		soez
	Ismael	nao	faraón	idea	arqueado	boxeo	campeón	loa	koala		oeste
	Rafael	Estanislao	Naomi		asambleario	camafeo		proa	loar		Noelia
	Maestre	Bilbao	Paola		aseado	fariseo		Ulloa	oasis		Villarroel
	Baena	Laos	zanahoria		balneario	museo		Ainhoa	toalla		
	Baeza	Venceslao			beato	europeo		Novoa	cloaca		
	aéreo				cereal	empleo		Ochoa			
					Galeano	feo					

-Words containing two vowels where one of them is “i” or “u”

V + i,u					
Stressed V + unstressed i, u					
Stressed V + unstressed i			Stressed V + unstressed u		
ái	éi	ói	áu	éu	óu
bonsái	alféizar	tiroides	Paula	Ceuta	Sousa
samurái	béisbol	sinusoide	Laura	feudo	Mouriño
káiser	dieciséis	asteroide	pausa	euro	Rouco
tráiler	géiser	convoy	tauro	deuda	
	jersey	coito	Claudia	neutro	
	aceite	hoy	flauta	fisioterapeuta	
	afeite	androide	aunque		
	deleite	boina	aula		
	peine	espermatozoide	fauna		

Unstressed V + (stressed and unstressed) i, u

Unstressed V +(stressed and unstessed) i					
ai		ei		oi	
aí	Unstressed ai	eí	Unstressed ei	oí	Unstressed oi
ahí	airoso	increíble	aceitoso	oído	boicot
ahínco	arcaizante	vehículo	aceituna	Eloísa	helicoidal
alcalaíno	bailable	seísmo	afeitado	oído	tiroideo
bilbaíno	caimán	proteína	deidad	heroína	sinusoidal
Caín	daiquiri	Andreína	descafeinado	egoísta	coincidencia
paraíso	faisán		deleitoso		Moisés
paracaídas	maizal		peineta		
raíz	paisano		reinado		
país	paisaje		voleibol		
Abigaíl	vaivén		Reinoso		
Anaís	vainilla		Deidamia		
Aída					

Unstressed V + (stressed or unstressed) u					
au		eu		ou*	
aú	Unstressed au	eú	Unstressed eu	oú	Unstressed ou
baúl	aflautado	reúma	ceutí	×	estadounidense
ataúd	araucano	transeúnte	endeudado	×	microuniverso
aún	audaz - audacia	feúcho	neutral	×	macrourbanización
Raúl	Paulina		seudónimo	×	
Saúl	auténtico		reunión	×	
	aurel		euforia	×	
	autarquia		neumático	×	
	autismo		mileurista	×	
	autóctono		feudal	×	
	auxiliar		europeo	×	
	autor		Eugenia	×	

* The vocalic sequence “ou” was discarded because there are few words containing it.

(unstressed and stressed) i, u + unstressed V											
(unstressed and stressed) i + unstressed V						(unstressed and stressed) u + unstressed V					
ia		ie		io		ua		ue		uo	
ía	ia	íe	ie	ío	io	úa	ua	úe	ue	úo	uo
María	socialista	israelíes	sociedad	río	accionario	actual	situación	tabúes	cuestión	búh o	virtuosismo
agonía	iniciativa	marroquíes	propiedad	frío	acondicionado	habitual	actuación	hindúes	antigüedad	dúo	duodécimo
mayoría	financiación	rien *[1]	propietario	vacío	evolucionista	anual	continuación	iglúes	ambigüedad	flúor	duodeno
compañía	asociación	jabalíes	inquieto	judío	internacional	intelectual	cuarenta	bambúes	pueril	actúo*[3]	fluorita
todavía	especialista	ceutíes	ansiedad	tío	violencia	usuario	evaluación	insinúe*[2]	buenísimo		fluorescente
policía	seria	saudíes	seriedad	mío	periodista	vestuario	puntuación		duelista		antiguo
economía	necesaria	marbellíes	variedad	desafío	profesional	espiritual	cuaderno		buenazo		residuo
energía	media	Díez	piedad	desvío	funcionario	virtual	ecuación		huesuda		monstruo
teoría	voluntaria		medieval	crío	provisional	manual	ecuador		huevera		mutuo
tecnología	criatura		diesel	envío	regional	puntual	individualismo				ambiguo
poesía	historia		inquietante	Calíope	mediodía	peruano	acuarela				asiduo
autonomía	obvia		vienés	Darío	guionista	ritual	guapísima				ingenuo
fotografía	propia		dieta	Rocío	racional	Eduardo	menstruación				arduo

fría	Marianela		obviedad	Ríos	violín	grúa	aguacate				continuo
alegría	familia		miedoso	Torío	camionero	púa	Juanita				oblicuo
Estefanía	memoria				terciopelo	rúa	lengua				Fructuoso
Elías	importancia				Dionisio	capicúa	Cuadrado				Lecuona
Lucía	consecuencia				Guiomar	cacatúa	cuartel				
García	experiencia				Violeta	tatuaje	antigua				
	justicia				Rioseco	ganzúa	paraguas				

*[1] The combination “íe” is very productive in Spanish, as many verb forms in the third person singular end in –íe: *sonríe, varíe, enfríe*. In plurar, the ending for the third person plural -íen is also productive: *chirríen, confíen, guíen*.

*[2] The combination “úe” is also quite productive in Spanish verbal endings: *desvirtúe, sitúe, evalúe*.

*[3] The vocalic sequence “úo” appear in some verbal endings for the first person singular: *insinúo, puntúo*.

Unstressed i, u + stressed V					
Unstressed i + stressed V			Unstressed u + stressed V		
iá (diphthong/hiatus)	ié	ió	uá	ué	uó
diario (d/h)	pie	situación	cuál	nuevo	cuota
piano (d/h)	también	información	cuándo	acuerdo	respetuoso
material (d)	quién	decisión	cuánto	cuenta	virtuoso
social (d)	miércoles	opinión	acuático	puesto	defectuoso
especial (d)	recién	televisión	ecuánime	juez	majestuoso
oficial (d)	hincapié	dirección	donjuán	muerte	monstruoso
Valencia (d)	higiénico	operación	cuádriceps	fuente	acuoso
violencia (d)	estiércol	investigación	escuálido	nuestro	secuoya
victoria (d)	soviético	reunión	zaguán	buen	
democracia (d)	tiempo	oposición	tatuaje	pueblo	
mundial (d/h)	bien	organización	Juárez	juego	
confianza (d/h)	siempre	cuestión	Suárez	hueso	
experiencia (d)	gobierno	formación	Marijuán	luego	

estudiante (d)	miembro	región	cuatro	fuerza	
Adrián (d)	suficiente	Diógenes	igual	muestra	
viaje (d/h)	cierto	Encarnación	sexual	respuesta	
Diana (d/h)	siete	Carrión	guardia	cuerpo	
fianza (d/h)	izquierda		lenguaje	vuelta	
	septiembre		suave	huevo	
	viernes		mensual	jueves	
	abierto		guapa	nueve	
	tierra		Atahualpa	fuego	
	miedo		Juan	puerta	
				huelga	
				cruel	
				sueco	

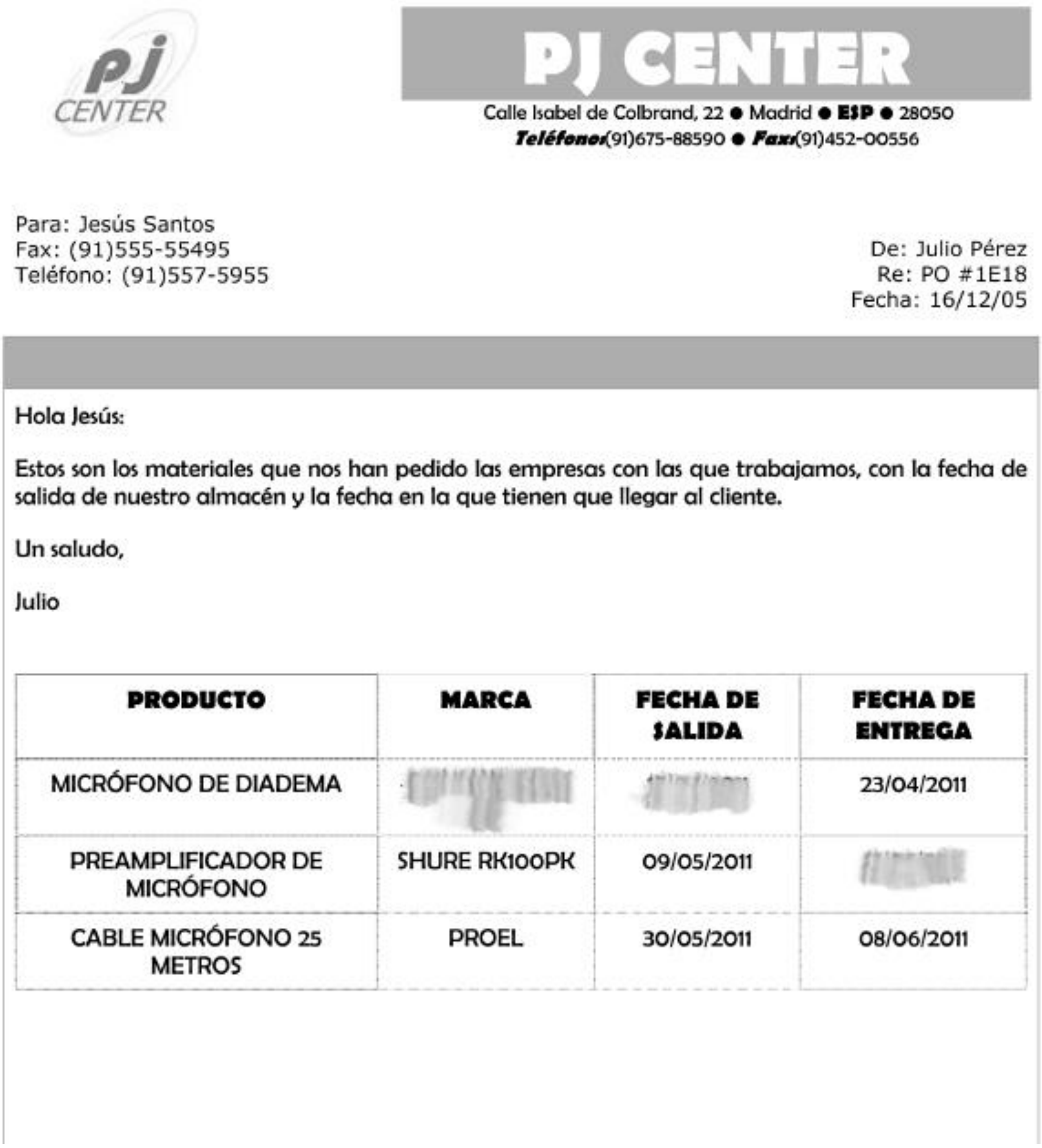
-Words containing the vowel combination of “i” and “u”

GROUP iu, ui					
Stressed				unstressed	
iu		ui		iu	ui
íu	iú	úi	uí		
NO	demiúrgico	NO	acuífero	ciudad	continuidad
	viuda		casuística	diurético	suicidio
	triunfo		jesuítico	viudez	ingenuidad
	miura		Ruiz	ciudadano	distribuidor
	oriundo		gratuito	triunfal	fluidez
	diurna		hinduista	premium	gratuidad
	veintiuno		beduino	medium	ruidoso
			juicio		ruinoso
			incluido		huidizo
			circuito		cuidado

			suizo		jesuita
			Suiza		Luisina
			genuino		ruiseñor
			atribuido		Ruiseco
			excluido		cuidador
			buitre		juicioso
			fortuito		
			ruina		

Appendix D: Fax samples for the second speaking task

Note. We include only the fax samples of speaker A.



PJ CENTER

Calle Isabel de Colbrand, 22 ● Madrid ● **ESP** ● 28050
Teléfono:(91)675-88590 ● ***Fax:***(91)452-00556

Para: Jesús Santos
Fax: (91)555-55495
Teléfono: (91)557-5955

De: Julio Pérez
Re: PO #1E18
Fecha: 16/12/05

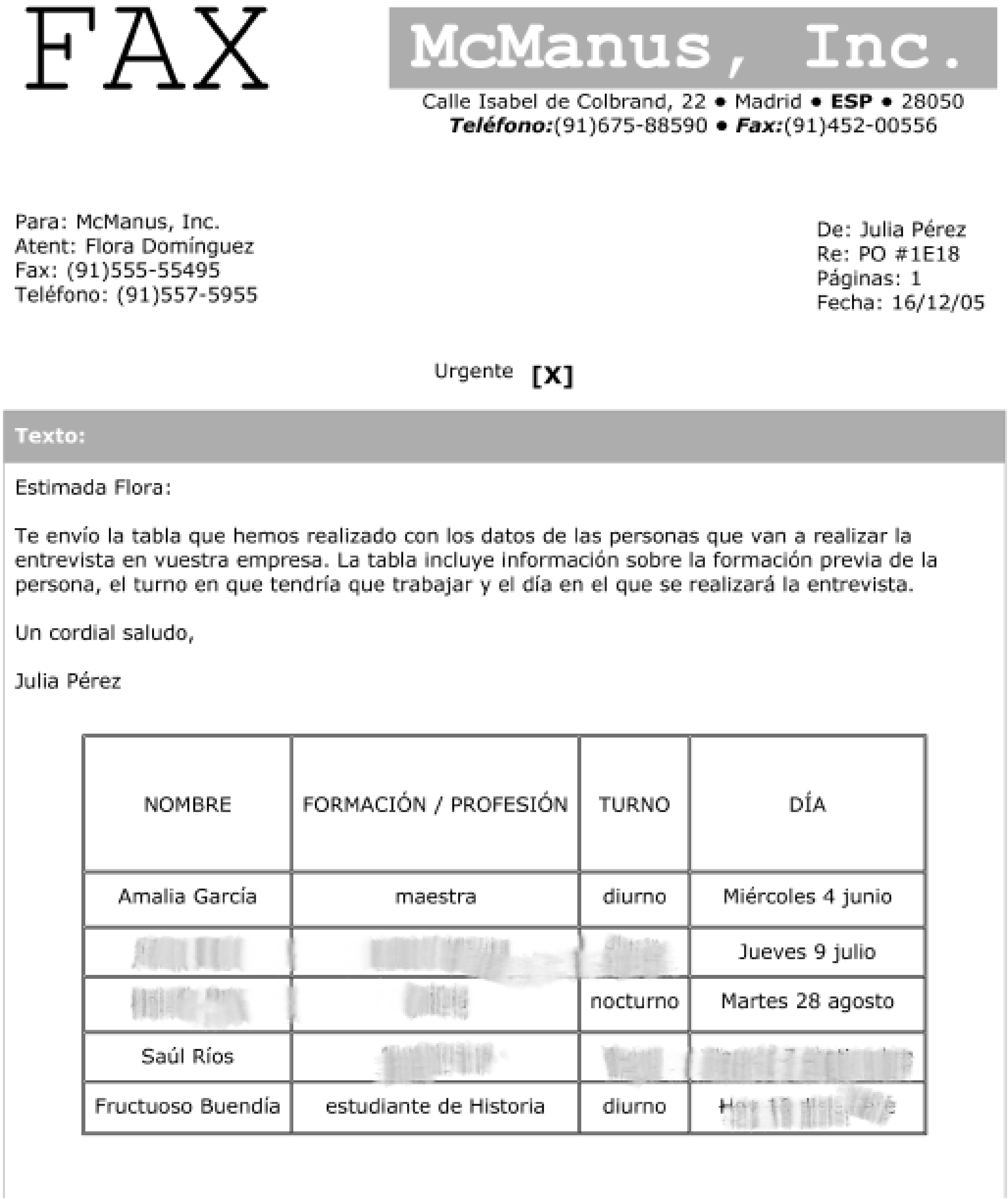
Hola Jesús:

Estos son los materiales que nos han pedido las empresas con las que trabajamos, con la fecha de salida de nuestro almacén y la fecha en la que tienen que llegar al cliente.

Un saludo,

Julio

PRODUCTO	MARCA	FECHA DE SALIDA	FECHA DE ENTREGA
MICRÓFONO DE DIADEMA	[redacted]	[redacted]	23/04/2011
PREAMPLIFICADOR DE MICRÓFONO	SHURE RK100PK	09/05/2011	[redacted]
CABLE MICRÓFONO 25 METROS	PROEL	30/05/2011	08/06/2011



FAX

McManus, Inc.

Calle Isabel de Colbrand, 22 • Madrid • **ESP** • 28050
Teléfono:(91)675-88590 • ***Fax:***(91)452-00556

Para: McManus, Inc.
Atent: Flora Domínguez
Fax: (91)555-55495
Teléfono: (91)557-5955

De: Julia Pérez
Re: PO #1E18
Páginas: 1
Fecha: 16/12/05

Urgente **[X]**

Texto:

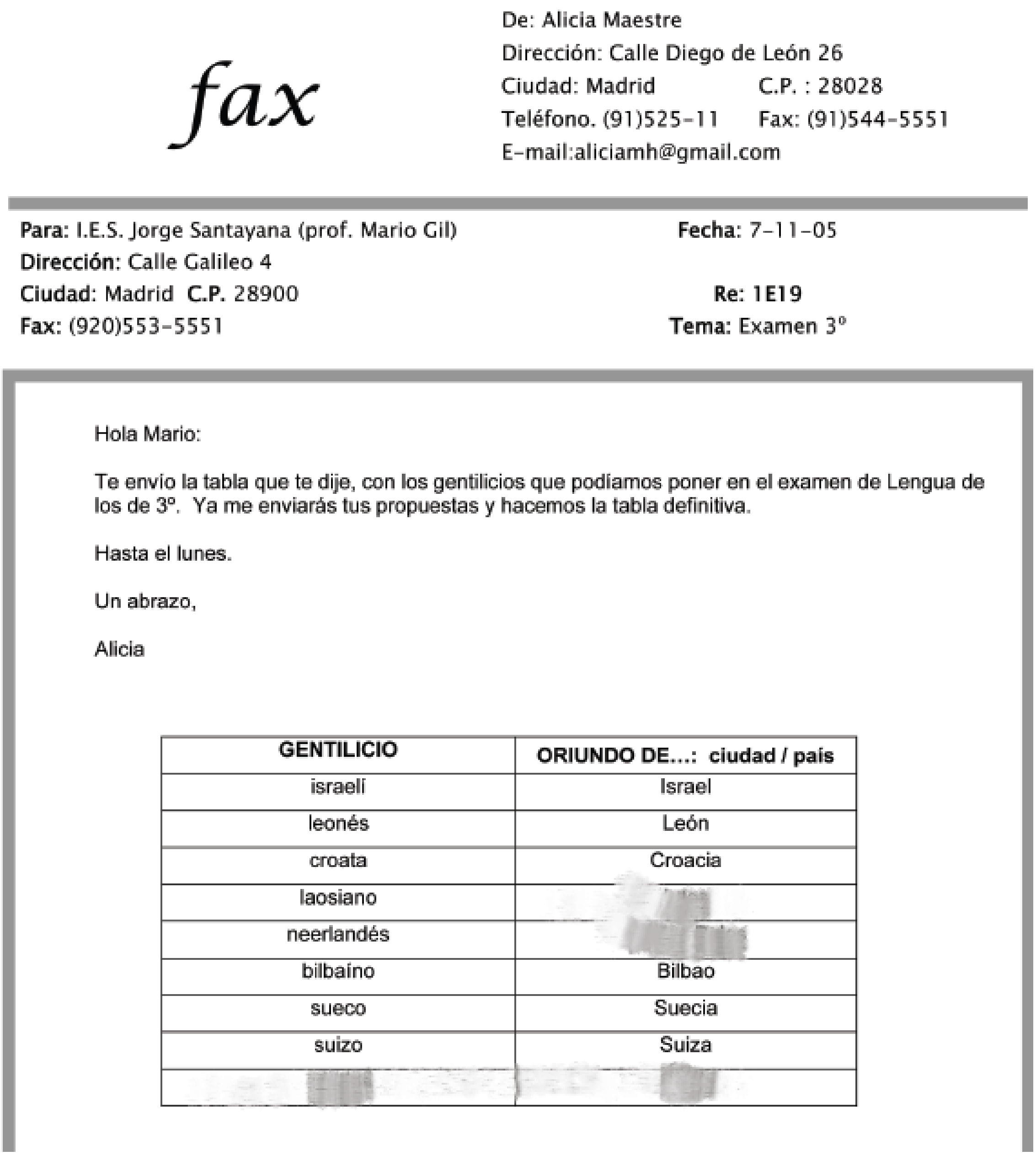
Estimada Flora:

Te envío la tabla que hemos realizado con los datos de las personas que van a realizar la entrevista en vuestra empresa. La tabla incluye información sobre la formación previa de la persona, el turno en que tendría que trabajar y el día en el que se realizará la entrevista.

Un cordial saludo,

Julia Pérez

NOMBRE	FORMACIÓN / PROFESIÓN	TURNO	DÍA
Amalia García	maestra	diurno	Miércoles 4 junio
[illegible]	[illegible]	[illegible]	Jueves 9 julio
[illegible]	[illegible]	nocturno	Martes 28 agosto
Saúl Ríos	[illegible]	[illegible]	[illegible]
Fructuoso Buendía	estudiante de Historia	diurno	[illegible]



fax

De: Alicia Maestre
Dirección: Calle Diego de León 26
Ciudad: Madrid C.P. : 28028
Teléfono. (91)525-11 Fax: (91)544-5551
E-mail:aliciamh@gmail.com

Para: I.E.S. Jorge Santayana (prof. Mario Gil) **Fecha**: 7-11-05
Dirección: Calle Galileo 4
Ciudad: Madrid **C.P.** 28900 **Re: 1E19**
Fax: (920)553-5551 **Tema:** Examen 3º

Hola Mario:

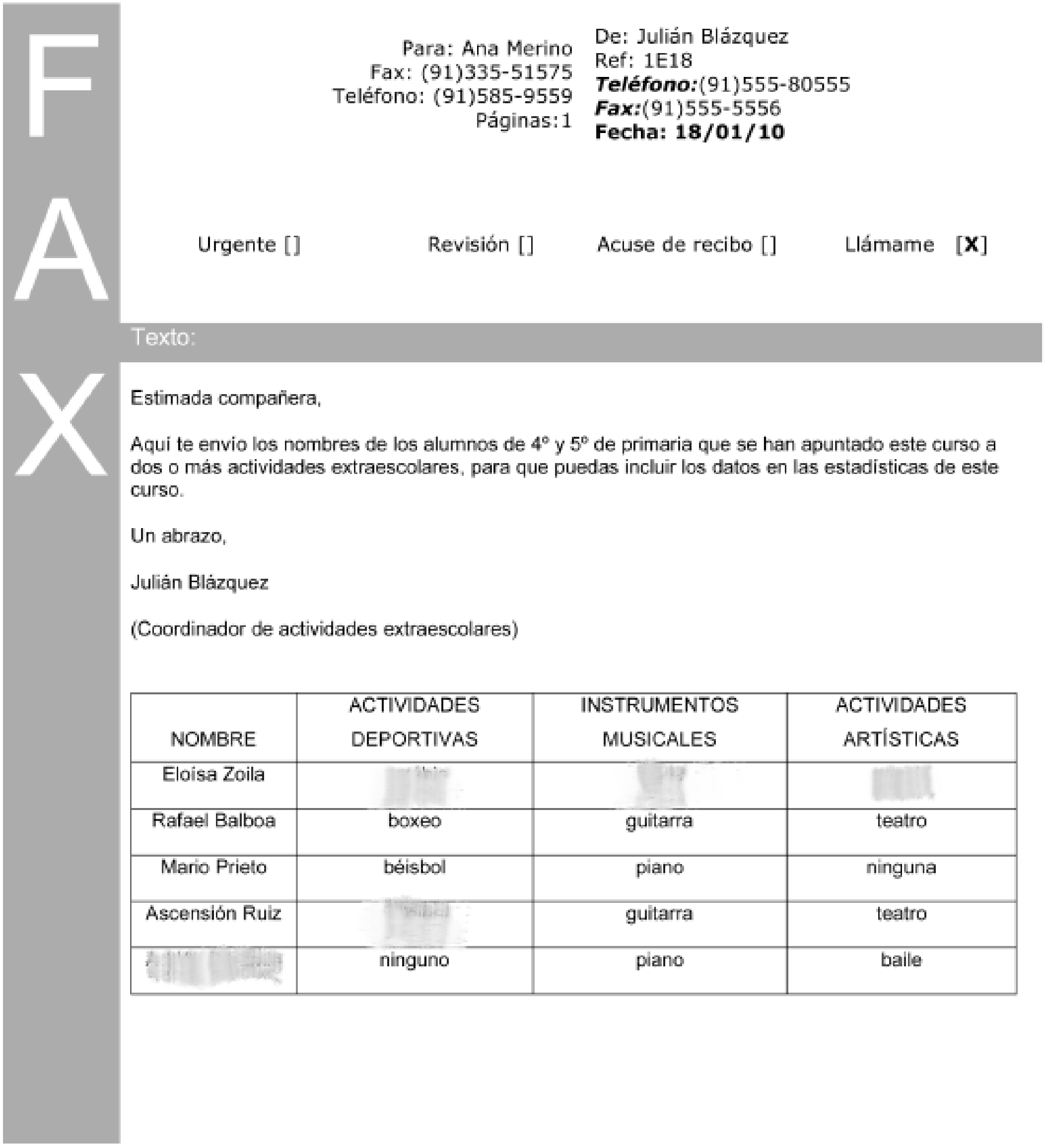
Te envío la tabla que te dije, con los gentilicios que podíamos poner en el examen de Lengua de los de 3º. Ya me enviarás tus propuestas y hacemos la tabla definitiva.

Hasta el lunes.

Un abrazo,

Alicia

GENTILICIO	**ORIUNDO DE…: ciudad / país**
israelí	Israel
leonés	León
croata	Croacia
laosiano	[illegible]
neerlandés	[illegible]
bilbaíno	Bilbao
sueco	Suecia
suizo	Suiza
[illegible]	[illegible]



F A X

Para: Ana Merino
Fax: (91)335-51575
Teléfono: (91)585-9559
Páginas:1

De: Julián Blázquez
Ref: 1E18
Teléfono:(91)555-80555
Fax:(91)555-5556
Fecha: 18/01/10

Urgente [] Revisión [] Acuse de recibo [] Llámame [**X**]

Texto:

Estimada compañera,

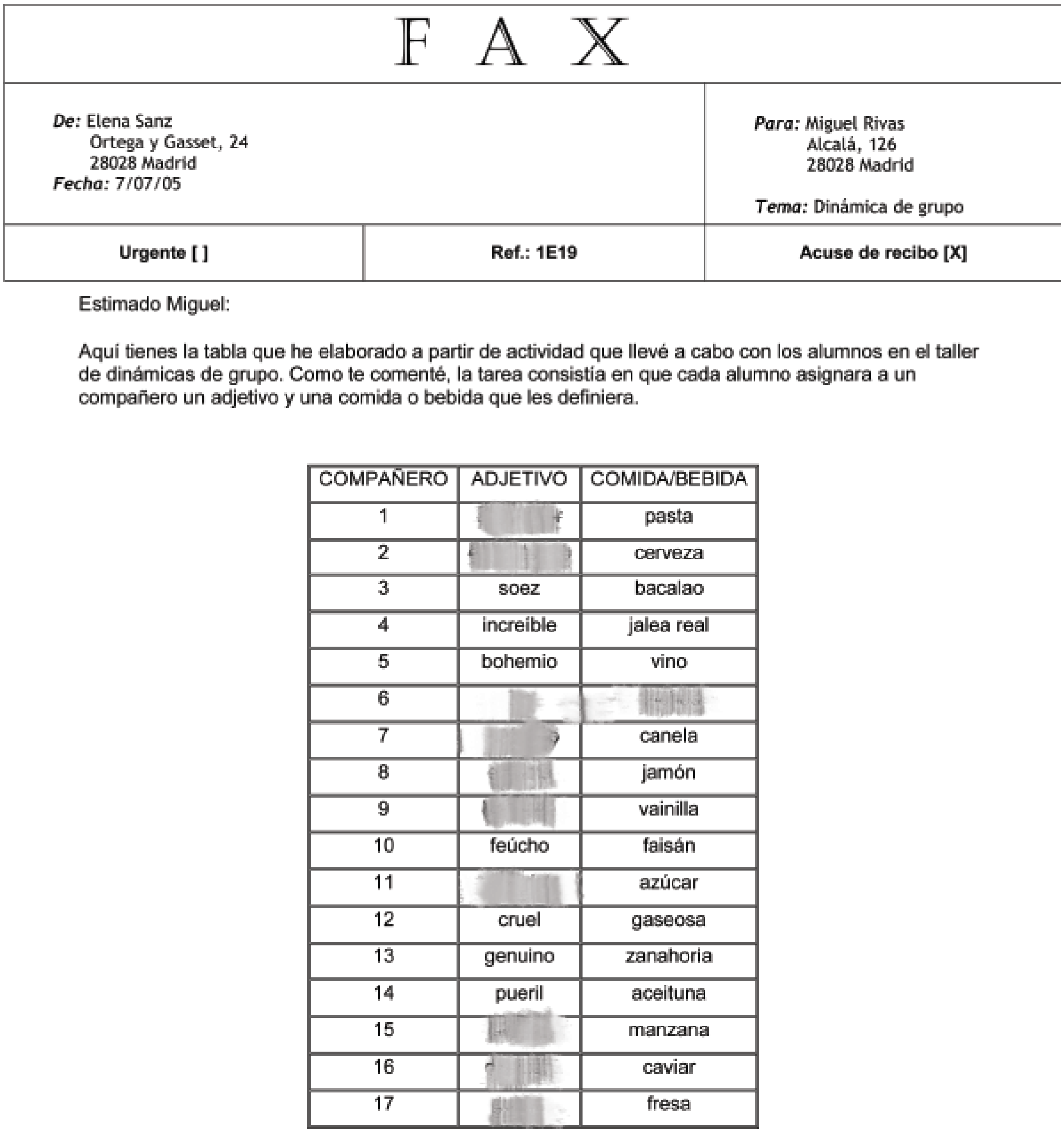
Aquí te envío los nombres de los alumnos de 4º y 5º de primaria que se han apuntado este curso a dos o más actividades extraescolares, para que puedas incluir los datos en las estadísticas de este curso.

Un abrazo,

Julián Blázquez

(Coordinador de actividades extraescolares)

NOMBRE	ACTIVIDADES DEPORTIVAS	INSTRUMENTOS MUSICALES	ACTIVIDADES ARTÍSTICAS
Eloísa Zoila	[redacted]	[redacted]	[redacted]
Rafael Balboa	boxeo	guitarra	teatro
Mario Prieto	béisbol	piano	ninguna
Ascensión Ruiz	[redacted]	guitarra	teatro
[redacted]	ninguno	piano	baile



F A X

De: Elena Sanz
Ortega y Gasset, 24
28028 Madrid
Fecha: 7/07/05

Para: Miguel Rivas
Alcalá, 126
28028 Madrid

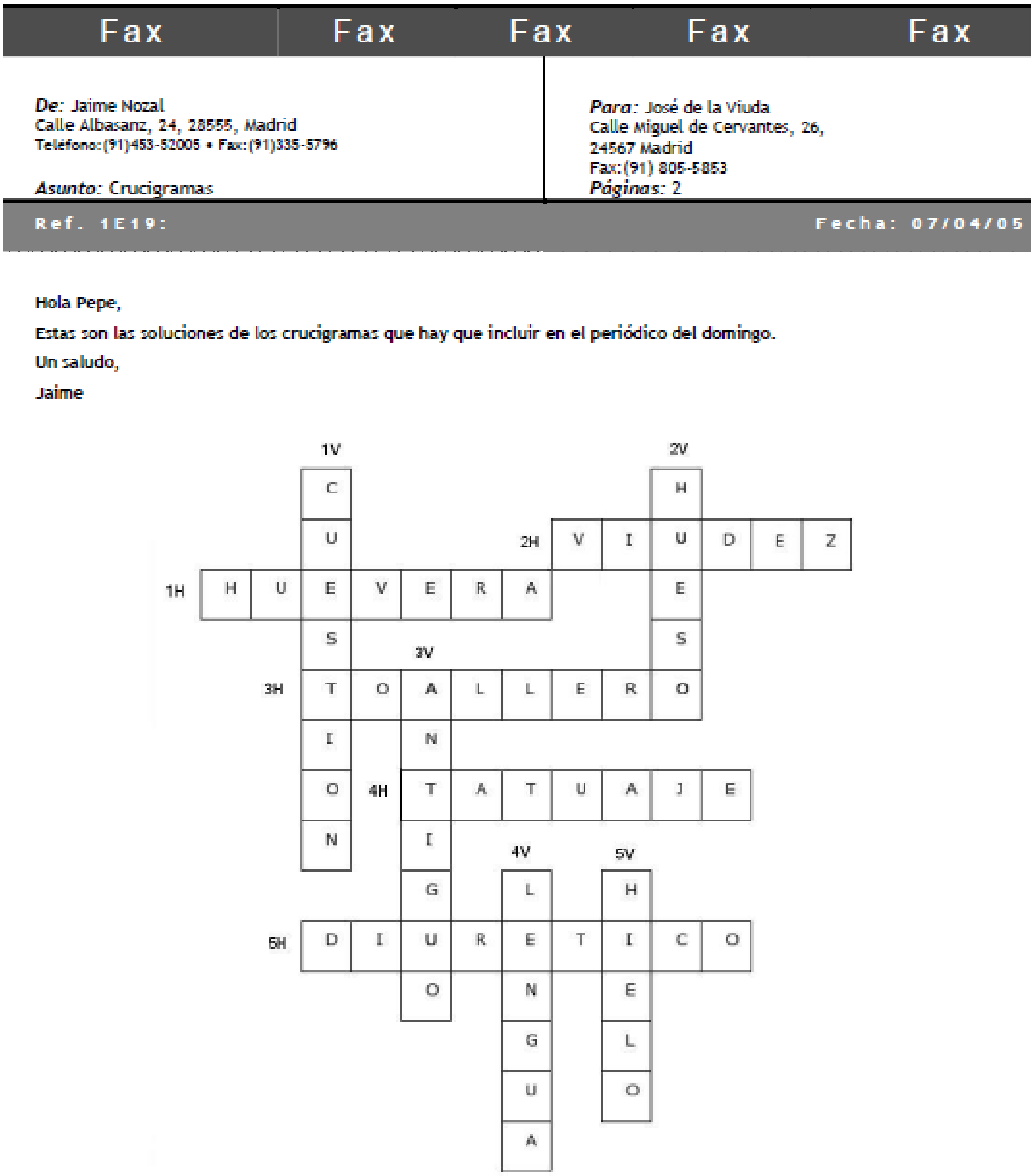
Tema: Dinámica de grupo

Urgente [ ]	**Ref.: 1E19**	**Acuse de recibo [X]**

Estimado Miguel:

Aquí tienes la tabla que he elaborado a partir de actividad que llevé a cabo con los alumnos en el taller de dinámicas de grupo. Como te comenté, la tarea consistía en que cada alumno asignara a un compañero un adjetivo y una comida o bebida que les definiera.

COMPAÑERO	ADJETIVO	COMIDA/BEBIDA
1	[redacted]	pasta
2	[redacted]	cerveza
3	soez	bacalao
4	increíble	jalea real
5	bohemio	vino
6	[redacted]	[redacted]
7	[redacted]	canela
8	[redacted]	jamón
9	[redacted]	vainilla
10	feúcho	faisán
11	[redacted]	azúcar
12	cruel	gaseosa
13	genuino	zanahoria
14	pueril	aceituna
15	[redacted]	manzana
16	[redacted]	caviar
17	[redacted]	fresa



Fax Fax Fax Fax Fax

De: Jaime Nozal
Calle Albasanz, 24, 28555, Madrid
Telefono:(91)453-52005 • Fax:(91)335-5796

Asunto: Crucigramas

Para: José de la Viuda
Calle Miguel de Cervantes, 26,
24567 Madrid
Fax:(91) 805-5853
Páginas: 2

Ref. 1E19:

Fecha: 07/04/05

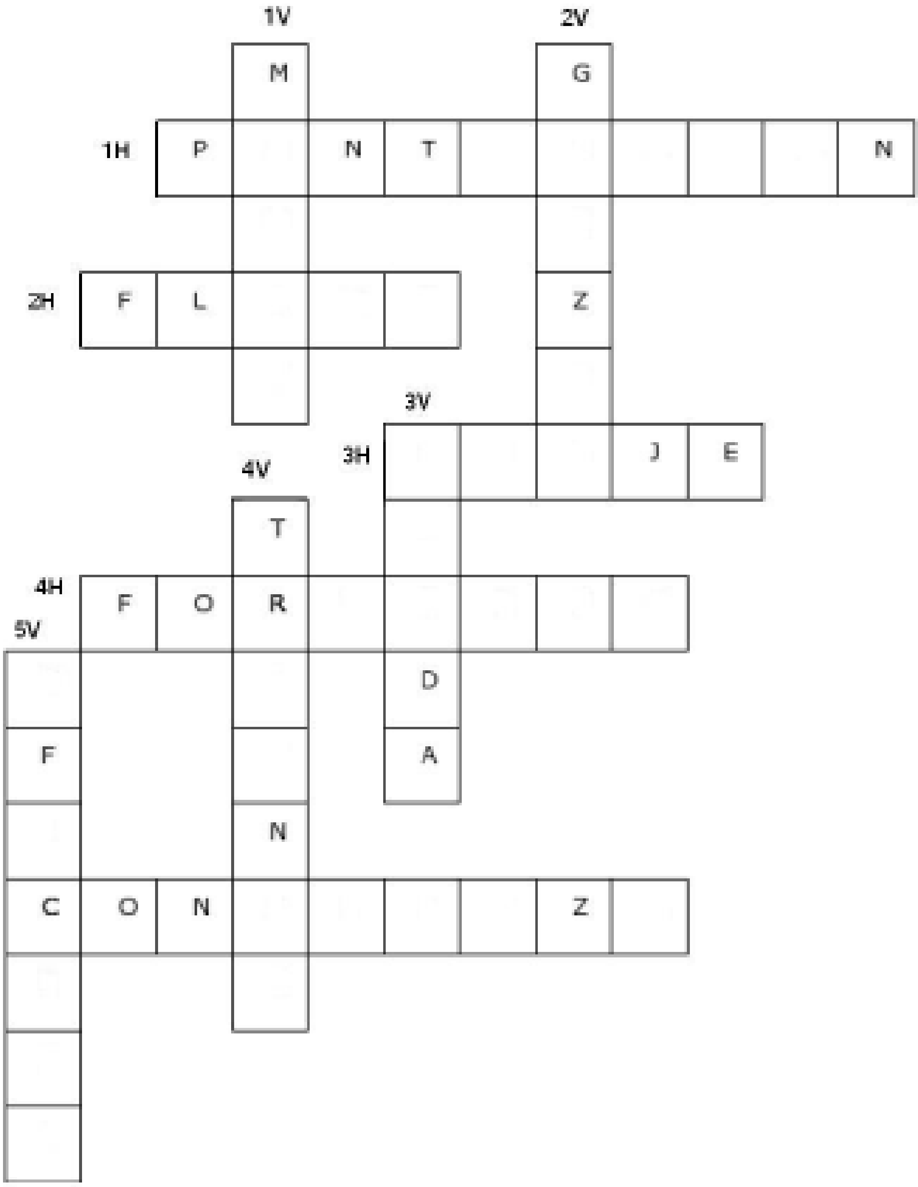
Hola Pepe,

Estas son las soluciones de los crucigramas que hay que incluir en el periódico del domingo.

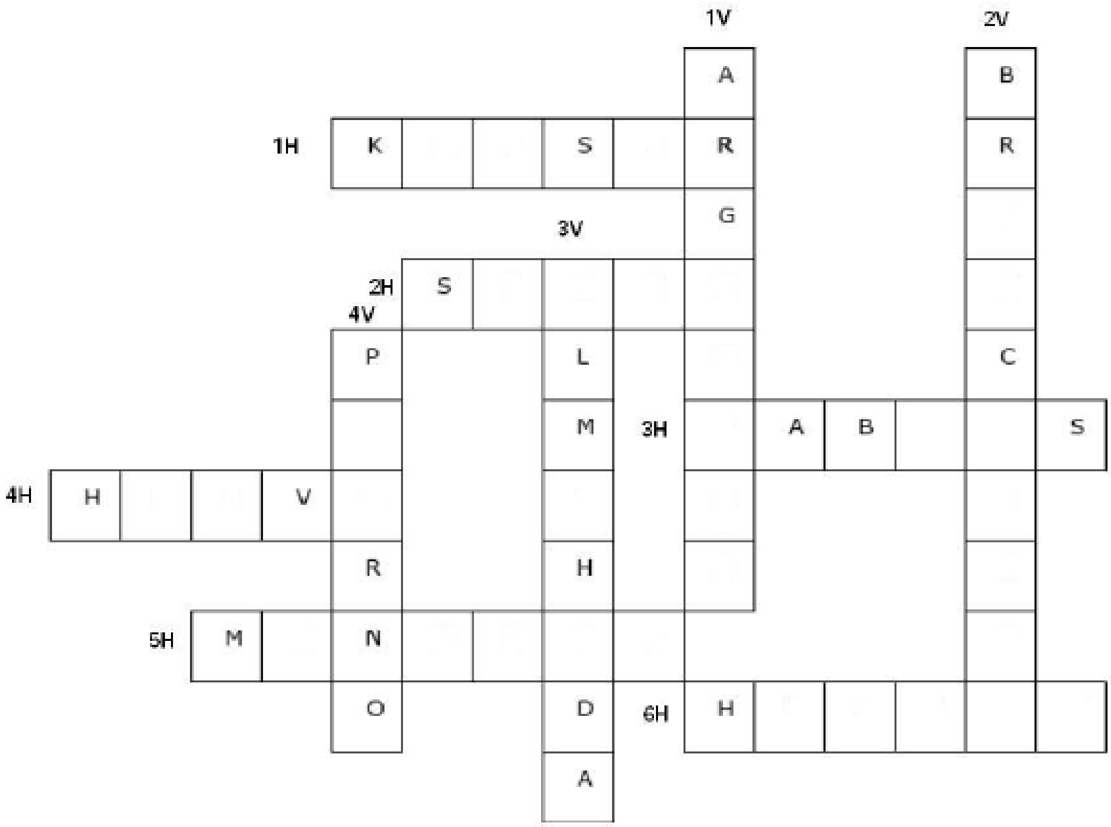
Un saludo,

Jaime

1V
C U E S T I O N
2V
H U E S O
2H
V I U D E Z
1H
H U E V E R A
3V
A N T I G U O
3H
T O A L L E R O
4H
T A T U A J E
4V
L E N G U A
5V
H I E L O
5H
D I U R E T I C O


1V
2V
M
G
1H
P
N
T
N
2H
F
L
Z
3V
3H
J
E
4V
T
4H
F
O
R
5V
D
F
A
N
5H
C
O
N
Z

							1V				
							B				
1H		S	E	I	S	M	O				
				2V			I				
2H	A	S	O	C	I	A	C	I	O	N	
				I			O				
		3V		U			T	4V			
		A		D				C			
3H		T	R	A	N	S	E	U	N	T	E
		A		D				O			
		H				4H	A	T	A	U	D
		U						A			
5H	V	A	C	I	O						
		L									
		P									
6H		A	N	S	I	E	D	A	D		


1V
2V
A
B
1H
K
S
R
R
G
3V
2H
S
4V
P
L
C
M
3H
A
B
S
4H
H
V
R
H
5H
M
N
O
D
6H
H
A

Appendix E: Three most relevant parameters per comparison (glottal source analysis)

Speaker Comparison	Type	Parameter 1	Parameter 2	Parameter 3
05-05	MZ	53	54	55
05-06	MZ	53	54	55
06-06	MZ	61	54	6
15-15	DZ	61	15	43
15-16	DZ	17	35	32
16-16	DZ	17	49	35
21-21	H	46	35	32
21-22	H	35	32	49
22-22	H	55	64	18
27-27	R	17	8	36
27-28	R	17	8	21
28-28	R	42	50	21
41-41	MZ	11	10	8
41-42	MZ	11	61	54
42-42	MZ	48	9	30
45-45	DZ	55	54	53
45-46	DZ	54	53	55
46-46	DZ	42	54	43
47-47	H	53	55	54
47-48	H	55	51	54
48-48	H	39	49	40
53-53	R	51	41	42
53-54	R	51	21	42
54-54	R	54	11	14

Appendix F: Twin studies in chronological order

Author	Year	Approach	n	♂ y ♀	Data collection method
Lundström	1948	Other[148]	202 twin pairs: 100 MZ pairs 102 DZ pairs	88 ♂ pairs 114 ♀ pairs	Twins recruited from schools in Stockholm
Gedda, Fiori, & Bruno	1960	Perceptual	24 twin pairs: 20 MZ pairs 4 DZ pairs	Half ♂ Half ♀	Twins recruited from the *Mendel Institute* in Rome
Alpert, Kutzberg, Pilot, & Friedhoff	1963	Acoustic	12 twin pairs: 6MZ pairs 6 DZ pairs	3 MZ, 4 DZ ♂pairs 3MZ, 2 DZ ♀pairs	Unspecified
Luchsinger & Arnold	1965	Perceptual[149]	40 twin pairs: 28 MZ pairs 12 DZ pairs	Unspecified	Unspecified
Flach, Schwickardi, & Steinert	1968	Acoustic	20 twin pairs:	Unspecified	Unspecified

[148] All of the approaches followed in the 34 studies reviewed circumscribe to any of the following four main topics: perceptual, acoustic, articulatory or automatic. Only Lundström (1948) and Spielman (2012) are designed with the label "other" since they do not belong to any phonetic domain, as we explained in Chapter 2.

[149] This book actually belongs to the clinical sciences. We classify it as "perceptual" since the reference to the identification of twins' voices on the telephone seems more appropriate here.

			10 MZ pairs 10 DZ pairs		
Cornut	1971	Acoustic	20 twin pairs: 13 MZ pairs 7 DZ pairs	Unspecified	Unspecified
Forrai & Gordos	1982	Acoustic	117 twin pairs: 70 MZ pairs 47 DZ pairs	29 MZ , 20 ♂ DZ 41 MZ, 27 ♀ DZ	Unspecified
Przybyla, Horii, & Crawford	1992	Acoustic	62 twin pairs: 53 MZ pairs 9 DZ pairs	11 MZ, 1 DZ ♂pairs 42 MZ, 8 DZ ♀ pairs	Through the *Midwest Twin Register* at the University of Kansas
Homayounpour & Chollet	1995	Perceptual, acoustic and automatic	9 MZ pairs 4 non-twin siblings	4 MZ ♂ pairs 5 MZ ♀ pairs	Unspecified
Nolan & Oh	1996	Acoustic	3 MZ twin pairs	2 ♂ pairs 1 ♀ pair	Unspecified
Johnson & Azara	2000	Perceptual	6 twin pairs: 5 MZ pairs 1 MZ pair	All female	Unspecified

Fuchs et al.	2000	Acoustic	31 MZ pairs	11 ♂ pairs 20 ♀ pairs	Twin association *Zwillingsclubs Werdau 1986 e.V*
Whiteside & Rixon	2000	Perceptual and acoustic	1 MZ pair	Male	Unspecified
Decoster, Van Gysel, Vercammen, & Debruyne	2000	Perceptual and Acoustic	15 MZ pairs	10 ♂ pairs 5 ♀ pairs	Twin members of *East Flanders Prospective Twin Survey* Advertisement in a local paper for students Snowball sampling method
Whiteside & Rixon	2001	Acoustic	1 MZ pair	Male	Unspecified
Yarmey, Yarmey, Yarmey, & Parliament	2001	Perceptual	1 MZ twin pair	Male	Unspecified[150]
Debruyne, Decoster, Van Gysel, & Vercammen	2002	Acoustic	60 twin pairs:[151] 30 MZ pairs 30 DZ pairs	All female	Same as in Decoster et al. (2000)
Whiteside and Rixon	2004	Acoustic	1 MZ pairs + their male siblings	Male	Unspecified

[150] It is to note that this study does not focus on twins but on the perceptual identification of familiar voices. That is why only one twin pair participates in this study (supposedly by chance) and no collection method is then described. Since the objective of the study is the recognition of familiar voices, the authors only state the following: "Following a convenience sampling procedure involving friends, colleagues, neighbours, etc., 68 men and women agreed to participate as speakers."

[151] It is not clear whether the twins are 60 or 30. Sometimes they refer to "30 pairs of MZ and 30 pairs of DZ" but some other times they mention "30 female MZ and 30 DZ twins", not "pairs".

Ryalls, Shaw, & Simon	2004	Acoustic	2 MZ twin pairs	All female	Unspecified
Scheffer, Bonastre, Ghio, & Teston	2004	Automatic	17 MZ pairs	7 ♂ pairs 10 ♀ pairs	Through the TV program ”Les Jumeaux : l’expérience inédite”
Van Lierde, Vinck, De Ley, Clement, & Van Cauwenberge	2005	Acoustic	45 MZ twin pairs	19 ♂ pairs 26 ♀ pairs	Most twins were members of *the East Flanders Prospective Twin Survey* and had been invited by the Flemish television transmitter to participate in an educational program.
Loakes	2006a	Acoustic	4 twin pairs: 3 MZ pairs 1 DZ pair	All male	Through the *Australian Twin Registry*
Loakes	2006b	Acoustic	4 twin pairs: 3 MZ pairs 1 DZ pair	All male	Through the *Australian Twin Registry*
Charlet & Peral	2007	Automatic	33 families	19 families with one son and one daughter, 10 with 2 sons and 4 with 2 daughters	Unspecified
Kinga	2007	Acoustic	3 MZ pairs 3 sister pairs	All female	Unspecified

Ariyaeeinia, Morrison, Malegaonkar, & Black	2008	Automatic	49 MZ pairs	9 ♂ pairs 40 ♀ pairs	Through the Centre for Twin Research and Genetic Epidemiology at St. Thomas' Hospital in London, UK.
Feiser	2009	Acoustic	10 sibling pairs	5 ♂ pairs 5 ♀ pairs	Unspecified
Kim	2009	Automatic	22 twin pairs: 17 MZ pairs 5 DZ pairs 1 triplet	All female	Unspecified
Künzel	2010	Automatic	35 MZ pairs	9 ♂ pairs 26 ♀ pairs	Author’s participation in a TV production called “Die Zwillings-Show” in which MZ twins competed to be “Germany’s most similar twins”.
San Segundo	2010c	Acoustic	3 brothers	Male	Ad hoc
Spielman, Brand, Buischi, & Bretz	2011	Other	9 twin pairs: 6 MZ pairs 3 DZ pairs	4 ♂ pairs 5 ♀ pairs	From the *Twins Institute for Genetics Research* at Montes Claros, Minas Gerais, Brazil

Weirich and Lancia	2011	Perceptual and acoustic	4 twin pairs: 2 MZ pairs 2 DZ paris	All female	Unspecified
Weirich	2011	Perceptual, acoustic and articulatory	7 twin pairs: 4 MZ pairs 3 DZ pairs	2 MZ, 1 DZ ♂ pairs 2 MZ, 2 DZ ♀ pairs	Unspecified
Feiser & Kleber	2012	Perceptual	5 pairs of brothers	Male	Unspecified
Cielo, Agustini, & Finger	2012	Perceptual and acoustic	2 MZ pairs	1 ♂ pair 1 ♀ pair	Unspecified
Leemann, Dellwo, & Kolly	2012	Acoustic	1 MZ pair 7 unrelated speakers	Male	Unspecified
San Segundo	2012	Acoustic	6 MZ pairs 4 DZ pairs	Male	Ad hoc
San Segundo & Gómez-Vilda	2013	Acoustic	7 MZ pairs 5 DZ pairs 4 brother pairs 4 unrelated-speaker pairs	Male	Ad hoc
San Segundo	2014	Perceptual	3 brothers	Male	Ad hoc