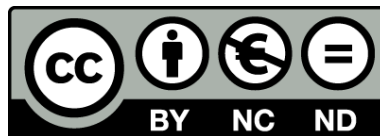




UNIVERSITAT DE
BARCELONA

Voice line-ups: Testing aural-perceptual recognition on native speakers of a foreign language

José Vicente Benavent Chàfer



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial – SenseObraDerivada 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial – SinObraDerivada 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial-NoDerivs 4.0. Spain License.**



UNIVERSITAT DE
BARCELONA

Voice line-ups: Testing aural-perceptual recognition on native speakers of a foreign language

José Vicente Benavent Chàfer

Tesi doctoral presentada per optar al grau de doctor

en el programa de doctorat

Ciència Cognitiva i Llenguatge

Departament de Filologia Catalana i Lingüística General

Universitat de Barcelona

Directors de la tesi:

Dra. Ana Ma. Fernández Planas i Dr. Ramon Cerdà Massó

Tutor:

Dr. Faustino Diéguez Vide

2019

‘In matters of truth and justice, there is no difference between large and small problems,
for issues concerning the treatment of people are all the same’

Albert Einstein

‘Nothing in life is to be feared, it is only to be understood. Now is the time to
understand more, so that we may fear less’

Marie Curie

Acknowledgements

The first word that springs to mind while writing these lines would be *ineffable*, a concept which refers to the inability to convey certain emotions to the fullest extent. With that in mind, acknowledgements are hereby written in the hope that at least a portion of gratitude may reach those involved in this ambitious project. Without further ado, let us dive into it.

First and foremost, I would like to express my deepest gratitude to my supervisors Dr. Ana María Fernández Planas and Dr. Ramon Cerdà Massó for the continuous support, guidance, and counseling received throughout my PhD studies. Besides being grateful for such a warm welcome, I could not help but be enthralled, not only by their undisputed scientific rigor, but by their kindhearted nature and noticeable modesty.

I would like to offer my special thanks to Dr. Carmen Gregori Signes and Dr. Rosana Dolón Herrero, whose intervention through recommendation letters allowed me to pursue higher education, thus delving deeper into the world of forensic linguistics.

I wish to acknowledge the helpful assistance provided by Dr. Wendy Elvira-García, Dr. Paolo Roseano, and Dr. María Jesús Machuca Ayuso through their invaluable insight and feedback offered during the research proposal's presentation. I cannot thank them enough for the time and patience dedicated to review and evaluate this PhD thesis as members of the jury. I am particularly grateful to Dr. Wendy Elvira-García and Dr. Paolo Roseano for their extra help in distributing this thesis' perception surveys amongst their students.

The substitute members Dr. Mireia Farrús and Dr. Juan M. Garrido Almiñana are also worthy of praise for their contribution and availability, should the occasion arise.

Speaking of which, I would like to express my great appreciation to the students who contributed voluntarily to the project by filling in the online perception questionnaires designed to this end. Not to mention the professors and staff members in general, who circulated said perception surveys via department e-mails. The list of people involved in said data-gathering process goes as follows: Dr. Robert Mayr, Dr. Jonathan Morris, Dr.

Mark Jones, Dr. Marco Tamburelli, Dr. Tess Fitzpatrick, Dr. Marta Crosby, Dr. Tim Hall, Dr. Lourdes Melcion, Dr. María Victoria Camacho Taboada, Dr. Juan Pablo Mora, Dr. Carmen Gregori Signes, Dr. Joan Pagès, Dr. Joana Salazar Noguera, Dr. Miguel Ángel Campos Pardillos, and Dr. Linus Jung. Without their contribution, this research would not have been possible. I owe them my greatest gratitude.

A special thanks is devoted to Dr. Robert Mayr, whose thought-provoking remarks on research ethics has inspired and shaped the code of ethics employed in the present thesis. Likewise, I would like to thank Dr. Tess Fitzpatrick for her useful feedback on the perception surveys' wording, which improved greatly the readability thereof.

Additionally, I greatly appreciate Dr. Sheila Queralt's advice on how to stratify the target population for statistical purposes, which was incredibly helpful at the later stages of this research.

I also wanted to thank Dr. Victoria Vázquez Rozas for granting me access to ESLORA corpus. Her kind and diligent manner has guided me through the process of choosing the most appropriate informants, so as to match the profiles needed for the purposes of this piece of work.

I would also like to express my gratitude to Dr. Min Wild, whose enthusiastic comments and always helpful suggestions have motivated me further, both during and after my Erasmus year.

I would like to dedicate a few words to the peers I have met during the many conferences I attended during my PhD studies. I greatly appreciate sharing views on different sub-fields within linguistics, which allows for a change in perspective, while also not losing sight of your main objectives.

Last but not least, I would like to thank my family, whose tolerance towards my ravings about forensic linguistics has been proven to be outside of this realm of existence. Their talks keep me grounded and prevent me from losing my sanity, or rather the remains of it, that is.

I cannot finish this section without thanking those who, despite their absence today, keep lighting up our paths, not from above, but from within.

Index

Acknowledgements	5
Resum	15
Resumen	17
Abstract	19
CHAPTER 1. INTRODUCTION.....	21
1.1. Thesis structure	24
1.2. Objects of study	26
1.3. Objectives	30
1.4. Hypotheses	35
CHAPTER 2. THEORETICAL FOUNDATIONS AND STATE-OF-THE-ART REVIEW	41
2.1. Theoretical frameworks	43
2.1.1. Variationist sociolinguistics	43
2.1.1.1. Sociolinguistic variation	48
2.1.1.2. Acoustic-phonetic variation	51
2.1.2. Forensic linguistics	55
2.1.2.1. Forensic phonetics	58
2.1.3. Voice line-ups/Voice parades	62
2.1.4. The psychology of earwitness identifications	70
2.1.4.1. Memory models	70
2.1.4.2. Memory and psychology	76
2.1.4.3. Memory and voice/face/context	78
2.2. Analytical proposal	82
CHAPTER 3. METHODOLOGY.....	85
3.1. Corpus	88
3.1.1. English corpus	88

3.1.2. Spanish corpus.....	89
3.1.3. Dutch corpus.....	90
3.2. Sample selection criteria.....	91
3.3. Jurors.....	93
3.3.1. British group.....	93
3.3.2. Spanish group.....	94
3.4. Code of ethics.....	94
3.5. Recordings.....	96
3.6. Novelty.....	97
3.7. Analyses.....	98
3.7.1. Perception surveys.....	98
3.7.1.1. Structure and design.....	99
3.7.2. Perception surveys-based analysis.....	106
3.7.3. Acoustic-phonetic analysis.....	121
3.7.3.1. Suprasegmental features.....	124
3.7.3.2. Segmental features.....	126

CHAPTER 4. RESULTS: PERCEPTION SURVEYS-BASED ANALYSIS... 131

4.1. Language familiarity.....	133
4.1.1. Identification.....	135
4.1.1.1. British group.....	135
4.1.1.2. Spanish group.....	142
4.1.1.3. British and Spanish group.....	148
4.1.2. Discrimination.....	154
4.1.2.1. British group.....	155
4.1.2.2. Spanish group.....	157
4.1.2.3. British and Spanish group.....	160
4.1.3. Summary of results.....	163
4.2. Discrimination or identification?.....	165
4.2.1. British group.....	166

4.2.2. Spanish group.....	169
4.2.3. Summary of results.....	171
4.3. Confidence levels.....	171
4.3.1. Identification.....	173
4.3.1.1. British group.....	173
4.3.1.2. Spanish group.....	174
4.3.1.3. British and Spanish group.....	176
4.3.2. Discrimination.....	181
4.3.2.1. British group.....	181
4.3.2.2. Spanish group.....	182
4.3.2.3. British and Spanish group.....	182
4.3.3. Summary of results.....	183
4.4. Age and gender.....	184
4.4.1. Identification.....	186
4.4.1.1. British group.....	186
4.4.1.2. Spanish group.....	189
4.4.1.3. British and Spanish group.....	193
4.4.2. Discrimination.....	197
4.4.2.1. British group.....	198
4.4.2.2. Spanish group.....	198
4.4.2.3. British and Spanish group.....	199
4.4.3. Summary of results.....	200
4.5. Cultural groups and linguistic environment.....	201
4.5.1. Cultural groups.....	202
4.5.1.1. Identification.....	203
4.5.1.2. Discrimination.....	204
4.5.2. Linguistic environment.....	205
4.5.2.1. Identification.....	205
4.5.2.2. Discrimination.....	207
4.5.3. Summary of results.....	208

4.6. Background noises and false alarms.....	209
4.6.1. Identification.....	211
4.6.1.1. British group.....	211
4.6.1.2. Spanish group.....	213
4.6.2. Discrimination.....	215
4.6.2.1. British group.....	215
4.6.2.2. Spanish group.....	216
4.6.3. Summary of results.....	217
4.7. Epilogue: Level of studies.....	218
4.7.1. Identification.....	219
4.7.1.1. British group.....	219
4.7.1.2. Spanish group.....	222
4.7.1.3. British and Spanish group.....	226
4.7.2. Discrimination.....	230
4.7.2.1. British group.....	230
4.7.2.2. Spanish group.....	231
4.7.2.3. British and Spanish group.....	232
4.7.3. Summary of results.....	233
CHAPTER 5. RESULTS: ACOUSTIC-PHONETIC ANALYSIS.....	235
5.1. Intravariability of suspects.....	238
5.1.1. Suprasegmental features.....	240
5.1.1.1. English voice samples.....	241
5.1.1.2. Spanish voice samples.....	242
5.1.1.3. Dutch voice samples.....	245
5.1.2. Segmental features.....	248
5.1.2.1. English voice samples.....	249
5.1.2.2. Spanish voice samples.....	253
5.1.2.3. Dutch voice samples.....	256
5.1.3. Summary of results.....	260

5.2. Intersubjectivity of foil speakers.....	263
5.2.1. Suprasegmental features.....	265
5.2.1.1. English voice samples.....	266
5.2.1.2. Spanish voice samples.....	268
5.2.1.3. Dutch voice samples.....	270
5.2.2. Segmental features.....	272
5.2.2.1. English voice samples.....	273
5.2.2.2. Spanish voice samples.....	285
5.2.2.3. Dutch voice samples.....	290
5.2.3. Summary of results.....	297
5.3. Acoustic-phonetic analysis or jurors' verdict?.....	300
5.3.1. English voice samples.....	302
5.3.2. Spanish voice samples.....	305
5.3.3. Dutch voice samples.....	308
5.3.4. Summary of results.....	311
CHAPTER 6. DISCUSSION.....	315
CHAPTER 7. CONCLUSIONS.....	325
CHAPTER 8. RECOMMENDATIONS FOR FUTURE RESEARCH.....	331
CHAPTER 9. BIBLIOGRAPHIC REFERENCES.....	335
CHAPTER 10. APPENDIXES.....	347
10.1. APPENDIX 1. English informants' (wildlife sound recordists) profiles.....	348
10.2. APPENDIX 2. Spanish corpus (ESLORA) informants' profiles.....	349
10.2.1. APPENDIX 2.1. ESLORA's signed agreement.....	350
10.3. APPENDIX 3. Dutch informants' profiles.....	351
10.4. APPENDIX 4. Voice line-ups: voice samples arrangement.....	352
10.5. APPENDIX 5. Perception surveys' structure and design.....	354
10.6. APPENDIX 6. Praat scripts.....	361
10.7. APPENDIX 7. Between-speaker variation of suprasegmental features in English voice samples.....	374

10.8. APPENDIX 8. Between-speaker variation of suprasegmental features in Spanish voice samples.....	377
10.9. APPENDIX 9. Between-speaker variation of suprasegmental features in Dutch voice samples.....	380
10.10. APPENDIX 10. Significant differences between the English suspect (SUSPECT Simon T. Elliott) and the voices used as distractors, both at the suprasegmental and segmental level.....	384
10.11. APPENDIX 11. Significant differences between the Spanish suspect (SUSPECT M12_020) and the voices used as distractors, both at the suprasegmental and segmental level.....	387
10.12. APPENDIX 12. Significant differences between the Dutch suspect (SUSPECT DVA8-F20L) and the voices used as distractors, both at the suprasegmental and segmental level.....	390
CHAPTER 11. LISTS OF FIGURES AND TABLES.....	393
11.1. List of figures.....	394
11.2. List of tables.....	398

Resum

La present tesi doctoral atén i examina el reconeixement i la percepció de parlants estrangers/nadius mitjançant l'ús de tres rodes de reconeixement auditives formades per un conjunt de dades multilingües (arxius d'àudio amb gravacions en anglès, espanyol i neerlandès). Aquest estudi pretén investigar i aclarir les relacions que existeixen entre el percentatge d'encerts i errors en tasques d'identificació i discriminació de locutors i els factors inherents al parlant, com el perfil sociolingüístic i els paràmetres acústics pertinents. Així mateix, es seleccionaren uns grups de participants (espanyols i britànics) que actuarien com jurats per respondre a les enquestes de percepció confeccionades per a aquesta fi.

Aquest estudi busca aprofundir la comprensió de les circumstàncies reals que envolten els procediments de reconeixement de parlants mitjançant l'ús de les rodes de reconeixement de locutors, tant des del punt de vista de l'oient com des de la del fontetista forense. Per això, la naturalesa de les dades emprades (durada reduïda de gravacions semi-espontànies) contrasta amb les condicions controlades que s'usaven fins ara en experiments d'aquest tipus. Des d'un punt de vista metodològic, aquesta és una de les contribucions principals de la present tesi, a més de ser un dels seus reptes, ja que pretén demostrar la viabilitat de l'anàlisi acústica en la discriminació de parlants malgrat les limitacions donades pel material analitzat.

Es conclou que no es trobaren dissimilituds significatives entre llengües familiars y desconegudes en cap dels dos grups de participants. Així i tot, les llengües apreses exhibiren un comportament impredecible. D'altra banda, l'anàlisi acústica causa una taxa d'error inferior a les produïdes pel jurat en proves d'identificació. No obstant això, aquests participants revelaren menys falses alarmes que l'enfocament de l'anàlisi acústica pel que fa a les tasques de discriminació, amb l'excepció de l'anàlisi de mostres angleses (amb una taxa d'error del 0%).

Tenint en compte l'anterior, es recomana seguir amb aquesta línia de recerca per poder verificar les afirmacions ja esmentades. De fet, els resultats obtinguts presenten limitacions en certa mesura, ja que la interdisciplinarietat del reconeixement de locutors estrangers i nadius suggereix la presència d'influències coexistents fora del nostre control

com els estats psicològics, la memòria i els factors mediambientals. A més, els resultats de les proves estadístiques no han estat tan contundents com es podria esperar. Malgrat això, aquesta tesi ens porta un pas més prop cap a la comprensió de les complexitats inherents a la comparació forense de veus en casos reals mitjançant l'anàlisi de parla semi-espontània, la informació de la qual és probablement més difícil d'analitzar que el que s'enregistra a les mostres de laboratori.

Paraules clau: percepció auditiva, reconeixement de locutors, fonètica forense, variació sociofonètica, percepció de la parla, rodes de reconeixement auditives.

Resumen

La presente tesis doctoral se centra en examinar el reconocimiento y percepción de hablantes extranjeros/nativos a través de tres ruedas de reconocimiento auditivas formadas a partir de un conjunto de datos multilingüe (archivos de audio con grabaciones en inglés, español y neerlandés). Este estudio pretende desentrañar las relaciones existentes entre el porcentaje de aciertos y errores en tareas de identificación y discriminación de locutores y los factores inherentes al hablante, como el perfil sociolingüístico y los parámetros acústicos pertinentes. Para ello, se seleccionaron varios grupos de participantes (españoles y británicos) que actuarían como jurados para responder a las encuestas de percepción confeccionadas para tal fin.

Este estudio aspira a ahondar en la comprensión de las circunstancias reales que rodean los procedimientos de reconocimiento de hablantes a través del uso de las ruedas de reconocimiento de locutores, tanto desde la perspectiva del oyente como desde la del fonetista forense. Por ello, la naturaleza de los datos usados (duración reducida de grabaciones semi-espontáneas) contrasta con las condiciones controladas hasta ahora empleadas en experimentos de este tipo. Desde un punto de vista metodológico, ésta es una de las principales contribuciones de la presente tesis, además de ser uno de sus retos, ya que pretende demostrar la viabilidad del análisis acústico en la discriminación de hablantes pese a las limitaciones dadas por el material analizado.

Se concluye que no se encontraron disimilitudes significativas entre lenguas familiares y desconocidas en ninguno de los grupos de participantes. Aun así, las lenguas aprendidas exhibieron un comportamiento impredecible. Por otro lado, el análisis acústico produjo una tasa de errores inferior a las producidas por el jurado en pruebas de identificación. Sin embargo, dichos participantes revelaron menos falsas alarmas que el enfoque del análisis acústico en cuanto a tareas de discriminación, con la excepción del análisis de muestras inglesas (con una tasa de error del 0%).

En virtud de lo expuesto, se recomienda seguir con esta línea de investigación para verificar dichas afirmaciones. De hecho, los resultados obtenidos presentan limitaciones en cierta medida, puesto que la interdisciplinariedad del reconocimiento de locutores extranjeros y nativos sugiere la presencia de influencias coexistentes fuera de nuestro

control como los estados psicológicos, la memoria y los factores medioambientales. Además, los resultados de las pruebas estadísticas no fueron tan contundentes como se podría esperar. A pesar de ello, esta tesis nos lleva un paso más cerca hacia la comprensión de las complejidades inherentes a la comparación forense de voces en casos reales mediante el análisis de habla semi-espontánea, cuya información es probablemente más difícil de analizar que la encontrada en grabaciones de laboratorio.

Palabras clave: percepción auditiva, reconocimiento de locutores, fonética forense, variación sociofonética, percepción del habla, ruedas de reconocimiento auditivas.

Abstract

The upcoming PhD thesis is aimed at testing foreign/native speaker perception and recognition through the conducting of three voice line-ups using a multilingual data set (English, Spanish, and Dutch audio files). By selecting groups of Spanish and British jurors to answer perception surveys, this study attempts to unravel the correlations of speaker-specific sociolinguistic factors and acoustic parameters impinging upon success/error rates in identification and discrimination tasks.

This study strives to gain a more in-depth understanding of the real-life circumstances at play during speaker recognition procedures through voice line-ups, both from the listeners and the forensic phonetician's side. In this vein, the nature of the data employed (reduced duration of voice samples, semi-spontaneous exchanges) contrasts with the ideal and controlled conditions hitherto used in experiments of this kind. From a methodological point of view, this is one of the main contributions of this work, besides being one of its challenges, since it aims to prove that differentiating speakers by means of acoustic-phonetic analysis is still plausible despite the limitations of the source material.

It is concluded that no significant relationships of dissimilarities are attested between familiar and unfamiliar languages in either group of participants. However, learned languages exhibit a rather unpredictable behaviour. On the other hand, acoustic-phonetic analyses are proven to yield less error rates than the jurors' responses gathered through identification tests. Nevertheless, jurors' scores in discrimination tasks reveal less false alarms than the ones shown in the acoustic-phonetic approach, with the exception of the English voice samples' analysis (0% error rates).

In light of the above, further research is naturally encouraged to verify such claims. These findings are indeed limited to some extent, given the interdisciplinary nature of foreign and native speaker recognition due to the presence of uncontrolled co-existing influences such as psychological states, the memory, and environmental factors. Furthermore, the statistical correlations found were not as statistically sound as one may expect. Despite that, this thesis brings us a step closer to better understand the intricacies of real-life forensic voice comparison through the analysis of semi-spontaneous speech, which is arguably harder to analyse than the samples recorded under laboratory conditions.

Keywords: auditory perception, recognition of speakers, forensic phonetics, sociophonetic variation, speech perception, voice line-ups.

CHAPTER 1

INTRODUCTION

Due to the many factors impinging on speech production and perception, speaker recognition tasks are often deemed problematic and controversial in judicial contexts, all the more considering the legal repercussions that may ensue upon the parties involved. When analysing speech, holistic approaches combining both manual and automated/semi-automated methods are advised so as to reduce the margin of error that the latter may cause. However, the data accessible to trained phoneticians and expert witnesses alike is somewhat limited, since the evidence produced (if any) typically contains footages of short duration with poor audio quality (Fernández Planas 2007: 50), which exacerbates the whole identification procedure. As an alternative route to yield evidence with probative value (either incriminatory or exculpatory), law enforcement officers must rely on the victim/witnesses' auditory memory to perform a speaker recognition test through conducting a set of voice line-ups based on the descriptions provided by the victims or witnesses themselves. This thesis is therefore addressing the need to optimise such tests by extracting the sociolinguistic features affecting speaker perception and recognition.

Chapter 1- Introduction

The underlying rationale lies in the prevention of ‘flawed identifications procedures (...) producing unreliable evidence’ (Broeders & van Amelsvoort 2001: 238) due to the wrong sequencing between visual and voice line-ups, and judicial sentences being based only on evidence gathered from eye/earwitness identification procedures. Furthermore, research on forensic phonetics has proven that a remarkable margin of error may arise in automatic and semi-automatic speaker recognition systems when confronted with adverse acoustic conditions, be it with telephone transmissions and background noises (Alexander et al. 2004), voice disguising through mouth masks, whispers, and raised/lowered pitch (Zhang & Tan 2008), or the fitness of the automatic systems to individualised phonological traits (González-Rodríguez 2014).

Even if researchers (Broeders & van Amelsvoort 1999, 2001; De Jong-Lendle et al. 2015, and Hollien 2012) have notably provided some guidance and valuable insights on the effects of sociolinguistic profiles and certain acoustic conditions upon speaker recognition, there is currently no standardised protocol in force to regulate the application of voice line-ups. Hence the reason for investigating this matter further. Not only this, but the aforementioned guidelines do not seem to consider a scenario whereby the victim/witness and the offender do not share the same sociolinguistic background. Given the unprecedented scale of globalisation in our society nowadays, this hypothetical situation should also be taken into account, and thus this study aims at discerning the success rates at recognition tests deriving from three distinct levels of familiarity with the exposed language, namely *familiar*, *learned*, and *unknown* languages.

Even though similar experiments have been carried out using various types of languages and familiarities (Köster et al. 1995, Köster & Schiller 1997, and Thompson 1987), it is noteworthy to ascertain whether the outcome is influenced by the experimental design adopted. In other words, the proposed voice line-ups’ set up attempts to more closely replicate a realistic scenario whereby exposure time to the stimuli is reduced in contrast with the traditional laboratory-controlled settings (see *chapter 3. Methodology* for a full discussion). Due to the apparent ethical boundaries that an experiment of this kind may rise, the selected Spanish and British jurors contributed to the project by filling in a series of online perception surveys through Google forms, instead of enacting a real-life situation whereby a voice line-up would be required.

1.1. THESIS STRUCTURE

The present thesis is structured around the two main analyses that, although differing in nature and execution, they both accommodate to offer a more detailed understanding of the phenomena at work during the recognition of an unfamiliar voice/speaker. The exploration of the dimensions analysed in each section is defined in chapter 4 (*Results: Perception surveys-based analysis*) and 5 (*Results: Acoustic-phonetic analysis*), whereas chapter 6 offers a discussion of the findings stemming from the aforementioned analytical stages.

In the introduction, however, (chapter 1), the objects of study (1.2.) are specified, as well as the objectives (1.3.) that this work attempts to achieve. Even though the objects of study are more thoroughly discussed (common applications, limitations) in the methodology section (chapter 3), it serves as an early introductory guide on the variables examined. As far as objectives are concerned, they intend to both contextualise the current state of affairs in this area and pinpoint the scope of the present work. This is further reinforced in the following section (1.4.), where specific hypotheses are formulated according to the objectives and needs previously mentioned.

Chapter 2 (*Theoretical foundations and state-of-the-art review*) begins with a brief introduction of the underlying theories around the main theme discussed (Voice line-ups). Point 2.1.1. (*Variationist sociolinguistics*) explores the core principles around language change and variation at two different levels (2.1.1.1. *Sociolinguistic variation*, and 2.1.1.2. *Acoustic-phonetic variation*). It then proceeds with the concerned field of expertise (2.1.2.1. *Forensic Phonetics*), not before putting forward an initial exploration of the relevant sub-fields within forensic linguistics (2.1.2.) besides authorship attribution. In point 2.1.3. (*Voice line-ups/Voice parades*), the possible outcomes of this test are illustrated, and subsequently legal and linguistic-phonetic considerations are discussed. The last section (2.1.4. *The psychology of earwitness identification*) provides a theoretical overview of the relevant memory models (2.1.4.1.), how psychological states shape the subject's memory and how to retrieve it (2.1.4.2. *Memory and psychology*), and the interferences that aspects such as voice, face or context may exert upon the traces of the memory (2.1.4.3. *Memory and voice/face/context*).

Chapter 3 is centered around the methodology that has hitherto tackled the issues and experiments employing voice line-ups. The source material used to compile the list of distractors/suspects in the voice line-up is described in point 3.1. (*Corpus*), according to the three familiarities (3.1.1. *English corpus*, 3.1.2. *Spanish corpus*, and 3.1.3. *Dutch corpus*). The next point (3.2. *Sample selection criteria*) deals with the sociolinguistic criteria adopted upon the selection of informants and extraction of excerpts from the audio files available across the three corpora. 3.3. *Jurors* describe the chosen participants that completed the perception surveys for the purposes of this research, namely the British group (3.3.1.) and the Spanish group (3.3.2.). As this investigation involves more than one cultural group, a code of ethics (3.4.) appears all the more mandatory to safeguard the interests and rights of the aforementioned jurors. Technical details of the recordings obtained from the corpora are specified thereafter (3.5. *Recordings*), whereas a description of the changes made in the methodology employed for the conducting of voice line-ups follows, including all the necessary observations and justifications (3.6. *Novelty*). In the last part (3.7. *Analyses*), some remarks on the data-gathering process are commented (3.7.1. *Perception surveys*) and their structure and design are also unveiled (3.7.1.1.). As a final note, 3.7.2. *Perception surveys-based analysis* mentions the sociolinguistic variables contemplated for the statistical analysis, whilst the second analytical phase (3.7.3. *Acoustic-phonetic analysis*) differentiates between the dimensions of speech being scrutinised: the suprasegmental (3.7.3.1.) and segmental (3.7.3.2.) features, with a subsequent statistical analysis as well.

Chapter 4 (*Results: Perception surveys-based analysis*) refers to the statistical tests and measures employed for hypothesis-testing. Specifically, it resorts to a set of chi-square tests of independence to figure out the correlations between the categorical variables *language* and *type of response* to discern the relationship between familiarity and success rates in identification scores. Secondly, a set of Kendall's tau-b correlation are run to examine whether there is a statistically significant relationship between CL (Confidence Levels) and speaker recognition test scores. Thirdly, a Wilcoxon signed-rank test inspects whether background noises impinge upon the proliferation of false alarms. Lastly, the statistical significance of sociolinguistic predictors (age, gender) is corroborated through a fixed effects model. Since the population sample could not reach significant numbers per each strata considered in the analysis of sociolinguistic predictors, an epilogue (4.1.)

Chapter 1- Introduction

shall cover the inclusion of studies (both *up until BA and MA/PhD*) and reveal what the outcome would be with this variable in the equation.

Chapter 5 (*Results: Acoustic-phonetic analysis*) alludes to acoustic-phonetic parameters, both segmental and suprasegmental features, to discriminate among the constituents of the voice line-up. Measuring the acoustic properties of the informants' voices is crucial to enable potential discrimination of speakers in similar environments with similar intonation patterns, and thus proving the robustness of these methods, much in contrast with the untrained ear. Nevertheless, the method is not exempt of errors, which leads to the subsequent considerations and limitations of this research.

Chapter 6 (*Discussion*) puts together the results from the previous analytical stages, and therefore contrasts the efficiency on identification/discrimination of speakers from the intended lay listener, or targeted participants (albeit trained, however slightly), and the acoustic-phonetic analysis.

Chapter 7 (*Conclusions*) draws on the empirical evidence found in this thesis and proceeds to formulate the challenges and limitations of said data by examining the findings obtained through every single hypothesis.

Chapter 8 (*Recommendations for future research*) elicits a series of suggestions for the application of voice line-ups, both from a theoretical and from a practical perspective. It also discusses the directions for further research, given the insight that the proposed methodological changes may have offered.

1.2. OBJECTS OF STUDY

This research's main objective is to conduct a perceptual study on voice line-ups from a theoretical perspective (to study how aural-perceptual recognition behaves in different scenarios). Thus, this work is divided into two main analytical stages according to its sub-objectives: to observe the sociolinguistic tendencies of the perceptual study itself (*4. Results: Perception surveys-based analysis*) and to analyse the stimuli employed for the voice line-ups with the purpose of explaining the results obtained in the previous analysis

(5. *Results: Acoustic-phonetic analysis*), while also discovering potential variables which may be useful in forensic voice comparisons (those that display high between-speaker variation, but low within-speaker variation).

Consequently, the two main parts of this research are clearly distinguishable in nature. The first half of the analytical phase revolves around the participants, or jurors. At this stage, the experiment is focused on gauging levels of successful identification/discrimination of speakers through the immediate response provided by the researched subjects. Perception surveys are, therefore, a data-gathering method that allows a rapid inspection on both the participants' sociolinguistic profiles and their responses to the stimuli presented.

Conversely, the second half of this analytical procedure consists primarily in extracting the acoustic properties of the recorded voices acting as stimuli in the previous stage. Specifically, it is sought to not only compare the informants' voices in terms of segmental and suprasegmental features, but also to compute the analysis in such a way that such features appear idiosyncratic, and, as a result, speakers' speech become distinctive, too. In other words, this stage is devoted to unravel those acoustic measures that contribute the most in distinguishing the selected informants' voices.

The objects of study proposed hereby are classified into two sub-sets according to the major analytical stages undertaken, the perception surveys-based analysis and the acoustic-phonetic analysis.

PERCEPTION SURVEYS-BASED ANALYSIS

- Gender: male/female.
- Age: 18-22/ Over 22.
- Studies: up until BA/ MA/PhD.
- Cultural group: Spanish/British.
- Sociolinguistic region¹: Spanish Monolingual, Bilingual/British Monolingual, Bilingual.
- Experimental conditions: target-absent with background noises/target-present without background noises.
- Language familiarity: familiar/learned/unknown.
- Confidence score: 1-10.

ACOUSTIC-PHONETIC ANALYSIS

Suprasegmental features

- Global pitch:
 - Mean pitch ($P\bar{x}$).
 - 25%, 50%, and 75% pitch.
- Global sound intensity:
 - Min. intensity ($I\downarrow$).
 - Max. intensity ($I\uparrow$).
 - Mean intensity ($I\bar{x}$).
- Pausing:
 - DurPaus.
 - N_paus/min.
 - Pause_%.
 - N_pause.
 - Speech rate.
 - Articulation rate.
 - ASD (Average Syllable Duration).

¹ Please note that only the environmental influence of a monolingual/bilingual setting is included hereby, rather than alluding to the jurors' individualized linguistic skills, given the impossibility of obtaining a representative sample given the wide array of possible combinations (dominant and secondary languages).

Segmental features
<ul style="list-style-type: none"> • Voiced plosives [b, d, g] and voiceless plosives [k, p, t]: <ul style="list-style-type: none"> • VOT (Voice Onset Time). • Release burst intensity. • Voiceless alveolar sibilant [s], and voiced alveolar sibilant [z] (only in English voice samples): <ul style="list-style-type: none"> • Spectral peak location. • Spectral COG. • Noise duration. • Noise amplitude. • F1-F3².

Table 1. Objects of study.

Table 1 illustrates the variables explored and breaks down each variable into their respective categories. In the first section, it is worthwhile to differentiate between those features inherent to the jurors' sociolinguistic profile characteristics (*gender, age, studies, Cultural group* and *sociolinguistic region*) and those pertaining to the experimental set up of the perception surveys (*experimental conditions, language familiarity, and confidence scores*). For a more detailed explanation on how the variables are handled and treated for each statistical test, consult point 3.7.2. *Perception surveys-based analysis*.

Moving to the acoustic-phonetic variables, suprasegmental features like the average pitch (and its quantiles) or sound intensity (min./max./mean values) are regularly used in forensic speaker recognition tests, and thus are seen as 'traditional' (Rose 2006: 173). Despite not being an officially coined term, *global pitch/intensity* refers to the max./min./mean values (and quantiles) gathered throughout a single voice sample, hence the suprasegmental nature of said variables. As for the pausing measures shown above in table 1 (duration of pauses, number of pauses per minute, percentage of pauses per excerpt, speech rate, articulation rate, and ASD), these can be extracted by using Praat scripts for an automated extraction, as proven by previous research (Cicres 2007; Lindh 2009). Consult section 3.7.3. (*Acoustic-phonetic analysis*) for a more detailed explanation

² Despite the seemingly contradictory addition of F1-F3 to the analysis of fricative consonants, the present thesis assumes Univaso et al.'s (2014) notion of formants (see 3.7.3.2. *Segmental features* for more details).

on the Praat scripts employed in the current thesis.

Since the corpus of voice recordings comprises three languages (Spanish, English, and Dutch) with differing phonemic units, research suggests that plosives (both voiced and voiceless) could be potential discriminating factors in such multilingual data sets at the segmental level, even if their realisations may vary across speakers (Wells 1997). Besides, the use of VOT is linked directly to the previous observation, as these acoustic measures can be interpreted whenever plosives occur in the spectrogram (see point 3.7.3. *Acoustic-phonetic analysis* for a detailed discussion). The voiceless alveolar sibilant [s] and its voiced counterpart [z] (only in English voice samples) are also added to the segmental analysis, since said sounds are found in a wide array of distinct languages (Gordon et al. 2002, Univaso et al. 2014).

1.3. OBJECTIVES

The concept of language change and variation pivots around the notion of *idiolect*, a unique stylistic variation in an individual's language use motivated by external and internal factors, or by a blend of both (Dittmar 1996: 111). This understanding of the language enables and justifies research on variationist sociolinguistics, and thus extends to related sub-fields such as forensic linguistics or forensic phonetics.

When it comes to compile the body of suspects and foils arranged in the voice line-up, it is indicated that such selections should be faithful to the voice description provided by the witnesses and/or victims (Broeders & van Amelsvoort 2001). Nevertheless, said input appears insufficient, since ‘it could be argued that while listeners were adequately equipped to assess whether a speaker was a native or non-native speaker of English, any information beyond this was unreliable’ (Tompkinson & Watt 2018: 35). As it seems that eliciting explicit identification cues is rendered ineffective in the judicial context, it appears imperative to reassess those procedures which provide probative evidence through the usage of implicit responses linked to the witness or victim’s traces of auditory memory, as it is the application of voice line-ups. Even then, it remains crucial to survey the self-perceived confidence level at the speaker recognition tasks and figure out if/how it correlates with the actual score of the perception test, since this is a key element

considered in real voice line-ups (Nolan 2001).

While this alternative fares relatively well in comparison with an investigative interview, it is not without its pitfalls. As a matter of fact, evoking these memories does not entail a fail-safe standardised process across individuals and, oftentimes, some traces cannot even be retrieved (see 2.1.4. *The psychology of eyewitness identifications*). What is more, accuracy and accessibility of memories appears conditioned by stress levels, emotional intensity, degree of involvement (Manzanero & Recio 2012: 21), and time elapsed (Acosta 2009: 2), amongst other uncontrolled factors. What can be controlled, however, are the so-called experimental conditions which are put forward in this case to compare the aural-perceptual abilities of two cultural groups, namely the Spanish and British group of jurors. In spite of not being able to investigate the effects of long-term memory due to constraints on the experimental design, it seeks to spot the extant relationships between success rates in speaker recognition tasks and the degree of familiarity with the language exposed, as well as the influence of noise disturbances in the recordings.

Even though the presentation of evidence in court apropos forensic comparison of voices has experienced a shift from merely a binary decision (the voice belongs to the suspect/it doesn't belong to the suspect) to a likelihood ratio (French and Harrison 2007), and thus restricting the probative value thereof, wrongful convictions still occur in cases where physical evidence (DNA, objects) is absent (Broeders & van Amelsvoort 2001: 238). Such scenarios are labeled as 'false positives', the conviction of an innocent citizen as a result of a wrong identification by the witness or victim (Braun 2016 :63). In the context of this research, it is sought to figure out whether the presence or absence of background noises hinders speaker recognition, and whether it is an influential factor in the emergence of false positives.

The very notion of *voice quality* seems ambiguous in itself when it comes to its description and use in the legal arena, and it could be argued that those elements playing a role in the perceptual domain are multidimensional (Gil & San Segundo 2014: 156). Following this line of thought, it appears hardly conceivable to pinpoint isolated variables and their contribution to the hearers' perception capabilities, since a plethora of conditions occur simultaneously while hearers encode the information around them. Particularly, it is remarkably challenging to account for those features beyond our control, such as

individualised patterns of alertness, natural predispositions towards noise, time of the day, etc. Even so, identification does not necessarily imply an individualised isolation of each co-occurring variable appearing during a verbal exchange, especially with phonological traits (Cerdà-Massó 2008: 62), since it would be virtually impossible to account for the complex network of correlates originated from observable variables and let alone those that remain unnoticed to our senses. By leaning towards the aspects which allow some degree of control, this thesis ventures into determining whether language familiarity, cultural group, sociolinguistic environment, background noises, age, level of studies, and gender impact on the hearer's proficiency at identifying/discriminating unfamiliar voices.

For the purposes of this research, the interface between sociolinguistic variation and phonemic features defines our sub-field of study: sociophonetics (Foulkes 2005). Similar to the notion of idiolect, two of the pivotal concepts follow in this sub-field which relate to intra-speaker and inter-speaker variability, or *within-speaker* and *between-speaker* variation in Rose's terms, accordingly (2002: 10). In its application to forensic sciences, and more concretely to the conducting of voice line-ups, intra-speaker variation (the contrasting of voice samples of the same speaker under different conditions) is expected to be low, as opposed to inter-speaker variability (comparing two voice samples of different speakers under similar acoustic conditions), which tends to score high in discrimination tests. Much in line with researchers who have verified such assumptions (Leemann et al. 2014, Rose 2002, and Stevens 1971), this study also attempts to gauge levels of intra/inter-speaker variability, only that it sets out to do so with three differentiated sets of linguistic data (i.e. Spanish, English, and Dutch), and thus disparate degrees of variability are expected to emerge across the three data sets. Cicres' (2007) study sets the precedent inasmuch as the same statistical procedure to calculate the discrimination power of acoustic variables studied is adopted, except that the acoustic units of measurement are changed in this case to best suit the characteristics of the resulting multilingual data set.

Additionally, an extra acoustic variable (intonation contour) is added to the acoustic-phonetic analysis so as to challenge the current notions on intra/inter-speaker variability. In this sense, the segmental and suprasegmental features of choice are tested against a condition whereby same speakers (the informants acting as suspects) display utterances with differing intonation contours (rising and falling intonations), whereas the second

Chapter 1- Introduction

condition calls into question whether the aforementioned acoustic units are able to distinguish between various speakers (informants acting as distractors) using similar intonation patterns.

It should be noted that this work's purview does not contemplate a descriptive account on the acoustic properties of the variables examined as the focal point, but rather employs them for an instrumental use (discover which phonetic units offer the best discriminatory power in a multilingual set up). Undoubtedly, findings could hint at implications of the empirical data at the conceptual level, but that is not, as a matter of fact, the main objective pursued.

<u>Nº</u>	<u>OBJECTIVES</u>
1	Compare two macrocultural regions' (Spanish and British) efficiency in foreign and native speaker recognition.
2	Investigate whether there is any correlation between success rates in speaker recognition tasks and the jurors' degree of confidence in carrying them out.
3	Examine the relationships of familiarity/unfamiliarity of the exposed language with speaker recognition (both identification and discrimination of speakers).
4	Figure out to what extent background noises hinder speaker recognition, and how it relates to familiar/unfamiliar languages.
5	Test whether the presence/absence of a background noises favours the production of false positives (i.e. the wrong identification of an innocent speaker/distractor as the culprit).
6	Discover the positive/negative correlations between jurors' profiles (gender, age, studies, etc.) and their success rates in speaker recognition tasks.
7	Prove that the intravariability of the suspects' voice samples with differing intonation contour (e.g. rising and falling intonation) and uncontrolled segmental phenomena is not statistically significant.
8	Prove that the intervariability of the foil speakers' voice with similar intonation patterns (rising intonation) and uncontrolled segmental phenomena is statistically significant.
9	Juxtapose the findings stemming from jurors' responses with the ones obtained through acoustic-phonetic analyses.

Table 2. Summary of the planned objectives.

As discussed before, the main objective of this research is centered around voice recognition and perception in a voice line-up setting. This is in turn divided into two specific sub-objectives: to observe how aural-perceptual recognition's tendencies are shaped by sociolinguistic features and/or experimental conditions (objectives 1-6), and to account for varying jurors' responses on the basis of the acoustic properties of the voice samples exposed to them (objective 9). Additionally, the latter analytical stage (acoustic-phonetic analysis) strives to unearth the most efficient segmental and suprasegmental

features in forensic voice comparison (objectives 7-8). Said objectives can be consulted in table 2 above.

The variables analysed within the suprasegmental domain refer to measures of pitch and intensity, which are commonly employed in phonetics and phonology research (Rose 2006: 173), besides including measures of pausing, which tend to be fruitful in cross-linguistic comparisons (Lindh 2009: 188). Regarding segmental units, the first group of variables revolves around voiced and voiceless plosives' VOT (Voice-Onset Time) values, whereas the second half of variables refer to measures concerned with the frication noise produced by [s] (and also incorporating [z] when analysing English voice samples).

Besides the objectives already mentioned in table 2, this thesis attempts to assess the feasibility of voice line-ups as methods employed in legal settings and warn, whenever appropriate, about the limitations thereof. Given the fact that technical factors might interfere with the resolution of voice line-ups, it should also not be neglected the human factor involved in the procedure. In fact, delving into the surrounding circumstances around a criminal act is essential to define the boundaries and limitations of witnesses and victims as reliable sources of information, lest it result in biased testimonies leading to miscarriages of justice.

1.4. HYPOTHESES

As a starting point, it is worth mentioning the tenets, or rather the assumptions governing this study:

- Speaker identification is feasible despite the adjustment of the acoustic properties of the voice, either conscious or unconscious, inherent to physiological traits or speech acts, among other factors (Cerdà-Massó 2011: 34).
- Notwithstanding the technological advances in forensic phonetics, errors still occur in automatic and semi-automated methods designed for speaker recognition practices (González-Rodríguez 2014).
- Unlike DNA evidence, the notion of *voiceprint* cannot guarantee the same degree of certainty in the courtroom (Rose 2002).

Chapter 1- Introduction

In the first two observations, human perception is advocated at the expense of automatic and semi-automated speaker recognition systems. As stated in *1.3. Objectives*, every single co-occurring variable in speech cannot be isolated and be subsequently subject to analysis. For this reason, the likelihood of humans to perceive, even if it is at the subconscious level, the parts of the whole is greater than the work of an automated algorithm. This is not to say that such software should be neglected entirely, but quite the contrary: As Nolan (2001: 9) puts it, naïve speaker recognition is not as effective as the one stemming from manual acoustic analyses, which allows for a higher degree of control and organisation with respect to the variables at hand.

Sociolinguistic and sociophonetic variation is expected irrespective of social strata, and just as the Labovian paradigm claims, language change and variation is systematic at all levels (Weinreich et al. 1968: 188). Be it because of distinctive phonological phenomena linked to regiolects, physical build of the vocal cords, or even the active accommodation of the speaker to the context, speech perception could end up being influenced by the multitude of ways in which production of speech may be altered. In spite of this, the stimuli chosen for the jurors in this experiment are balanced in terms of situational context, sociolinguistic background, gender, and age. On the jurors' side, it is investigated whether a reduced familiarity with the language exposed hinders identification accuracy (*Hypothesis 1*), and whether discrimination of speakers is more feasible than identification tasks (*Hypothesis 2*) (Hollien 2002, Köster & Schiller 1995; 1997, and Thompson 1987).

The following hypotheses (3-6) deal with the potential sociolinguistic predictors or experimental conditions that facilitate or hinder speaker recognition. Hypothesis 3 directs its attention at the reliability of earwitnesses, as it seeks to determine whether their self-perceived confidence level matches the actual score of the test. Inferential statistics follow to calculate the influence of sociolinguistic predictors like age, gender (*Hypothesis 4*), cultural groups and linguistic environments (*Hypothesis 5*) upon the success rates found in identification/discrimination tests, with a later addition of *studies* in the epilogue (4.7). Due to its relevance in the legal arena, hypothesis 6 examines whether the addition of background noises is correlated with a higher production of false alarms (Alexander et al. 2004). Ultimately, the first section of hypotheses' purpose is dedicated to further refine

the already existing guidelines on voice line-ups (Broeders & van Amelsvoort 1999; 2001, De Jong-Lendle et al. 2015, and Hollien 2012).

PERCEPTION SURVEYS-BASED ANALYSIS

Hypothesis 1

‘Aural-perceptual recognition is enhanced as the familiarity of the juror with the language exposed also increases’.

Hypothesis 2

‘Jurors are more proficient in discrimination tests than in identification tasks’.

Hypothesis 3

‘A heightened self-perceived confidence level at speaker recognition tasks has a positive effect on the voice line-up’s outcome’.

Hypothesis 4

‘The efficiency at speaker recognition is conditioned by age and gender’.

Hypothesis 5

‘Speaker recognition capabilities are not influenced by cultural groups (Spanish or British) nor by linguistic environment (monolingual or bilingual)’.

Hypothesis 6

‘Background noises hinder voice recognition, thus resulting in a higher frequency of false alarms’.

ACOUSTIC-PHONETIC ANALYSIS

Hypothesis 7

‘Intravariability of the suspects’ voice samples with differing intonation contour (rising and falling intonation) and uncontrolled segmental phenomena is not statistically significant’.

Hypothesis 8

‘Intervariability of the foil speakers’ voice samples with similar intonation patterns (rising intonation) and uncontrolled segmental phenomena is statistically significant’.

Hypothesis 9

‘Foreign and native speaker recognition using acoustic-phonetic analysis is more accurate than the lay listener’s (jurors) judgement’.

Table 3. Hypotheses considered for each analytical stage.

The third assumption commented above touches on the misconceptions around the notion of *voiceprint*. While constructing a profile centered around the parameters of the voice may be revealing, it does not guarantee a successful speaker recognition ‘in the way we believe fingerprints allow us to discriminate every individual’ (Nolan 2001: 2). To this end, research has been searching the most prominent acoustic variables (Cicres 2007) to increase the accuracy and ultimately the validity of the evidence provided by acoustic-phonetic analyses. As shown in table 3 above, hypotheses 7-9 are designed to meet this end. More specifically, hypotheses 7 (within-speaker variability with differing intonation patterns) and 8 (between-speaker variability with similar intonation contour) aim to test the robustness of the segmental and suprasegmental features selected in conditions far from the ideal controlled laboratory settings (semi-spontaneous recordings whose production of segmental units remains uncontrolled), and therefore representing a more realistic scenario.

The last hypothesis (9) compares the ability of the trained expert (and the use of software devoted to acoustic-phonetic analysis like Praat) against the intuition of the average lay

Chapter 1- Introduction

listener (the participants that completed the perception surveys acting as jurors). Previous research (Nolan 2001: 9) has suggested that the experts' knowledge on the subject matter and the available resources grants them a significant advantage in this regard. Nevertheless, it seems a matter of interest to discover if this tendency still prevails when more than one language is used, above all considering the ensuing degrees of familiarity with the linguistic input (familiar, learned, and unknown).

CHAPTER 2

THEORETICAL FOUNDATIONS AND STATE-OF-THE-ART REVIEW

The upcoming chapter is devoted to explore the theoretical frameworks (2.1.) concerned with the application of voice line-ups, ranging from the most general levels (2.1.1. *Variationist sociolinguistics*) to the deepest layers of analysis (2.1.2.1. *Forensic phonetics*). In the former, relevant definitions and concepts related to sociolinguistic variation (2.1.1.1.) shall be elaborated to account for between-speakers and within-speaker variation in either written or audio format (2.1.1.2. *Acoustic-phonetic variation*). Moving to forensic linguistics (2.1.2.), this point referring to the sub-discipline in applied linguistics conceived as the interface between the language and the law shall discuss its scope and its contribution to legal disputes, with a more detailed focus on disputed audio material thereafter (2.1.2.1. *Forensic phonetics*). Additionally, this literature review does not only include linguistic-related theories, but legal, psychological, and cognitive aspects are also drawn in the following sections (2.1.3. *Voice line-ups/Voice parades*, and 2.1.4 *The psychology of earwitness identifications*) to create a more comprehensive picture as for the complex interplay of features involved in forensic speaker comparison, foreign/native speaker perception and recognition, and more specifically, the application

of voice line-ups. Upon gathering various theoretical and technical perspectives on the subject matter, point 2.2. (*Analytical proposal*) puts forward the procedure to be followed in the present thesis and sets the boundaries thereof.

2.1. THEORETICAL FRAMEWORKS

The following section is divided into four sub-sections which account for the main supporting theoretical basis and principles consulted for the making of this thesis, namely the relevant theories on sociolinguistic and acoustic-phonetic variation, and thus providing explanations and layers of said linguistic changes, as well as dealing with the sub-fields of interest (forensic linguistics and forensic phonetics), their scope, previous work, and overall regulations on the subject matter. It is also discussed the current state of the art approaches concerned with the pragmatic application of voice line-ups or parades, both from the technical side and the psychological dimensions impinging on the procedure.

2.1.1. Variationist sociolinguistics

In its broader sense, sociolinguistics is the science in charge of unveiling the on-going relationships between sociological phenomena and linguistic features and, where possible, establishes causal links between language usage and society (Coulmas 1997: 2). Up until the late 1960s, linguists treated internal (word order, sentence and word stress, etc) and external factors (age, gender, education level, etc) as separate entities in language change and variation. Nevertheless, the so-called Labovian paradigm demonstrated not only that language change is systematic at all levels (ranging from the generic language standard to the unique individual usage or idiolect), but also that both internal and external factors motivate and shape linguistic variation (Weinreich et al. 1968: 188). Despite being interrelated, Labov (1982: 52) claims that this correlation does not necessarily entail causation between them.

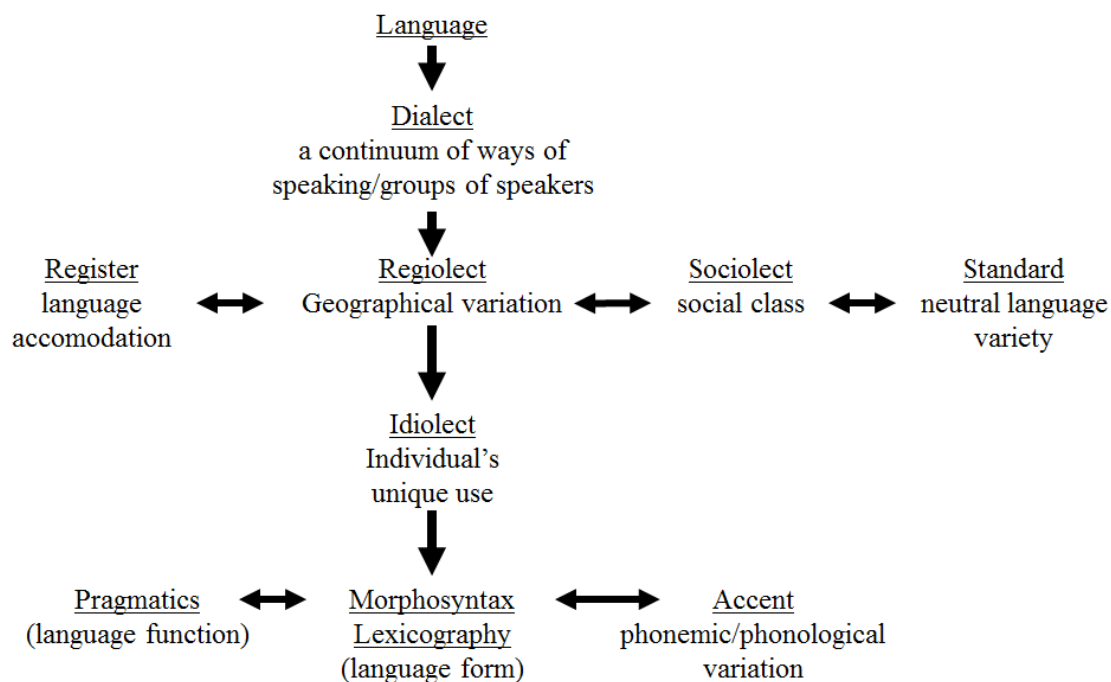


Figure 1. Basic levels of stylistic variation. Adapted from Schultz (2007: 54).

The most fundamental notion of language change and variation conceives it as a compound of various dialects which, in turn, are formed by a set of individualised idiolects. Even though *dialect* could also be referred to as an equivalent term for *colloquial* language, it is employed here as a means to allude to the linguistic continuum of variation illustrated in figure 1. That being said, geographical boundaries are inherently enclosing distinctive regional varieties of the language (regiolect), even though physical spaces do not necessarily draw a clear-cut line around linguistic expressions. What is more, the term isogloss (or heterogloss) does pinpoint characteristics across different geographical areas, but standardisation of such traits could be rendered troublesome given the influence of migratory movements (globalisation) and the language contact established with neighbouring languages (Chambers 2004: 370). Hence, dialectologists differentiate between physical geographical spaces (Euclidean space), social spaces and perceived spaces (Britain 2004: 604) This may be linked with the phenomenon in which speakers lose their sense of belonging, or loss of cultural identity, just as their authentic linguistic code is now shared and extended beyond their place of origin. Leaving the issue of language and cultural identity aside, Patrick (2004) asserts that, even with the inclusion of heterogeneous groups of migrants within the speech community, sociolinguistic patterns can still be evaluated within larger units, provided that such idiosyncrasies are ‘systematically related’ (p. 592).

Next up comes the notion of sociolect, which is broadly seen as the diversity of language usage across social classes, but also includes ethnicities, culture, and economic status. It would appear that the stratification of society is reflected through its use of language, as well as enacting power relationships through affiliating or disaffiliating to the respective collective identity by using distinctive sociolinguistic markers (in-group or out-group references). Besides, Fought (2004) also notes distinctive linguistic attitudes deriving from *borrowing* linguistic traits inherent to the in-group speech community, ranging from cultural integration to promoting ethnical/cultural stereotypes (455). In short, this linguistic meticulousness the speaker adopts in order to align with certain members of society could involve a voluntary change of register.

As Schilling-estes (2004) explains, registers are ‘highly ritualised, routinised varieties, often associated with performance or artistic display of some kind’ (375). Regardless of the nature of the exchange, a defining feature for *register* is the speaker’s *role* in the interaction. This set of projected expectations on the speaker create a behavioural pattern to be followed, and all the parties involved act (and speak) accordingly. The range of situational contexts range from informal, colloquial exchanges (conversational tone) to more formal or even technical varieties of the language, as a job interview or a highly specialised speech would entail.

The notion of standard variety of the language is typically the one which is devoted to public spaces and fostered in educational environments. Albeit not necessarily ungrammatical³, non-standard varieties of the languages may appear counter-intuitive by the already established language standards, and sharing knowledge is further exacerbated since the average interlocutor would lack the referential material that the non-standard language user is employing, just as Henry (2004) notes when reporting the difficulties encountered when working on Belfast English.

One may be tempted to view the standard variety as the ruling force of dialects, but it may, in reality, influence and be influenced in turn by the other three components within the same layer (register, regiolect, and sociolect). In fact, Nolan (2001: 9) conceives how

³ That is, as long as linguistic properties are faithfully reflecting the mannerisms of said variety (lexis, phonetic realisations, etc.), even if it appears ungrammatical from the standardised language variety’s point of view.

sociolects interact with urban dialects (often implying a change in register), and how it relates to the superposed standard from the regional variety.

All the aforementioned levels of variation are embedded within the idiolect (a distinctive use of the language for every individual). As for what linguistic properties are being shaped according to the surrounding circumstances at play, three main elements are foregrounded: those related to language function, form, and production. In the realm of pragmatics, discursive purposes are motivated by illocutionary acts for specific purposes in definite contexts (persuade, command, etc.), even though definitions and applications of discourse variation still require further work through replicable studies to find out the extent to which known and unknown variables affect such practices (Macaulay 2004:298). As for the language form, morphosyntax and lexicography are also shaped by the already mentioned levels of language variation. In spite of being partially conditioned by external physiological features (age, vocal apparatus, illness, etc.), the speaker's accent can also be modulated at will to best suit the interaction's requirements by virtue of the organs involved in speech production's 'plasticity' (Nolan 2001: 2).

As Bell (1984: 167) investigates, there are at least three possible emerging responses when a speaker is confronted with a differing stylistic variety of the language:

- A shift of the speaker's style is adopted in order to accommodate to the addressee's personal characteristics.
- A stylistic change and accommodation ensue after examining the addressee's speech.
- The speaker seeks other linguistic variables (lexicographic/prosodic, etc.) in the addressee's discourse and shifts his/her style accordingly.

Schilling-estes (2004) argues that, even if it appears that the individual's idiolect is influenced by external societal forces, speakers make use of style shifts actively as a means to renegotiate interpersonal relationships, and re-direct the immediate situational context to its desired outcome (p. 378). A natural predisposition and willingness towards linguistic accommodation from the speaker's side is assumed in the three modalities of style shifts above, although contentious linguistic attitudes for competing varieties of the language may arise in less ideal situations. In such confrontations, there may be more

than one factor at play, like linguistic prestige and linguistic self-loathing, contextual constraints (register, communicative events), psychological factors (exhaustion), and even projected misconceptions about the other speaker's culture, ethnicity, or country of origin.

At any rate, Schilling-estes (2004) determines that such stylistic variations are associated more commonly to intra-speaker variation (the individual's linguistic adjustment) than to inter-speaker variation (the variation of linguistic formulae across speakers) (p. 375). The former is also named as *within-speaker variation*, whereas the latter is commonly referred to as *between-speaker variation* by Rose (2002: 10). The overall trend in sociolinguistics research is to account for the variations of the same speaker across distinct contexts (assuming that differing registers and linguistic varieties would emerge as a consequence), and to pinpoint distinctive linguistic features of individual speakers who experience similar communicative events. As a matter of fact, this dynamic notion of idiolect remains a focal point in the present study, since its implied individual distinctiveness enables the discrimination and identification of speakers, and, as previously noted, said stylistic variation permeates through all linguistic areas, namely at the syntactical, semantical, discursive, pragmatic, and phonetic levels (Dittmar 1996: 111).

One more aspect to consider here is the maintenance, disposal, or adaptation of linguistic forms over time. As Hazen (2011) discusses, language change is not only conceptualised through synchronic (language description represented at a given point in time) lenses, but is extended to diachronic (investigating how the language evolves across time periods) approaches, since language change itself is not represented by cumulative linguistic traits alone, but also by the development thereof (p. 33). This matter is especially crucial in forensic phonetics studies, where the validation of the subject's *voiceprint* in court cases is called into question, as acoustic properties of the voice are ever-changing due to physiological processes (see 2.1.2.1. *Forensic phonetics* for full details). While writing skills are not as susceptible to aging as phonetic traits, they may also experience variation over time due to the incorporation of new lexis and mannerisms through the speakers' lifetime, be it through personal or educational experiences.

Once the main foundations of language change and variation are clarified, the next points proceed to elaborate further on the research around sociolinguistic and acoustic-phonetic variation.

2.1.1.1. Sociolinguistic variation

As discussed in the previous point, linguistic formulae are bound to be shaped by sociolinguistic parameters, regardless of the channel of communication being used (tactile, visual, auditory, etc.). Language change is also motivated by internal processes (language structure) and social dynamics, thus yielding a complex pair of correlations between gradual changes on linguistic structures influencing their social significance, and vice versa (Weinreich et al. 1968: 186). Due to its inevitable societal nature, Hickey (2014) notes a series of noteworthy issues in sociolinguistics:

Variable	Issue
Social networks	It has a greater impact on language use and change than social class (hardly measurable).
Dissociation	Conscious or unconscious stylistic shift to differentiate yourself from others (normally low-prestige language users accommodate to a more socially prestigious variety).
Gender differences	How recognisable and consistent these differences are and whether they are localised or not (is standardisation of this variable even possible?).
Solidarity and politeness	The notion of <i>face</i> (not losing one's social status through maintaining politeness formulae).
Second language acquisition	How do social factors impinge or facilitate the quality and scope of second language acquisition?
Education	To what extent are governmental policies on education influencing on the linguistic content learned in school premises? How do children socialise in this context?

Table 4. Sociolinguistic issues worth addressing according to Hickey (2014: 20).

From the list shown in table 4 above, only more issues arise when considering the grand scheme of sociolinguistic interactions. Perhaps the use of social media could render fruitful analyses on social networks, although this would neglect spontaneous face-to-face interactions. As for dissociation and gender differences, they might be useful in some specified contexts with representative samples, but they would hint at tendencies of the target population rather than reporting on a widespread social phenomenon due to evident differences across geographical regions, social strata, etc. Not only physical constraints, but also the notion of culture and intercultural communication hamper the accountability of the last three variables: Representations of politeness and solidarity vary significantly in form (ranging from indirect to more direct linguistic formulae) and functionality (not losing face, personal or other gains alike), let alone the underlying social protocols established in different countries (differing degrees of perceived impoliteness through the breach of unspoken rules constrained by culture and situational context). As for second language acquisition and education, the cultural component is again crucial in explaining language change and variation, with uncontrollable features such as parental control and influence, the individualised child's social interactions (amongst peers, learner-teacher, speaker-strangers, etc.), or even sociolinguistic environments (multicultural societies or more homogeneous communities) with their own sociocultural sub-strata.

It is certainly not unthinkable to examine the aforementioned factors in isolation and extract potential linguistic patterns but, as a matter of fact, accounting for the complex net of interrelations would be an arduous task, to say the least, since there may be virtually as many possible combinations as the number of existing speakers and conceivable communicative contexts. An example of selecting isolated variables for scrutiny may be Dong's (2014) study on gender differences in utterance-choosing. In said piece of work, it is asserted that females' linguistic repertoire is typically focused on harmonising with non-confrontational strategies, whilst males' display greater assertivity and certainty in their speech (94-95). Some observations could be very well defined as overall gender-driven linguistic tendencies concerning the levels of social involvement. Nevertheless, these considerations should be interpreted with caution, inasmuch as they are constrained by the communicative context, culture, country, degrees of social awareness, gender roles, and time periods, among other external factors.

In order to measure the degree of linguistic change or stability over time, there are essentially two approaches to survey the target population: the apparent-time (synchronic) and real-time (diachronic) approach. The former investigates the use of certain linguistic forms at a given point in time used by different generations, whereas the latter surveys the same intended social stratus at different points in time. Both perspectives are not without their inconvenients, which are addressed hereby:

- **Apparent time.** It has been broadly discussed about the representativeness of data coming from apparent-time research, since said studies cannot be deemed as diachronic descriptions of language changes. This is to say that they would be indicative of linguistic variations occurring due to generational gaps, rather than evaluating how the language has evolved over time. (Bailey 2004: 314). This involves more problematic aspects of the examined sociolect and its influences, such as previous experiences, time periods, the evolution of linguistic policies thereof, etc.
- **Real time.** Albeit time-consuming, researchers may either use some pre-existing data for the sake of comparison or start the process of surveying an entire speech community (and being resurveyed after consecutive time periods) (Bailey 2004: 325). Again, this would be the ideal approach to measure language variation, if it was not for the uncertainty caused by the engaging of external human respondents in a lengthy process: Informants may commit initially but decline after some time, or they could even have undergone drastic linguistic changes due to personal experiences (rather than due to a natural language phenomenon).

Time is also a crucial matter in determining the preservation or death of minority languages when co-existing with other varieties, besides the institutional and social processes of normalisation and normativisation. Competing linguistic varieties may result in language attrition, leading to either minority languages being differentiated or assimilated progressively (Sankoff 2004: 656). Despite considering overall linguistic structures (internal factors) and large societal forces, the seemingly minor role of individuals (and small social groups) should not be overlooked, since such members of the speech community do contribute to the proliferation of discursive practices and, in the end, to the relevant linguistic outcomes emerging from language contact (*ibid*: 659).

It should also not be neglected that, with the emergence of new technologies and social media, new forms of communicating are being proliferated as a result (like the use of emojis, internet memes, etc.). The appearance of new mediums poses bigger challenges, as it adds up more layers interacting with how the language is developing and, oftentimes, gives rise to neologisms related to this domain (abbreviations, shortened forms, etc.), which creates new dimensions for multimodal analysis. Such shifts in interactional patterns and lexis may imply, as suggested above, changes in the society at large, which renders a more complex picture in terms of language variation and societal change. Insofar as feasible, the use of linguistic cues for pragmatic purposes (as in forensic linguistics/phonetics research) may incorporate such emerging features for analytical purposes, thus yielding richer reports on the suspect's idiolect, as it would be the case in authorship attribution or in forensic speaker comparison/recognition.

2.1.1.2. Acoustic-phonetic variation

As exposed in the previous section, acoustic-phonetic variation is also susceptible to changes in time (diachronic and synchronic approaches, aging-related physiological processes, or even internal structural linguistic changes), sociolinguistic factors (gender, socioeconomic/sociocultural class, etc.), external factors (environmental conditions), and even the pragmatics behind speech acts (various intonation patterns oriented towards certain goals).

Regarding gender differences, females generally have smaller vocal cords than males, which renders higher rates of cord vibration and thus higher F0 values than those encountered in their male counterparts (Rose 2002: 37). Even though the ambivalent concept of *voice quality* may be alluding to *phonation types* (any of the various kinds of activities in the glottis aiming at producing sound) or the perceptual features that permeate through the speaker's speech (Gil & San Segundo 2014: 156), it encompasses both laryngeal characteristics such as creaky, breathy, harsh, and pressed voice; and supralaryngeal settings such as (open and closed) nasality, lip rounding and jaw lowering (Jessen 2010: 388). It is assumed that such organic traits alongside socio-cultural factors may end up influencing average F0 and formant structures (Jessen 2010: 391).

Kiparsky (2015: 8) poses some questions apropos the challenges found in historical phonology when accounting for acoustic-phonetic variation:

- **The constrains problem:** Is the change naturally induced, or is it a random effect? Can it be predicted?
- **The regularity problem:** Are patterns of change sporadic?
- **The implementation problem:** Does the change occur progressively or is it abrupt?

As a response to these concerns, the existing literature on acoustic-phonetic variation addresses the topic:

Linguistic change begins when the generalisation of particular alternation in a given subgroup of the speech community assumes direction and takes on the character of orderly differentiation.

The generalisation of linguistic change throughout linguistic structure is neither uniform nor instantaneous; it involves the covariation of associated changes over substantial periods of time, and is reflected in the diffusion of isoglosses over areas of geographical space.

Linguistic change is transmitted within the community as a whole; it is not confined to discrete steps within the family. Whatever discontinuities are found in linguistic change are the products of specific discontinuities within the community, rather than inevitable products of the generational gap between parent and child (Weinreich et al. 1968: 187-188).

As implied previously, language change is not induced, but rather appears to replicate further the divergences occurring in the speech community from a bottom-up approach. As the study on prosodic variability in uptalk instantiations by Warren (2017) investigates, New Zealand English (NZE) speakers adopt innovative (overlapping phonetic realisations with the near vowel, [iə]) and conservative (open phonetic realisation, [eə]) strategies in statements and questions, and that the interpretations of said prosodic changes reckons on the speakers' sociophonetic cues. With this in mind, this

English variety may develop differently from others' because, as a matter of fact, the pragmatics of uptalk may also change across linguistic sub-types.

As for the regularity problem, it is worth mentioning the Great Vowel Shift as a starting point of discussion:

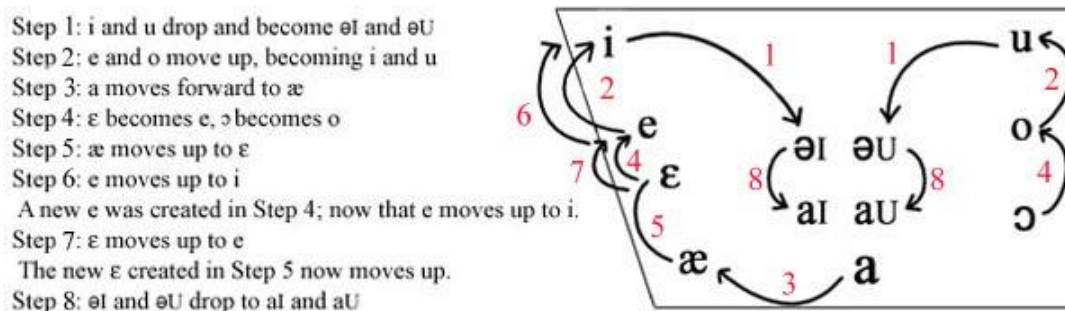


Figure 2. Great Vowel Shift explained (Menzer 2000).

The figure shown above represents the systematic phonetic progression undergone between the long vowels in Middle English (14th century) until evolving into the more familiar sounds of Early Modern English (18th century) and Present-Day English (Giancarlo 2001: 27). The eight steps imply the rising of such vowels higher up in their place of articulation inside the mouth, whereas those which were already at the top became diphthongs. The understanding of this paradigm has been subject to debate, thus questioning its origins, scope, accuracy of the phonological phenomena observed, and its own interpretation (*ibid*: 28). What is certain, however, is that these changes were not applied immediately nor happened in an ordered sequence, but rather were assimilated as generations went by, just as nowadays older generations retain linguistic traits while youngsters adopt new ways of speaking. Leaving this paradigm aside, future changes at the phonetic (pronunciation) and phonological (sound system structure) level may not be as notorious as in the Great Vowel Shift's case. What is more, English extended use as a *lingua franca* could even shape its own conceptualisation in the future and lead to more diverse varieties of vowel realisations, in a similar way to the coexisting phonemic realisations exhibited between 1500-1900 shown in figure 3 below:

me.		1500	1600	1700	1800	1900	
/a:/	æ:	—————					name
	ɛ:	=====					
	e:	-----					
/ɛ:/	ɛ:	—————					sea
	e:	-----					
	i:	=====					
/e:/	i:	=====					see
/i:/	i:	=====					time
	ɪ	-----					
/ɔ:/	ɔ:	—————					boat
	o:	=====					
/o:/	u:	=====					boot
/u:/	ʊ	=====					foul
	u	-----					
/ai/	ui/æi/ɛi	-----					way
	ɛ:	=====					
	e:	-----					
/au/	ɒ	-----					cause
	ɔ:	=====					
/ɔu/	o:	=====					blow

Figure 3. Conservative and progressive pronunciations according to Dobson (1968). Transitional periods are signalled with segmented lines.

As for the last issue commented, it seems that language change does go through all social strata and speech communities, although a multiplicity of scenarios could be reproduced when examining stratified and clearly differentiated clusters for the migrant population. In language contact situations, learners of a second language are reportedly prone to phonological interference or transfer from the native dominant language (Sankoff 2004: 644), but even then, there is the possibility that a speaker may be ‘bidialectal’ when they possess a proficient active command of two distinct languages (Nolan 2001: 9), which would complicate the matter for speaker recognition tests.

In reporting the variability and subtle appreciations on speakers’ acoustic-phonetic diversity, Dumas (1990) work on a criminal law context could prove the innocence of a defendant based on stress patterns (*police*: [pə’lis] and [’pəlis]), and the appearance of divergences in segmental pronunciations (either as fully represented diphthongs [ai] or through lengthened monophthongs [a:] in words like *l*), which were indicative of two distinct American regiolects (the standard variety as opposed to southern rural area’s speech) (p. 345-347). The issue around the truthfulness of said phonemic realisations could be raised in legal disputes, given the suspects’ reluctance to collaborate with law

enforcement officers. However, checking for consistency in their style-shifting tendencies is crucial, since the typical sound produced is unlikely to be replaced by the intended manipulated phonetic unit in each and every instantiation. In this regard, measuring the range of F1-F4 values, or Long-Term Formant Distribution (LTF), was first proposed by Nolan & Grigoras (2005), and is claimed to be effective in detecting speaking styles across individuals.

2.1.2. Forensic linguistics

In its very nature, the overarching term of applied linguistics calls for the use of linguistic knowledge to solve real life issues. Forensic linguistics emerges as a sub-discipline stemming from applied linguistics which is centered on the resolution of linguistic disputes in the legal arena, even though the ensuing findings could be of particular interest for historians, sociologists, and psychologists alike (Olsson 2008: 3). Even if the interdisciplinary field of forensic linguistics is often described as the interface between language and the law, it also encompasses sub-themes such as judicial system dynamics, the rhetoric of the expert witness, issues on accessibility and comprehensibility of legal documents, trademark disputes, authorship attribution, speaker profiling, the investigation of copyright infringement cases, detection of plagiarism, and analysing warning labels, among others (Johnson & Coulthard 2010: 7).

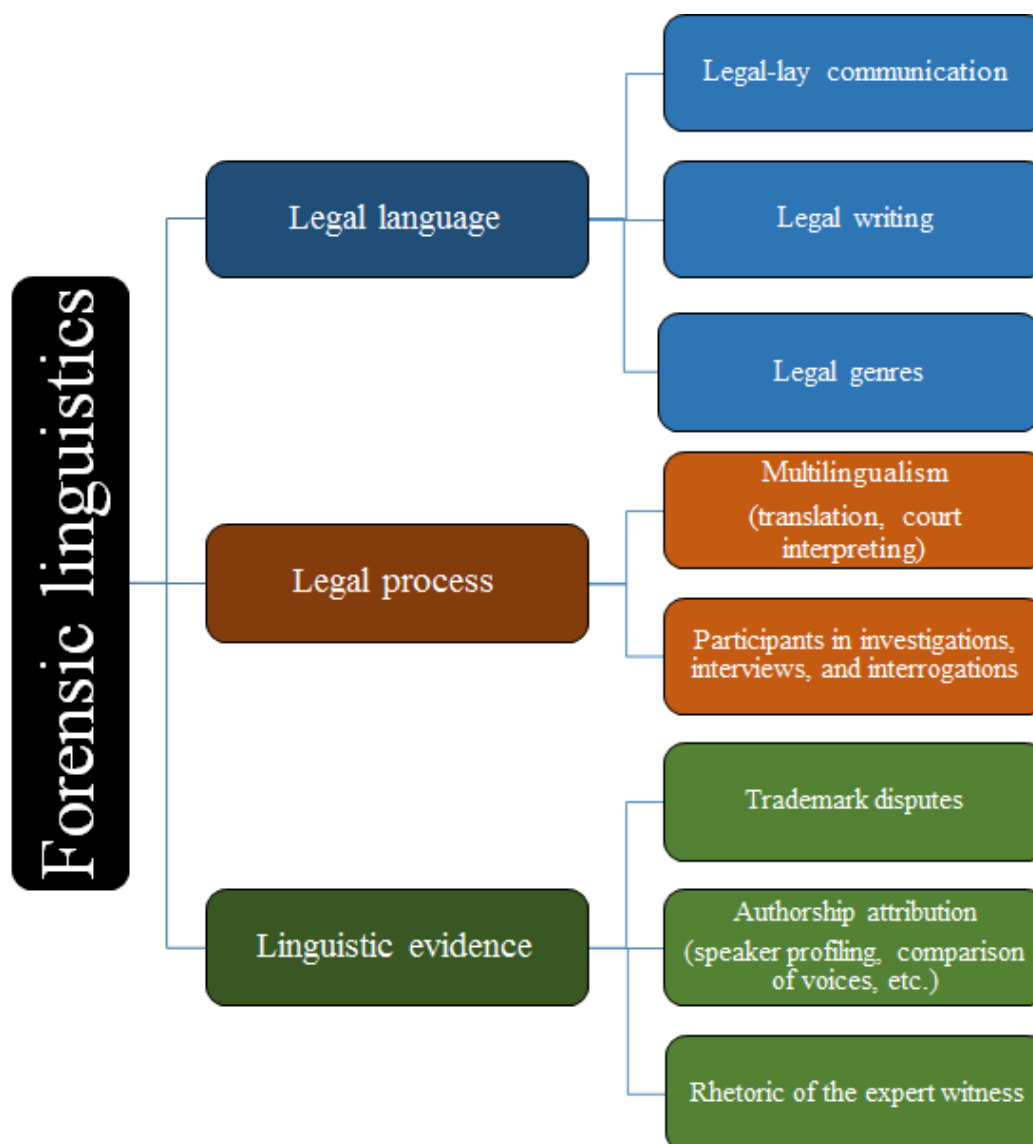


Figure 4. Main areas explored in forensic linguistics.

As broad as it may appear, current practices in Spain are oriented mainly towards authorship attribution (detection of plagiarism, forensic comparison of voices, speaker profiling), description of judicial and legal genres, and court interpreting/translation in multilingual settings (Jiménez et al. 2014: 35).

When analysing legal language, the main objective is to breach the gap between lay people and the verbose, and at times ambiguous nature of legalese. This stage overlaps with studies focused on the legal process (court interpreting/translation) and linguistic evidence (reports from the expert witness aimed at reassessing the comprehensibility of legal actors/documents). In this sub-section, linguistic empowerment is foregrounded, especially to groups who cannot self-represent adequately in judicial contexts due to

extenuating circumstances (intellectually handicapped individuals, children, rape victims, etc.) (Johnson & Coulthard 2010: 5). Legal writing, as the category implies, is directed at assessing the comprehensibility of written documents such as reports, legislations, policies, etc. The hardships at this point lie in the necessity to convey the legal implications of said documents to the fullest extent without unintentionally leaving out part of their meaning, lest the average citizen should not be affected (Johnson & Coulthard 2010: 3). Lastly, examining the intricacies of legal genres attempts to tease out the underlying expectations of each sub-genre by unravelling interactional dynamics, the use of persuasion, the meaning-making devices employed, and evaluate how such features affect the outcome of the trial/judicial process.

As for the investigation of legal processes, the inadequacy of some unqualified interpreters in court calls for an awareness in this regard, given the unjust treatment non-native speakers may receive from a wrongly conveyed message, and this is further extended to governmental agencies and immigration departments (Johnson & Coulthard 2010: 3). The next point covers the procedures at place that regulate interactions between law enforcement officers and participants in investigative interviews or interrogations. With this in mind, it is sought to guarantee and safeguard the rights of those involved in the process through looking at linguistic manipulation techniques such as coercive behaviours or leading narratives/questioning. Again, special care is devoted to vulnerable witnesses/victims (Johnson & Coulthard 2010: 5)

Moving to the realm of linguistic evidence, trademark disputes may argue about certain written, semiotic, or even phonological features (Cerdà-Massó 2008) in the advertising of products to claim the originality of one registered brand against the alleged imitator, leading to possible litigations or avoided altogether with a monetary settlement agreement. When it comes to authorship attribution, recognition tasks can be performed by using either written or auditive material from the suspect to identify. It typically entails the comparison of undisputed texts/voice samples with disputed documents or recordings, with the purpose of deciphering whether the disputed material corresponds in any way to the undisputed speaker. As for speaker profiling, LADO (Language Analysis for Determination of Origin) stands as an example of how ‘descriptive methods’ (Johnson & Coulthard 2010: 5) stemming from dialectological studies may help to recognise the ethnic origin of speakers, as it is applied to asylum seekers in this particular case. As a

final note in this regard, the rhetoric of the expert witness does not only contain its verbatim testimony on the account per se, but encompasses all types of created evidence with probative value, such as the inspection of warning labels on consumer products or even the assessment of semiotic landscapes (traffic signs), even if such contributions lack the absolute certainty that the judicial context expects (Johnson & Coulthard 2010: 5).

As hinted in point 2.1.1.1. (*Sociolinguistic variation*), the emergence of new mediums of communication opens up larger opportunities for verbal/written exchanges, as well as for fraudulent or illegal activities under the veil of anonymity. However, the judicial system updates its legislation and investigational procedures in compliance with technological advances. In this vein, SMS text messages are seen as admissible evidence in court, with the retrieval of additional information from the sender upon forensic analysis, like country of origin, mobile phone company, and the exact date and time of the message's reception (Hellín 2014: 363). This brand-new area of expertise may pose further issues on manipulation of evidence, therefore questioning the validity thereof. However, cell phones shall not be the only medium considered, but legislations may also incorporate specific regulations on IM (Instant Messaging) platforms or social media networking sites in due time.

2.1.2.1. Forensic phonetics

Forensic phonetics is commonly conceived as the use of linguistic alongside phonetic/phonological knowledge to solve legal issues, therefore it also encompasses tasks such as speaker profiling, forensic comparison of speakers (voice recognition/author attribution), authentication of recordings, and phonetic transcriptions (Johnson & Coulthard 2010: 381-394). In the speaker recognition section, the linguist hinges on segmental and suprasegmental features to calculate rates of inter/intra-speaker variation. As one of the main notions applied to this end, the fundamental frequency (F0) refers to the vibrational force involved in the speech production process which reportedly signals '*speaker-specific behaviour[s]*' (Loakes 2006: 205).

Concerning disputed audio material, research on forensic phonetics has proven that a remarkable margin of error may arise in automatic and semi-automatic speaker

recognition systems when confronted with adverse acoustic conditions, be it with telephone transmissions and background noises (Alexander et al. 2004), voice disguising through mouth masks, whispers, and raised/lowered pitch (Zhang & Tan 2008), or the fitness of the automatic systems to individualised phonological traits (González-Rodríguez 2014). In advocating the role of human aural perception in speaker recognition tasks, this thesis attempts to unearth the preeminent parameters impinging upon foreign/native speech perception. Thus, Spanish and English jurors have been exposed to audio files in familiar, learned, and unfamiliar languages through the conducting of a voice line-up and a perception survey to gauge their identification and discrimination capabilities as well as their confidence in doing so.

Apart from the surveying data-gathering method, the present study is structured according to the two analytical procedures employed for the processing of said data, where the perception surveys-based analysis measures the degrees of correlation and variance across sociolinguistic variables (age, gender, education level, addition of stimuli, etc.), and the acoustic analysis inspects and assesses the variability of suprasegmental and segmental features among the voice samples presented at the voice line-up. Additionally, the perception surveys-based analysis heeds to the aforementioned sociolinguistic variables to control for differences in the jurors' success rates at identification/discrimination tasks, whereas the latter analytical stage gauges idiosyncrasies (segmental and suprasegmental phenomena) embedded to voice samples which may explain jurors' performance at identification/discrimination tasks. However, forensic phonetics research in this domain is not exempted from errors, as the literature explores hereafter.

Braun (1995: 11-14) warns that F0 can be altered through physiological (age, illness), technical (the setup of recording devices), and psychological factors (excitedness, background noises, time of the day, etc.) which could end up in distorted audio material. Not only this, but Nolan (1983: 11) encourages adopting precautionary measures in acoustic analysis by extracting phonetic units that present high inter-speaker variability and low intra-speaker variability, are easy to extract, are recurrent and stable throughout the voice sample, and are resistant against masking and voice disguising. Besides that, Rose (2002: 53) reminds linguists acting as expert witnesses in court that their reports are considered in conjunction with other pieces of evidence and that it is the jurors/judge's

duty to assess them altogether. In this respect, reports should not present absolutist claims but rather a probabilistic calculation of whether the suspect's voice sample belongs to the defendant's voice: the so-called 'likelihood ratio' (French and Harrison 2007). An example of cautious reporting of results in speaker identification tests by the IAFPA (International Association of Forensic Phonetics and Acoustics) is illustrated in table 5 below:

	Most positive
5	'I personally feel <i>quite satisfied</i> that X is the author'.
4	'It is in my view <i>very likely</i> that X is the author'.
3	'It is in my view <i>likely</i> that X is the author'.
2	'It is in my view <i>fairly likely</i> that X is the author'.
1	'It is in my view <i>rather more likely than not</i> that X is the author'.
0	'It is not possible to express an opinion'.
-1	'It is in my view <i>rather more likely than not</i> that X is not the author'.
-2	'It is in my view <i>fairly likely</i> that X is not the author'.
-3	'It is in my view <i>likely</i> that X is not the author'.
-4	'It is in my view <i>very likely</i> that X is not the author'.
-5	'I personally feel <i>quite satisfied</i> that X is the not author'.
	Most negative

Table 5. Scale of opinions in reporting authorship identification results (Coulthard 2010: 480).

As for the Bayesian statistical model itself, it refers to the probability (P) of the evidence displayed (E) to comply with H1 (parameters are consistently showing a reassuring degree of similarity between the suspect and the voice sample), as opposed to the probability of resulting evidence supporting H2 (acoustic parameters are found in the large population) (Nolan 2001: 14). It can be summarised with the following formula:

$$\frac{P(E | H1)}{P(E | H2)}$$

Figure 5. Bayesian statistical model on likelihood ratios according to Nolan (2001: 14).

In other words, identification of speakers does not only entail comparing the appointed voice samples for examination but refer to the overall probability of that particular voice

to be matched with the suspects', given the acoustic-phonetic idiosyncrasies of the general population, as figure 5 explains. Furthermore, the misleading concept of *voiceprint* appears to be perceived just as *fingerprint* in terms of reliability, and thus the general public seems prone to internalise this unfounded belief (Nolan 2001: 2). Said wrong assumption is not only linked to the validity of the proof, but also generalisations about the efficiency of practitioners or linguistic phoneticians acting as expert witnesses may lead to the '*infallibility trap*' (Rose 2002: 53), a cognitive bias whereby the lay person is inclined to believe in foolproof mechanisms revealing the *truth*, given the expertise of the authorities operating the system.

On the other hand, specialised research groups pertaining to the *Guardia Civil* and *Policía Científica* in Spain make use of the resources available to refine automatic speaker recognition methods, granting enough reliability to send the resulting piece of evidence to court (Morrison 2009: 304). The Polytechnic University of Madrid also developed a biometric-based automatic speaker recognition software (SIBMATI⁴), whilst other programmes alike such as BATVOX, BS3 (Biometric Speaker Spotting System), ASIS (Automatic Speaker Identification System), or FASR (Forensic Automatic Speaker Recognition Program) are widely employed in other laboratories across the country and beyond (Jiménez et al. 2014: 37).

Despite being a fairly stable unit of phonetic measurement, Prieto (2002: 28) asserts that the high and low threshold of frequencies may be intentionally modified for pragmatic applications, although this fact turns F0 into a variable endowed with discriminatory potential for such individualised linguistic behaviours. As a feasible variable in speaker recognition tests, the '*long-term fundamental frequency distribution*' (Baldwin & French 1990: 45) assists in discrimination tasks by displaying 1-3 minutes of naturally occurring speech, which is further corroborated by other studies (Loakes 2006, Baldwin & French 1990: 47). Nevertheless, it should be reminded that recordings made in real-life situations may not display such ideal length, aside from rendering dubious audio quality (Fernández Planas 2007: 50) due to environmental noises, overlapping of speech, etc. This brings in turn added technical difficulties for proper speaker recognition, besides from individualised tendencies of hearing/perceiving phonetic traits of familiar speech

⁴ Sistema de Identificación Biométrica Multimodal Aplicado a las Tecnologías de la Información (Multimodal Biometric Identification System Applied to Information Technologies).

communities, rather than identifying the speaker behind the voice displayed (similarity of voices in contrast with actual identification).

In this regard, the following sections shall consider influential factors on speaker perception and recognition in voice line-ups, and how (if) these can be circumvented through procedural regulations and technical adjustments.

2.1.3. Voice line-ups/Voice parades

With the purpose of aiding the judicial system, speaker recognition tests such as voice line-ups/parades have been largely used and accepted (with occasional controversies) in cases where the victim/witness could not maintain a visual contact with the suspect/offender, but perceived his/her voice (San Segundo 2014). Just as in the case of speaker comparison reports carried out by expert witnesses, the validation of the proof presented in court depends on the quality, duration, and nature of the recordings obtained (Delgado 2014: 210). Among the approaches adopted in perceptual recognition tests, there is the technical approach and the naïve approach, where the former is performed by trained experts in forensic phonetics whereas the latter is carried out by non-experts (Künzel 1995: 74). The procedure consists of '*putting together* an audio tape which contains recordings of a number of speakers, including the suspect' (Butcher 1996: 97). After the witness/victim is instructed on the procedure, he/she is requested to identify the suspect, although discriminating the mock speakers is equally decisive, for it prevents miscarriages of justice from occurring. Besides technical/naïve pairs, an additional distinction is made between identifying familiar and unfamiliar voices/speakers:

In some cases, such as robbery or rape, the witness may also be the victim. In these situations, it makes a difference, both scientifically and legally, whether or not the witness knew the offender from before the crime. In the former situation, the required task for the witness is called familiar-speaker identification and in the latter unfamiliar-speaker identification. Familiar-speaker identification enters the evidential process in the form of a regular witness statement. whether such a witness statement is reliable or whether adverse conditions occurred that cast doubt on its reliability (Jessen 2010: 379).

In this experiment, only unfamiliar speaker recognition is considered, and thus a separate set of conditions and safety precautions are adopted in this domain, as opposed to familiar-speaker recognition. As a preventive measure, research on the optimal conditions for the conducting of voice line-ups reveals that recordings should last no more than 45 seconds, that they should reproduce a text independent from the words uttered by the suspect, and that the voice should reproduce the emotion expressed at the time of the incident (Rodríguez Bravo et al. 2003: 33). In the present adopted method, however, the optimal time is restrained to 20 seconds for each recording since a longer exposition could increase the jurors' fatigue, thus yielding unreliable data. As for the informants who provided the recorded data set, their conversations range from semi-directed interviews to spontaneous exchanges, which fulfils the second condition. In this thesis, the picking of foils and suspects also complies with the emotion-based criteria since all recordings were conducted in the same room under the same conditions, which equates every speaker inasmuch as they undergo similar emotional states.

For clarification purposes, the possible outcomes of a voice line-up are graphically represented in figure 6 below:

		Listener's Decision	
		yes	no
Correct Answer	yes	Hit	Miss
	no	False Alarm	Correct Rejection

Figure 6. Possible outcomes of a speaker identification experiment (Braun 2016 :63).

From the possibilities depicted in Figure 5, they can be further classified according to the type of task at hand, namely an identification or a discrimination task:

- Identification:
 - **Hit** (*true positive*). The juror/witness correctly identifies the intended suspect in a voice line-up.
 - **False Alarm** (*false positive*). The juror/witness wrongly identifies a foil speaker as the suspect.

- Discrimination:
 - **Correct Rejection** (*true negative*). The juror/witness correctly acknowledges that the suspect to identify is absent from the voice line-up.
 - **Miss** (*false negative*). The juror/witness wrongly assumes that the suspect to identify is absent from the voice line-up even though the intended suspect is present (Braun 2016 :63).

From the list above, false positives pose the most severe threat to the judicial system and public safety. Indeed, as Broeders & van Amelsvoort (2001) point out, a 90% of recent wrongly convicted American cases involved dubious incriminatory evidence resulting from biased voice line-ups (p. 238). Flawed procedures may reckon on a wrong sequencing of identification tests and/or the burden of proof placed upon the witness/juror/victim. For instance, the first aspect could render skewed results if the criminal possesses a distinctive physical trait which is coincidentally shared with a foil suspect (i.e. tattoos, visual impairment, limping, etc), which would facilitate a biased response in the visual line-up. Furthermore, the burden of proof may exacerbate this since the witness' predisposition to identify a suspect is high due to the expectations of his/her situational context, even if the actual criminal is absent from the visual/voice line-up. In the light of these considerations, some researchers (Broeders & van Amelsvoort 1999, 2001; De Jong-Lendle et al. 2015, and Hollien 2012) do provide guidelines to ensure fairness in voice line-ups. A summary of the main conditions is listed hereby:

- The voice line-up size consists of 5-7 voices (including the suspect's).
- All voice samples last for less than 20 seconds.
- All foil speakers & suspects share traits such as age, gender, type of dialect/accent, education level, culture, and socio-economic background.
- All foils/suspects' voices are recorded under the same acoustic conditions and circumstances within their group (e.g. employing the same hardware and space).
- At least, 1 foil's voice is similar enough to the suspect's.
- At least, 1 foil's voice is dissimilar enough to the suspect's.
- No disruptive behaviours are shown in the recordings (i.e. alcohol intoxication, extreme tiredness, etc.).
- Familiarity of the voice: To avoid secondary identification (i.e. that the voice was heard beforehand by the witness), the voice samples are not familiar to the witness/juror.
- There are no family relationships attested within the foils/suspects' group.
- The jurors briefing includes clear instructions on the procedure to follow.
- The jurors are warned about the possibility of an absent suspect in the line-up.

As for the possible approaches on speaker identifications, voice line-ups can be arranged in three distinct manners, according to Hollien (2002):

- **The simultaneous single-trial line-up.** This procedure consists of placing an audio-tape containing a set of distractors and the suspect to identify in random order. The witness/victim is expected to hear the whole line-up once only without interruptions (p. 62).
- **The multiple-trial, simultaneous line-up.** The same situation is reproduced in this procedure, only that this time it allows the earwitness to hear the recorded samples multiple times (the recordings' order changes randomly for each attempt) (*ibid*).

- **The sequential method.**

Figure 5.1

A graphic display of the 'sequential' approach to earwitness identification. A is the witness, B = a tape recorder, C = the tape recordings, D is the administrator, E = the videocam, F = the TV monitors and G shows observers.

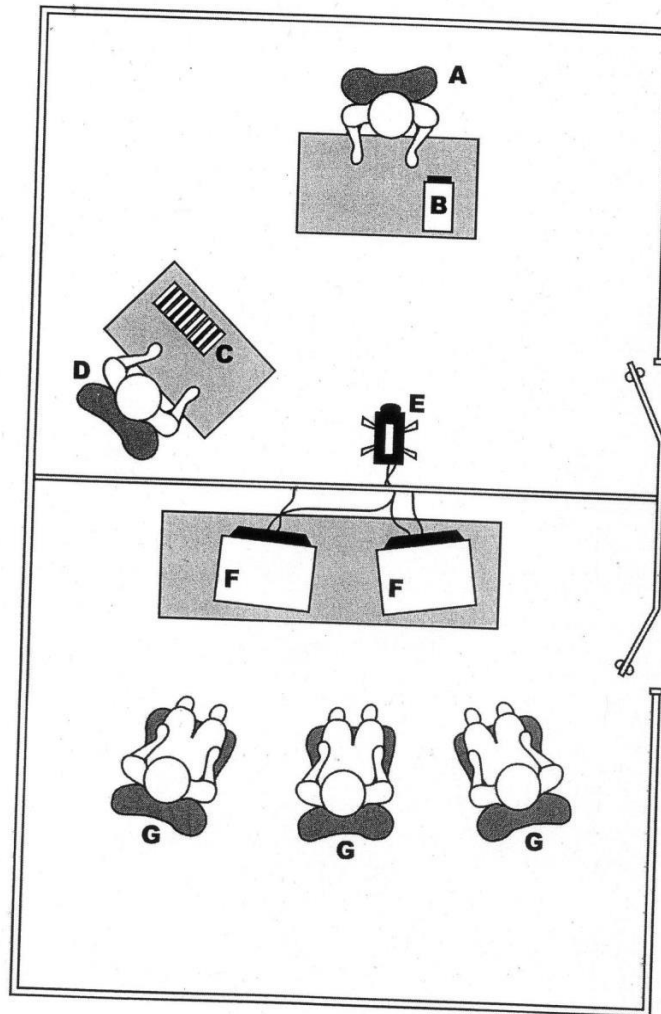


Figure 7. The sequential method according to Hollien (2002: 63).

As figure 7 shows, the sequential method is equipped with two distinct rooms. In the first one, the witness is placed before the administrator, whom is asked about which tape should be listened and in which order. Thus, the earwitness is able to control the sequencing of audio tapes, and whether he/she wishes to replay some of them. At the end of the procedure, the witness is not obliged to make a decision. In the other room, interested parties in the lawsuit may be observing the witness' behaviour, whose session has been video-taped, should it be needed for future consultation (*ibid*: 63).

Amongst the three possible options, the sequential method appears to be the most favourable for the earwitness, but the question of its validity lingers and is still being

debated whether if such degree of control over the evidence may condition the individual to issue a rushed judgement (as opposed to the single-trial line-up, which would be more representative of a real-case scenario). As for the selected method for the purposes of the present experimental voice line-up, the sequential method is chosen and adjusted to the intricacies of online environments, as opposed to the traditional face-to-face line-up (see 3.7.1. *Perception surveys* for full details).

As for delaying the whole procedure, an overall tendency to erroneously identify a suspect arises after a short span of time (24h), even when none is present. Regarding female hearers, Manzanero & Barón (2017) report that the rate of false alarms increases drastically when the target speaker to identify is also female, and adds that such voice line-ups experiments are conducted in optimal conditions (hearing, no distracting factors, easily recognisable voices, and the fact that the participants were aware of the aim of the study, which was to identify an unfamiliar voice) (p. 59). On the other hand, Papcun et al. (1989) detects a general trend on an enhanced accuracy in the subject's identification as his/her certainty increases. Conversely, Kerstholt et al. (2004) argue that the earwitnesses' confidence scores do not influence the accuracy in their judgement, and thus advise caution in taking the earwitness' confidence as an effective evidence validation criterion. The variable CL (*Confidence Level*) is considered in this study to test the aforementioned assumptions.

As for the fairness of the line-up from a forensic phonetics perspective, research suggests to plot the acoustic differences of the foils to the accused in order to avoid a potential dissimilarity bias, which could redirect the jurors' judgement towards a specific voice. Even so, 'it is impossible to state how similar is similar enough' (Yarmey 1995: 808).

Nevertheless, it is reminded that the procedures leading to voice recognition are controlled and calibrated by humans, and so the forensic phonetician is bound to deliberately make subjective choices. However, Hollien et al. (2016: 18) warn about the limitations on emotional, health states and cognitive biases on the selection process, on how the standards for the practise are met, and on the potential external variables that influence the process, which could contaminate the outcome of the report/experiment. Besides the unwitting influence of the researchers themselves, the auditory line-up can

also be spoiled at the moment when the suspect is being recorded in a real-life case scenario, as Jessen explains:

Another form of uncooperative behaviour occurs when a suspect agrees to a recording, but then tries to disguise his voice in an apparent or subtle way. In such a case, the expert has to decide from a forensic-phonetic perspective whether this evidence can still be used. The methodology used in speaker comparisons involves a wide variety of both auditory and acoustic parameters (Jessen 2010: 379).

In Tomkinson & Watt (2018), it is warned about the restrictions of untrained listeners to accurately describe an unfamiliar heard voice, making it even harder for unknown accents. Broeders and van Amelsvoort (2001) also indicate that the selection of foils should be faithful to the voice description given by the witness. In the light of recent research, however, this referential material may not be a reliable source to set up a voice line-up, as the witness' inaccurate description may discard possible suspects while including less likely subjects. Tomkinson & Watt (2018) confirm this and note that voice description systems should be optimised further to elicit reliable information of evidential value from the earwitnesses (p. 21).

The particularity of this research is that it deals with multilingual data extracted from various sources, namely Spanish, English, and Dutch (only for the Spanish jurors' test) data, which in theory would reduce the success rates in identification/discrimination tests for jurors unacquainted with the exposed language. The issues on perception of a foreign dialect or language are highlighted by Hollien (2002) and Goldstein et al. (1981) who state that regional dialects do not compromise the voice line-up nor the overall performance of the lay listener, with the exception of a lower success rate in identifying Chinese alongside white/black American speakers, which is further aggravated when reducing the hearing sample from one full-sentence to just one word. With this exception, it seems that accented and unaccented foreign speech does not seem to vary significantly from hearing a native speaker, as far as the short-term memory's use in a laboratory setting is concerned. However, the outcome could be exacerbated in situations where the activation of the long-term memory is required (such as a voice line-up), just as Yarmey (1995) notes, ethnic groups are more prone to perceive speakers outside their community as more 'homogeneous or similar to each other' (p. 799), whilst discerning clear

idiosyncrasies of those in-group individuals. This phenomenon is further tested by Mullennix et al. (2011) under the label of ‘voice typicality’, who acknowledge its influence on voice recognition ‘just as face typicality affects facial recognition’ (p. 33). On another note, familiar identification does not only involve the jurors’ familiarity with the language or accent, but also includes the relationship with the actual voice and speaker, as commented previously. In such cases, ‘decreases in identification accuracy will correlate with reductions in familiarity’ (Hollien 2002: 32).

Furthermore, Thompson (1987); Köster et al. (1995); and Köster & Schiller (1997) report detrimental effects on speaker recognition and identification when listeners are confronted with foreign speech, by exploring the relationships between German, Spanish, Chinese, and English input. Additionally, Hollien (2002) contrasts the efficiency of lay listeners against the trained phonetician's ear, the latter displaying better overall results than the former. The fact that lay hearers are compelled to make a 'swift judgment' (2002: 37) in the voice line-up implies that the experts have the upperhand in this respect, since they are equipped with the required means (materials, expertise, and time) to carry out a thorough analysis, and thus leading to a more accurate verdict than the lay listener's immediate response.

Yet again, a follow-up research of the previous studies (Schiller et al. 1997) used German speakers, monolingual English speakers, and English speakers with some knowledge of German to determine whether the removal of linguistic information (i.e. telephone transmission) along the familiarity/unfamiliarity of the language spoken plays a role in speaker recognition. Results proved that the speakers' sensitivity was not affected, and varying responses were not accounted as statistically significant. In fact, Broeders et al. assert that identifying a foreign suspect's voice seems productive as long as the suspect's background matches that of the foils' (2002: 111). Concerning the present thesis, there are no drawbacks with the extracted recordings in this respect, since they were collected from the same database and thus interviewees (both foils & suspects) share features such as first language, socio-economic background, and education levels.

In more extreme cases, Sebastian et al. (2013) establish that recognition of identical monozygotic twins rendered perceptually distinguishable voices despite its difficulties, whereas Loakes’ research (2003) on the speech patterns of identical and non-identical

twins hinted at an easier discrimination of their speech by means of acoustic parameters, but a worsened auditory recognition. Given the uncertainties arising when conducting acoustic-phonetic analyses of the voice (whether some properties are intentionally modified by the individual or conditioned by his/her vocal tract), it is advised to proceed with caution in judicial contexts (Gil & San Segundo 2014: 156). As a matter of fact, standards in proof validation and their influence upon the resolution of litigations have changed from absolutist claims to the incorporation of a likelihood ratio (French and Harrison 2007), and even a set of conditions or rather guidelines have been proposed to regulate the expert witness' reports (Willis 2009). Similarly, evidence originating from voice line-ups is considered and its influence restricted, whenever necessary, whilst prioritising the value of biometric data (DNA-related evidence).

2.1.4. The psychology of earwitness identifications

The psychological implications of earwitness identifications, albeit not central to the main theme of the present thesis, do have an undeniable impact on speaker recognition. In the following section, a review of the relevant research concerning the topic shall indicate and restrict, whenever necessary, the potential findings originating from a linguistic-based study, such as the one hereby proposed. It shall refer to the current models of the memory with a theoretical approach, leading up to the intersection between psychological states and the memory, whilst specific aspects of the context itself (voice/face/context) are examined thereafter.

2.1.4.1. Memory models

The average earwitnesses' inability to produce an accurate description of unfamiliar voices by their lack of specific vocabulary to describe acoustic items (Tomkinson & Watt 2018) is not the only hindrance to successful speaker recognition, but the intrinsic tendency of the human memory to focus on the codification of the message conveyed at the expense of leaving out acoustic information also plays a role. As Nolan (2001) warns, '[The memory] is selective and stores information in a processed and encoded manner. And not all that is stored can be retrieved accurately at will' (Nolan 2001: 5). Overall, basic models of the memory comprise the encoding, storing and subsequent retrieval of

information from the long-term memory through rehearsing the information in the working memory.

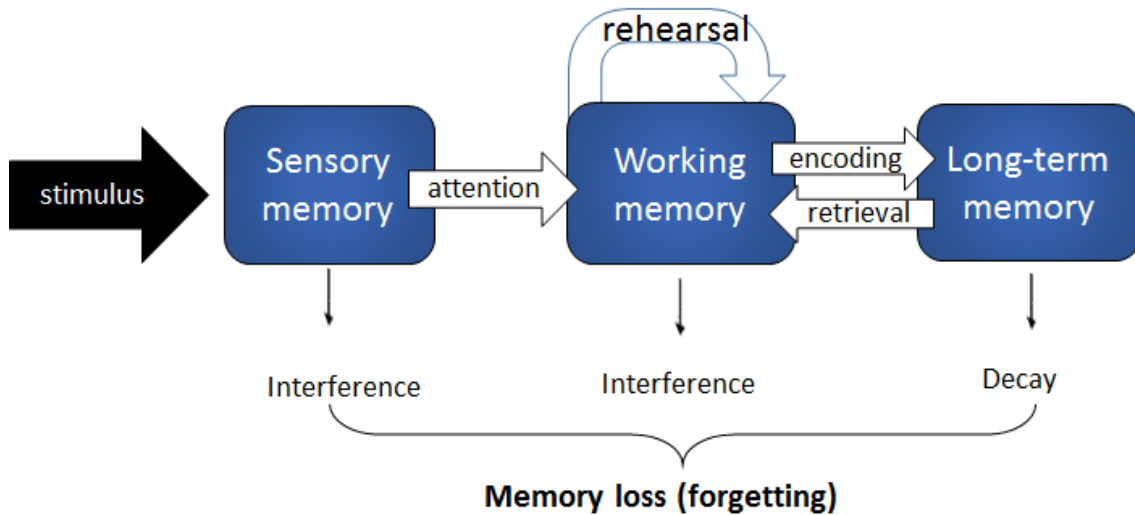


Figure 8. Encoding and storage model of the memory. Adapted from Anderson (2014: 127).

As shown in figure 8, the process of acquiring new memories goes through a lineal chain of interdependent processes, and chances of partly filtering out the information gathered are contemplated at each stage. Through the human auditory system, external stimuli enter our sensory memory, which can be obstructed by environmental noises, psychological states, physiological conditions, etc. Once the subject gains awareness and pays heed to the input, it proceeds to the working memory, where the targeted content undergoes a rehearsal process to enable the long-term memory encoding (whilst part of it is forgotten due to external factors). Lastly, the input is assimilated in the long-term memory and unlocks the possibility of retrieving traces of the information absorbed, but once more, the stored information is subject to decay over time and due to other factors related to the learning process like sleeping patterns, and times of the day, among others (Anderson 2014: 158).

Insofar as non-words and unfamiliar words are concerned, it seems that the short-term memory is irretrievably linked to the phonological representations stored in the long-term memory to carry out a ‘pattern completion’ to recognise the word (Hulme et al. 1991: 700). This connectionist conception on speech perception appears akin to the reconstruction of sequential events in the memory, since it lends itself to claim that the

brain reconstructs phonemic information when confronted with missing values coming from unknown codes or languages, just as a witness reconstructs his/her memory with self-generated details for every recall made. Nevertheless, the rehearsal of these unknown lexical items in isolation neglects pivotal aspects in natural speech processing such as the immediate phonetic context (how the phonemes interact with the upcoming and preceding sounds), sentence stress and intonation, and the semantic code (deciphering their meaning to relate them to the hearer's already known semantic net). Whether the aforementioned aspects activate or hinder the encoding, storage and retrieval of the speech signal within the memory is unattested for unknown/unfamiliar lexical items.

As for the types of existing memories that derive from this procedure, they are classified as follows:

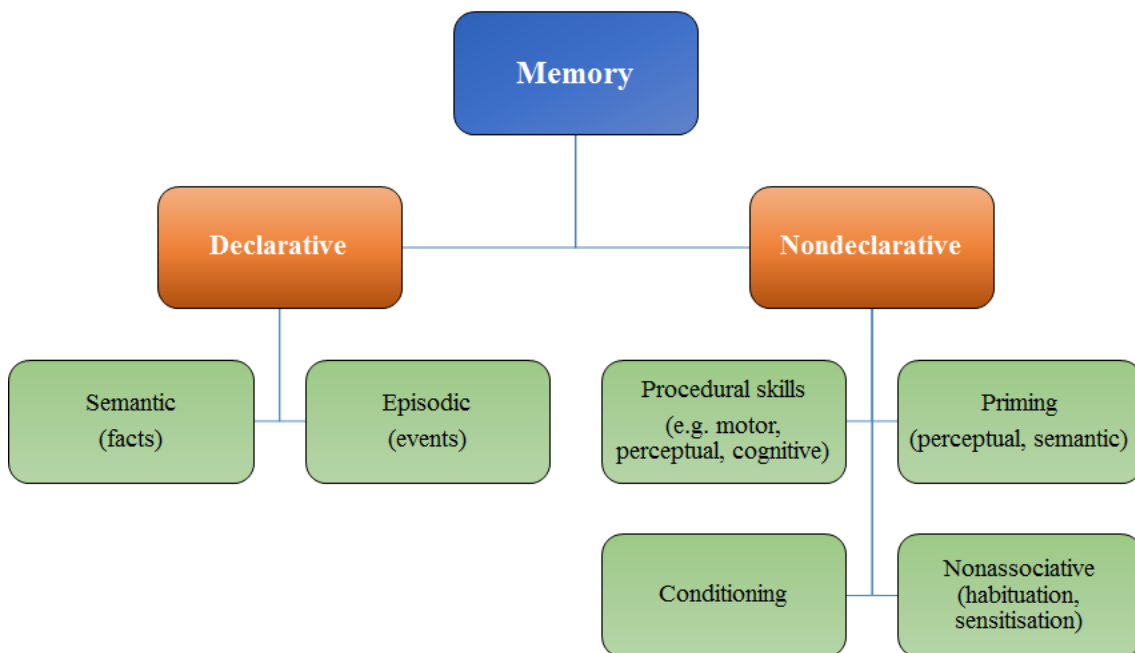


Figure 9. Varieties of memories. Adapted from Squire (1987: 170).

As figure 9 illustrates, declarative memories, or explicit memories, relate to the events (episodic) and facts (semantic) that can be consciously recalled. In this domain, there are the concepts that relate to general knowledge whose main distinctive trait is the meaning (semantic information), and those that imply a certain autobiographical time and place (episodic information) (Manzanero 2006: 407).

Nondeclarative memories, on the other hand, are those which are recalled implicitly, such as the skills acquired through practice and repetition through the regulation of routines and ultimately establishing codes of conduct (procedural) (Manzanero 2006: 407). As for priming, conditioning and nonassociative memories, these are concerned with learning processes through procedural and behavioural changes: Priming is the exposition of stimuli (perceptual, semantic, auditive, etc.) prior to identifying the intended object or word to be learned. Conditioning refers to the automatisisation of behaviours given the repeated reinforcements and desired rewards. Lastly, habituation takes place when reinforcements are not enough to provoke a response from the subject, whilst sensitisation alludes to the opposite effect, an over-exaggerated reaction to the stimuli that is being repeated over time (Anderson 2014: 179).

Anderson and Bower's (1974) Holographic Associative Memory (HAM) and Tulving's (1983) General Abstract Processing System (GAPS) lay the basic foundations which relate to the encoding of information and the degrees of cognitive processing. When it comes to memory retrieval, research hints at the binomial pair of controlled recall and automatic recall, which are influenced by the type of information to be retrieved, the objective for the retrieval, the types of tasks employed for such ends, the type of cognitive processes, and the kind of experiences the subject undergoes (Manzanero 2006: 405), as table 6 illustrates below:

Types of memory recall	
Controlled	Automatic
Type of information	
<ul style="list-style-type: none"> • Perceptual information (stimuli) • Conceptual information (introduced by the context) • Space-time and autobiographic information 	<ul style="list-style-type: none"> • No context • Sensory information only • Procedural information
Aim	
<ul style="list-style-type: none"> • Retrieve episodic information (the information and its context) • Regain a lost memory (recovery) 	<ul style="list-style-type: none"> • Undertake a task whereby a certain information needs to be applied • Recovery is not actively sought
Task	
<ul style="list-style-type: none"> • Explicit- the subject attempts to recover his/her memories deliberately, implying both consciousness and intention. • Context and instructions provided 	<ul style="list-style-type: none"> • Implicit- no consciousness on the memory-retrieval process • Familiarity and fluency on the conceptual and perceptual level (semantic memory)
Processes	
<ul style="list-style-type: none"> • Elaboration • Tedious, analytic • Integration of context-information • Synergistic echphory (recovery of a memory through a trigger) • Conceptually guided 	<ul style="list-style-type: none"> • Activation and fluency • Guided by: sensory information coming from stimuli • Perceptually guided
Type of experience	
<ul style="list-style-type: none"> • ‘remember’ 	<ul style="list-style-type: none"> • ‘knowing’ • ‘implicit response’

Table 6. Types of memory recall according to Manzanero (2006: 405-407).

In controlled recall, there is a perceptive (stimuli) and a conceptual (semantic information integrated by the context) component involved, while these memories are in turn imbued by autobiographic and spatial-time circumstances. In automatic recall, there is a lack of context, since it just provides sensory information.

The controlled recall’s objective is to retrieve episodic information, its context and the recovery is foregrounded as the main end. The automatic’s objective is not to retrieve information, but to undertake an implicit task whereby it requires certain information to

successfully complete it (e.g. the retrieval of information enables the completion of another task, like perceptual or behavioural studies). The subject is aware of his skills, but not of the issue of memory loss nor of its retrieval (after successful treatment).

As for the processes involved, the controlled recall constitutes the elaboration of a stringent analytic active recovery, which is guided and restricted by the context (Tulving, 1983). Conversely, the automated recall is not elaborated but activated by the undertaking of a specific task. It is guided by the stimuli influencing the subject's experience. Hence, it could be argued that the former relates to the conceptual domain whereas the latter is purely perceptual (Manzanero 2006: 406).

Rajaram (1993) also suggests three distinct types of experiences/responses: 'remembering', 'knowing', and 'implicit response'. In controlled recall, there is the 'remembering' where the subject is aware of the information he/she is retrieving as a lost trace related to a previous context, placed in a specific time and place, hence the 'autonoetic consciousness' (Tulving 1985: 4). The automatic recall is, on the one hand, 'knowing', which is when the subject is not aware of the information being related to his past, but is aware that he/she possesses this knowledge, or 'noetic consciousness' (Tulving 1985: 4). In other words, they are aware of the information but not of its context. Lastly, in the 'implicit response' there is no awareness neither of the information nor of its context, thus implying a non-knowing state or 'anoetic consciousness' (*ibid*). Therefore, it can be concluded that the stored information within the memory relates to differing cognitive processes, and that every type of retrieval entails different experiences, with varying degrees of 'automatisation' (Manzanero 2006: 407). In fact, Manzanero (2006) asserts that the memory-retrieval system is represented by a continuum (as the one shown in *figure 10*), which ranges from the most automated recalls to the most active and cognitively demanding processes (p. 405), which is represented hereby:

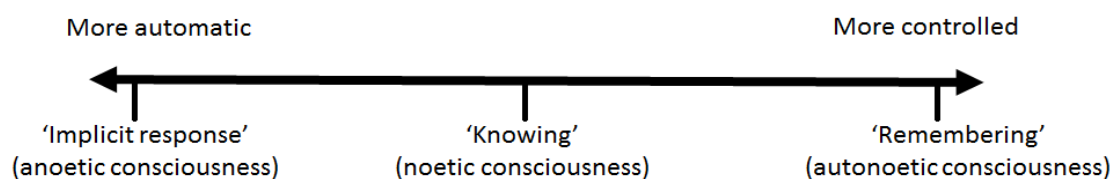


Figure 10. Continuum on types of retrieval according to Manzanero (2006: 405).

2.1.4.2. *Memory and psychology*

Acosta clarifies that the stored information on the witness/victim's brain is susceptible to subjectivity, emotional states, cognitive processes and the time elapsed between the crime and the testifying process, even if the witness or victim is readily willing to contribute to the investigation (2009: 2). Additionally, some psychological variables at play during the time of the incident affecting the memory are the duration, degree of violence, lighting, the sequencing of events suffered, stress levels, gender, age, expectations, previous practices before the recall, psychological state (before, during, and after the event), and whether the person is detailed-oriented or easily distracted. All of the aforementioned features may hamper the witness' ability to perceive or recall the incident, and can even distort their own perceptions about the crime (2009: 4).

As Ibáñez (1979) states, the social background of the witness/victim in relation to the offender cannot be neglected, since previous experiences and expectations on the suspects' social stratum are bound to emerge as a result of the categorisation of in-group and out-group conceptualisation (p. 80). Another example of how subjectivity affect not only what is perceived but also how it is reported are the creation of false memories due to suggestive interview techniques (especially when dealing with traumatic events and experiences). In this scenario, the interviewee tends to make inferences as for what is expected from them in said psychological context with legal implications (Anderson 2014: 165), and thus unconsciously create and believe new false memories, often conditioned by the interviewer's misleading wording (e.g. *You did see him, did you not?*).

According to Arce & Papillon (2002: 404), some individuals wish to forget after suffering an anxiety-induced traumatic experience, and so they end up forgetting about it. On another note, some other people seem proficient at constructing objects and creating events that never took place. The emotional burden involved in such happenings may explain why some witnesses perceive something that others neglect, and why others categorise elements that were non-existent, whether it serves as a deception technique or not. (Ibáñez, 1979). It should be contemplated when evaluating a witness' testimony the contamination around a false testimony induced by a misinterpretation of the events, one-

sided perception, a short attention span or by non-intentional conditioning (how suggestible a witness may be) (Köhnken et al. 2015: 15).

Johnson & Raye's concept of Reality Monitoring (1981) distinguishes between memories originated internally and externally. They differ inasmuch as they reveal differing information, as the internal is more focused on cognitive processes whereas the external collects sensory and contextual information. This distinction could hint at the internal sabotaging of the memory (creation of false memories). The time elapsed is crucial when determining the veracity of perceived information coming from external memories, as they tend to be more inaccurate as time goes by. Furthermore, the suspicion on information stemming from internal processes tends to increase over time as well, since the witness could seize the time available to elaborate an imagined event on a false testimony (Arce & Fariña 2006: 70). In fact, Clifford et al. (1981) establish that especially the 'voice memory under delay conditions is not very good, and that as the delay in testing increases so the certainty concerning the validity of testimony should decrease' (p. 208).

When it comes to testify, let us not forget that the witness may resort to the simulation of physical (dizziness, weakness, delirium, cephalic pain) or even psychological disorders (psychopathy, schizophrenia), even though the latter is rarer and harder to perform for its increased demands of energy and the high cognitive load it requires (Acosta 2009: 9).

In order to facilitate memory retrieval, some general techniques are put forward:

- **Mental positioning** or recreation of the physical and personal circumstances at the time of the incident. This method comprises the remembrance of emotional states (evoking), sequential elements and perceptual features (Acosta 2009: 3).
- **Free recall of partial memory.** The witness is not interrupted, questioned and his/her narrative is encouraged, which may unleash associated memories. As Manzanero (2006) puts it, these tasks may activate an automatic recall through the conceptual and perceptual familiarity that impregnate the witness' semantic memory (p. 29). However, the tendency to remember the first (primacy effect) and last (recency effect) items in an ordered sequence may leave out important information required for investigative police interviews (Hulme et al. 1991: 686).

- **Change of perspective.** This technique relies on the witness/victim's ability to change perspectives and experience the event from the aggressor's side. (Acosta 2009: 4)
- **Inverse memory recall.** It seeks to recover small details by alternating the order of events (*ibid*).

Needless to clarify that not every technique applies equally to every individual, and, consequently, interviewers must adapt to the interviewee's cognitive and linguistic skills. It is worthwhile to point at the specifications on children and vulnerable witnesses, who are typically treated with special care with non-coercive behaviours (like avoiding exerting pressure by repeating the same questions or labeling/interpreting what the witness said) (Arce & Fariña 2006: 54).

2.1.4.3. Memory and voice/face/context

In the realm of speaker identification, the memory acting upon voice recall and face recall seem to differ in terms of their performance and functionality. The FOE (Face Overshadowing Effect) proves that an initial exposition to the assailant's face is theorised to interfere with the identification task, thus enhancing the witness' abilities when the face is absent and the length of exposure to the unknown perpetrator's input is expanded (Cook & Wilding 2001: 617). This 'involuntary attention' (Cook & Wilding 2001: 627) exerted onto the face over the voice of the individual appears to suggest that the former is oriented towards identification whereas the latter is likely to lean on the interpretation of the code. Hence, here lies the importance of preventing biased sequencings of voice line-ups, where the suspects' faces are shown first and a voice examination is undertaken afterwards.

Not only the face, but also the lexicon employed and the familiarity towards the voice (as seen in 2.1.3. *Voice line-ups/Voice parades*) have an effect upon memory traces and their retrieval. The strength of a memory and its activation depends upon how well the retrieval cue matches the initial encoding at the moment of the exposure, since 'these surface details are not lost or discarded during the encoding process' (Goh 2005: 42). This means

that using similar codes (lexical items) and the same voices may increase the activation levels of an otherwise inaccessible memory, especially in long-term memories.

Other than the sheer perceptual configuration of the individual, the contextual situation of a voice line-up may raise the suggestibility of the witness for a necessary identification, even when it may not be convincing (the enacting of the witness' role). If combined with face exposure, it seems that the introduction of profile information about the suspect and the event may be detrimental to the earwitness' identification attempt, as it might imply a 'preferential learning' where the voice input is neglected and thus his/her performance is negatively affected as a consequence (Cook & Wilding 1997: 540).

In many cases, however, additional elements may interfere with the witness' identification ability. A summary of the main studied dimensions can be consulted in table 7 below:

Variables that influence the witness' identification ability			
Variables to assess		System-related variables	
From the incident	From the witness	Of the process	Of the line-up
<ul style="list-style-type: none"> • Sensory and perceptual factors • Duration • Familiarity • Aggravating circumstances • Number of assailants • Use of violence • Weapon employed 	<ul style="list-style-type: none"> • Gender • Age • Ethnicity • Previous training/experience • Expectations and belief systems • Anxiety • The witness' role 	<ul style="list-style-type: none"> • Time-lag effects • Post-event information • Photography/Recorded material • Previous descriptions • Facial composite/voice description 	<ul style="list-style-type: none"> • Arrangement of the line-up • Number of speakers • Selection of foil speakers • Presentation method • Instructions delivered

Table 7. Variables that influence the witness' identification ability. Adapted from Manzanero & González (2015: 132).

First of all, psychologists should evaluate the nature of the incident and the ensuing memories generated. This is to say that memories originated during traumatic events are essentially different from episodic/autobiographic ones, and thus it influences how the encoding, storage and possible retrieval is carried out. The defining factors on recall during such memories are stress levels, intensity of the emotions experienced, and degree of involvement with the crime, which entails differences in regards to accuracy and accessibility of memories.

Trauma can be explained on the basis of the psychological and physical effects the crime/aggression left on the victims' psyche and how the subsequent emotional disturbance interferes with their daily life (Manzanero & Recio 2012: 21). Even though it could be argued that the same traumatic event can be fragmented or either be remembered vividly by the victim (*ibid*), some circumstances like the use of violence and the degree thereof, number of assailants, use of weapons, previous familiarity with the offender, and the duration of the crime (one-time offence or protracted crime) do render aggravated and unpredictable conditions for memory retrieval. As for the emotional side, stress levels tend to produce intense, persistent and vivid memories, but also deteriorate the attention span and recall (*ibid*).

Concerning the witnesses or victims themselves, it appears that the intensity of the emotions associated with the crime and the degree of involvement are key. Autobiographic events with an emotional involvement are remembered in more detail than the mundane events with low emotional involvement. Also, suffering the event is categorically different from just witnessing it in terms of memory storage and recall (and possible retrieval). Being the sufferer leads to fragmented, confusing, more intense associated emotions, and presumably more accessible (since there is a tendency to re-experiment the event and reflect on it) in contrast with the witnesses' memories, due to an obvious lesser degree of emotional involvement (*ibid*). As explained in 2.1.4.2. (*Memory and psychology*), subjectivity of the witness and victims themselves also influence the quality of their ability to identify the suspects. A list of the aforementioned variables include age, gender, ethnicity, belief systems, previous experiences, anxiety, and internal sabotage.

On the other hand, issues related to the procedure itself may arise as well. Here a distinction must be drawn between voice and visual line-ups, since it has been found that the auditory and visual memory differ inasmuch as the former is critically degraded over time, whereas the latter's responses do not deviate significantly under the same conditions (Hollien 2002: 61) After applying short retention intervals to ensure that the earwitness has not retained the voice information stored within the working memory, the ability to identify speakers seems nullified (Manzanero & Barón 2017: 59). The detrimental effect that this delay exerts on the auditory memory has been reported, measured, and corroborated further in retention intervals up to 5 months (Papcun et al. 1989, Yarmey 1995), although this loss does not become noticeably dramatic until 24 hours have elapsed from the initial exposure (Legge et al. 1984). There are, however, notable exceptions, like the experiment conducted by Kerstholt et al. (2004), which concluded that the answers given in a target-absent voice line-up after a week of the initial exposure were more precise, and contained less false alarms than those answering the voice line-up right away, provided that the exposure duration was longer (30s-70s). Consequently, one does not only consider the contextual information impinging upon memory span and retention intervals, but the quality and the duration of the input are deemed necessary, too. In this regard, the audio material used for this thesis is balanced in terms of duration and audio quality.

Aside from those elements inherent to the process, some external forces may be present, too, such as cognitive biases (e.g. police officer's guided narrative or questioning in investigative interviews), and noise disturbances, whereas the validity of the witness/victim's descriptive testimony could be misleading, as discussed in 2.1.3. (*Voice line-ups/Voice parades*).

Lastly, a number of technical flaws could be detected in the line-up's set up, such as the unbalanced selection of foils and suspects (listing dissimilar voices in such a way that only a few stand out), the number of speakers selected (too few individuals is seen as unfair for the suspect and the other components of the line-up, whereas choosing too many speakers is deemed as demanding for the witness/victim), how the recorded voices are presented to the witness/victim, and how they are instructed on the procedure (*see 2.1.3. Voice line-ups/Voice parades* for full details).

2.2. ANALYTICAL PROPOSAL

After a thorough inspection upon the existing literature on variationist theories on sociolinguistics, forensic linguistics, forensic speaker recognition, and the psychology of the earwitness, limitations on accounting for the complexity of interdependent factors become apparent. Thus, this thesis identifies four separate subtypes of elements: controlled, partially controlled, uncontrolled, and unknown factors.

Those controlled factors typically refer to the selected informants' profiles, characteristics like age, gender, level of studies, sociolinguistic environment, and the creating of stimuli under similar conditions. For those which are partially controlled, they refer to the volunteers who participated in this experiment (or jurors), as their sociolinguistic profiles cannot be balanced entirely (the researcher may allow some degree of flexibility to gather an even distribution of cases for *age*, for example. However, it could be argued that this intentional intervention does not provide a *faithful* representation of the target population, but rather one that is convenient to the study⁵), as well alluding to the particularities of such perception surveys (absence of traumatic experiences at the time of completing the surveys). Nevertheless, the online format of said surveys entail the emergence of uncontrolled factors, such as the time of the day, lighting, noise conditions, the time of exposure to the stimuli⁶, etc.). As for unknown factors, these may derive from previous preconceptions around the theme (language and the law), the research process itself, or even past experiences related to this area.

The main purpose of this work revolves around the premise around the aural-perceptual inabilities, or rather the difficulties arising for hearers, whose reduced language familiarity with the target speaker renders less reliable judgments apropos legal standards within foreign and native speech recognition. This state of the affairs is tackled through a two-fold approach, a first section dealing with sociolinguistic features and how they facilitate or exacerbate identification and discrimination of speakers (*chapter 4. Results:*

⁵ The word *convenient* is used here to refer to the ideal conditions required in order to provide statistically sound results with an adequately and evenly stratified sample (see point 3.7.2. *Perception surveys-based analysis* for a full discussion).

⁶ A disclaimer is added to the surveys, stating that replaying recordings and going back to previous identification tests is not allowed, but, given the online medium and anonymous nature thereof, little control can be exerted in this domain.

Perception surveys-based analysis), whilst the second part (chapter 5) involves an acoustic-phonetic analysis in which segmental and suprasegmental's discriminatory power shall be contrasted to bring out the most influential elements in a multilingual data set (English, Spanish, and Dutch audio material).

Even though the second analytical stage also entails the conducting of statistical measures, the first section (*Results: Perception-surveys based analysis*) relies exclusively on statistical tests, and thus includes a wider variety of them to best suit the research questions and objectives planned (see 3.7.2. *Perception surveys-based analysis* for more details). In this section, population trends are drawn in relation to the interplay between success rates in speaker recognition tests and sociolinguistic variables and, where possible, inferential statistic measures shall identify potential sociolinguistic predictors.

The second section examines the audio material employed in the construction of said experimental voice line-ups. In this respect, source material is consulted to gauge levels of inter-speaker and intra-speaker variation amongst the listed speakers. In so doing, between-speaker variation is measured by means of relevant segmental and suprasegmental features (Cicres 2007), according to the literature found in studies employing diverse linguistic data sets (see 3.7.3. *Acoustic-phonetic analysis* for more details). As far as within-speaker variation is concerned, the same acoustic parameters as in the previous step are considered, only that this time same-speaker speech samples contain differing intonation contours, with the purpose of testing the range of variation within a single individual and discovering whether this distinction is enough to set speakers apart from each other at the perceptual level. Ultimately, it is sought to compare acoustic-phonetic analyses with the average speaker's intuition, and decipher what aspects affect their efficiency at foreign and native speaker recognition tasks.

The real-life implication of this study would be to incorporate more suggestions to the guidelines regulating voice line-ups (Broeders & van Amelsvoort 1999, 2001; De Jong-Lendle et al. 2015, and Hollien 2012) through the experimentation of various sociolinguistic and acoustic conditions. Incidentally, it is also worth revisiting the concepts around the validity of such probative evidence in cases where a proof of this kind becomes definite in the judge's/jurors' verdict. For this reason, limitations on the effectiveness of such methods (both the line-up as a standalone method, and the testimony

Chapter 2- Theoretical foundations and state-of-the-art review

of a bystander/victim originating from it) shall be considered at the conclusion of this thesis, due to the evident legal repercussions this may entail, lest the procedure should be flawed in some respects.

CHAPTER 3

METHODOLOGY

The upcoming chapter provides an exhaustive account about the methodological aspects surrounding voice line-ups by referring both to the specificities of this particular research and the subsequent modifications made from traditional aural-perceptual tests.

Firstly, the online sources consulted from which informants' voices were obtained are discussed in 3.1. (*Corpus*). Once the voice samples were successfully extracted, a procedure involving audio cropping through Camtasia Studio began, and the resulting excerpts were employed to build up the voice line-up's body of foil speakers (distractors) and suspects (see 3.2. *Sample selection criteria* for a full discussion). When the requirements to create the perception tests were fulfilled (rendering audio files, arrangement of voices, deciding which informant's voice acts as the intended suspect, etc.), they were in turn split into three distinct language perception tests, whose voice samples come from different corpora, namely the English (3.1.1.), Spanish (3.1.2.), and Dutch (3.1.3.) corpus. Such multilingual data set is designed to test varying degrees of language familiarities (familiar, learned, and unknown language) against two groups of

jurors (3.3.): The British (3.3.1.) and the Spanish (3.3.2.) group, with their respective sub-groups (monolingual and bilingual linguistic environments).

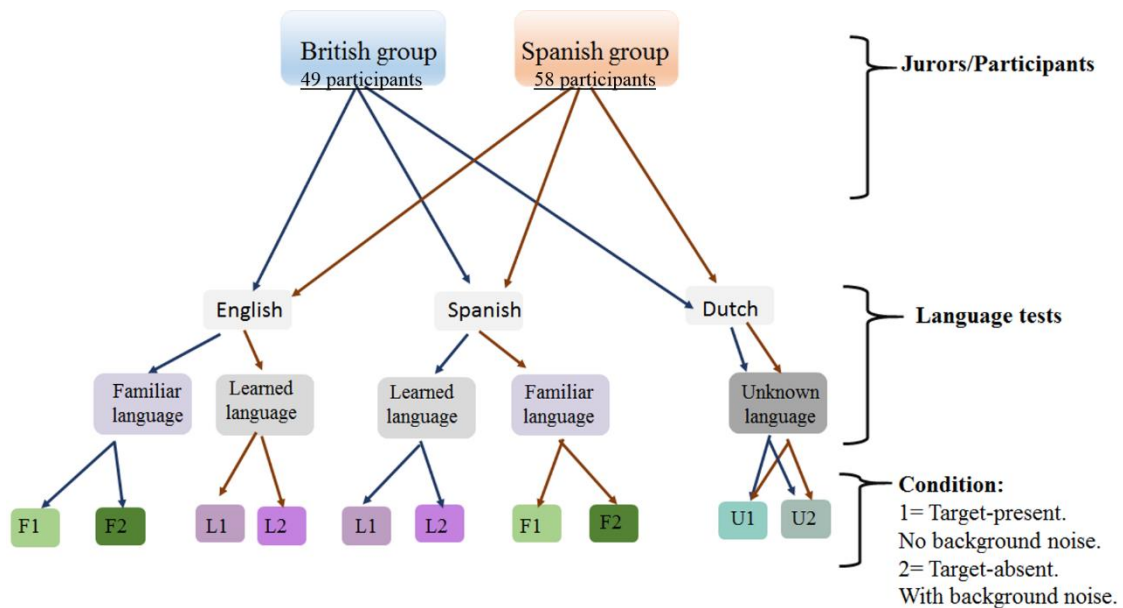


Figure 11. The experiment: jurors, language tests, and experimental conditions.

The relationships between groups of jurors, input (language tests), and language familiarity to said input can be observed in figure 11 above. Additionally, two experimental conditions (target- present with clear sound, and target-absent with background noises) have been created per language test, which attempts to shed light on the complex issue of perception and recognition being presumably hindered by background noises and target-absent conditions.

Since the proposed research activity implies the involvement of university students (jurors) coming from British and Spanish universities, a code of ethics (3.4.) is compiled apropos the regulations and measures adopted on data management, confidentiality, anonymity, withdrawal, and participants' consent. Given the fact that this is a forensic phonetics experiment, some technicalities concerned with audio encoding need to be addressed in a separate sub-section (3.5. *Recordings*). This thesis does base its founding principles on relevant publications dealing with the application, regulation, and results of experiments involving voice line-ups. However, it must be noted that some methodological changes have been incorporated for the sake of discovering new

possibilities in the field of foreign and native speaker recognition, and they are discussed in 3.5. (*Novelty*).

Moving to the analytical part (3.7. Analyses), some adjustments to the perception surveys are commented in 3.7.1., lest should the set up compromise the validity of the findings. Thereafter follows a thorough examination of the perception surveys' structure (3.7.1.1.), including the definition, order, and coding of the relevant variables provided therein. After clarifying the intricacies and modus operandi of such a data-gathering method, it shall ensue a statistical analysis of the variables gathered through perception surveys (3.7.2.), and thus this point tackles each and every hypothesis formulation, ranging from previous literature findings and statistical measures to data handling and established hierarchies of analysed strata. The complementary acoustic-phonetic analysis (3.7.3.) employs the software Praat (Boersma & Weenink 2019) to extract voice parameters, both suprasegmental (3.7.3.1.) and segmental features (3.7.3.2.), which intend to spot significant variables with high discriminatory power, which will be calculated with statistical measures, too.

3.1. CORPUS

In the making of this corpus, audio files in English, Spanish, and Dutch conversations were extracted from various websites. In the next sections, the foundations and principles behind sample selection, informants' profile, and controlled sociolinguistic factors are specified for each source.

3.1.1. English corpus

The English-spoken corpus was extracted from the British Library Sound Archive (2016). Conversely to the Spanish and Dutch corpus, this website's content was not created for research purposes, but it is conceived as a public storage system that allows users to access its reference materials provided that there is no copyright infringement and the source material is referenced appropriately.

The chosen collection of voice samples offers a series of recordings with British wildlife sound recordists, ranging from researchers to hobbyists. In this case, profession and educational background are controlled factors and should expect little variation. As for the interview itself, it follows the same structure across all guest speakers (childhood, early influences, recording experiences, academic research, emerging technologies, etc.). Factors such as proximity and status could be equated to *acquaintance* and *inferior to interviewer* (the interviewee has less control over the interview), accordingly. The reason being that the interviewer is Mark Peter Wright, a British sound artist whose research interests converge with the interviewees', but the interview format projects expectations on them in terms of allowed communicative practices.

The language produced by the interviewees is not technical and hence can be understood by the average lay listener, even if it may appear otherwise due to the topics discussed (work and research). As for age groups, most of the informants are aged above 55 years old, while only five of them are aged from 47 to 53, hence establishing different generations. The five aforementioned informants' voices were selected for the voice line-up, and including two of the other group (above 55 years old) to comply with the requirements of the line-up (providing at least one voice which is noticeably dissimilar from the suspect's). Concerning the age difference between interviewee and interviewer, it could be described as *interviewee is older than the interviewer* in all cases, because the interviewer Mark Peter Wright was 34-37 at the time of the interviews (2013-2016). The interviewee's profiles can be consulted in *Appendix 1*.

3.1.2. Spanish corpus

To create the Spanish voice line-up, voice samples were obtained from ESLORA corpus, which registers semi-spontaneous exchanges recorded in Galicia between 2007 and 2015 (Vázquez 2014: 1). This corpus is in line with PRESEEA's (2014) (Project for the Sociolinguistic Study of Spanish from Spain and America) guidelines concerned with stratifying voices according to gender, age groups, and level of studies, amongst other factors, which gives rise to distinguishable sociolects. Specifically, ESLORA offers two types of interactions in its recordings: semi-directed interviews and spontaneous conversations. For the purposes of this research, the former is preferred, as it allows for a

certain degree of control over the content being displayed in the speech sample, namely the order of the topics being discussed.

Given that the data set was not made publicly available at the time of the request, access thereof was granted after a brief exchange with relevant staff members in charge of managing ESLORA's corpus, and after being required to sign a written statement in which the researcher involved with this PhD project commits to both a) use the material provided for an already defined research activity (to compile the body of speakers in the Spanish language perception test, with a subsequent acoustic-phonetic analysis to unveil possible factors with discriminatory potential, in this case) and b) to never share or disclose it to third parties beyond the reach of its intended piece of work (see *Appendix 2.1.*)

After discussing the possibilities for the intended target population, several recordings matching the established criteria (informants sharing sociolinguistic background, with similar age group and level of studies) were sent, and thus the target group was defined as females aged 19 to 34 (which is further redefined afterwards, see 3.2. *Sample selection criteria* for more details), with university and medium levels of study (see *Appendix 2* for a detailed overview). Even though these are semi-directed interviews, the age difference between interviewee and interviewer does not tend to be too dissimilar. Occasionally, there are instances where both interlocutors in the exchange share sociolinguistic profiles, as it is M13_016 and M13_016_hab2's case, who are both female, aged 20 with university studies. For this reason, the proximity between interviewee and interviewer could be described at least at the level of *acquaintance* or *peer*, in spite of the fact that both are fulfilling their assigned role in the interaction.

3.1.3. Dutch corpus

Apropos the last corpus, all the data has been extracted from the IFADV corpus, which is a 'visual version of the friendly Face-to-Face dialogs of the Spoken Dutch Corpus' (Van Son et al. 2008: 1). In this free smaller version, the corpus compiled a total of 34 speakers, namely 10 males ranging from 21 to 72 years old, and 24 females within the 12-62 age group (*ibid*: 2). As females outnumber males in this corpus, the target speakers selected

to perform as foils and/or suspects are females aged 18-28 (IFADV 2007). This data set is aimed at British and Spanish jurors for testing their aural perception on a completely unknown language (since English might not be unknown for the Spanish group in all cases due to the increasing influence of English as a lingua franca).

In contrast with the previous sources, this set is composed of spontaneous conversations between two informants in front of a camera, without supervision of an interviewer, and thus enhancing the spontaneity of the exchange. Besides that, all participants are either friends or colleagues, which renders a high degree of proximity and a relationship of equals in this context. Since these exchanges mimic everyday talk, restraints on conversation topics do not apply, which causes a more unpredictable sequencing of the content than the one appearing in the previous corpora's semi-directed interviews (Van Son et al. 2008: 2). To ease the researcher's task and avoid potential misunderstandings with the Dutch language, IFADV (2007) offers a summary of the conversational content in English as well as an orthographical transcription of the words uttered by both speakers involved in the oral exchange. The informants' profile can be consulted in *Appendix 3*.

3.2. SAMPLE SELECTION CRITERIA

For the selection of foils and the suspect in the English corpus, the voice line-up's composition revolves around the 5 youngest speakers (47-53), whereas informants belonging to older generations (above 55 years old) as used as distractors to balance the test with a discernable dissimilar voice. As the sociolinguistic standard for this group, it must be noted that the area of London reunites most of the speakers in the list. Only speakers located further away from the British capital are considered as distractors whose dialect differs from the suspect's to a certain extent, since one of this kind is required to ensure fairness in the voice line-up. As for the embedding of recordings in the line-up itself, short sequences of speech with rising intonation (either instantiated by tag/tail questions or uptalk) assemble the group of foils, whereas sentences with falling intonation are employed in the suspect's introduction.

Upon close inspection of the voice samples provided in the Spanish corpus, it has been decided to include all the informants that share core sociolinguistic features (females, 20-

32 years old, with medium to university level of studies) and all samples that did not match the criteria were discarded. The only exception here, in comparison with the other two corpora, occurs in choosing an interviewer (instead of all voice samples belonging to interviewees) in one occasion as a foil speaker (in the 2nd language test with background noises) due to her perceptual similarities with the target mentioned earlier, since other female informants' voices were too distinguishable (and the line-up already includes two speakers with clear differentiations from the suspect's voice). As noted above, the Spanish suspect's voice sample includes an excerpt with falling intonation, whereas the body of distractors displays instances of rising intonation (and the sentence of choice for the suspect is changed when presented alongside the rest of recordings in the 1st language perception test).

As the Dutch corpus does not pose major issues in terms of heterogeneity in speaker profiles, the 8 candidates are selected for the voice line-ups (only that some appear in the 1st phase and others in the 2nd test with background noise). For this corpus, portions of the discourse containing uptalk (or high rising terminal) are selected to create the body of foil speakers. When presenting the suspect to identify, a sentence with descending intonation (e.g. a declarative sentence) is selected with the aim of contrasting differing suprasegmental features.

As the literature notes, it is undeniably troublesome to draw the line between similar and dissimilar voices, lest it should bias the whole identification procedure (Yarmey 1995: 808). In this regard, the applied criteria in the cropping of informants' audio files (to render the resulting recording in the line-up) are put forward hereby. Firstly, chosen excerpts must be clear and should not contain distortions. Interruptions, overlapping of speech, false starts, etc. When compiling the body of foil speakers, it has been ensured that the instances of rising intonation exhibited in the recordings is differentiated enough from the suspect's descending intonation.

Another aspect that is theorised to be influential in speaker recognition is exposure duration of the input to identify (Cook & Wilding 2001: 617). In order to avoid undesired biases, all voice samples displayed in each line-up last for a similar amount of time (under 20 seconds). The specifics of how these voices were arranged in every voice line-up (for both experimental conditions) in each group can be consulted in *Appendix 4*. The above-

mentioned sample length is only relevant at this stage of the research for its intended purpose (gauging degrees of aural-perceptual recognition capabilities across specified groups of jurors). However, an extended sample length is considered in the second half of this thesis (see 3.7.3. *Acoustic-phonetic analysis* for more details) in order to get a more comprehensive view on the informants' voice parameters.

3.3. JURORS

With the purpose of testing aural-perceptual recognition on various groups of speakers with differentiated sociolinguistic backgrounds, jurors have been selected from Bangor, Cardiff, Swansea, Southampton, Winchester, Roehampton (British group), València, Barcelona, Girona, Seville, and Granada universities (Spanish group). The procedure started with contacting relevant administrators and staff members from each university department. Once an agreement was reached, the created perception surveys were distributed amongst the university students of said academic institutions via department e-mail (see 3.4. *Code of ethics* for more information).

3.3.1. British group

As noticed in the previous section, the British group's distribution of participating universities entails a further sub-categorisation insofar as sociolinguistic environment is concerned, namely the monolingual and the bilingual groups. Starting with the monolingual type, Winchester, Southampton, and Roehampton allude to the region of South East England, whereas Swansea, Bangor, and Cardiff represent the bilingualism of English/Welsh in Wales. Despite the fact that such sociolinguistic areas should in theory mirror the target populations' linguistic skills, true bilingual societies are at times unattainable due to the increasing influx of people with contrasting sets of linguistic skills. Also, minority languages like Welsh, where a 2001 Census determined that a 20.8% of the population in Wales could speak Welsh (IWA 2001: 1), could compromise this notion of real bilingual communities. Rather than looking at the jurors' linguistic repertoire, monolingual and bilingual groups are conceptualised by means of the present study as linguistic environments which shape jurors' hearing acuity. In this regard, the data provided by each sub-group (monolingual and bilingual environment) includes all the

cities belonging to said categories, instead of looking at each of them individually. Thus, the 49 respondents who participated in this study in the British group are split up as follows: monolingual (28) and bilingual (21) groups.

3.3.2. Spanish group

As argued above, the Spanish group does make the distinction between monolingual and bilingual linguistic environments. In this case, the monolingual side refers to respondents coming from Andalusian universities (Seville and Granada), whereas the bilingual stratum surveys Spain's east coast communities with Spanish/Catalan proficiencies: both the Valencian Community (València), and Catalonia (Barcelona and Girona). In contrast with Welsh' language usage, the Spanish case is relatively successful in integrating the co-official language (Catalan), with a 48.88% of the overall population speaking it in the Valencian Community (IVE 2001: 5), and amounting to a 73.43% of its use amongst youngsters around the same age range (2-14 years old) in Catalonia, according to a 2001 Census (IDESCAT 2001). However, due to the reasons mentioned in the previous section (3.3.1. *British group*), this situation is far from reflecting a perfect command of both languages in the target population. Similarly, no distinctions are made between individual cities, but their data is taken as a whole for each sub-group. As a result, the 58 Spanish participants are distributed between monolingual (33) and bilingual (25) groups.

3.4. CODE OF ETHICS

The following code of ethics has been created by following the founding principles sustaining the *Recommendations on Good Practice in Applied Linguistics* (BAAL 2016).

Prior information

The introduction to the survey covers:

- An explanation on what type of data informants are consenting to provide.
- Information about the general grounds and aims of the study.

Informed consent

Information is purposefully given on:

- The purposes of the on-going research
- The time needed to complete the task at hand
- Briefing on the task: structure, development, and expected responses.
- Specific observations for the task: warning about the possibility of an absent suspect and clarifying that the main goal is to identify the suspects, rather than understanding the languages they are speaking.
- Access to the data provided (researchers involved in said thesis)
- Implied consent: The surveys have been circulated through department e-mails coming from university staff members. As such, participants may opt either to take part voluntarily or ignore the message altogether. Additionally, a disclaimer clarifies this implied consent right before the participants submit their responses:

Disclaimer: by clicking on 'SUBMIT', you consent to the usage of the data provided for research purposes only. Said data will be kept anonymous at all times.

Confidentiality

The survey is designed to elicit generic personal information which renders the participating subject almost unidentifiable (e.g. Age, gender, level of studies, L1, etc.). The core of the survey displays a series of voice line-ups which demands minimal input from the subject (question types such as multiple choice and scales from one to ten), which again makes the informant remotely identifiable.

Anonymity

There is a dedicated note addressing this issue at the beginning of the survey: '- Your responses will be kept anonymous.' Before submitting the responses once they have all been collected, the informant is once again reminded of it with a disclaimer: 'Said data will be kept anonymous at all times'.

Withdrawal

This feature is not permitted in the present research. The following observations shall justify this methodological decision:

- Enabling withdrawal would break the confidentiality/anonymity feature, since the researcher would have to identify the informant's data first before removing it from the sample. In this respect, confidentiality/anonymity take priority over withdrawal.
- The identification process to enable such withdrawal would be troublesome both for the researcher and the participant involved. The participant would have to provide the exact time and date at which he/she completed the survey, as well as providing all the given responses in the right order. (Example: an informant sends an e-mail with the following information: Female, aged 18-22, University studies BA, English L1. According to the sample, this information amounts to the 75% of total respondents and thus makes identification a real issue.).

Data management

- Data storage and potential destruction: All the data is stored in a private Google Drive account. Only the main researcher is granted access to it. The account will be deleted once the data provided for the study is not needed anymore.
- Said Google Drive account has been created solely for the purposes of this research, which prevents the mixing of personal and research data.
- Data anonymisation: Since the survey only elicits generic input, no more caution is required to anonymise the extracted data.

3.5. RECORDINGS

The resulting recorded material is set at different sound qualities since each research group employed different hardware (Olympus DS-40 tape recorder and a ST XQ built-in stereo microphone for the Spanish corpus; unspecified for the English data; and Samson QV head-set microphones for the Dutch corpus). However, the software Camtasia Studio

8.1.2 (TechSmith 2013) has been used to crop and collect the needed excerpts for the conducting of two perception tests per language (both with and without background noise), and to equate the audio encoding settings to 44.100kHz, Mono, 128kBits/sec in every audio file.

3.6. NOVELTY

As seen in 3.1. (*Corpus*), the stimuli employed for perception surveys has been gathered from various sources, instead of creating ad hoc voice samples, just as previous research (Mullennix et al. 2011, Roebuck & Wilding 1993) has done. It could be argued that this methodological decision does not allow for the researcher's control over the phonemic variables of interest, but it is precisely through this change that the very purpose of the experiment develops in contrast with previous studies. In other words, it is sought here to replicate an aural-perceptual recognition scenario where acoustic conditions are closer to those in naturally-occurring speech, and thus researcher's control over the utterances produced is irretrievably reduced. This methodological proposal is therefore investigating more efficient ways in extracting features concerned with inter- and intra-speaker variability of less controlled samples, as opposed to instructing informants to produce certain lexical items (either in isolation or in context).

To add up an additional dimension for analysis, the employed corpora contains multilingual input (English, Spanish, and Dutch), which could potentially identify speaker recognition principles that establish common ground amongst the mentioned languages. Avoiding skewed results is also a concern of this research, and thus two distinct group of jurors (British and Spanish) have been incorporated with their own sub-groups (monolingual and bilingual environments). In this regard, the main object study is participants' language familiarity (familiar, learned, and unknown) with the input exposed, rather than their implicit knowledge of the languages themselves (Köster & Schiller 1997).

Admittedly, Kerstholt et al. (2004) do inspect the effect of acoustic environments, and thus conclude that such effects on identification accuracy are not relevant. However, the present study does not only consider the acoustic conditions surrounding the voice itself

(background noises), but also adds the influence of speakers' immediate linguistic environment, to discover how/if hearers' aural-perceptual skills are shaped by monolingual or bilingual settings (through elements like speech community practices or semiotic landscapes, for example).

3.7. ANALYSES

Here follows an explanation of the data collection methods employed and the data processing procedures adopted. To gather the intended data for subsequent analysis, online perception surveys (3.7.1) are created (and distributed afterwards) through Google forms with a Google Drive account (Google 2019). In interpreting jurors' responses, the statistical software IBM SPSS statistics 25 (IBM Corp. 2017) is consulted to test whether statistically significant correlations occur between the obtained variables (see 3.7.2. *Perception surveys-based analysis* for further information). Lastly, the speech analysis software Praat 6.0.25 (Boersma & Weenink 2019) unveils similarities/dissimilarities in suprasegmental and segmental phenomena among the informants' recorded voices (see 3.7.3. *Acoustic-phonetic analysis* for full details).

3.7.1. Perception surveys

The created online perception surveys adopt a between-subject experimental design. That is, it revolves around two different groups' scores with the researcher intended modification of a given variable (Rasinger 2013: 41). In this case, background noises will be varying across the three perception tests (two for the English jurors) as the juror may develop an 'artefact', which is the spoiling of results by respondents' reaction to the task and not to the stimuli itself, which in itself poses significant issues around the validity of the data obtained (*ibid*: 43). In order not to favour or hinder the perception of a determined group over the others, the sound clip's theme played at the background is the same for every test (i.e. sounds of rainfall with differing intensities according to assumed difficulty levels). Thus, selected sounds extracted from the British Sound Archive are assigned to each language perception test as follows: F2 (gentle rain becoming heavier), L2 (rainfall-coastal rainforest), and U2 (rain and thunder- heavy rain becomes lighter). Despite this, and as already noted in 3.5. *Recordings*, every audio file's encoding settings puts them

on equal terms. The arrangement of every voice sample employed per voice line-up (and their duration) can be consulted in *Appendix 4*.

3.7.1.1. Structure and design

In order to collect the jurors' information required for the subsequent statistical analysis, two templates have been created by adjusting the questioning to the needs of both groups of respondents (Spanish jurors and British jurors). Both language perception surveys are identical, with the minor exception of Spanish and English voice line-ups being interchanged depending on the degree of familiarity with the intended group of jurors, ordered from most to least familiar languages (British jurors are exposed to English input first, whereas the Spanish group is tested on their mother tongue as well). Because of such similarities, the British group's survey template is used as a guideline⁷, and thus screenshots of it can be consulted in *Appendix 5*.

- **Presentation:** The first page provides a brief introduction to the PhD project's topic, followed by instructions on how to undertake the perception tests, and a disclaimer dealing with confidentiality and data protection (see 3.4. *Code of ethics* for full details).

- **Profile:** This section accounts for the participant's characteristics that may impinge upon aural perception and recognition:
 - Gender
 - Age
 - Education level
 - Familiarity with linguistics and/or phonetics
 - Whether the informant has received musical training
 - Bilingual Spanish-Catalan/English-Welsh or monolingual Spanish/British
 - Languages spoken and acquired level

⁷ Online perception surveys can be consulted through the following links: Spanish (<https://forms.gle/CK4RbZb6Nmah2fFN6>) and English survey (<https://forms.gle/4ueiZxNk6tUHtdiAA>). Please note that, unlike their original versions, no required fields are shown in this questionnaire for an easier navigation through the sections of interest.

- Languages' comprehension level

- **The Spanish suspect:** The cropped audio file of the suspect to identify is shown here.
 - Test 1- Voice line-up 1: 6 speakers without background noise:
 - a) Identify the suspect and grade of certainty.
 - b) Identify the least similar voice to the suspect and grade of certainty.

 - Test 1- Voice line-up 2: 6 speakers with background noise:
 - a) Identify the suspect and grade of certainty.
 - b) Identify the least similar voice to the suspect and grade of certainty.

- **The English suspect:** A sample of the selected English suspect is played here.
 - Test 2- Voice line-up 1: 6 speakers without background noise:
 - a) Identify the suspect and grade of certainty.
 - b) Identify the least similar voice to the suspect and grade of certainty.

 - Test 2- Voice line-up 2: 6 speakers with background noise:
 - a) Identify the suspect and grade of certainty.
 - b) Identify the least similar voice to the suspect and grade of certainty.

- **The Dutch suspect:** The voice sample to identify is displayed here.
 - Test 3- Voice line-up 1: 6 speakers without background noise:
 - a) Identify the suspect and grade of certainty.
 - b) Identify the least similar voice to the suspect and grade of certainty.

 - Test 3- Voice line-up 2: 6 speakers with background noise:
 - a) Identify the suspect and grade of certainty.
 - b) Identify the least similar voice to the suspect and grade of certainty.

- **Thank you page.**

After revealing the backbone of the perception surveys, it is worth noting a few observations from it. The sociolinguistic variables *age* and *level of studies* may suggest a ratio scale variable. However, for the purposes of this study, they are deemed as ordinal dichotomous variables, both in *age* (1= 18-22, 2= Over 22) and *level of studies* (1= Up to BA, 2= MA/PhD). Due to convoluted nature of prediction models (accounting for too many variables may wrongly indicate statistical significances where there are none), it is decided to exclude the variables *familiarity with linguistics*, *musical training*, and the individualised levels of linguistic proficiencies, despite the fact that linguistic environments are still considered.

As for the aural-perception language tests, it should be mentioned that the chosen voice samples that introduce the suspect to identify contain speech with flat or descending intonation patterns, whereas all voices presented in the voice line-up itself include instances of rising intonation or uptalk. This serves as an acoustic criterion to differentiate suspects' speech from foils', and thus adds more variety to the body of speakers at the line-up, which gets closer to a more realistic scenario (as opposed to rehearsing pre-selected linguistic items or reciting passages in a certain manner). Needless to mention that the sentences employed in voice samples introducing the suspect and the ones employed when the actual suspect appears in the voice line-up during the 1st experimental condition (target-present) are different for the sake of the perception test's fairness. Participants are explicitly instructed to avoid going back and forth between the six voice line-ups created with the attempt to modify their choices. After completing the online survey, an option is made available to the respondent in the *thank you* page to review their scores through every language test and experimental condition, should they wish to do so.

Language test	Experimental condition	Identification/Discrimination scores (points per correct answer)			
		British group (49 participants)		Spanish group (58 participants)	
			Score		Score
Spanish test Familiar/Learned	1. Target-present (no background noise)	L1	1	F1	1
	2. Target-absent (with background noise)	L2	1	F2	1
English test Familiar/Learned	1. Target-present (no background noise)	F1	1	L1	1
	2. Target-absent (with background noise)	F2	1	L2	1
Dutch test Unknown	1. Target-present (no background noise)	U1	1	U1	1
	2. Target-absent (with background noise)	U2	1	U2	1
Total			6		6

Table 8. Perception language tests and experimental conditions tested on groups of jurors, along with the resulting test score per correct answer.

Table 8 above illustrates the combination of linguistic input (Spanish, English, and Dutch) being exposed to both group of jurors (British and Spanish), as well as the foreseen experimental conditions (1st: target-present without background noises, and 2nd: target-absent with background noise disturbances). As a result, six different aural-perception tests have been compiled to gauge the participants' human auditory skills. When sorting perception tests according to language familiarity and experimental condition, a set of codes have been assigned, depending on theoretical levels of difficulty, from lowest to highest: F1 (Familiar language, 1st condition), F2 (Familiar language, 2nd condition), L1 (Learned language, 1st condition), L2 (Learned language, 2nd condition), U1 (Unknown language, 1st condition), and U2 (Unknown language, 2nd condition). Admittedly, the

same weight is given for every language test (1 point per correct answer), regardless of its difficulty. The resulting variable (overall.score) is then ranked on a scale of 0 to 6, which yields a number for overall success rates in speaker recognition tasks. Moreover, said scale is calculated to assess the two sides of speaker recognition, namely identification and discrimination. Identification tasks are those appearing in the subsection *a*) (Identify the suspect and grade of certainty) above, whereas discrimination tests conform to the letter *b*) (Identify the least similar voice to the suspect and grade of certainty). A summary of the variables provided through the means of online perception surveys is listed hereby, in table 9:

Variables	Options	Codes
Profile		
Age	18-22	1
	Over 22	2
Gender	Male	1
	Female	2
Education level (studies)	Up to BA	1
	MA/PhD	2
Linguistic environment	Monolingual	1
	Bilingual	2
Cultural groups (country)	British	1
	Spanish	2

Tests		
Overall scores	-	0-6
Confidence level (CL)	-	1-10
Identification tests	False alarm	1
(1 st condition):	Miss	1
F1, L1, U1	Hit	2
Identification tests	False alarm	1
(2 nd condition):	Correct rejection	2
F2, L2, U2		
Discrimination tests	False alarm	1
(1 st condition):	Correct rejection	2
F1. Dis, L1. Dis, U1. Dis		
Discrimination tests	Correct rejection	2
(2 nd condition):		
F2. Dis, L2. Dis, U1. Dis		
Excluded from analysis		
Musical training	No	1
	Yes	2
	No previous knowledge	0
Familiarity with linguistics/phonetics	With linguistics	1
	With phonetics	1
	With linguistics and phonetics	2
	A1	1
	A2	2
Language proficiencies	B1	3
(understanding/hearing)	B2	4
	C1	5
	C2	6

Table 9. Variables provided by online perception surveys and their subsequent coding for statistical processing.

As inferred from table 9 above, this research has resorted to dummy coding as a workaround for employing the aforementioned variables in statistical measures which require numerical data in order for them to be computable. There is, however, an

exception in which such data transformation is not needed, as it is the usage of chi-square tests in the first hypothesis' case (see 3.7.2. *Perception surveys-based analysis*).

As for the sociolinguistic factors (profile), these variables' coding does not necessarily entail a hierarchical order, but refer to features which differ categorically (like male/female, British/Spanish, monolingual/bilingual), although it could be argued that *age* (1= 18-22, 2= Over 22) and *studies*' (1= Up to BA, 2= MA/PhD) assigned values do comply with a relationship of higher and lower ranks. Cultural groups, or rather the respective countries from which the data was gathered, are classified in two areas depending on their linguistic environments: monolingual and bilingual. In the Spanish' case, the former considers students from Andalucian universities, whereas the latter refers to bilingual Catalan/Spanish domains (Valencian Community and Catalonia). British jurors, on the other hand, fall into two sociolinguistic and geographical regions: Wales (Swansea, Bangor, and Cardiff), and South East England (Winchester, Roehampton, and Southampton).

Regarding the following variables, their heading *test* encompasses all the information relative to the aural-perception tests themselves. As explained previously, overall scores refer to an indicative measure reflecting the respondents' speaker recognition capabilities through a 0-6 scale. When completing each and every single identification and discrimination test, confidence levels (CL) are also considered, and measured in a scale from the least certain (1) to the most certain (10) attitudes in speaker recognition tasks. Moving on to specific identification and discrimination tests, it is remarkable that some of the options appearing on the table above do not necessarily conform to the model of voice parades' outcomes shown in 2.1.3. *Voice line-ups/Voice parades*, which might be misleading when classifying them into clear-cut categories.

In this thesis, identification and discrimination tests are differentiated by means of what is being requested from the jurors, namely identifying the suspect (identification) and identifying the least similar voice to the suspect's (discrimination). Even though one may be inclined to associate identification tests with target-present (1st condition) scenarios and discrimination tests with target-absent (2nd condition) settings, they display clear distinctions in this research and jurors' responses also vary between them. The most notable difference is the possibility to select *none of the above* in identification tests,

whilst the discrimination side does not allow said option. For this reason, the first perception tests (target-present) in identification include *miss* besides the expected *hit* and *false alarm*. When moving to the second aural-perception tests (target-absent) in this domain, only *false alarm* and *correct rejection* appear, since missing the target (wrongly assuming that the suspect is absent from the line-up) is, by definition, out of the picture. Similarly, discrimination tests in the first experimental condition (target-present) display *false alarm* and *correct rejection* as possible outcomes, since their purpose is to spot the most dissimilar speaker from the suspect (and fail to do so when selecting the suspect himself/herself). The perception tests guaranteeing a 100% of success rates are, by their own nature, discrimination second perception tests (target-absent). This particular case removes the chances for *false alarm* (due to the suspect being absent) and *miss* (since choosing *none of the above* is not an option), and thus participants are bound to select a voice other than the suspect's (correct rejection).

Lastly, and as noted previously, the variables which have been registered through perception surveys but could not be considered due to practical issues are *musical training* (1-2 range), *familiarity with linguistics and/or phonetics* (0-2 range), and *individual language proficiencies* (1-6 scale).

3.7.2. Perception surveys-based analysis

Once the data has been gathered and compiled in an Excel spreadsheet table, the ensuing data set is imported into SPSS. This software is capable of carrying out descriptive as well as inferential statistic tests (which shall be discussed shortly after), apart from creating a variety of data visualisation tools such as graphs, plots, and summaries of data in general. Even though statistical tests described below adapt to the specified research question's needs, it is generally assumed to take sociolinguistic factors and experimental conditions as independent variables which account for the dependent variable's variance (that is, the scores produced by the jurors). The explored statistical relationships are explored in more depth below but, before delving into it, the perception surveys-based analysis' structure is defined to help the reader navigate through each and every hypothesis/research question formulated in section 1.4 (*Hypotheses*):

Hypothesis 1: Language familiarity

Identification

British group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Spanish group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

British and Spanish group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Discrimination

British group

1st perception tests (target-present, without background noises).

Spanish group

1st perception tests (target-present, without background noises).

British and Spanish group

1st perception tests (target-present, without background noises).

Hypothesis 2: Discrimination or identification?

British group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Spanish group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Hypothesis 3: Confidence levels

Identification

British group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Spanish group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

British and Spanish group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Discrimination

British group

1st perception tests (target-present, without background noises).

Spanish group

1st perception tests (target-present, without background noises).

British and Spanish group

1st perception tests (target-present, without background noises).

Hypothesis 4: Age and gender

Identification

British group

Overall scores

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Spanish group

Overall scores

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

British and Spanish group

Overall scores

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Discrimination

British group

Overall scores

1st perception tests (target-present, without background noises).

Spanish group

Overall scores

1st perception tests (target-present, without background noises).

British and Spanish group

Overall scores

1st perception tests (target-present, without background noises).

Hypothesis 5: Cultural groups and linguistic environment

Cultural groups

Identification

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Discrimination

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Linguistic environment

Identification

British group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Spanish group

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Discrimination

British group

1st perception tests (target-present, without background noises).

Spanish group

1st perception tests (target-present, without background noises).

Hypothesis 6: Background noises and false alarms

Identification

British group

Spanish group

Discrimination

British group

Spanish group

Epilogue: Level of studies

Identification

British group

Overall scores

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Spanish group

Overall scores

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

British and Spanish group

Overall scores

1st perception tests (target-present, without background noises).

2nd perception tests (target-absent, with background noises).

Discrimination

British group

Overall scores

1st perception tests (target-present, without background noises).

Spanish group

Overall scores

1st perception tests (target-present, without background noises).

British and Spanish group

Overall scores

1st perception tests (target-present, without background noises).

Once the overall structure of the statistics section is unveiled, similar patterns can be discerned in terms of hierarchical order across the formulated hypotheses. As a general rule, identification language tests shall precede discrimination tasks, and the target population's exploration is arranged in alphabetical order (British group before Spanish jurors). Nevertheless, both group of participants are considered together in several research questions, depending on the requirements thereof. The specificities posed for each hypothesis in regard to content order and required statistical measures are described in detail hereafter.

Hypothesis 1: Language familiarity

As a reminder, the first hypothesis aims to discover whether 'aural-perceptual recognition is enhanced as the familiarity of the juror with the language exposed also increases' (*1.4. Hypotheses*), which takes into account previous research findings:

- 'Unfamiliarity with the target language affects the ability to recognize a speaker' (Köster & Schiller 1995: 181).
- 'If no linguistic information on the target language is understood, recognition results are poorer' (Köster & Schiller 1997: 25).

The adopted form of hypothesis testing is broken down into two steps to address this research question: a first glimpse of the variables' distribution through a Friedman two-way analysis of variance by ranks, and a post-hoc analysis which comprises a set of chi-square tests, relying on contingency tables and Phi coefficient/Cramer's V values for a richer understanding of language familiarity's correlations with language test scores, or the lack thereof.

Firstly, the Friedman two-way analysis of variance by ranks test is the non-parametric alternative to an ANOVA for related or dependent samples, which calculates 'whether the rank totals for each condition/treatment differ significantly from the values which would be expected by chance' (Pereira et al. 2015: 2638). In this case, the assessed conditions are the categorical language familiarity levels (familiar, learned, and unknown) influencing the dichotomous response types (1= false alarm/miss, 2= hit/correct rejection). In order to comply with the requirement of samples being related,

these tests are computed separately, both for cultural groups (British and Spanish), and experimental conditions (1st and 2nd perception tests). Hence, the variables shall be coded accordingly: British- *language1*response1*, *language2*response2*; and Spanish- *language1*response1*, and *language2*response2*.

After checking for potential correlations between language tests and types of responses, chi-square tests follow with a cross-tabulation of the aforementioned categorical variables. To attest their significance within this table, a great emphasis is placed on adjusted residuals whose critical values (Z-score) exceed the -1.96/1.96 range (for a 95% confidence level) and surpass the expected count, which in turn are deemed statistically significant (with a subsequent Bonferroni correction adjusting said confidence intervals). As Cabin & Mitchell (2000) point out, Bonferroni corrections are applied in occasions where ‘two or more tests [...] address a common null hypothesis’ (246), as it is the case with the current research question: three distinct language tests (familiar, learned, and unknown languages) testing their relevance on speaker recognition scores.

A standard rule of thumb apropos the validity of Pearson’s chi-square tests for cross-tabulations larger than 2x2 is that each observation should be independent of all the others (i.e. one observation per subject) and that ‘no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater’ (Yates et al. 1999: 734). Also, Phi’s coefficient (ϕ) is used in 2x2 contingency tables, while Cramer’s V is preferable in tables with more rows and/or columns. The pre-existing conventions for describing and interpreting the magnitude of association in contingency tables are described in table 10 below:

Value of ϕ or Cramer’s V	Description
0.00- 0.10	Negligible association
0.10- 0.20	Weak association
0.20- 0.40	Moderate association
0.40- 0.60	Relatively strong association
0.60- 0.80	Strong association
0.80- 1.00	Very strong association

Table 10. Types of association depending on ϕ or Cramer’s V values. Adapted from Rea & Parker (1992: 203).

After compiling the necessary information from a Friedman two-way analysis of variance by ranks, chi-square tests and their cross-tabulations, and Phi's coefficient/Cramer's V, the analysis shall proceed to observe the next stratum in the experiment with the order established above: Identification tasks come first and discrimination tests are placed afterwards. Cultural groups examined in alphabetical order (British, Spanish) followed by a more generalised approach (British and Spanish group analysis), and first experimental conditions (target-present without background noises) prevailing over the second ones (target-absent with background noises). A notable exception for the latter aspect occurs while tackling discrimination aural-perception tests. In such particular cases, the second experimental condition is not considered for the lack of variance in its values, given the test's nature which equates all values to correct rejections (as the suspect is absent from the voice line-up and thus cannot be selected).

Hypothesis 2: Discrimination or identification?

This research question ventures into unraveling whether discrimination and identification tasks differ in terms of efficiency across the cultural groups of jurors surveyed. The quotations below shall be consulted to explain the current state of affairs further:

- 'In real line-ups the reluctance to incriminate an innocent person is probably more than balanced by the suggestion that the police must have rather strong leads to go to the trouble of a voice line-up' (Thompson 1987: 126).
- 'While familiar voices invoke a discrimination process which can be likened to a pattern recognition task, the perception of unfamiliar voices taps into one which involves feature analysis' (Hollien 2002: 59).
- 'Discrimination of an unfamiliar voice heard during a short period of time is possible, [but] the ability to identify it after a short retention interval is null' (Manzanero & Barón 2017: 59).

As cited above, cognitive biases may overtake the jurors' reluctance towards speaker recognition tasks (Thomson 1987: 126), which might end up increasing false alarm rates. However, it has not been studied whether the explicit instructions from the researcher's side (to identify the suspect or to identify the non-suspects) could influence the outcome. Also, literature focuses on setting apart familiar from unfamiliar voice recognition, the

latter leading to higher likelihood for errors, since naïve speakers are not trained to perceive and extract such acoustic-phonetic features accurately and make voice judgements as a result (Hollien 2002: 59). It is also reminded of the negative effects that delay has on hearers' retention intervals, along with the possible psychological traumas stemming from the involvement in criminal acts, which hamper speaker recognition abilities (Manzanero & Barón 2017: 59).

It is certain that previous experiments (Kerstholt et al. 2004, Smith & Baguley 2014) have researched both target-present and target-absent conditions, but this thesis contribution also lies in implementing and exploring an additional dimension of analysis through the information retrieved from the jurors. Given the ambiguity of terms employed which refer to target-absent conditions and discrimination tasks as identical entities (see 3.7.1.1. *Structure and design*), identification and discrimination tests are differentiated here according to their purpose in the current research: either to identify the suspect in the lineup or the absence thereof (identification), or to identify the least similar voice to the suspect's (discrimination).

To inspect this research question, descriptive statistics shall be drawn insofar as mean test scores become relevant. After a brief observation, a Wilcoxon signed-rank tests shall consider whether the score tests observed for each language test pair and stratum are significantly different or not. Due to the fact that this is a non-parametric statistical measure which examines related samples, the spreadsheet arranges the data in such a way that it follows the same order as established above: British group identification-discrimination tests comparison in the 1st experimental condition, followed by the 2nd perception tests, Spanish inspection of the same pair-wise language tests comparisons in the 1st tests, and a final account on the 2nd experimental condition. In contrast with hypothesis 1, test scores do not need to be grouped together in this research question, since a Wilcoxon signed-rank test allowing for unrelated samples (British and Spanish) to converge would lead to questionable results.

Hypothesis 3: Confidence levels

This research question explores whether the jurors' self-perceived confidence level affects their speaker recognition capabilities in any way (either positively or negatively). Naturally, it is based on previous research:

- 'The witness identified the suspect with a high degree of confidence [...] and a conviction resulted' (Nolan 2001: 7).
- 'The confidence score of the witness had no predictive value for the accuracy of his or her judgment' (Kerstholt et al. 2004: 335).
- 'For both *heard previously* and *not heard previously* responses, there was a trend toward increasing accuracy as a function of increasing listener certainty' (Papcun et al. 1989: 913).

As the quotes above demonstrate, there are two contrasting views when establishing whether self-perceived confidence levels affect the outcome of language perception tests, and thus this issue should be addressed given its relevance in real-life court cases, as Nolan (2001) notes above. As for the stance adopted in hypothesis 3, it has been decided to follow Papcun et al.'s (1989) results as a starting point, since it foresees both familiar (heard previously) and unfamiliar (not heard previously) voices, even though only the latter are considered in the present study. It should be reminded, though, that forensic linguistics research attempts to isolate each variable to account for their contribution to speaker recognition, but, as Clifford (1980: 390) claims, other factors (psychological, physiological) are at play and thus it is advised caution when believing the witness' or victim's testimony.

To address this issue, a Kendall's tau test is employed in order to unearth possible correlations between confidence levels (CL) and test scores (both identification and discrimination in all experimental conditions and all cultural groups). Kendall's rank correlation coefficient 'evaluates the degree of similarity between two sets of ranks given to a same set of objects' (Abdi 2007: 1), and is also considered the non-parametric alternative to Pearson correlation coefficient. As discussed in 3.7.1.1. (*Structure and design*), the variable *CL* makes use of a 10-point Likert scale which ranges from *not confident* (1) to *highly confident* (10). Test scores, too, are deemed ordinal variables due

to the fact that lower scores (1) represent false alarm/miss, whereas higher scores (2) stand for hits/correct rejections. Thus, the arrangement of paired items (each participant provides one *CL* and one test score per language test) fulfils the conditions for statistical analysis.

All the possible combinations of scenarios are explored in the following order: type of test (identification over discrimination), groups of jurors (British, Spanish, and British and Spanish), and experimental conditions (target-present before target-absent). It must be noted, however, that target-absent tests are not considered in discrimination tasks, since they exhibit no variance in their values and thus *CL* cannot account for it.

Hypothesis 4. Age and gender

- ‘The influence of the listeners’ age on the performance in speaker recognition remains rather unclear’ (Schiller & Köster 1996: 181).
- ‘It stands out that when it comes to women trying to identify a woman’s voice, false alarms reach a rate of 100% when the line-up does not show the target voice’ (Manzanero & Barón 1996: 59).

Notwithstanding the unclear influence that *age* exerts over success rates, it is revealed that false identifications (false alarm) increase in accordance with age, whilst false rejections (miss) are more common in younger hearers (Schiller & Köster 1996: 182). As the generational gap widens, another study (Ohman et al. 2013) discerned a slight increase in children’s (11-13 years old) success rates when performing aural-perceptual tests immediately, in contrast with adults’ responses. In the present study, the two error types (false positive/false alarm and false negative/miss) are grouped together so that the pair of items renders dichotomous variables pointing at either success (2) or failure (1).

Despite females’ significant same-gender tendency for false alarms in target-absent conditions, Manzanero & Barón (1996: 59) clarify that their experiment did not find any significant correlation between gender (considering both jurors’ and informants’) and the outcome of the line-up itself, which is reinforced by the findings of previous studies (Yarmey 1995). In this regard, there does not seem to be detrimental features inherent to the jurors’ perception, but issues seem to arise in the procedure (how the voice parade is

set up) and underlying cognitive biases (jurors' tendency to select a culprit regardless of their presence/absence), which appear to condition the results to a higher extent.

To address this research question, a linear fixed effects model is used to spot the extant relationships in our data set in terms of a function:

test scores~ age+gender

In the function above, *test scores* stand for the dependent variable, whereas *age* and *gender* (and possibly their interaction term: *age*gender*) are treated as fixed effects which account for the variance of the former. It could be very well decided to conduct a linear mixed effects model instead, but the current research design does not contemplate multiple responses per subject (only 1 score per language test and experimental condition at a time) which removes the need of inter-dependent values for random effects (Winter 2013: 2).

Before reaching the figures and statistical measures concerned with statistical significance, a brief account on descriptive statistics shall consult the distribution of age and gender across the target population to get a clearer picture of how it is stratified. Once this stage is cleared, p-values and estimates of age and gender (and their interaction term whenever appropriate) shall pinpoint relevant interactions amongst the studied variables.

As for the order in which each specific scenario is scrutinised, it follows the same hierarchy as established in previous hypotheses regarding type of test (identification and discrimination), culture groups (British, Spanish, and British and Spanish), and experimental conditions (1st and 2nd). A notable exception occurs here in the type of language test performed, and thus prediction models are drawn first for overall scores (all scores for each cultural group irrespective of language familiarity), with a subsequent inspection on each individual language test (familiar, learned, and unknown). As for discrimination tests, they follow the same pattern, except that the target-absent test (2nd experimental condition) is removed from the equation due to its null variance, much in line with the first hypothesis (language familiarity).

Hypothesis 5: Cultural groups and linguistic environment

The fifth hypothesis seeks to find out if cultural groups (British or Spanish) alongside linguistic environments (monolingual and bilingual) influence the participants' speaker recognition skills. Its formulation (that cultural groups and linguistic environments do not impinge upon speaker recognition) is based on the following quotes:

- 'There does not seem to be evidence that recognition performance is correlated with typological difference' (Köster & Schiller 1997: 181).
- 'The strengths and weaknesses [...] exhibited by the human auditory system are then discussed, as are elements (acoustic and otherwise) related to the environment and the nature of the speaker' (Hollien 2002: 22).

As inferred, Köster & Schiller (1997) theorise that exposure to voices coming from languages categorically different from the hearers' linguistic background could possibly hamper their performance in recognition tasks, just as the differences between hearer-target language increase. However, this theory is refuted after proving that a group of Spanish, German, English, and Chinese subjects did not show apparent differences in their aural-perceptual skills.

As for linguistic environments, Hollien (2002) highlight acoustic interferences on the speech signal for the most part (background noises), but also assess the speaker's training on perceiving and remembering voices. Even though linguistic environment is a feature more related to language acquisition studies, it is worthwhile to consider it alongside cultural groups for the sake of broadening our views on speaker perception and recognition.

As for the variables themselves, four possible combinations of cultural groups and linguistic environments are drawn in total: British monolingual, British bilingual, Spanish monolingual, and Spanish bilingual. Far from a perspective on language acquisition, this research does not intend to measure each individual's linguistic skills, but rather takes into account the environment in which jurors are immersed. This is, of course, not without its concerns on whether a monolingual/bilingual setting mirrors in actuality a speaker's language proficiency living in segregated sociolinguistic areas. However, this study shall

consider linguistic environment only as an external influence shaping the hearer's speaker perception and recognition skills through continuous exposure to the speech community's input, instead of claiming a modification of the speakers' linguistic production per se.

Opposed to hypothesis 2, this scenario looks specifically at test scores originating from divergent groups, since the data set allows such mixture. To address this, a Mann-Whitney U test is employed, which is similar to the Wilcoxon signed-rank tests, only that the former deals with independent or unrelated samples and thus complies with how the data set is arranged in this research question. In reporting this test's results, it is added an additional statistical measure: The absolute value of the Pearson product-moment coefficient (r), which is similar to Phi's coefficient and Cramer's V employed in hypothesis 1. It describes the magnitude (strength) of the relationship between variables, and thus the coefficient's sign (- or +) indicates the direction of said relationship (Cohen 1988: 75-107). Its power can be classified according to the resulting value obtained:

- Small effect size: $r = 0.10$
- Medium effect size: $r = 0.30$
- Large effect size: $r = 0.50$

This measure of effect size, or r , is calculated by dividing Z by the square root of N ($r = Z / \sqrt{N}$), as Field & Hole (2003: 235) assert.

Hypothesis 6: Background noises and false alarms

The sixth hypothesis claims that background noises and acoustic distortions may have detrimental effects on speaker recognition capabilities, which could lead to higher counts of false alarms. This stance is sustained by the following principle:

- 'A noisy environment will both mask and otherwise degrade speech' (Hollien 2002: 47).

Not only environmental factors come into play in this domain, but also technical elements (tape recorder's input or audio quality) may play a role, too. Audio material degraded in

such a way may in turn negatively affect the hearer's judgement. Nevertheless, the voice samples gathered in this study have been encoded with the same audio qualities (see 3.5. Recordings for further details). As for wrongly identifying a speaker as the culprit, Kerstholt et al. (2006) report high chances of scoring false alarms (in contrast with success rates) even in the absence of noise disturbances, since the jurors 'cannot approach the [speaker recognition] task unbiased' (Nolan 2001: 8). It is surmised, and therefore put forward for hypothesis testing, that the addition of noises may facilitate the appearance of false alarms.

To approach the matter, language tests are sorted by experimental condition (target-present with no noise disturbances, and target-absent condition with background noises) in a way that the data set allows the researcher to compare noiseless voice line-ups (F1, L1, and U1) with their noisy counterparts (F2, L2, and U2). This language test pairwise comparison is carried out through several Wilcoxon signed-rank tests, much in line with Hypothesis 2 (identification or discrimination?). In order to avoid uneven distribution of values amongst the two experimental conditions, it has been decided to code the line-up's outcome as either success or failure. Hence 1st condition tests options (Hit/Miss/False alarm) are reduced to a binary response variable (1= Miss/False alarm, 2= Hit), while 2nd language tests are coded similarly (1= False alarm, 2= Correct rejection).

As specified in the structure above, identification tasks precede discrimination ones. As for culture group's order, both the British and Spanish group are examined simultaneously, although results shall be discussed starting with the British group. As inferred, no generalised account (British and Spanish group) is required in this case.

Epilogue: Level of studies

This epilogue is a follow-up study of hypothesis 4 (*age and gender*), which comes to include the jurors' level of studies as well (hence hypothesis 4.1.). The reason for separating such statistical models lies in the fact that adding a third variable could end up compromising the representativeness of the sample and, consequently, the results themselves. This overly stratified model would display large distances between groups, and some strata may not even be represented. As such, it is best practice to account for both models, should they provide meaningful insights, or even be complementary in their

analyses. The rationale behind *studies*' addition comes from Nolan's (2001) commentary on technical speaker identification, which surpasses naïve speaker recognition due to the trained phonetician's 'advantage in bringing to consciousness, and being able to organise, evaluate, and communicate, delicate distinctions of pronunciation' (p. 9). The expert's opinion is presumably less prone to errors in general, although they do face their own challenges when testifying in court (see 2.1.2.1. *Forensic phonetics* for more details).

As for the statistic measure utilised, a linear fixed effects model is employed here, but the resulting formula is somewhat modified:

test scores~ age+gender+studies

With the addition of *studies* to the formula, more combinations of fixed effects yield various interaction terms (*age*gender*, *age*studies*, *gender*studies*, and *age*gender*studies*). However, the aforementioned interaction variables will only be incorporated to the model provided that they are statistically significant at the 0.05 level. As for the hierarchical order, it follows the same trend as the one shown in hypothesis 4 (*age and gender*).

3.7.3. Acoustic-phonetic analysis

The current acoustic-phonetic analysis displays methodological differences with respect to the previous research tradition. Starting with acoustic-phonetic analyses aiming at investigating levels of inter- and intra-speaker variability through voice parameters, the recorded samples selected for analysis normally involve controlled segmental and/or suprasegmental variables, regardless of the corpus' nature (Cicres 2007). As for the present thesis' objects of study, the considered length of voice samples in this study amounts to roughly 3 minutes of audio per subject (or pairs of speakers), but the excerpts used as stimuli in the voice line-up were considerably shortened (4-14 sec.). When creating individual samples out of the aforementioned long sound objects, each informants' speech was isolated by removing any other speaker in the interaction. In this regard, researcher's control over the produced acoustic-phonetic units by the speaker is given up in return for recreating a realistic scenario (in conditions close to naturally-

occurring speech) whereby phoneticians cannot possibly influence disputed and undisputed pieces of evidence (voice samples in this case), as opposed to the scenario where the constituents of the line-up read aloud a pre-defined text (which typically contains a balanced sample of the segmental variables of interest).

From the aural-perception side, previous literature detects an enhancement of identification accuracy when known sentences to the juror are used as stimuli (Hollien 2002: 23), while some studies use recorded materials for the voice line-up involving reading pre-selected passages out loud in a target language (Köster & Schiller 1997). Nevertheless, the above would imply a higher degree of control over the informants' voice samples provided, and thus the exchange would not be as genuine as a semi-spontaneous interaction.

As for the acoustic-phonetic analysis undertaken here, the proposed analytical procedure involves the identification of acoustic-phonetic features and their subsequent extraction. In some cases, specific voice parameters are preferably extracted by means of built-in Praat commands and scripts designed to certain ends. However, as the two upcoming subsections show, several phonetic units of measurement were manually extracted (e.g. VOT).

One of the adjustments needed for the present thesis is the enhancement of some audio files (see *Appendix 4* for full details) through Camtasia Studio. Had it been left unaltered, there would have been clear differentiations amongst certain constituents of the line-up in perceptual terms, which would have created a disadvantageous scenario for some of them. This situation would threaten the fairness of the line-up itself and would ultimately contradict the principles outlined in the guidelines of voice line-ups (Broeders & van Amelsvoort 1999, 2001; De Jong-Lendle et al. 2015, and Hollien 2012). Nevertheless, such an intervention might have changed some acoustic-phonetic units' values in comparison with their original version with lower audio volume levels, which makes the subsequent forensic voice comparison all the more challenging (since, in theory, said voice samples have become less distinguishable and more similar to each other).

Similar to the response-based approach adopted in perception surveys, the length of voice samples for acoustic-phonetic analysis remains under 20 seconds. Ideally, the audio files'

duration should range from 1 to 3 minutes (or longer) for the sake of extracting a representative sample of the speakers' regular linguistic habits (Baldwin & French 1990: 45, Nolan 1983: 13). However, this thesis aims to provide the same conditions for both the perceptual study (jurors' judgement) and the trained ear's setting (examining informants' recordings through specialised software) for comparison purposes. In this sense, informants' speech samples are limited to instances of uptalk/rising (distractors) and falling intonation (suspects). As a methodological tool with the aim of reducing the so-called researcher bias, a series of Praat scripts (see *Appendix 6* for full details) are run so as to ensure the obtainment of objective results:

- Syllable Nuclei v2.praat: Extracts the following suprasegmental variables related to measures of pausing: duration of pauses, number of pauses per minute, percentage of pauses per excerpt, speech rate, articulation rate, and ASD (Average Syllable Duration). The script was originally created by De Jong & Wempe (2008), but it has been further improved by Hugo Quené, Ingrid Persoon, & Nivja de Jong (17/09/2010).
- draw_pitch_histogram_from_sound.praat: Extracts F0 basic statistics like min./max./mean pitch, and 25%-75% quantiles. Additionally, it draws a histogram according to every pitch point found in the audio file, and those are saved separately in a plain text file (Lennes 2013).
- zero-crossing-and-spectral-moments (v. 1.3.).praat: This script is purposefully designed for the analysis of fricative consonants. As its name suggests, it gathers a set of measures concerned with zero crossings and spectral moments, while also including the duration of labeled intervals, max. frequencies, and min./max./mean intensities, among others (Elvira-García 2014).
- get F1, F2, F3 (averages).praat: This Praat script calculates the average values of F1-F3 within all the intervals specified in the files (TextGrid and audio files) inside a folder (Kawahara 2010).

Regarding acoustic-phonetic segmentation, the SAMPA transcription offered by the UCL puts forward a generic guideline on graphemes indexing phonetic units applicable to six different languages (English, Danish, Spanish, German, French, and Swedish) while also displaying language-specific units, including Dutch, English, and Spanish, among others

(Wells 1997). As for the Dutch data set employed for this thesis, the IFADV corpus (2007) already contains a set of aligned txt files which provide phonetic transcriptions of the recordings available at the word and phonetic level.

A few special cases where segmental information appears omitted or merged with the immediate linguistic environment are worth mentioning here as common phonological phenomena emerging in casual speech. For instance, Simon T. Elliott's elision of [d] at the end of the word *sound* is motivated by the following word's (*different*) initial consonant. Hence, the resulting phonetic transcription: ['saʊn 'dɪfənt]. Additionally, note that *different* is pronounced ['dɪfənt] instead of ['dɪfrənt], thus dropping the [r] in the middle. This observation brings us to the next point, which is that the speaker's actual pronunciation will not always necessarily match what is conceptualised to be the normative pronunciation of the language, as it is the RP (Received Pronunciation) for English. Consequently, the phonetic transcriptions provided in this study reflect such idiosyncrasies appearing in naturally occurring speech at the phonetic level (tier). On the other hand, the tier at the word level provides the orthographically correct version of the lexical item, as it would be *sound* and *different* in the example discussed above.

3.7.3.1. Suprasegmental features

The main contrastive units of analysis contemplated for this study in the suprasegmental realm relate to prosodic features pivoting around the notion of intonation. By modelling the user's pitch, intonation creates a supra-lexical effect which may signal syntactic information (e.g. statements, questions, etc.) or emotional attitudes (anger, disappointment, etc.) (Rose 2002: 150). In line with Lindh's voice line-up (2009), this thesis also considers suprasegmental features related to speech tempo, for they constitute effective parameters in identifying and discriminating speakers (Künzel 1997). The set of studied variables is put together hereby:

- Global pitch (Hz). This category takes into account the 25%, 50%, and 75% quantiles of pitch, and the mean pitch value (\overline{P}) found throughout the recording. The script created by Lennes (2013) was run to gather the values of the aforementioned variables. It has been decided to neglect the max./min. pitch

values that said script provides, given the inaccuracies that may emerge within this semi-spontaneous data set. Hence the reason for choosing 25%-75% quantiles.

- Global sound intensity (dB) comprises intensity peaks/max. intensity (I_{\uparrow}), valleys/min. intensity (I_{\downarrow}), and the mean value (\bar{I}) of intensity values registered in the whole excerpt. In this case, a manual selection and extraction of values seems more appropriate, since the researcher is able to overlook unnecessary, but nearly unavoidable environmental noises whose influence may be reflected in the output window.
- Pausing: Despite referring to unvoiced fragments, this broad category also includes measures related to phonation/silent times (speech rate, articulation rate, and ASD), aside from those units concerned with the pauses per se (Lindh 2009: 188). Numerical values can be extracted from the audio files through the Praat script *Syllable Nuclei v2* (De Jong & Wempe 2008), thus creating the following variables as a result:
 - DurPaus: Pauses duration per minute.
 - N_Paus/min: Number of pauses per minute.
 - Pause_ %: Percentage of pauses per excerpt.
 - N_paus: Number of pauses per excerpt.
 - Speech rate: As De Jong & Wempe (2009) assert, it is a measure of fluency which is commonly applied in studies related to second language acquisition (p. 385), and it is calculated with the following formula: n° of syllables/total duration of the excerpt.
 - Articulation rate: Calculates produced syllables per second (n° of syllables/phonation time). This unit's influence has also been recognised in intonation patterns such as tonal alignment (Cicres 2007).
 - ASD (Average Syllable Duration): As the term suggests, this unit measures the average duration per syllable ($\text{phonation time}/n^{\circ}$ of syllables).

Mietta Lennes' (2013) script on F0 measures defines a threshold of 80-400 Hz by default. However, it could be argued that such range could be modified for a higher accuracy in the extraction of acoustic units of measurement, given that the English group contains male voices, whereas females are recorded in the Spanish and Dutch corpora. As a result,

recordings for males have been re-adjusted with a 75-300 Hz range, while females' threshold is set at 100-500 Hz, as Boersma (2019) suggests. Besides that, it is worth mentioning that only mean values are considered for the purposes of this research, which leaves out the extra features that said script provides, like individual pitch points and visual representations thereof through histograms.

3.7.3.2. Segmental features

Plosive consonants are heeded to in the acoustic-phonetic analysis related to segmental phenomena. Stops or plosives are oral occlusives, which is indicative of an obstruction of the airflow in the vocal tract. Unlike the high variance of vowels and diphthongs encountered across languages and dialects, consonant plosives are relatively standardised in all languages as voiced ([b, d, g]) and voiceless plosives ([k, p, t]), with the minor exception of an exclusive application of [g] in loan-words in the Dutch phonetic system (e.g. goal [go:l]) (Wells 1997).

Specifically, the concept VOT (Voice-Onset Time) refers to the time elapsed between the release of a stop consonant at the obstruction point and the beginning of voicing through a periodic vibration in the larynx (Yao 2007: 183). Studies like Whiteside et al.'s (2004) research on British speakers suggest differences across gender groups with their voiceless plosives VOT due to physiological differences and sociophonetic factors. Aspects such as speaker-related factors like age, gender, speaking rate, place of production, and lung volume; and non-speaker-related factors like phonetic context or environment (Yao 2007) can be found as potential correlates of VOT.

The values of the variables concerned with VOT shown below were gathered manually due to the fact that an automatic extraction would imply the addition of an extra tier with further segmentations (marking boundaries between the obstruction and the release burst of the consonant, and narrowing down the interval for the release burst itself) which could compromise the readability of an otherwise straightforward TextGrid. The considered acoustic cues for differentiating the individualised use of plosives are listed hereby:

- VOT (Voice Onset Time). Time difference between obstruction and release burst.
- The intensity of the release burst (UCL 2003).

VOT values are normally positive for voiceless plosives [k, p, t], whereas [b, d, g] display negative values (or near-zero values) in the Spanish language (Soto-Barba 1999: 128). In English, voiced plosives' VOT values follow the same trend, albeit with higher variability than in the Spanish language (Clegg & Fails 2017: 253). Dutch realisations of [b, d] tend to occur earlier than what is reported for English speakers (even though it still displays negative values), whilst [g] is rather realised by the voiceless velar stop [k], as explained by Lisker & Abramson (1964: 391). Again, the Dutch language displays a minor exception in the case of [g] in this respect, which is concerned with the use of loan-words (Wells 1997).

Regarding the voiceless plosives, [p, t]'s VOT is slightly longer in Dutch (10-15ms) than in (Puerto Rican) Spanish (4-9ms), and yet [k] has the opposite effect (Cho & Ladefoged 1999: 208). Contrastively, English VOT values are typically longer due to the occasional presence of aspiration in [k, p, t], but ultimately such differences are conditioned by other factors such as 'aerodynamics, articulatory movement velocity, and differences in the mass of the articulators' (Cho & Ladefoged 1999: 209).

Additionally, the voiceless alveolar sibilant [s] is also added due to its greater overall discriminatory power compared to other consonants in Argentinian-Spanish (Univaso et al. 2014: 120), and in endangered languages such as Hupa, Scottish Gaelic, Aleut or Apache, among others (Gordon et al. 2002). Also, its voiced counterpart [z] is incorporated to the acoustic-phonetic analysis of English voice samples, due to the shortage of [s] in some of the speakers within said group of informants. Despite acknowledging that the realisations of [s] are influenced by temporal variation and coarticulation (Koenig et al. 2013:1180), amongst other factors, the current data set does not allow for differentiations between fricative types depending on their position within a lexical item (initial, mid, and final position). Rather, all segmental units are accounted for irrespective of their immediate linguistic environment. Here are listed the chosen parameters for analysis:

- Spectral peak location. It measures the highest point in frication noise. This variable has been reported to signal speaker-specific behaviour and is also dependent on the properties of the following vowel (Jongman et al. 2000: 1253).
- Spectral center of gravity (Spectral COG). It is typically applied in descriptive studies of fricative characteristics by weighting the overall noise produced (the higher the COG values, the higher the likelihood of the sound being placed at the front of the mouth) (Styler 2017: 29).
- Noise duration. It also has been proven to be effective in distinguishing distinctive phonetic traits (Jongman et al. 2000: 1255).
- Noise amplitude. It measures the mean frication noise in dB (Jongman et al. 2000: 1254).
- Formants F1-F2-F3⁸ also appear to yield promising results in speaker discrimination tasks, especially F1 and F3 (Univaso et al. 2014: 115). Rather than examining individual points of data with a Long-Term Formant Distribution (LTF) approach (Nolan & Grigoras 2005), this thesis attempts to extract the means of the aforementioned formants, since monitoring their progression may be troublesome given the unrepresentativeness that short voice samples entail.

A specific script (Elvira-García 2014) was run to calculate fricatives' spectral COG, noise duration (interval duration), and noise amplitude (mean intensity). F1, F2, and F3 mean values were extracted through Kawahara's (2010) script. Regarding spectral peak location, the procedure specified by Jongman et al. (2000: 1255) was adopted, which uses a 25ms Hamming window with the standard pre-emphasis of 6.0 (db/octave) along with LPC (Linear Predictive Coding) and FFT (Fast Fourier Transform) spectral slices to estimate where the high-frequency peaks may be.

⁸ Please note that the notion of formants in this case adopts Univaso et al.'s (2014) definition and is thus conceived as a concentration of energy in the spectrogram. This distinction is crucial to be made since fricative consonants are inharmonic by definition (Martínez 2007: 69), which contradicts the very purpose of a formant (to measure the acoustic resonances of the vocal tract).

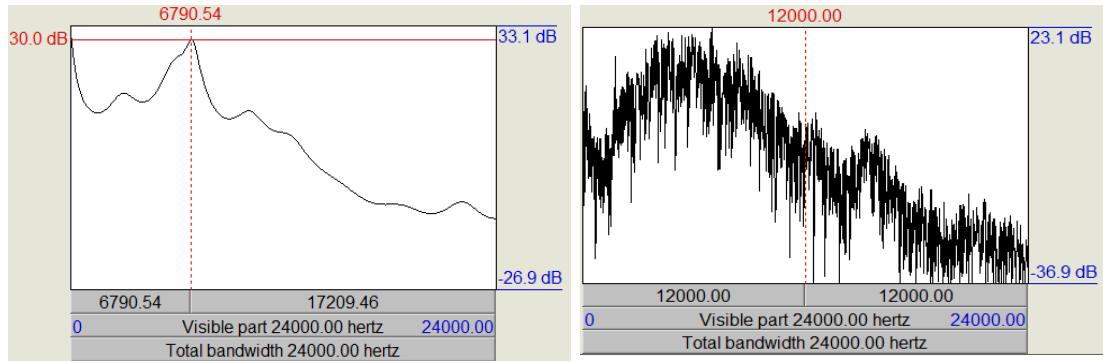


Figure 12. LPC slice (left side) and FFT slice (right side) of [s] in the word *Huis*, Dutch speaker (DVA10-F19P).

Instead of cropping the slices through the excerpt (Styler 2017: 25), Praat objects' interface is consulted to extract individual segmental units ([s] and [z] in this case), which allows for pairwise comparisons between the created spectral slices and the original spectrogram to best guess the highest peak of frication noise, as it typically occurs in the middle of the consonant (Jongman et al. 2000: 1255). As discerned in figure 12, the peaks do not necessarily match with utmost accuracy, but the tendency may remain around the midpoint of the frication noise.

CHAPTER 4

RESULTS:

PERCEPTION SURVEYS-BASED

ANALYSIS

The next chapter refers back to the formulated hypotheses (1.4.) concerned with the data originated from perception surveys. Therefore, this section covers from hypothesis 1 to the epilogue dedicated to the jurors' level of studies. The sub-chapters start with the formulation of the null hypothesis (H_0) and the alternative hypothesis (H_x), an orientating summary of the cultural groups, experimental conditions and types of recognition tasks (and their hierarchies), a reminder of the studied variables, and the distribution thereof within the target population, whenever appropriate. The considered statistical measures are mentioned thereafter, with tables and figures displaying significant correlations amongst the variables analysed. After the statistical processing, results for every research question are discussed and shown separately, as each scenario requires different statistical tests to answer said hypotheses. Therefore, they appear in the following order: language familiarity (4.1.), discrimination or identification? (4.2.), confidence levels (4.3.), age and gender (4.4.), cultural groups and linguistic environment (4.5.), background noises and false alarms (4.6.), and the epilogue on jurors' level of studies (4.7.).

4.1. LANGUAGE FAMILIARITY

The first hypothesis examines the relationships of familiarity/unfamiliarity of the exposed language with jurors' speaker recognition capabilities (both identification and discrimination of speakers). Hence, the null hypothesis (H_0) is formulated alongside the research hypothesis (H_1):

- H_0 : Jurors' familiarity with the language exposed does not affect hearers' aural-perceptual recognition capabilities.
- H_1 : Aural-perceptual recognition is enhanced as the familiarity of the juror with the language exposed also increases.

As explained previously in the methodology section (3.7.2. *Perception surveys-based analysis*), the established order to address this research question begins with identification tests, cultural groups (British, Spanish, and British and Spanish group), and experimental conditions (1st target-present with no noise disturbances, and 2nd target-absent with background noises condition). After each section is cleared, the process shall restart from the very beginning, analysing discrimination tasks this time around.

Let us now refer to the analysed variables in this particular research question in the table below:

Categorical variable	Categories	Code
language1	Familiar	1
language2	Learned	2
All.language1		
All.language2	Unknown	3
Dis.language		
response1	False alarm	1
All.response1	Miss	1
	Hit	2
response2	False alarm ⁹	1
All.response2	Correct rejection	2
Dis.response		

Table 11. Categorical variables and their assigned codes for hypothesis 1.

For clarification purposes, the variables *language1/2* and *response1/2* listed in table 11 both refer to identification tasks, whereas *Dis.language* and *Dis.response* allude to discrimination tests. As for the inclusion of *All* before the variable, it reflects to the scenario where perception scores are put altogether (British and Spanish group analysis), whereas removing it (*language/response*) refers to separated analysis (either British or Spanish groups alone). To simplify the matter, discrimination tasks do not incorporate *All* when both British and Spanish groups are being considered altogether, but employ the generic label *Dis.language* and *Dis.response*, given the little variability of values that exists between comparing isolated cultural groups and accounting them together.

As could be inferred, the numbers placed after *language/response* reveal the experimental conditions in which they take place: 1 (target-present with no background noises) and 2 (target-absent with background noises). Hence, it is seen that language familiarity is invariably the same regardless of the language test. In this case, dummy coding arranged the relationship of familiarities according to their increasing difficulty: familiar (1), learned (2), and unknown (3) language. As for the types of responses in *(All)response1*, identifying the target is coded as 2, whereas missing or producing a false alarm is coded

⁹ Due to its design, second language (target-absent) discrimination tasks cannot yield false alarms, as explained in 3.7.1.1. (*Structure and design*). Only in this specific case, correct rejections would be the only possible choice.

as *I*. Likewise, the target-absent condition codes 2 as a correct rejection, and *I* as a false alarm. It must be noted at this point that target-absent identification tasks (*(All)response2*) and the first discrimination tests (target-present) offer the same possibilities in terms of voice line-up's outcome.

Since this hypothesis is undertaken through two distinct statistical processes, it should be stressed that only the aforementioned dummy coding is used for the first test (Friedman two-way analysis of variance by ranks test), whereas the ordinal categorical variables worded as in the second column (table 11) are included in the second analytical stage (chi-square tests, phi coefficients, and contingency tables).

4.1.1. Identification

This section explores the relationships of language familiarity with types of responses and outcomes of a voice line-up through identification tasks (identifying both the suspect and/if the suspect is absent from the line-up). As pointed out already, it follows the established order for cultural groups (British, Spanish, and British and Spanish) and experimental conditions (target-present with no noise disturbances, and target-absent with background noises).

4.1.1.1. British group

To start with the British group's 1st experimental condition (target-present with clear sound), the analysis shall proceed to the first step to, conducting a Friedman two-way analysis of variance by ranks test.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of language1 and response1 are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Table 12. Friedman’s two-way analysis on *language1* and *response1* for hypothesis 1 (British group, identification tests).

After discerning a significant correlation in table 12, the null hypothesis cannot be retained, and thus is rejected for the moment. A post-hoc analysis is performed hereafter to explore this interaction of variables even further.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	13.263 ^a	4	.010
Likelihood Ratio	13.460	4	.009
N of Valid Cases	147		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 8.00.

Table 13. Chi-square test for hypothesis 1 (British group, identification tests, 1st exp. condition).

As table 13 demonstrates, a chi-square test of independence has found a significant correlation between the categorical variables *language1* and *response1* ($\chi^2(4, N = 147) = 13.263, p < 0.05$).

Symmetric Measures^c

		Value	Approximate Significance
Nominal by Nominal	Phi	.300	.010
	Cramer's V	.212	.010
N of Valid Cases		147	

c. Correlation statistics are available for numeric data only.

Table 14. Cramer's V and Phi coefficient for hypothesis 1 (British group, identification tests, 1st exp. condition).

Since the resulting table is greater than 2x2, Cramer's V appearing in table 14 is selected to assess the magnitude of such correlation, which reflects a moderate association (0.21) of the values involved.

response1 * language1 Crosstabulation

		language1			Total	
		familiar	learned	unknown		
response1	False alarm	Count	23 _a	10 _b	23 _a	56
		Expected Count	18.7	18.7	18.7	56.0
		% within response1	41.1%	17.9%	41.1%	100.0%
		Adjusted Residual	1.6	-3.1	1.6	
	Hit	Count	21 _a	25 _a	21 _a	67
		Expected Count	22.3	22.3	22.3	67.0
		% within response1	31.3%	37.3%	31.3%	100.0%
		Adjusted Residual	-.5	.9	-.5	
	Miss	Count	5 _a	14 _a	5 _a	24
		Expected Count	8.0	8.0	8.0	24.0
		% within response1	20.8%	58.3%	20.8%	100.0%
		Adjusted Residual	-1.4	2.8	-1.4	
Total	Count	49	49	49	147	
	Expected Count	49.0	49.0	49.0	147.0	
	% within response1	33.3%	33.3%	33.3%	100.0%	

Each subscript letter denotes a subset of language1 categories whose column proportions do not differ significantly from each other at the .05 level.

Table 15. Contingency table for hypothesis 1 (British group, identification tests, 1st exp. condition). Statistically significant adjusted residuals are marked in bold ($\alpha = 0.05$).

A contingency table is generated in table 15, which reveals the found interactions between the sub-categories within *response1* and *language1*. After applying a Bonferroni correction for the pairwise comparisons, SPSS reveals that false alarms in the learned language are statistically different (-3.1) from the familiar (1.6) and unknown (1.6) language tests, while hits do not differ substantially across the three language tests.

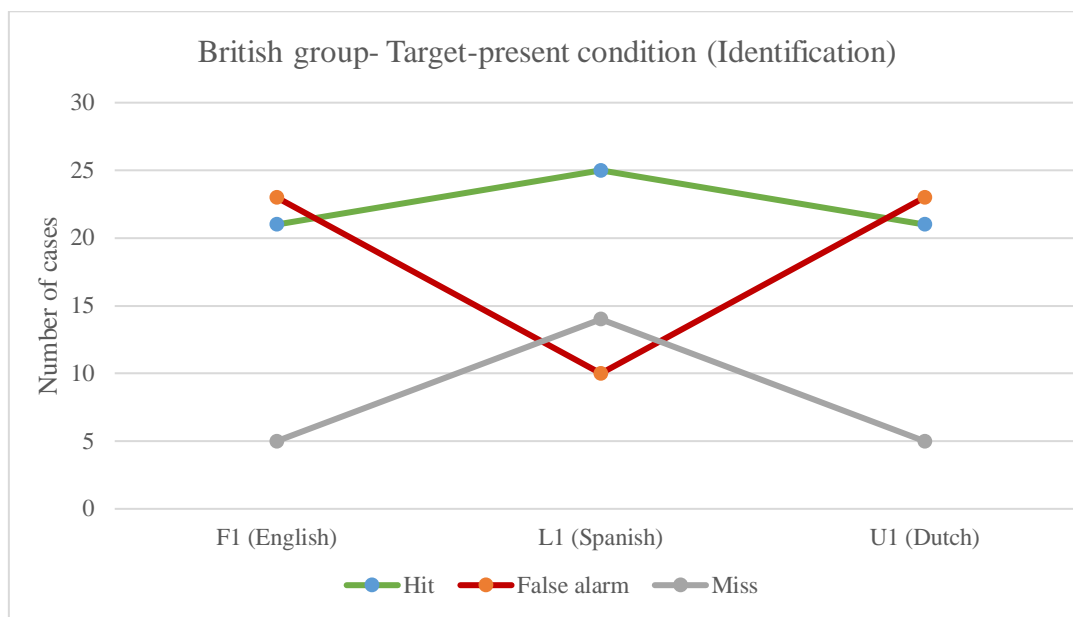


Figure 13. British group's count of response types across three language familiarities in the first experimental condition (target-present, identification tests).

As shown in figure 13, L1's (first learned language test) count of false alarms is significantly inferior to those in F1 (first familiar language test) and U1 (first unknown language test).

Lastly, the miss category appears to differ substantially in learned languages (2.8) from familiar (-1.4) and unknown (-1.4) languages, but a Bonferroni correction confirms that such differences amongst these pairs are not relevant enough in statistical terms, even if the adjusted standardised residual values highlighted in the contingency table surpass the critical value range of -1.96/1.96 (at the 0.05 alpha level of significance). Hence, it is concluded that the amount of false alarms generated in the learned language test are far below the expected values.

Moving to the second experimental condition (target-absent with background noises), a Friedman two-way analysis of variance by ranks test is performed.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of language2 and response2 are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	,000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Table 16. Friedman’s two-way analysis on *language2* and *response2* for hypothesis 1 (British group, identification tests).

After checking for *language2* and *response2*’s distribution of values through table 16’s results, the null hypothesis cannot be accepted. A post-hoc analysis investigates this correlation further.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	6.214 ^a	2	.045
Likelihood Ratio	6.599	2	.037
N of Valid Cases	147		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.33.

Table 17. Chi-square test for hypothesis 1 (British group, identification tests, 2nd exp. condition).

As seen in table 17, a chi-square test of independence has found a significant correlation between the categorical variables *language2* and *response2* (χ^2 (2, N = 147) = 6.214, $p < 0.05$).

Symmetric Measures^c

		Value	Approximate Significance
Nominal by Nominal	Phi	.206	.045
	Cramer's V	.206	.045
N of Valid Cases		147	

c. Correlation statistics are available for numeric data only.

Table 18. Cramer's V and Phi coefficient for hypothesis 1 (British group, identification tests, 2nd exp. condition).

To assess the magnitude of such correlation, the subsequent Phi coefficient and Cramer's V value displayed in table 18 reflect a moderate association (0.206) amongst the analysed variables.

responses2 * language2 Crosstabulation

			language2			
			familiar	learned	unknown	Total
response2	Correct rejection	Count	11 _{a, b}	15 _b	5 _a	31
		Expected Count	10.3	10.3	10.3	31.0
		% within response2	35.5%	48.4%	16.1%	100.0%
		Adjusted Residual	.3	2.0	-2.3	
	False alarm	Count	38 _{a, b}	34 _b	44 _a	116
		Expected Count	38.7	38.7	38.7	116.0
		% within response2	32.8%	29.3%	37.9%	100.0%
		Adjusted Residual	-.3	-2.0	2.3	
Total	Count	49	49	49	147	
	Expected Count	49.0	49.0	49.0	147.0	
	% within response2	33.3%	33.3%	33.3%	100.0%	

Each subscript letter denotes a subset of language2 categories whose column proportions do not differ significantly from each other at the .05 level.

Table 19. Contingency table for hypothesis 1 (British group, identification tests, 2nd exp. condition). Statistically significant adjusted residuals are marked in bold ($\alpha = 0.05$).

After the Bonferroni correction is applied in table 19, correct rejections and false alarms are deemed statistically different in the learned-unknown pairs, whilst familiar-learned and familiar-unknown language tests display no significant differences. Albeit subtle, the tendency for correct rejection to occur within the familiar category is favoured slightly (0.3), whereas false alarms' critical value is slightly below the expected (-0.3). The same trend can be observed in learned language tests, although this time it does reach significance with critical values surpassing -1.96/1.96, thus yielding more correct rejections (2.0) and less false alarms (-2.0) than the expected count. For the unknown test results, however, it tends to diminish the appearance of correct rejections (-2.3) while favouring the rate of false alarms (2.3).

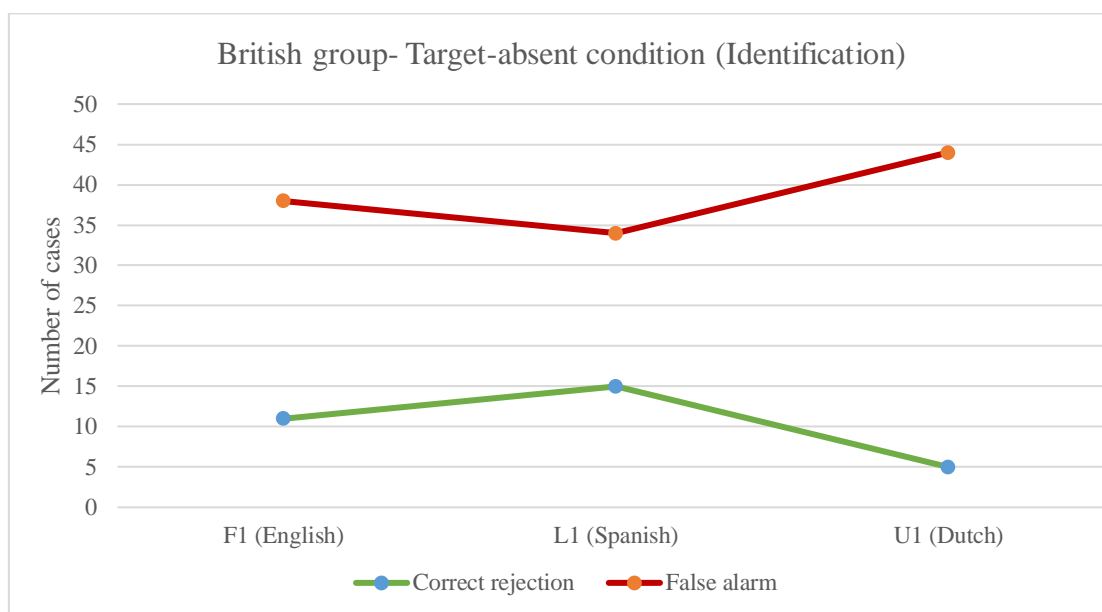


Figure 14. British group's count of response types across three language familiarities in the second experimental condition (target-absent, identification tests).

In this regard, unknown language tests seem to produce the least auspicious outcome with values exceeding the critical value (Correct rejection: -2.3, False alarm: 2.3). On the other hand, this condition is reversed in the learned language test, and therefore correct rejections are enhanced (2.0) whilst false alarms are minimised (-2.0), as illustrated in figure 14. As for the familiar language test, its values fall somewhere in between the other two perception tasks, as indicated by sharing the same letter (*a* and *b*) in the crosstabulation above.

4.1.1.2. Spanish group

After completing the analysis for British first and second perception tests, the Spanish group of jurors follow, and thus this sub-section proceeds to unearth the possible correlations between language familiarity and voice line-up’s outcome in identification tasks. As discussed, this analysis deals first with the 1st experimental condition (target-present with no background noises). In this regard, a Friedman two-way analysis of variance by ranks test constitutes the first analytical step.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of language1 and response1 are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Table 20. Friedman’s two-way analysis on *language1* and *response1* for hypothesis 1 (Spanish group, identification tests).

After glancing at table 20, it seems that the variables *language1* and *response1* are indeed correlated, much in line with British group’s findings. However, the post-hoc analysis should unveil whether the relationships found in the previous group of jurors is equivalent to the one being investigated here.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	8.672 ^a	4	.070
Likelihood Ratio	8.772	4	.067
N of Valid Cases	174		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 6.00.

Table 21. Chi-square test for hypothesis 1 (Spanish group, identification tests, 1st exp. condition).

As reported in table 21, a chi-square test of independence has found a near-significant trend between the categorical variables *language1* and *response1* ($\chi^2(4, N = 174) = 8.672$, $p = 0.07$).

Symmetric Measures^c

		Value	Approximate Significance
Nominal by Nominal	Phi	.223	.070
	Cramer's V	.158	.070
N of Valid Cases		174	

c. Correlation statistics are available for numeric data only.

Table 22. Cramer's V and Phi coefficient for hypothesis 1 (Spanish group, identification tests, 1st exp. condition).

Since the crosstabulation is larger than a 2x2 table, Cramer's V assesses the magnitude of such correlation, which yields a value reflecting a weak association (0.15) of the aforementioned variables, as seen in table 22 above.

response1 * language1 Crosstabulation

		language1			Total	
		familiar	learned	unknown		
response1	False alarm	Count	21 _a	34 _b	21 _a	76
		Expected Count	25.3	25.3	25.3	76.0
		% within response1	27.6%	44.7%	27.6%	100.0%
		Adjusted Residual	-1.4	2.8	-1.4	
	Hit	Count	31 _a	18 _b	31 _a	80
		Expected Count	26.7	26.7	26.7	80.0
		% within response1	38.8%	22.5%	38.8%	100.0%
		Adjusted Residual	1.4	-2.8	1.4	
	Miss	Count	6 _a	6 _a	6 _a	18
		Expected Count	6.0	6.0	6.0	18.0
		% within response1	33.3%	33.3%	33.3%	100.0%
		Adjusted Residual	.0	.0	.0	
Total	Count	58	58	58	174	
	Expected Count	58.0	58.0	58.0	174.0	
	% within response1	33.3%	33.3%	33.3%	100.0%	

Each subscript letter denotes a subset of language1 categories whose column proportions do not differ significantly from each other at the .05 level.

Table 23. Contingency table for hypothesis 1 (Spanish group, identification tests, 1st exp. condition). Statistically significant adjusted residuals are marked in bold ($\alpha = 0.05$).

After applying the Bonferroni correction in table 23, it appears remarkable that hits and false alarms are statistically different between the learned-familiar and learned-unknown pairs. The critical values exceeding the -1.96/1.96 threshold reflect an increased tendency in the learned language for false alarm rates (2.8) and a noticeable reduction concerning hit rates (-2.8).

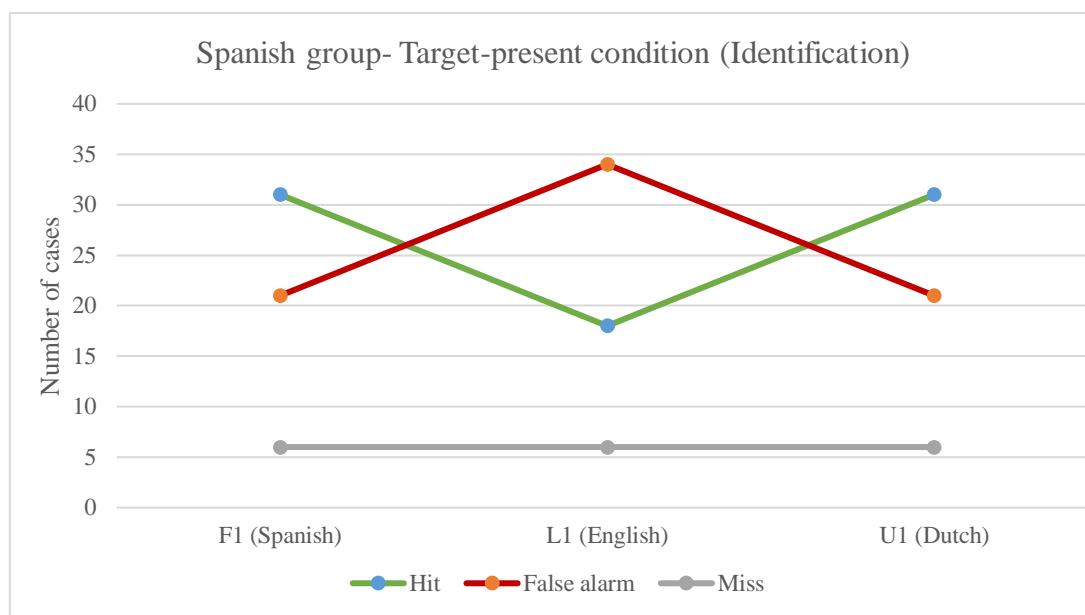


Figure 15. Spanish group’s count of response types across three language familiarities in the first experimental condition (target-present, identification tests).

The opposite applies to familiar and unknown language tests, which seem prone to produce higher hit rates (both scoring 1.4), thus reducing false alarm rates with identical negative scores (-1.4.). As for the probability to miss the target, it does not display significant differences among the three language tests, which is represented in figure 15 above.

After scrutinising the extant relationships between *language1* and *response1* in the Spanish group identification tasks, these variables are also examined in the second tests (target-absent with background noises). As a starting point, a Friedman two-way analysis of variance by ranks test is performed.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of language2 and response2 are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Table 24. Friedman’s two-way analysis on *language2* and *response2* for hypothesis 1 (Spanish group, identification tests).

From table 24 above, it would appear that in the target-absent scenario, *language2* and *response2* are also correlated. The following post-hoc analysis looks at such relationship in more depth.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	.447 ^a	2	.800
Likelihood Ratio	.451	2	.798
N of Valid Cases	174		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 13.67.

Table 25. Chi-square test for hypothesis 1 (Spanish group, identification tests, 2nd exp. condition).

A chi-square test of independence has not found a significant correlation between the categorical variables *language2* and *response2* (χ^2 (2, N = 174) = 0.447, $p > 0.05$), as table 25 above demonstrates.

Symmetric Measures^c

		Value	Approximate Significance
Nominal by Nominal	Phi	.051	.800
	Cramer's V	.051	.800
N of Valid Cases		174	

c. Correlation statistics are available for numeric data only.

Table 26. Cramer's V and Phi coefficient for hypothesis 1 (Spanish group, identification tests, 2nd exp. condition).

In light of these results, the Phi coefficient and Cramer's V reflect an expected negligible association between the categorical variables (0.05), which is reflected in table 26.

response2 * language2 Crosstabulation

			language2			
			familiar	learned	unknown	Total
response2	Correct rejection	Count	15 _a	12 _a	14 _a	41
		Expected Count	13.7	13.7	13.7	41.0
		% within response2	36.6%	29.3%	34.1%	100.0%
		Adjusted Residual	.5	-.6	.1	
	False alarm	Count	43 _a	46 _a	44 _a	133
		Expected Count	44.3	44.3	44.3	133.0
		% within response2	32.3%	34.6%	33.1%	100.0%
		Adjusted Residual	-.5	.6	-.1	
Total		Count	58	58	58	174
		Expected Count	58.0	58.0	58.0	174.0
		% within response2	33.3%	33.3%	33.3%	100.0%

Each subscript letter denotes a subset of language2 categories whose column proportions do not differ significantly from each other at the .05 level.

Table 27. Contingency table for hypothesis 1 (Spanish group, identification tests, 2nd exp. condition).

As inferred by looking at table 27, the number of correct rejections and false alarms across the three language groups is not statistically significant. There is no correlation between familiarity of the language and type of response (adjusted residuals are far from the -1.96/1.96 range).

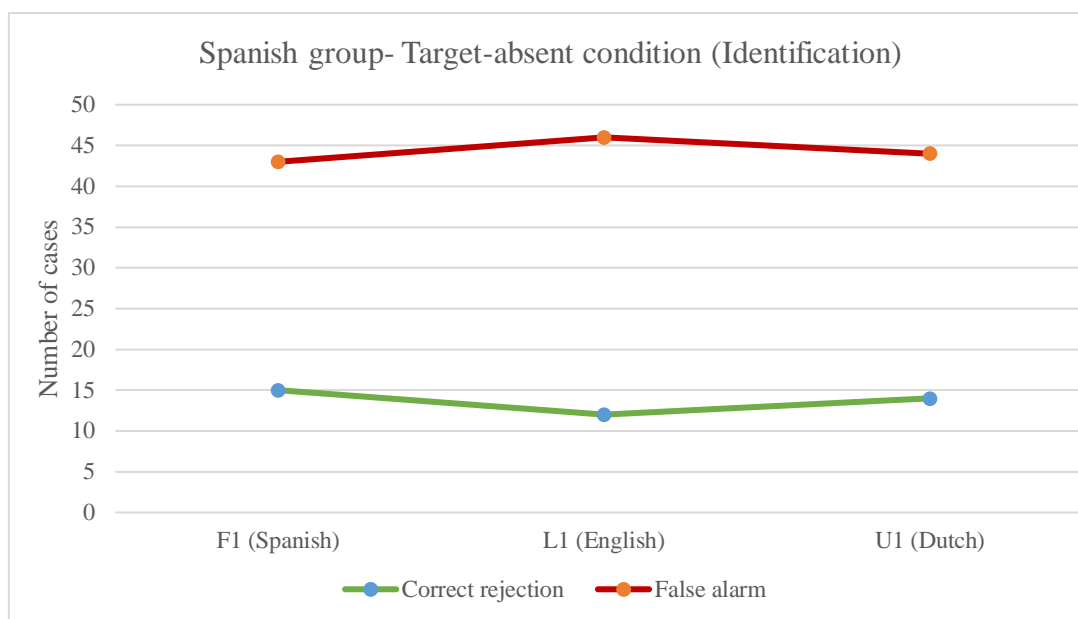


Figure 16. Spanish group’s count of response types across three language familiarities in the second experimental condition (target-absent, identification tests).

However, correct rejections are more common in the familiar (adjusted Residual=0.5), and in the unknown language tests (adjusted Residual=0.1), while the learned language test gets a negative value (adjusted Residual=-0.6). This trend can be easily recognised by looking at figure 16.

4.1.1.3. *British and Spanish group*

After analysing British and Spanish group’s test scores (response types) separately, this last sub-section merges the results from both groups to provide an overview of speaker recognition abilities in identification tasks, regardless of the juror’s cultural group. Likewise, target-present (1st tests) conditions shall precede target-absent (2nd tests) scenarios. As is customary, a Friedman two-way analysis of variance by ranks test is consulted as an initial step.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of All.response1 and All.language1 are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	,000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Table 28. Friedman’s two-way analysis on *All.language1* and *All.response1* for hypothesis 1 (British and Spanish group, identification tests).

The hypothesis test summary described in table 28 above considers that the shown distributions of *All-response1* and *All.language1* are statistically differentiated. To discover how significant this correlation between language familiarity and type of response is in the first identification tests, a post-hoc analysis is undertaken.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	4.959 ^a	4	.292
Likelihood Ratio	4.783	4	.310
N of Valid Cases	321		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 14.00.

Table 29. Chi-square test for hypothesis 1 (British and Spanish group, identification tests, 1st exp. condition).

As shown in table 29 above, a chi-square test of independence could not find a significant correlation between the categorical variables *All.language1* and *All.response1* (χ^2 (4, N = 321) = 4.959, $p > 0.05$).

Symmetric Measures^c

		Value	Approximate Significance
Nominal by Nominal	Phi	.124	.292
	Cramer's V	.088	.292
N of Valid Cases		321	

c. Correlation statistics are available for numeric data only.

Table 30. Cramer’s V and Phi coefficient for hypothesis 1 (British and Spanish group, identification tests, 1st exp. condition).

Predictably, Cramer’s V value in table 30 reflects a negligible association (0.08). Due to the apparent contradiction between Friedman’s test and chi-square values, it is worth referring to contingency tables with the aim of improving our understanding of the correlations at hand.

All.responses1 * All.languages1 Crosstabulation

		All.language1			Total	
		familiar	learned	unknown		
All.response1	False alarm	Count	44 _a	44 _a	44 _a	132
		Expected Count	44.0	44.0	44.0	132.0
		% within All.response1	33.3%	33.3%	33.3%	100.0%
		Adjusted Residual	.0	.0	.0	
	Hit	Count	52 _a	43 _a	52 _a	147
		Expected Count	49.0	49.0	49.0	147.0
		% within All.response1	35.4%	29.3%	35.4%	100.0%
		Adjusted Residual	.7	-1.4	.7	
	Miss	Count	11 _a	20 _a	11 _a	42
		Expected Count	14.0	14.0	14.0	42.0
		% within All.response1	26.2%	47.6%	26.2%	100.0%
		Adjusted Residual	-1.1	2.1	-1.1	
Total	Count	107	107	107	321	
	Expected Count	107.0	107.0	107.0	321.0	
	% within All.responses1	33.3%	33.3%	33.3%	100.0%	

Each subscript letter denotes a subset of All.language1 categories whose column proportions do not differ significantly from each other at the .05 level.

Table 31. Contingency table for hypothesis 1 (British and Spanish group, identification tests, 1st exp. condition). Statistically significant adjusted residuals are marked in bold ($\alpha = 0.05$).

By inspecting the crosstabulation shown in table 31 above, there is no statistical distinction of response types across the three language tests, as expected due to the non-significant p-value. The weak Cramer's V association is discernible, however, in the learned language, where hits are placed below the expected values (-1.4), at the expense of missing the target more frequently (2.1). Even if missing exceeds the set critical values (1.96/-1.96), these differences become non-significant after applying the Bonferroni correction, as implied in the table above by sharing the same subscript letter (a). As for false alarm rates, their count is identical across the three language tests.

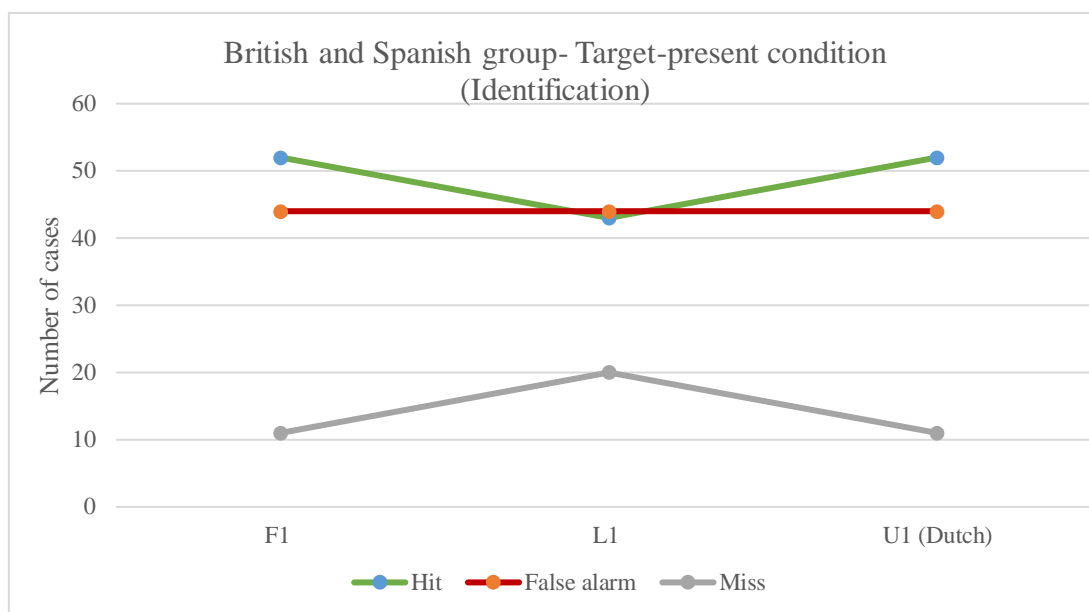


Figure 17. British and Spanish group's count of response types across three language familiarities in the first experimental condition (target-present, identification tests).

To get a clearer picture of the correlations shown in the contingency table, figure 17 illustrates quite clearly how miss rates are above the average, if we take familiar (F1) and unknown (U1) language tests as the baseline. In spite of perceiving subtle differences between learned language tests and familiar/unknown perception surveys in terms of failing or succeeding at the identification task, their correlation does not amount to statistical significance.

Nevertheless, the study turns its attention now towards the second experimental condition (target-absent with background noises) to prove whether the same pattern observed in the first test results' is repeated when putting British and Spanish group's results together.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of <i>All.language2</i> and <i>All.response2</i> are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	,000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Table 32. Friedman’s two-way analysis on *All.language2* and *All.response2* for hypothesis 1 (British and Spanish group, identification tests).

After conducting a Friedman two-way analysis of variance by ranks test, rejecting the null hypothesis seems necessary insofar as the distributions of *All.language2* and *All.response2* are concerned. Once this is confirmed by consulting table 32 above, let us proceed to the second analytical phase of this condition with a post-hoc analysis.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2.041 ^a	2	.360
Likelihood Ratio	2.099	2	.350
N of Valid Cases	321		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 24.00.

Table 33. Chi-square test for hypothesis 1 (British and Spanish group, identification tests, 2nd exp. condition).

As seen in table 33, a chi-square test of independence has not found a significant correlation between the categorical variables *All.language2* and *All.response2* (χ^2 (2, N = 321) = 2.041, $p > 0.05$).

Symmetric Measures^c

		Value	Approximate Significance
Nominal by Nominal	Phi	.080	.360
	Cramer's V	.080	.360
N of Valid Cases		321	

c. Correlation statistics are available for numeric data only.

Table 34. Cramer’s V and Phi coefficient for hypothesis 1 (British and Spanish group, identification tests, 2nd exp. condition).

As a result, table 34 above on symmetric measures reports a scarce Cramer’s V value, which reflects a negligible association (0.08).

All.response2 * All.language2 Crosstabulation

		All.language2			Total	
		familiar	learned	unknown		
All.response2	Correct rejection	Count	26 _a	27 _a	19 _a	72
		Expected Count	24.0	24.0	24.0	72.0
		% within All.response2	36.1%	37.5%	26.4%	100.0%
		Adjusted Residual	.6	.9	-1.4	
	False alarm	Count	81 _a	80 _a	88 _a	249
		Expected Count	83.0	83.0	83.0	249.0
		% within All.response2	32.5%	32.1%	35.3%	100.0%
		Adjusted Residual	-.6	-.9	1.4	
Total	Count	107	107	107	321	
	Expected Count	107.0	107.0	107.0	321.0	
	% within All.response2	33.3%	33.3%	33.3%	100.0%	

Each subscript letter denotes a subset of All.language2 categories whose column proportions do not differ significantly from each other at the .05 level.

Table 35. Contingency table for hypothesis 1 (British and Spanish group, identification tests, 2nd exp. condition).

As expected due to the non-significant p-value, table 35 shows no statistical distinction on the response type across the three language tests. The negligible association stemming from Cramer’s V is seen on the deviation of unknown language test scores in relation to the other two perception tests.

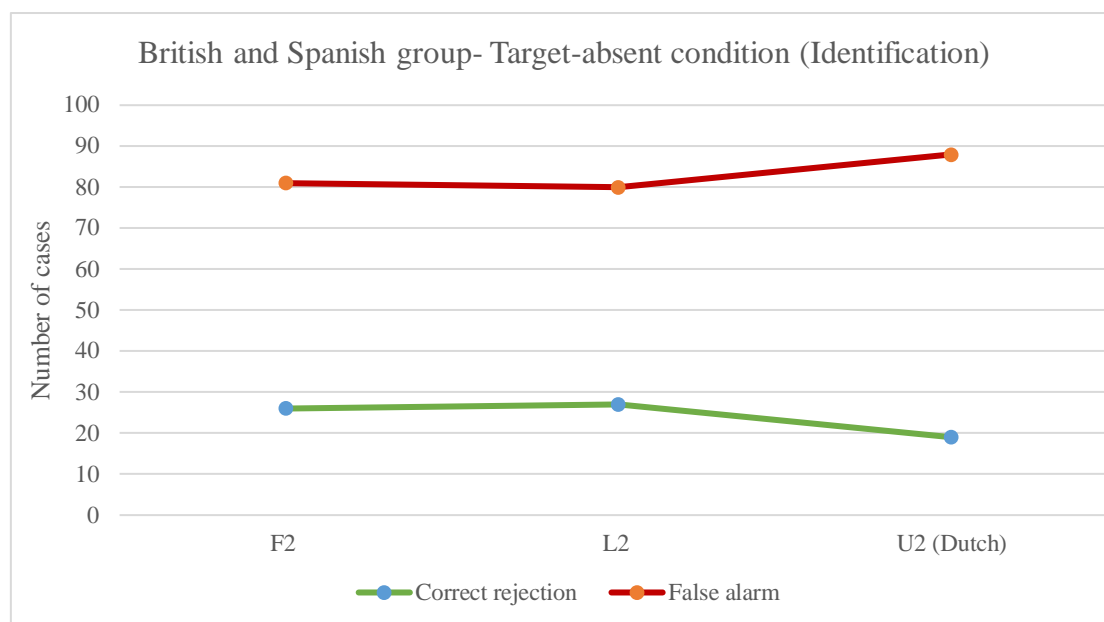


Figure 18. British and Spanish group’s count of response types across three language familiarities in the second experimental condition (target-absent, identification tests).

Albeit non-significant, the target-absent tests tend to increase correct rejection rates from the learned (0.9) and familiar (0.6) languages, while those decrease when perceiving unknown linguistic input (-1.4), as figure 18 illustrates. It is concluded hereby that neither the first (target-present) nor the second (target-absent) experimental condition yield significant correlations between language familiarity and voice line-up’s outcome when putting together the results for both cultural groups in identification tasks.

4.1.2. Discrimination

In a similar vein to identification tasks’ analysis, this sub-section seeks to find out whether language familiarity affects a voice line-up’s outcome. Unlike identification tests, discrimination tasks request participants to look for the most dissimilar voice to the suspect’s. Due to the fact that second language perception discrimination tests (target-absent) do not allow for false alarms to occur, only the first experimental condition (target-present with clear sound) is considered throughout this sub-section. As for cultural group’s order, it follows the same structure as in the previous analysis: British, Spanish, and British and Spanish.

4.1.2.1. British group

Let us start with a Friedman two-way analysis of variance by ranks test on British group’s language familiarity (*Dis.language*) and response types (*Dis.response*) in discrimination tests.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of <i>Dis.response</i> and <i>Dis.language</i> are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	,680	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Table 36. Friedman’s two-way analysis on *Dis.language* and *Dis.response* for hypothesis 1 (British group, discrimination tests).

At this stage, the null hypothesis can be retained, as it seems that language familiarity and responses (correct rejections and false alarms) are not influencing one another, as discerned in table 36 above. To explore this matter further, the crosstabulation of values shall be consulted below, along with chi-square tests and Cramer’s V:

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	5.383 ^a	2	.068
Likelihood Ratio	6.166	2	.046
N of Valid Cases	147		

a. 3 cells (50.0%) have expected count less than 5. The minimum expected count is 1.67.

Table 37. Chi-square test for hypothesis 1 (British group, discrimination tests, 1st exp. condition).

In this case, the chi-square test of independence reflects a near-significant p-value (χ^2 (2, N = 147) = 5.383, p= 0.068). However, its validity is compromised since there are 3 cells (50%) displaying less numbers than the expected count (5), as table 37 reports.

Symmetric Measures^c

		Value	Approximate Significance
Nominal by Nominal	Phi	.191	.068
	Cramer's V	.191	.068
N of Valid Cases		147	

c. Correlation statistics are available for numeric data only.

Table 38. Cramer's V and Phi coefficient for hypothesis 1 (British group, discrimination tests, 1st exp. condition).

Additionally, Cramer's V reports a weak association (0.19) amongst the variables *Dis.language* and *Dis.response*, according to the resulting output seen in table 38.

Dis.response * Dis.language Crosstabulation

		Dis.language			Total	
		familiar	learned	unknown		
Dis.response	Correct rejection	Count	45 _a	49 _a	48 _a	142
		Expected Count	47.3	47.3	47.3	142.0
		% within Dis.response	31.7%	34.5%	33.8%	100.0%
		Adjusted Residual	-2.3	1.6	.6	
		Adjusted Residual	2.3	-1.6	-.6	
Dis.response	False alarm	Count	4 _a	0 _a	1 _a	5
		Expected Count	1.7	1.7	1.7	5.0
		% within Dis.response	80.0%	0.0%	20.0%	100.0%
		Adjusted Residual	2.3	-1.6	-.6	
		Adjusted Residual	2.3	-1.6	-.6	
Total	Count	49	49	49	147	
	Expected Count	49.0	49.0	49.0	147.0	
	% within Total	33.3%	33.3%	33.3%	100.0%	
	Dis.response					

Each subscript letter denotes a subset of Dis.language categories whose column proportions do not differ significantly from each other at the .05 level.

Table 39. Contingency table for hypothesis 1 (British group, discrimination tests, 1st exp. condition). Statistically significant adjusted residuals are marked in bold ($\alpha = 0.05$).

As seen in table 39 above, the reason for a near-significant p-value lies in the higher adjusted residuals for the familiar language, displaying less correct rejections (-2.3) and more false alarms (2.3) in comparison with the other tests. As discerned in the line-graph

below (figure 19), such variation of values in the familiar language test are minimal, and thus do not reflect significant differences amongst the perception tests in terms of succeeding or failing the discrimination task.

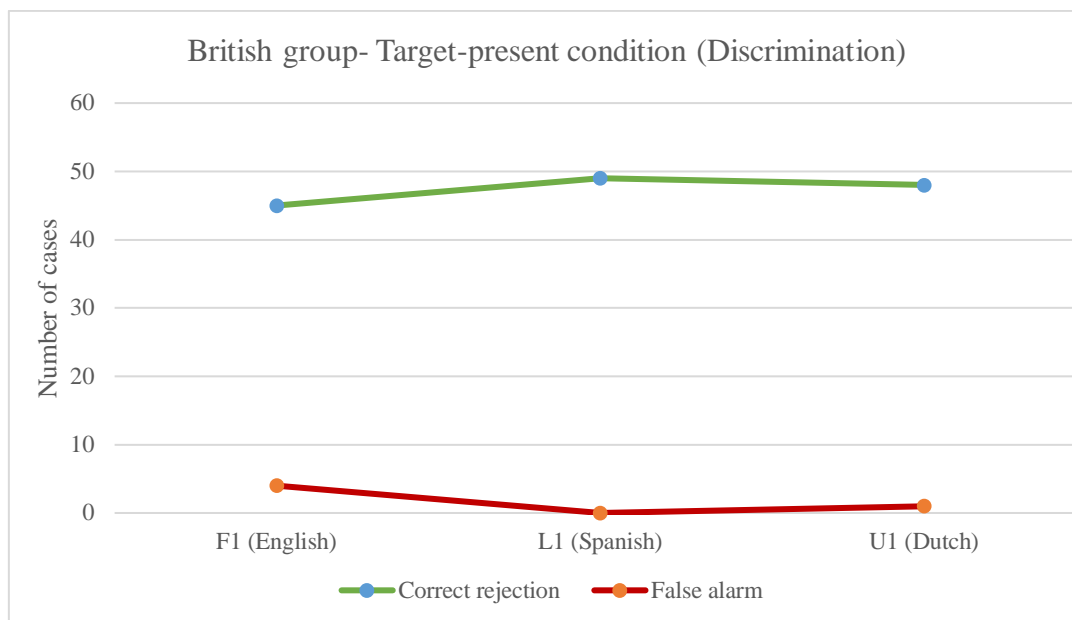


Figure 19. British group’s count of response types across three language familiarities in the first experimental condition (target-present, discrimination tests).

Nevertheless, the low significance level alongside the breach of chi-square assumptions (that each individual cell expected value should be larger than 5) invalidate the previous observations. As for the second experimental condition, this calculation is skipped since all responses would inevitably lead to correct rejections (absent suspect).

4.1.2.2. Spanish group

Similar to the British analysis on discrimination tests, the Spanish group also considers the first experimental condition (target-present) alone, and thus its analytical phase starts with a Friedman two-way analysis of variance by ranks test.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of <i>Dis.language</i> and <i>Dis.response</i> are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.780	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Table 40. Friedman’s two-way analysis on *Dis.language* and *Dis.response* for hypothesis 1 (Spanish group, discrimination tests).

The hypothesis test summary displayed in table 40 above found no statistically significant relationships between language familiarity (*Dis.language*) and type of response (*Dis.response*) in discrimination tests. Consequently, it appears safe to retain the null hypothesis.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	.512 ^a	2	.774
Likelihood Ratio	.483	2	.785
N of Valid Cases	174		

a. 3 cells (50.0%) have expected count less than 5. The minimum expected count is 1.33.

Table 41. Chi-square test for hypothesis 1 (Spanish group, discrimination tests, 1st exp. condition).

Quite expectedly, chi-square p-values appearing in table 41 are non-significant ($\chi^2(2, N = 174) = 0.512, p > 0.05$), while its validity is once again compromised given that the cells with less than 5 expected counts surpass the 20% of total cells in the table dedicated to unveil interactions between the variables at hand (Yates et al. 1999: 734).

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.054	.774
	Cramer's V	.054	.774
N of Valid Cases		174	

c. Correlation statistics are available for numeric data only.

Table 42. Cramer's V and Phi coefficient for hypothesis 1 (Spanish group, discrimination tests, 1st exp. condition).

Unsurprisingly, table 42 above reports a negligible association (0.05) according to Cramer's V.

Dis.response * Dis.language Crosstabulation

		Dis.language			Total	
		familiar	learned	unknown		
Dis.response	Correct rejection	Count	56 _a	57 _a	57 _a	170
		Expected Count	56.7	56.7	56.7	170.0
		% within Dis.response	32.9%	33.5%	33.5%	100.0%
		Adjusted Residual	-.7	.4	.4	
	False alarm	Count	2 _a	1 _a	1 _a	4
Expected Count		1.3	1.3	1.3	4.0	
% within Dis.response		50.0%	25.0%	25.0%	100.0%	
Adjusted Residual		.7	-.4	-.4		
Total	Count	58	58	58	174	
	Expected Count	58.0	58.0	58.0	174.0	
	% within Dis.response	33.3%	33.3%	33.3%	100.0%	
	Adjusted Residual					

Each subscript letter denotes a subset of Dis.language categories whose column proportions do not differ significantly from each other at the .05 level.

Table 43. Contingency table for hypothesis 1 (Spanish group, discrimination tests, 1st exp. condition).

As noticed in table 43 above, and much in line with the results exhibited in the British group's sub-section, false alarms expected count amounts to less than 5 in each and every language test. Conversely, adjusted residuals are, in the Spanish' case, not significant enough to ascertain a relevant statistical relationship between *Dis.language* and

Dis.response. As a matter of fact, the figure below shows the little variance that such low adjusted residuals exhibit.

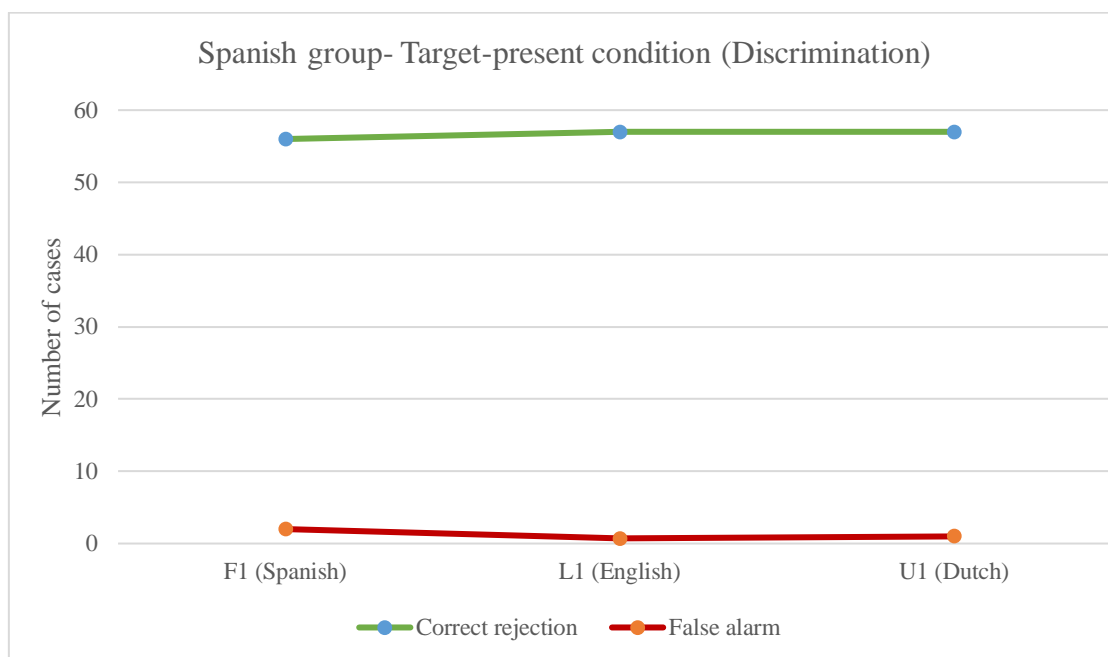


Figure 20. Spanish group's count of response types across three language familiarities in the first experimental condition (target-present, discrimination tests).

Separate accounts on types of response and language familiarity do not seem to yield statistical significance, as figure 20 shows. However, high expected counts on correct rejections do emerge across the two sociocultural groups. In this regard, the analysis shall proceed to explore the aforementioned variables taking into consideration both the Spanish and British groups' scores altogether.

4.1.2.3. *British and Spanish group*

A Friedman two-way analysis of variance by ranks test calculates whether language familiarity and voice line-up's outcome are correlated when putting together the results from the British and Spanish group.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of <i>Dis.language</i> and <i>Dis.response</i> are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	,628	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Table 44. Friedman’s two-way analysis on *Dis.language* and *Dis.response* for hypothesis 1 (British and Spanish group, discrimination tests).

The first analytical stage did not reveal any statistically significant correlation between the variables of interest, as table 44 shows. A post-hoc analysis reveals here the chi-square p-values, Cramer’s V, and a crosstabulation of each sub-category included within the variables *Dis.language* and *Dis.response*.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	4.801 ^a	2	.091
Likelihood Ratio	4.634	2	.099
N of Valid Cases	321		

a. 3 cells (50.0%) have expected count less than 5. The minimum expected count is 3.00.

Table 45. Chi-square test for hypothesis 1 (British and Spanish group, discrimination tests, 1st exp. condition).

As table 45 reveals, the chi-square test of independence conducted found a near-significant correlation between the categorical variables inspected (χ^2 (2, N = 321) = 4.801, p= 0.09).

Symmetric Measures^c

		Value	Approximate Significance
Nominal by Nominal	Phi	.122	.091
	Cramer's V	.122	.091
N of Valid Cases		321	

c. Correlation statistics are available for numeric data only.

Table 46. Cramer's V and Phi coefficient for hypothesis 1 (British and Spanish group, discrimination tests, 1st exp. condition).

Cramer's V, on the other hand, reports a weak association (0.12), just as table 46 demonstrates. The crosstabulation below illustrates such correlations in more detail.

Dis.response * Dis.language Crosstabulation

		Dis.language			Total	
		familiar	learned	unknown		
Dis.response	Correct rejection	Count	101 _a	106 _a	105 _a	312
		Expected Count	104.0	104.0	104.0	312.0
		% within Dis.response	32.4%	34.0%	33.7%	100.0%
		Adjusted Residual	-2.2	1.4	.7	
	False alarm	Count	6 _a	1 _a	2 _a	9
		Expected Count	3.0	3.0	3.0	9.0
		% within Dis.response	66.7%	11.1%	22.2%	100.0%
		Adjusted Residual	2.2	-1.4	-.7	
	Total	Count	107	107	107	321
Expected Count		107.0	107.0	107.0	321.0	
% within Dis.response		33.3%	33.3%	33.3%	100.0%	
Dis.response						

Each subscript letter denotes a subset of Dis.language categories whose column proportions do not differ significantly from each other at the .05 level.

Table 47. Contingency table for hypothesis 1 (British and Spanish group, discrimination tests, 1st exp. condition). Statistically significant adjusted residuals are marked in bold ($\alpha = 0.05$).

As in the British jurors' case, the resulting p-value is close to significance ($p = 0.09$) due to the influence of familiar language's adjusted residuals, whose values fall below the expected with correct rejections (-2.2) and exceeding counts of false alarms (2.2), as seen

in table 47 above. However, the numbers do not seem significant enough when putting them into perspective, as figure 21 below demonstrates.

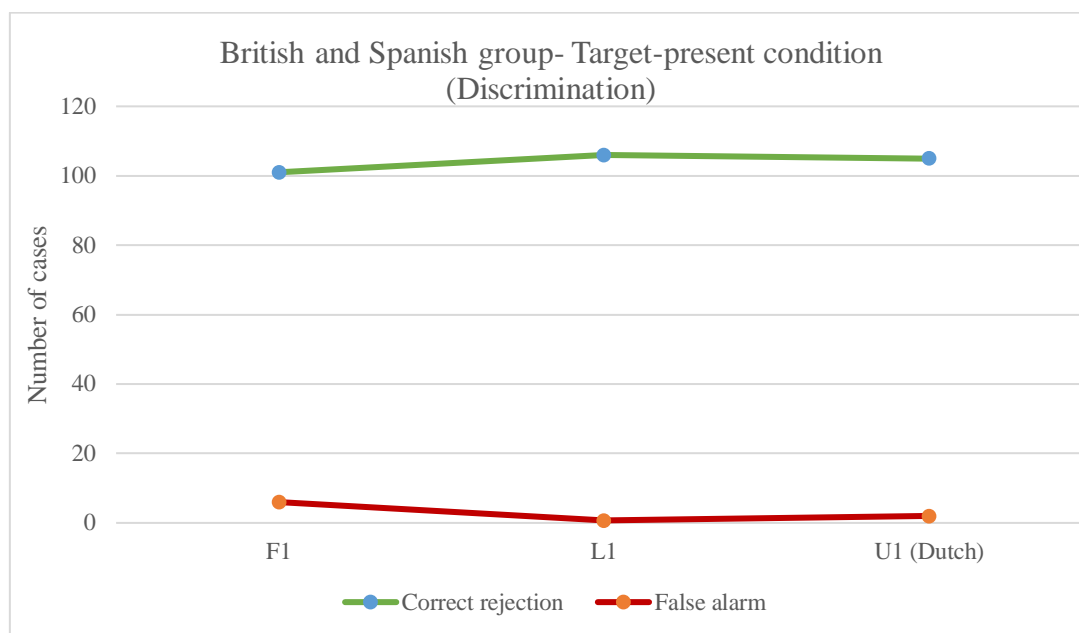


Figure 21. British and Spanish group’s count of response types across three language familiarities in the first experimental condition (target-present, discrimination tests).

However, as stated previously, this correlation’s validity is nullified owing to non-significant p-values and the violation of chi-square assumptions (only the 20% of total cells displayed can exhibit expected counts less than 5).

4.1.3. Summary of results

To conclude, this last section gathers the results exposed in the past sections to answer the first formulated hypothesis (aural-perceptual recognition is enhanced as the familiarity of the juror with the language exposed also increases). Said statement implies a linear relationship between language familiarity and type of response, whereby familiar languages produce the most fruitful results and speaker recognition is exacerbated the most in unknown language tests.

When combining the results from both groups (British and Spanish groups), none of the two experimental conditions (target-present without background noise, and target-absent with background noise) appear to yield statistically significant results in identification

tests. Despite this, some idiosyncrasies become noticeable in the data set when it is split according to each group of participants, Spanish and British group:

- In the **British- 1st test (target-present)**, statistical differences arise in false alarm rates (the learned language test produces them to a lesser extent than familiar and unknown language tests) in return for a slight increase in hitting and missing, albeit not statistically significant.
- In the **British- 2nd test (target-absent)**, the distribution of correct rejections and false alarms of the learned language test differs in relation to the unknown language test. The former facilitates correct rejections and produces less false alarms, whereas the opposite is true for unknown input. Values obtained from the familiar language test fall in between these two categories without differing significantly from either of them.
- In the **Spanish 1st test (target-present)**, L1(English) results produced more false alarms than familiar and unknown language tests, which are both equated in terms of success/failure.
- In the **Spanish 2nd test (target-absent)**, the same trend as above is observed (similar percentages of false alarms and correct rejections for familiar and unknown language tests, whilst the learned language test produces more false alarms and less correct rejections), although this time the variance of the data is minimal and therefore not statistically significant.

Out of the four possible identification tests, only three of them rejected the null hypothesis (which assumes that there is no relationship between language familiarity and type of response), whereas the Spanish 2nd test produced non-significant variation in the data. As for the other three tests, the learned language test scores differed significantly in contrast with the results generated by the familiar and unknown language tests. This observed variance within the studied categorical variables, however, does not follow the same orientation in either group. For instance, British group's learned language test presents less cases of false alarms than expected, both in the 1st and 2nd condition, whereas the opposite trend applies to the 1st and 2nd learned language tests in the Spanish group, even if the 2nd test is not statistically significant in this particular case. As for the discrimination tasks, examining both groups of jurors in conjunction and/or in isolation

rendered no statistically significant correlations, due to the low variance of values exhibited.

In conclusion, the statistical measures adopted found no linear relationship between familiar and unknown language, but pointed at the learned language test scores, which deviated significantly from the other two tests. In this regard, how familiar/unfamiliar the linguistic input might be to the hearer should not necessarily be a predictor of his/her performance at recognition tasks. This finding seems to indicate that hearers' responses are more influenced by the languages they are learning (intermediate exposure) rather than their native tongue or a completely unknown language. A possible explanation for this could be the differentiation between acquisition (or non-acquisition, in the case of unknown languages) and learning, as the learner's auditory schemata and linguistic expectations are developing differently across individuals (pacing, oral/writing skills, etc.) with the purpose of enhancing one's proficiency in said language, thus leading to disparate responses. At any rate, future research on speaker recognition is needed to address this language acquisition/learning distinction and provide supporting evidence with the purpose of verifying this claim.

4.2. DISCRIMINATION OR IDENTIFICATION?

As the first hypothesis concludes, visual representations of data may have hinted at potential distinctions concerned with the distribution of values shown between discrimination and identification tasks. It is through the second hypothesis that such relationships are explored. Hence, both the null hypothesis (H_0) and the alternative hypothesis (H_2) are formulated hereby:

- H_0 : Jurors perform equally well, regardless of the type of recognition task presented.
- H_2 : Jurors are more proficient in discrimination tests than in identification tasks.

At this point, the order in which the diverse strata chosen for analysis are explored complies with the pre-established hierarchy explained in point 3.7.2. (*Perception surveys-based analysis*): British groups' contrastive account on identification and discrimination

tasks examines target-present conditions first, followed by the target-absent condition. Once this analysis concludes, the same pattern shall be adopted to analyse the Spanish group.

Type of task	Experimental condition	Language tests	Outcome	Code
Identification tasks	target-present	F1	False alarm	1
		L1	Miss	1
		U1	Hit	2
	target-absent	F2	False alarm	1
		L2	Correct rejection	2
		U2		
Discrimination tasks	target-present	F1.Dis	False alarm	1
		L1.Dis	Correct rejection	2
		U1.Dis		
	target-absent	F2.Dis	Correct rejection	2
		L2.Dis		
		U2.Dis		

Table 48. Numerical variables assigned for each language test in identification and discrimination tasks for hypothesis 2.

As noticed by glancing at table 48 above, the same criterion is used to assign dummy codes to the possible voice line-up's outcomes, namely 1 for failure and 2 for a success condition. What differentiates this set up from hypothesis 1's, however, is that this research design computes the values of each language test separately (F1, L1, U1, F2, L2, U2...etc.), instead of grouping them all under the same column as categorical variables. As discussed already, this data set makes use of the dummy coding listed above, while the categorical variables (voice line-ups' outcomes) are mentioned only for reference.

4.2.1. British group

As a starting point, the first set of online perception surveys (target-present) considered for this analysis is selected to discern any difference between identification and

discrimination tasks within the British group of jurors (n=49). To get an initial observation of the distribution of values across all the language tests involved in this comparison, a table on descriptive statistics is consulted.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
F1.Dis	49	1.92	.277	1	2	2.00	2.00	2.00
L1.Dis	49	2.00	.000	2	2	2.00	2.00	2.00
U1.Dis	49	1.98	.143	1	2	2.00	2.00	2.00
F1	49	1.43	.500	1	2	1.00	1.00	2.00
L1	49	1.51	.505	1	2	1.00	2.00	2.00
U1	49	1.43	.500	1	2	1.00	1.00	2.00

Table 49. Descriptive statistics of each language test in identification and discrimination tasks for hypothesis 2 (British group, 1st exp. condition).

The most remarkable observation from table 49 above is the fact that every mean score obtained through discrimination tasks (F1.Dis, L1.Dis, and U1.Dis) exceeds their equivalents in identification tests (F1, L1, and U1). Another important aspect drawn from this is the low standard deviation emerging in discrimination tasks, which confirms the little variance of values shown while solving the first hypothesis.

Test Statistics^a

	F1 - F1.Dis	L1 - L1.Dis	U1 - U1.Dis
Z	-4.707 ^b	-4.899 ^b	-5.196 ^b
Asymp. Sig. (2-tailed)	.000	.000	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Table 50. Wilcoxon-signed ranks test for each pair of language test in identification and discrimination tasks for hypothesis 2 (British group, 1st exp. condition). Statistically significant values are marked in bold ($\alpha = 0.05$).

As the p-values described in table 50 through the Wilcoxon signed-rank test suggest, there are statistically significant differences between identification tests and their discrimination counterparts. This phenomenon applies irrespective of the linguistic input perceived.

As for the second half of online perception surveys (target-absent), here follows a descriptive statistics account for the types of tests investigated.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
F2.Dis	49	2.00	.000	2	2	2.00	2.00	2.00
L2.Dis	49	2.00	.000	2	2	2.00	2.00	2.00
U2.Dis	49	2.00	.000	2	2	2.00	2.00	2.00
F2	49	1.22	.422	1	2	1.00	1.00	1.00
L2	49	1.31	.466	1	2	1.00	1.00	2.00
U2	49	1.10	.306	1	2	1.00	1.00	1.00

Table 51. Descriptive statistics of each language test in identification and discrimination tasks for hypothesis 2 (British group, 2nd exp. condition).

Just as in the target-present condition, table 51 shows that discrimination tasks mean scores outnumber those in identification tasks. It is because of the specific set up of this condition (target-absent) that jurors cannot select the suspect, which is why there is no existing standard deviation for discrimination tests, and their mean scores amount to the maximum 2.00 that can be obtained.

Test Statistics^a

	F2 - F2.Dis	L2 - L2.Dis	U2 - U2.Dis
Z	-6.164 ^b	-5.831 ^b	-6.633 ^b
Asymp. Sig. (2-tailed)	.000	.000	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Table 52. Wilcoxon-signed ranks test for each pair of language test in identification and discrimination tasks for hypothesis 2 (British group, 2nd exp. condition). Statistically significant values are marked in bold ($\alpha = 0.05$).

Again, the Wilcoxon signed-ranks test reinforces the situation previously seen during the target-present condition: identification tasks' performance is significantly inferior to discrimination tests, as indicated in table 52 above.

4.2.2. Spanish group

Moving on to the Spanish group (n=58), it proceeds to inspect the first experimental condition (target-present) which again seeks differences in the distribution of values amongst identification and discrimination tasks. Similarly, the analytical stage shall initiate with a table on descriptive statistics measures.

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimum	Maximum	25th	Percentiles 50th (Median)	75th
F1.Dis	58	1.97	.184	1	2	2.00	2.00	2.00
L1.Dis	58	1.98	.131	1	2	2.00	2.00	2.00
U1.Dis	58	1.98	.131	1	2	2.00	2.00	2.00
F1	58	1.53	.503	1	2	1.00	2.00	2.00
L1	58	1.31	.467	1	2	1.00	1.00	2.00
U1	58	1.53	.503	1	2	1.00	2.00	2.00

Table 53. Descriptive statistics of each language test in identification and discrimination tasks for hypothesis 2 (Spanish group, 1st exp. condition).

Just as witnessed in the British group, the Spanish case appears to obey the same underlying principles that set apart identification from discrimination tests, as observed in table 53. This is to say that mean scores in discrimination tasks are considerably higher than identification tasks, whose scores vary more due to greater standard deviations in relation to the ones emerging in the discrimination department.

Test Statistics^a

	F1 - F1.Dis	L1 - L1.Dis	U1 - U1.Dis
Z	-5.000 ^b	-6.245 ^b	-5.099 ^b
Asymp. Sig. (2-tailed)	.000	.000	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Table 54. Wilcoxon-signed ranks test for each pair of language test in identification and discrimination tasks for hypothesis 2 (Spanish group, 1st exp. condition). Statistically significant values are marked in bold ($\alpha = 0.05$).

Unexpectedly, the Wilcoxon signed-ranks test shown in table 54 also confirms the premise that discrimination tests perform significantly better than identification ones, given the tiny p-value that ensues. Again, such differences apply to all language tests employed (familiar, learned, and unknown languages).

Once the first analysis is cleared, this sub-section explores whether the same patterns appear at the second experimental condition (target-absent). To this end, a table on descriptive statistics (table 55) is drawn and consulted hereby.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
F2.Dis	58	2.00	.000	2	2	2.00	2.00	2.00
L2.Dis	58	2.00	.000	2	2	2.00	2.00	2.00
U2.Dis	58	2.00	.000	2	2	2.00	2.00	2.00
F2	58	1.26	.442	1	2	1.00	1.00	2.00
L2	58	1.21	.409	1	2	1.00	1.00	1.00
U2	58	1.24	.432	1	2	1.00	1.00	1.25

Table 55. Descriptive statistics of each language test in identification and discrimination tasks for hypothesis 2 (Spanish group, 2nd exp. condition).

Similar to the British case (target-absent) mentioned before, discrimination tasks display perfect scores with zero variance, whilst identification tasks' scores are much lower in comparison, with relatively higher standard deviations.

Test Statistics^a			
	F2 - F2.Dis	L2 - L2.Dis	U2 - U2.Dis
Z	-6.557 ^b	-6.782 ^b	-6.633 ^b
Asymp. Sig. (2-tailed)	.000	.000	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Table 56. Wilcoxon-signed ranks test for each pair of language test in identification and discrimination tasks for hypothesis 2 (Spanish group, 2nd exp. condition). Statistically significant values are marked in bold ($\alpha = 0.05$).

Once again, the Wilcoxon signed-ranks test's p-values shown in table 56 report statistically significant differences between the distribution of scores within identification and discrimination tests, the former performing worse than the latter.

4.2.3. Summary of results

In light of the foregoing, the null hypothesis can be rejected (that jurors' performance does not vary across types of recognition tasks), and thus the alternative hypothesis is accepted (that discrimination tests fare better than identification tasks). Besides, it is concluded that this tendency applies regardless of the group surveyed (British and Spanish jurors), and experimental condition (target-present without background noises, and target-absent with noise disturbances).

4.3. CONFIDENCE LEVELS

As for the third hypothesis, it aims to figure out whether the relationship between jurors' self-perceived confidence in identification/discrimination tasks improves their actual scores. To this end, the null hypothesis (H_0) and the alternative hypothesis (H_3) are formulated below:

- H_0 : The degree of self-perceived confidence level does not have a discernible effect on the voice line-up's outcome.
- H_3 : A heightened self-perceived confidence level at speaker recognition tasks has a positive effect on the voice line-up's outcome.

As discussed during the methodological section (3.7.2. *Perception surveys-based analysis*), this research question follows the same generic hierarchy established: Identification tasks first, followed by discrimination tasks. Immediately after both groups of jurors are analysed in isolation, and a grouped account of both British and Spanish jurors is used as a summary. Lastly, the first experimental condition (target-present) precedes the second one (target-absent).

Variable	Outcome	Code
Test scores	False alarm	1
	Miss	1
	Hit	2
	False alarm	1
	Correct rejection	2
Confidence levels	Not confident	1
		2
		3
		4
		5
		6
		7
		8
		9
	Highly confident	10

Table 57. Numerical variables considered in hypothesis 3: test scores and confidence levels.

Regarding the variables required for hypothesis testing, these are listed in table 57 above. Similar to hypothesis 2 design, test scores from every perception test is included, according to type of task (identification and discrimination) and experimental condition (target-present and target-absent). For the sake of clarity, only the possible outcomes are illustrated in table 57, and thus breaking down each language test based on the aforementioned categories is avoided.

On the other hand, confidence levels (hereinafter referred to as CL) employ a Likert scale which measures the hearer's degree of certainty at identification/discrimination tasks after a decision is made. Kendall's tau tests use both test scores (1-2 points) and CL's 10-point scales to find possible correlations between such pairs (i.e. F1-CL. F1).

4.3.1. Identification

To start with identification tasks, a series of Kendall's tau tests have been conducted to determine whether the discrete numeric variable CL (confidence levels) has an effect on the ordinal variable test scores (for example, F1= 1st familiar language test's outcome, F2= 2nd familiar language test's outcome, etc.). It is decided to proceed with the British group of jurors first, and Spanish group's analysis shall follow shortly after. Results are summarised in an analysis that puts together the results from both British and Spanish jurors.

4.3.1.1. British group

A Kendall's tau-b correlation coefficient was run to determine the relationship between CL and language test scores amongst 49 participants. The results are shown in the list below:

- F1- CL. F1 ($\tau_b = -0.065$, $p = 0.607$).
- L1- CL. L1 ($\tau_b = -0.100$, $p = 0.425$).
- U1- CL. U1 ($\tau_b = -0.058$, $p = 0.639$).
- F2- CL. F2 ($\tau_b = 0.033$, $p = 0.790$).
- L2- CL. L2 ($\tau_b = 0.383$, $p = 0.002^{10}$).
- U2- CL. U2 ($\tau_b = -0.131$, $p = 0.293$).

As observed, there was a strong, positive correlation between CL and test scores obtained in the second test for the learned language (L2), which was statistically significant. The other language tests yielded no statistically significant correlation with the CL variable.

¹⁰ This correlation is significant at the 0.01 level (2-tailed).

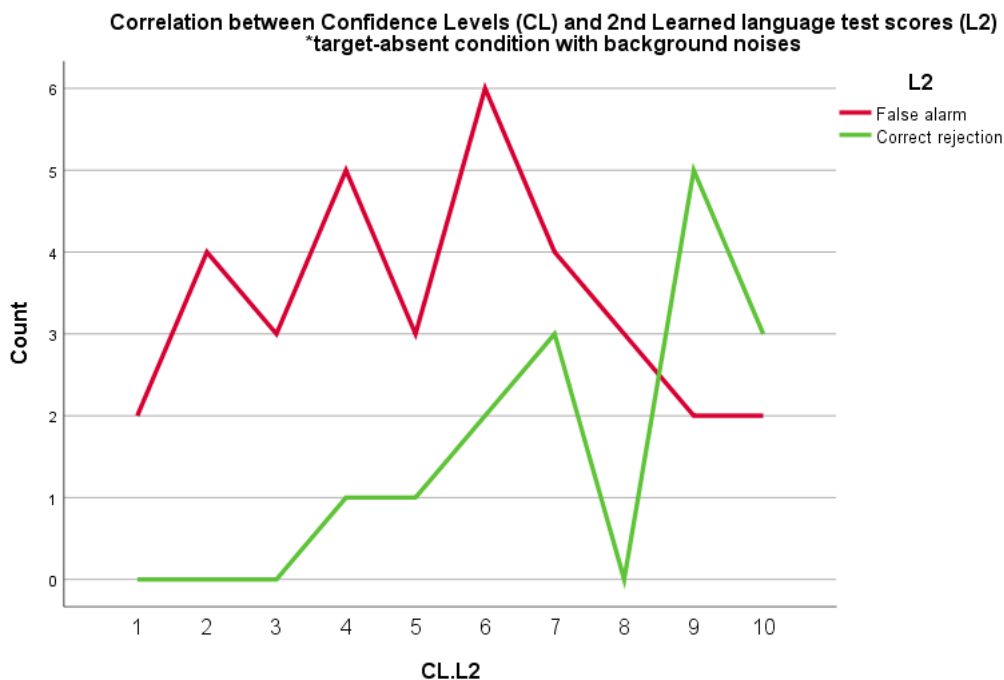


Figure 22. Multiple line graph on the L2-CL. L2 correlation in the British group (identification tests).

As seen in figure 22, there is a positive correlation between the self-perceived confidence level at undertaking the perception survey and the actual score obtained in the second learned language test in the British group. False alarms tend to increase at the lowest confidence levels, whereas correct rejections occur more frequently on the opposite end. Nevertheless, it is not until the 9-point confidence level that the probability of success becomes prominent.

4.3.1.2. Spanish group

A Kendall's tau-b correlation coefficient was conducted to determine the relationship between CL and language test scores amongst 58 jurors. The results including both experimental conditions are included in the following list:

- F1- CL. F1 ($\tau_b = 0.209$, $p = 0.072$).
- L1- CL. L1 ($\tau_b = 0.157$, $p = 0.175$).
- U1- CL. U1 ($\tau_b = 0.236$, $p = 0.040^{11}$).
- F2- CL. F2 ($\tau_b = 0.108$, $p = 0.347$).

¹¹ This correlation is significant at the 0.05 level (2-tailed).

- L2- CL. L2 ($\tau_b = -0.020$, $p = 0.862$).
- U2- CL. U2 ($\tau_b = -0.078$, $p = 0.497$).

There is a positive correlation between CL and test scores obtained in the first unknown language test (U1), and a near-significant influence ($p = 0.072$) following the same trend in the first the familiar language test (F1). The rest of language tests yielded no statistically significant correlations with CL.

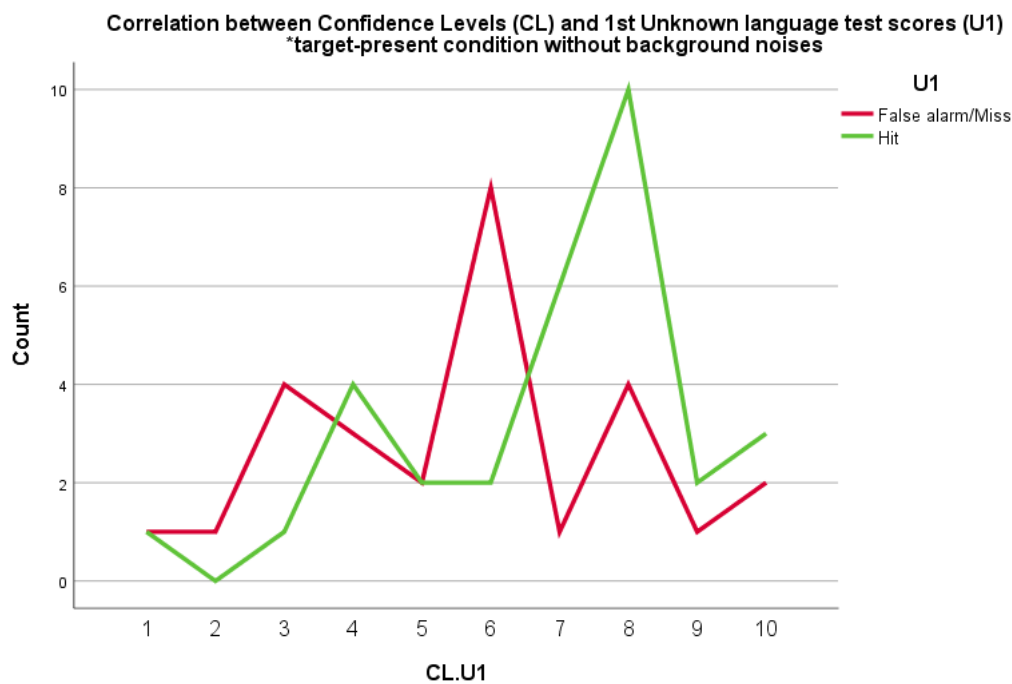


Figure 23. Multiple line graph on the U1-CL. U1 correlation in the Spanish group (identification tests).

In contrast with the British group, the first unknown language test (target-present without background conditions) is significantly correlated with confidence levels ($p = 0.040$) in the Spanish group. However, figure 23 attests that the overall orientation of the data remains the same, and thus a positive correlation originates between enhanced confidence levels and higher hit rates, reaching its identification potential at the highest self-perceived confidence levels (8-10).

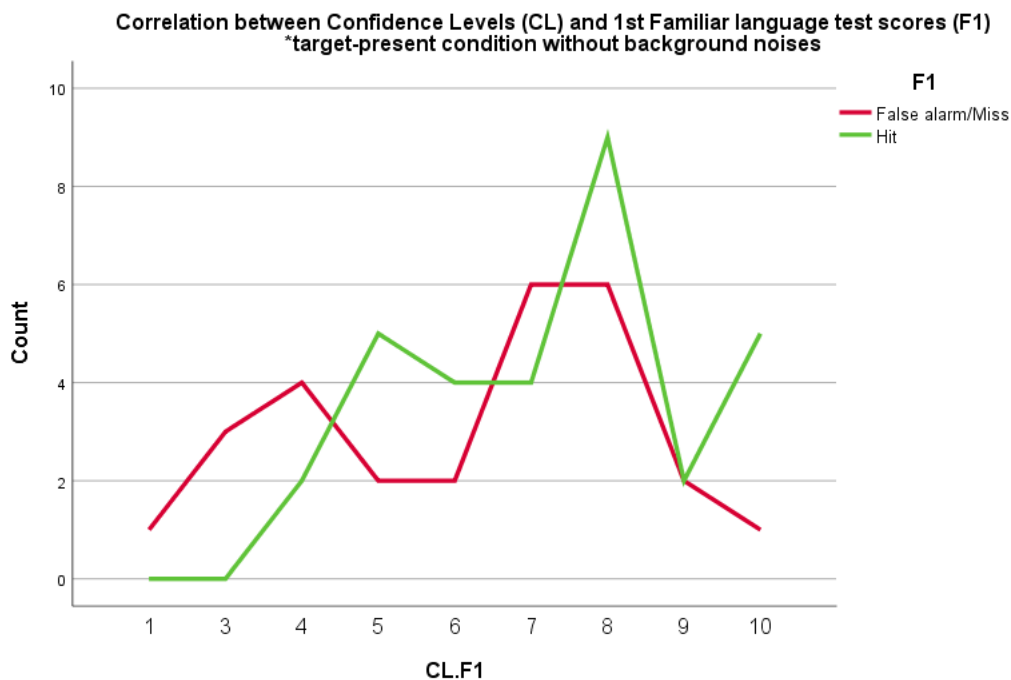


Figure 24. Multiple line graph on the F1-CL. F1 correlation in the Spanish group (identification tests).

Despite being close to significance ($p= 0.072$), the first familiar language test (F1) does follow the same pattern as in the previous cases exposed above, as figure 24 illustrates. What is more, the most auspicious results are also in line with the Spanish U1 and British L2 results, therefore deeming the 8-10 confidence points as the most reliable units within the CL (confidence levels) variable.

4.3.1.3. British and Spanish group

In order to get a wider picture of how language test scores and self-perceived confidence levels are distributed in identification tasks amongst cultural groups and language tests, two tables gathering descriptive statistics measures are compiled for British (table 58) and Spanish (table 59) jurors.

British group						
	Language tests					
	1 st condition			2 nd condition		
Mean	F1	L1	U1	F2	L2	U2
Test score	1.43	1.51	1.43	1.22	1.31	1.10
Confidence levels	6.80	6.92	5.63	6.04	6.06	4.24

Table 58. Mean test scores and confidence levels in the British group (identification tasks).

Spanish group						
	Language tests					
	1 st condition			2 nd condition		
Mean	F1	L1	U1	F2	L2	U2
Test score	1.53	1.31	1.53	1.26	1.21	1.24
Confidence levels	6.74	6.53	6.29	6.50	5.98	5.24

Table 59. Mean test scores and confidence levels in the Spanish group (identification tasks).

Insofar as descriptive statistics is concerned, mean test scores and confidence levels reveal two trends in the target population, irrespective of the group surveyed: 1) mean test scores diminish in the second condition, in contrast with their noiseless counterparts, and 2) the juror's self-perceived confidence is undermined in the second experimental condition as well.

Albeit not necessarily a statistically sound conclusion, it appears that the highest values for both mean test scores and confidence levels appear in the learned language (British group), whereas Spanish jurors display the best results in the familiar language tests (both first and second condition). In this respect, addressing whether such correlations are significant or not shall be dealt with hereafter. To understand how both group of jurors' responses blend together, a combined account of confidence levels and mean test scores is provided hereby.

A Kendall's tau-b correlation coefficient is calculated to determine the relationship between CL and language test scores amongst 107 participants. The list below illustrates the results for each language test and experimental condition:

- F1- CL. F1 ($\tau_b = 0.142$, $p = 0.095$).
- L1- CL. L1 ($\tau_b = 0.046$, $p = 0.589$).
- U1- CL. U1 ($\tau_b = 0.160$, $p = 0.056$).
- F2- CL. F2 ($\tau_b = 0.084$, $p = 0.317$).
- L2- CL. L2 ($\tau_b = 0.188$, $p = 0.025^{12}$).
- U2- CL. U2 ($\tau_b = -0.062$, $p = 0.457$).

¹² This correlation is significant at the 0.05 level (2-tailed).

There was a strong, positive correlation between CL and test scores obtained in the second test for the learned language (L2), which is statistically significant. The other language tests were not correlated with the CL variable:

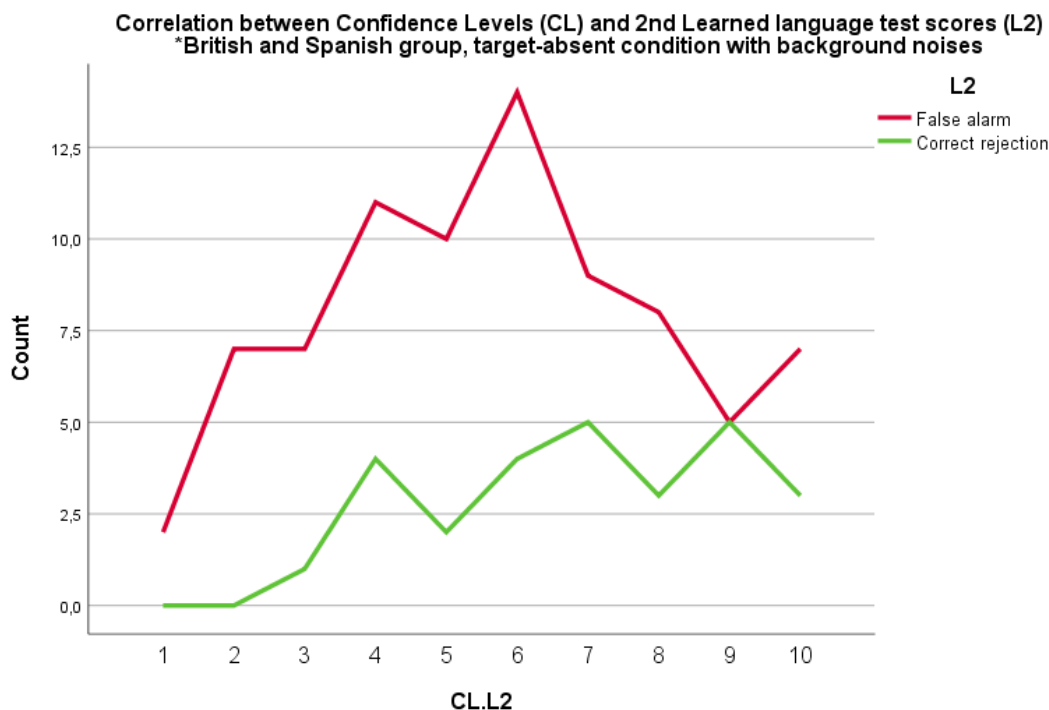


Figure 25. Multiple line graph on the L2-CL. L2 correlation in the British and Spanish group (identification tests).

In contrast with the British-only group comparison, the correlation between CL and L2 scores becomes weaker when adding the influence of the Spanish group to the equation. Nevertheless, there is still a positive correlation ($\tau_b = 0.188$, $p = 0.025$), whereby the most optimal outcome in figure 25 seems to be located between the 7-9 CL points, thus increasing the rate of correct rejections while reducing the chances for false alarms. It is also noticeable that, despite reaching statistical significance, correct rejections do not outnumber false alarms anywhere in the graph when combining both the British and Spanish groups' results, as opposed to considering L2 and CL. L2 correlation in the British group analysis, whose ratio of correct rejections peaks at the highest confidence levels (9-10).

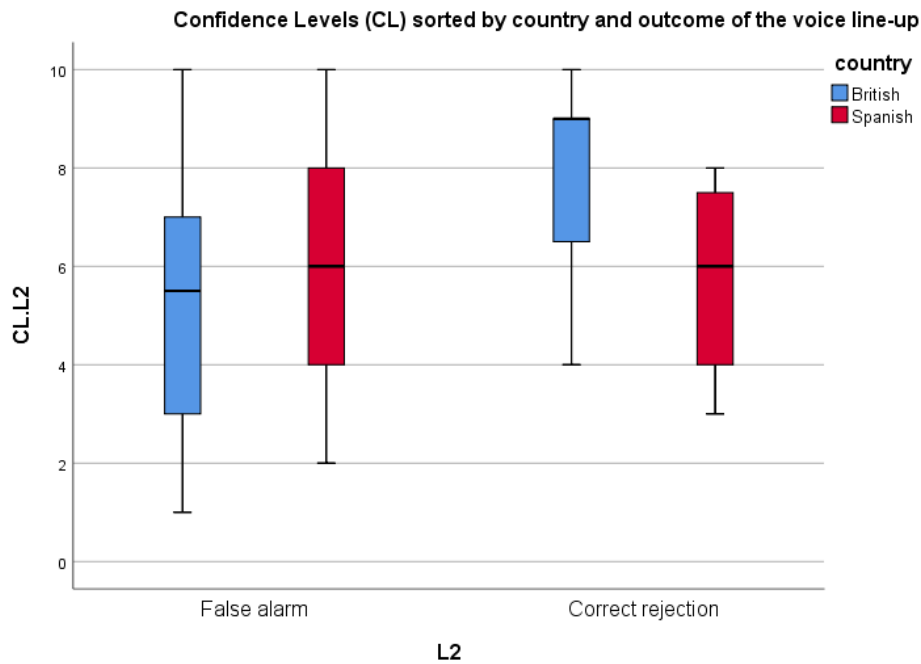


Figure 26. Boxplot on the L2-CL. L2 correlation in the British and Spanish group (identification tests).

To isolate the influence of each group of jurors, the boxplots in figure 26 above illustrates the two trends observed in the target population. On the one hand, the Spanish group confidence level remains nearly unaltered regardless of the perception survey’s outcome. British participants, however, do undergo discernible changes in their self-perceived confidence level on the task, leading to a higher range (7-9) of values in successful trials.

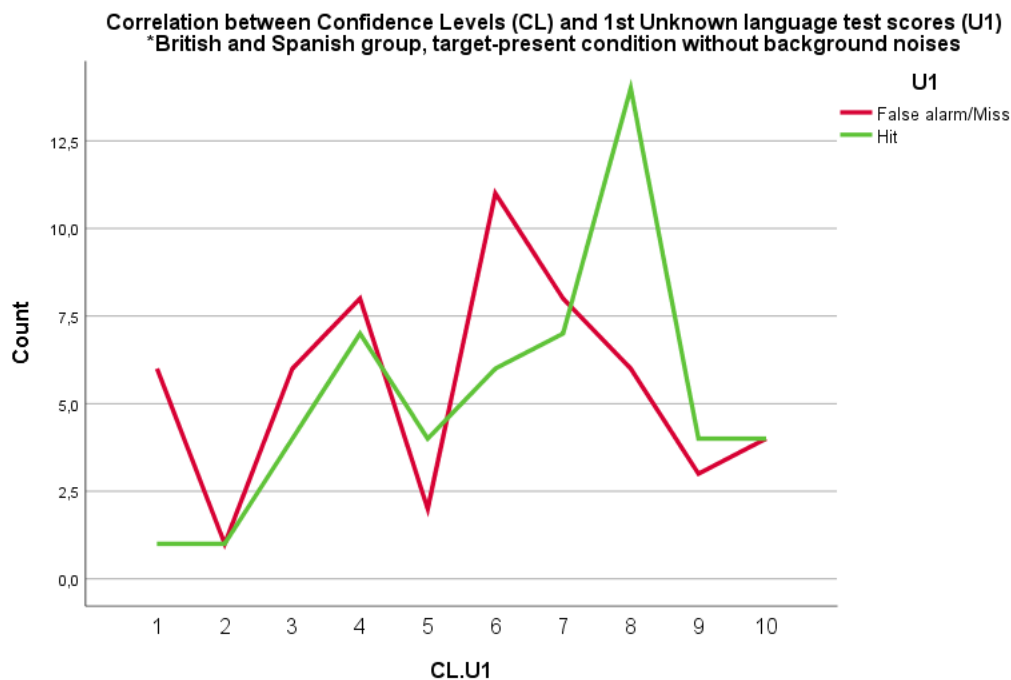


Figure 27. Multiple line graph on the U1-CL. U1 correlation in the British and Spanish group (identification tests).

It is also worth mentioning the case for the first unknown language first test (U1), whose correlation with CL is close to reaching statistical significance ($r_b = 0.160$, $p = 0.056$). In the line graph above (figure 27), a positive correlation is discernible with higher occurrences of hits pivoting around 8-10 confidence levels, much in line with U1's Spanish-only group results.

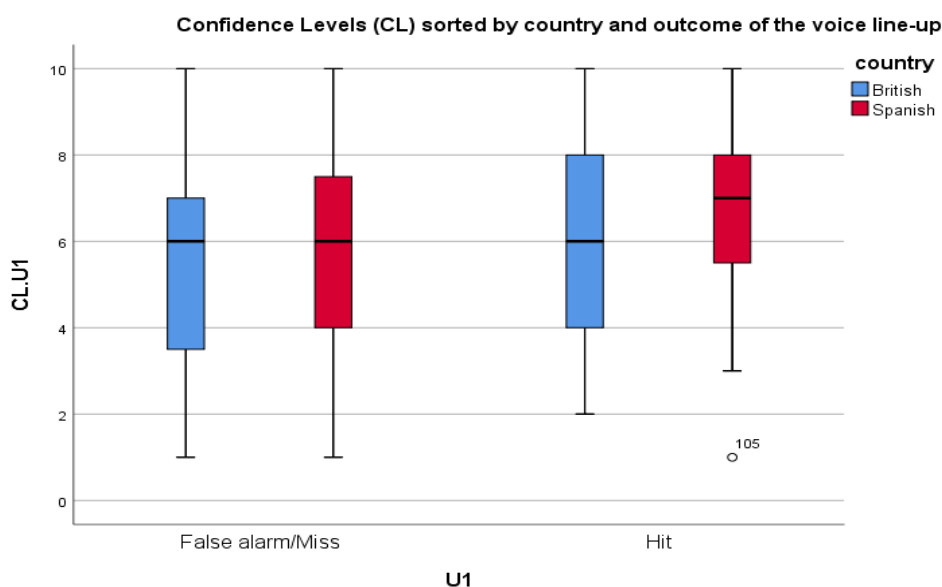


Figure 28. Boxplot on the U1-CL. U1 correlation in the British and Spanish group (identification tests).

As done in the previous case, the boxplots shown in figure 28 are illustrating the differences between juror groups in regard with their confidence levels and identification task's outcome. In this case, and contrary to L2's results, it is the Spanish group the one that excels at hit rates in the unknown language test (1st experimental condition) when increasing the self-perceived confidence levels (7-8 point range).

4.3.2. Discrimination

Similar to identification tasks' analysis, several Kendall's tau tests were undertaken to discern the possible correlations between the discrete numeric variable CL (confidence levels) and the ordinal variable test scores. As explained previously, the order to inspect each cultural group is arranged as follows: British, Spanish, and British and Spanish groups.

4.3.2.1. British group

A Kendall's tau correlation coefficient has been conducted to detect possible correlations between CL (confidence levels) and language test scores in 49 participants. As mentioned in 3.7.2. (*Perception surveys-based analysis*), language tests including the target-absent condition are not considered here due to their lack of variance in their values. The list below displays said correlations and their levels of significance:

- F1. Dis- CL. F1. Dis ($\tau_b = -0.008$, $p = 0.956$).
- L1. Dis- CL. L1. Dis ($\tau_b = .X$, $p = .X^{13}$).
- U1. Dis- CL. U1. Dis ($\tau_b = -0.085$, $p = 0.497$).

No relevant correlations were found amongst CL and language test scores, since the latter exhibits little variation. However, self-perceived confidence is hindered the most at the unknown language test (with a higher negative coefficient), presumably for the uncertainty that such input may generate on the juror's decision.

¹³ Cannot be computed because at least one of the variables is constant.

4.3.2.2. Spanish group

A Kendall’s tau correlation coefficient has been undertaken to find out if the numeric variable CL (confidence levels) has a discernible effect on language test scores in 58 participants. Hereby follows a list of the values obtained for the first experimental condition:

- F1. Dis- CL. F1. Dis ($\tau_b = -0.015$, $p = 0.896$).
- L1. Dis- CL. L1. Dis ($\tau_b = 0.115$, $p = 0.328$).
- U1. Dis- CL. U1. Dis ($\tau_b = -0.077$, $p = 0.506$).

Little variation seems to appear in mean test scores, which in turn reflects non-significant p-values. What is consistent, however, is the jurors’ decrease in self-perceived confidence when exposed to unknown linguistic input, even if it does not amount to statistical significance.

4.3.2.3. British and Spanish group

To explore the inexistent correlations between test scores and CL, the two tables below resort to descriptive statistics measures with the purpose of improving our understanding of the subject matter:

British group						
	Language tests					
	1 st condition			2 nd condition		
Mean	F1	L1	U1	F2	L2	U2
Test score	1.92	2	1.98	2	2	2
Confidence levels	7.45	7.14	6.27	6.92	6.53	5.02

Table 60. Mean test scores and confidence levels in the British group (discrimination tasks).

Spanish group						
	Language tests					
	1 st condition			2 nd condition		
Mean	F1	L1	U1	F2	L2	U2
Test score	1.97	1.98	1.98	2	2	2
Confidence levels	7.90	7.91	6.79	7.57	7.17	5.78

Table 61. Mean test scores and confidence levels in the Spanish group (discrimination tasks).

As seen in tables 60 and 61 above, little variation is expected in language perception test's outcome, as suggested by their high mean test scores (> 1.90). In this respect, most of respondents' decisions were correct rejections in the first experimental condition, in both groups of jurors. It is noteworthy that, despite guaranteeing a perfect 2.00 score in the target-absent condition, the mean CL appear to diminish in contrast with their noiseless counterparts. One final observation is that British group's jurors appear more confident when exposed to English input (in both experimental conditions), whereas Spanish participants performed the English test (1st condition) with more confidence (L1), but this changes to the Spanish language in the second test with background noises (F2).

4.3.3. Summary of results

After scrutinising each language perception test in every stratum, this study proceeds to conclude the third hypothesis with a summary.

Firstly, in identification tasks, only one target-absent test was productive from the British jurors' side. In this scenario, the learned language test (Spanish) is found to be correlated with participants' self-perceived confidence levels (CL.L2), thus increasing correct rejections at 9-10 CL points. On the other hand, the Spanish group found a positive correlation for the target-present unknown language test (U1 and CL.U1), which is translated pragmatically into higher hit rates and less false alarm/miss rates at the highest confidence levels (8-10). As a side note, it would appear that the target-present familiar language is close to significance ($p= 0.072$), which coincidentally shares the same trend as in U1's test scores.

When combining both group of jurors' test scores, the second learned language becomes relevant again, although it exhibits the only case found where false alarms outnumber correct rejections throughout the entire 10-point Likert scale. This is likely caused by the addition of Spanish group's L2's test scores (whose correlation with CL.L2 was non-significant), as the boxplot in the previous section demonstrates. Despite this, the correlation is still significant overall in this grouped account. The instances where correct rejections get the closest to false alarms appear at the 7-9 CLpoints. Conversely, the unknown target-present test's (U1) correlation with CL.U1 is found nearly-significant

($p= 0.056$). Again, the boxplot illustrates how confidence levels increase at the instances of hits only for the Spanish group, reaching its maximum potential at 8-10 CL points.

As for discrimination tasks, it is found that their scarce variance concerned with test scores does not allow for an insightful exploration of their correlations with CL. Even though it is not statistically significant, CL's means appear to diminish at target-absent tests, which could be indicative of the uncertainty that the background noises exert on the hearer.

To conclude, the third hypothesis (a heightened self-perceived confidence level at speaker recognition tasks has a positive effect on the voice line-up's outcome) is accepted only on some specific cases (British L2, and Spanish U1), while the null hypothesis is retained in the rest of scenarios. As noticed, no apparent pattern seems to outline the hearer's predisposition to match their CL and actual test scores, since it appears unrelated to cultural groups, language familiarity, and even experimental conditions (the presence/absence of the target and the interference of background noises). Either way, confidence levels should be considered to predict speaker recognition tests, if at all, when they reach the highest points in the Likert scale (7-10). Even then, caution must be exercised in relation to this variable's validity.

4.4. AGE AND GENDER

Moving to the fourth hypothesis, it attempts to unearth whether success rates and false alarms in speaker recognition tasks are influenced by the hearer's gender and/or age. With this in mind, the null hypothesis (H_0) and the alternative hypothesis (H_4) are formulated below:

- H_0 : The efficiency at speaker recognition is not conditioned by age and gender.
- H_4 : The efficiency at speaker recognition is conditioned by age and gender.

Unsurprisingly, the variables employed for this particular hypothesis are *test scores* (dependent variable), and *age* and *gender* (independent variables, or predictors). They are coded according to the following table:

Variable		Options	Code
Test scores	Overall.scores	-	0-6
	Identification tasks (target-present): F1, L1, U1	False alarm	1
		Miss	1
		Hit	2
	Identification tasks (target-absent): F2, L2, U2 Discrimination tasks (target-present): F1.Dis, L1.Dis, U1.Dis	False alarm	1
		Correct rejection	2
Gender	Male	1	
	Female	2	
Age	18-22	1	
	Over 22	2	

Table 62. Dependent and independent variables considered for the fourth hypothesis.

As noticed in table 62 above, the first set of variables are concerned with test scores. In this domain, the first variable in the list is *overall.scores*, which computes the total amount of points collected from the six identification tests (F1, F2, L1, L2, U1, U2), and thus its score ranges from 0 to 6 depending on the juror’s success rates. Below this level, each individual identification and discrimination test is broken down into the possibilities offered for each one of them (where 1 stands for failure and 2 stands for success in said speaker recognition tasks), with the exclusion of discrimination tests for the target-absent condition (due to their lack of variance in their test score values).

As for the last two variables, they represent the independent variables acting as predictors. *Age* reflects an ordinal scale, with interval variables reflecting higher/lower values in accordance with their assigned code (18-22=1, and Over 22=2). Conversely, *gender* does not follow the same criterion, and is only coded as a binary code without existing hierarchies (Male=1, Female=2).

As a guiding principle, the linear fixed effects model effect shall include interaction terms (i.e. *gender*age*) in the model, provided that they are significant in accounting for the variance of the dependent variable (test scores). Otherwise, only main effects acting as independent variables shall be considered. The reason for this is that SPSS F-tests yield varying p-values when a main effect is involved in an interaction, especially if this interaction effect is meaningful (Murray 1998: 293).

As customary, the order established to undertake this analysis puts identification tasks first, followed by discrimination tasks. As for cultural groups, they follow the alphabetical order, from most specific to most generic accounts (British, Spanish, and British and Spanish). After this, overall scores take priority and, after looking for potential correlations in this area, it shall proceed to the first (target-present) and second (target-absent) experimental condition. A notable exception occurs in discrimination tests, which do not consider target-absent tests.

4.4.1. Identification

The first section considers both target-present language test (F1, L1, U1) and target-absent (F2, L2, U2) identification tasks to discover whether age and gender are influential in each scenario and, if possible, whether it could predict the scores of said aural-perception tests. Although not without considering the overall scores obtained first for the analysis. As commented already, the order for exploring the selected cultural groups goes as follows: British, Spanish, and British and Spanish group.

4.4.1.1. British group

Before proceeding to the statistical analysis itself, providing a brief overview on the target population's age and gender seems advisable, so that the reader can get a closer look at the juror's profiles in relation to said sub-categories:

Age	Gender		Total	
	Male	Female		
18-22	7	24	31	49
+22	3	15	18	
Total	10	39		
	49			

Table 63. Distribution of age and gender across British jurors.

As perceived in table 19, there seems to be an unbalanced design where females (39) outnumber males (10) in every age group, which could compromise the validity of statistical tests due to some sub-categories being underrepresented. Albeit not as uneven as gender, age's distribution is somewhat skewed as well, with higher participants aged 18-22 (31) and less jurors whose age exceeds 22 (18). However, this distribution is justifiable inasmuch as it mirrors the typical profiles encountered in our target population: university students. Logically, undergraduate studies will typically include the 18-22 age range and, as the figures show, females seem higher in number in linguistics-related degrees.

Moving to the statistics model, the main effects *gender* and *age* were entered alongside their interaction term (*gender*age*). However, the latter did not reach the established level of significance ($\alpha = 0.05$) in any of the proposed scenarios, and therefore was discarded from the model. Hence, the resulting operations included main effects only.

After computing a linear fixed effects model, it is shown that *age* and *gender* do not appear as relevant predictors for overall scores. For this reason, the analysis shall proceed to account for the aforementioned predictors' influence on each individual test scores undertaken in this research study, namely F1, L1, U1, and F2, L2, U2.

Only one of the aural-perception language tests above rendered statistically significant results for one of the predictors. The table below illustrates the case for L2's scores being correlated with age.

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.517990	.108960	46	13.932	.000	1.298666	1.737315
[age=1]	-.300068	.133192	46	-2.253	.029	-.568170	-.031966
[age=2]	0 ^b	0
[gender=1]	-.107943	.159317	46	-.678	.501	-.428632	.212746
[gender=2]	0 ^b	0

a. Dependent Variable: L2.

b. This parameter is set to zero because it is redundant.

Table 64. Estimates of age and gender in L2’s scores (British group, identification tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

Since categorical variables were coded dichotomously for *gender* (Male=1, Female=2) and *age* (18-22=1, +22=2), the estimates shown in table 64 above do not seem to be meaningful at first glance. Upon closer inspection, however, it is seen that the estimate for *age1* carries a negative value (-0.3), which could be interpreted as the dependent’s variable’s (L2 test scores) values decreasing as *age* also decreases.

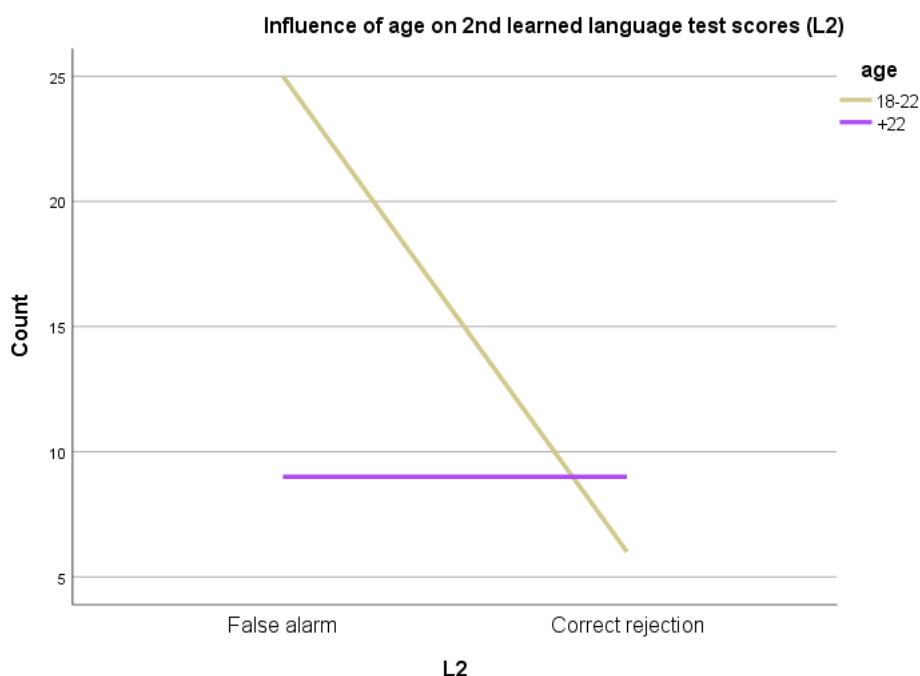


Figure 29. Multiple line graph of age’s influence upon L2’s scores in the British group (identification tests).

When plotting age's influence on L2's responses, figure 29 confirms the first premise. Therefore, it seems that younger participants (18-22) decrease test scores (false alarm=1, correct rejection=2) at the alpha level of 0.05 ($p= 0.029$). Also, notice that the 95% confidence interval (henceforth called CI) is not too wide (from -0.5 to -0.03), given the relatively low standard error (0.13). A remarkable feature observed in the multiple line graph above is the stability of responses in the older group of jurors (over 22), and thus false alarms are prominently diminished in this slightly older group in contrast with the first one.

4.4.1.2. Spanish group

As done in the previous point investigating the British group, this sub-section also initiates its analytical procedure through a descriptive account on Spanish participants' age and gender groups:

Age	Gender		Total	
	Male	Female		
18-22	7	32	39	58
+22	9	10	19	
Total	16	42		
	58			

Table 65. Distribution of age and gender across Spanish jurors.

Similar to the British jurors' distribution, Spanish female participants outnumber males in the younger age group (18-22), as noticed in table 65. However, this distance is reduced in older participants (+22). As discussed above, this alleged unevenness is representative of the tendencies observed within the target population and thus should not compromise the reliability of statistical results.

Since *age* and *gender* do not seem to be relevant predictors when considering *overall.scores* as the dependent variable, each combination of language test (familiar, learned, and unknown) and experimental condition (target-present without background noises, and target-absent with background noises) is explored hereafter. The first learned language test (L1) identified *gender* as a potential predictor.

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.266675	.121436	55	10.431	.000	1.023313	1.510038
[age=1]	-.068761	.131279	55	-.524	.603	-.331851	.194329
[age=2]	0 ^b	0
[gender=1]	.325908	.137855	55	2.364	.022	.049641	.602175
[gender=2]	0 ^b	0

a. Dependent Variable: L1.

b. This parameter is set to zero because it is redundant.

Table 66. Estimates of age and gender in L1's scores (Spanish group, identification tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

It should be reminded that the both the dependent variable L1 and the independent variable *gender* are coded in a dichotomous manner for the purposes of conducting the linear fixed effects model. The positive estimate shown in table 66 for *gender1* (Male) appears to suggest that this sub-category enhances L1 results. This is further corroborated with a small p-value (0.022) exhibiting a low standard error (0.13). Also, the 95% CI bounds remain on positive values (0.05-0.6), thus indicating higher values lingering around 1 (Male), as opposed to 2 (Female).

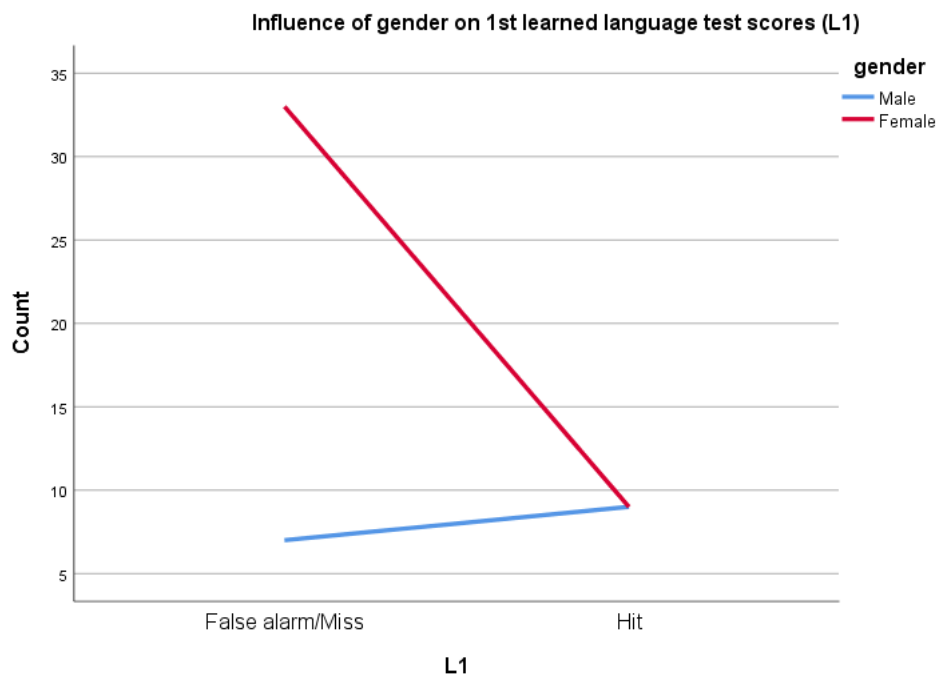


Figure 30. Multiple line graph of gender’s influence upon L1’s scores in the Spanish group (identification tests).

Said correlation can be readily observed in figure 30 above, where males’ responses do not seem to vary significantly, whilst females’ false alarms appear prominently higher than their chances to produce hits.

Besides L1’s case, the familiar (F1) and unknown (U1) language tests do not spot existing statistical relationships between *age* and/or *gender* and test scores, as far as the first experimental condition (target-present without background noises) is concerned.

As for the second perception tests, the familiar language (F2) identifies relevant sociolinguistic predictors, again pivoting around *gender*:

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.134947	.118052	55	9.614	.000	.898366	1.371528
[age=1]	.072882	.127621	55	.571	.570	-.182877	.328641
[age=2]	0 ^b	0
[gender=1]	.270667	.134013	55	2.020	.048	.002099	.539235
[gender=2]	0 ^b	0

a. Dependent Variable: F2.

b. This parameter is set to zero because it is redundant.

Table 67. Estimates of age and gender in F2's scores (Spanish group, identification tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

The resulting p-value obtained after computing the linear fixed effects model ($p = 0.048$) renders this correlation significant at the established significance level ($\alpha = 0.05$). The 95% CI revolves around 0.002 and 0.5 with a low standard error (0.13), as table 67 shows. These values are, in turn, reflecting a positive estimate for the dummy variable *gender1*, whose stratum (males) appears to produce a more positive outcome on the speaker recognition tests, as figure 31 illustrates:

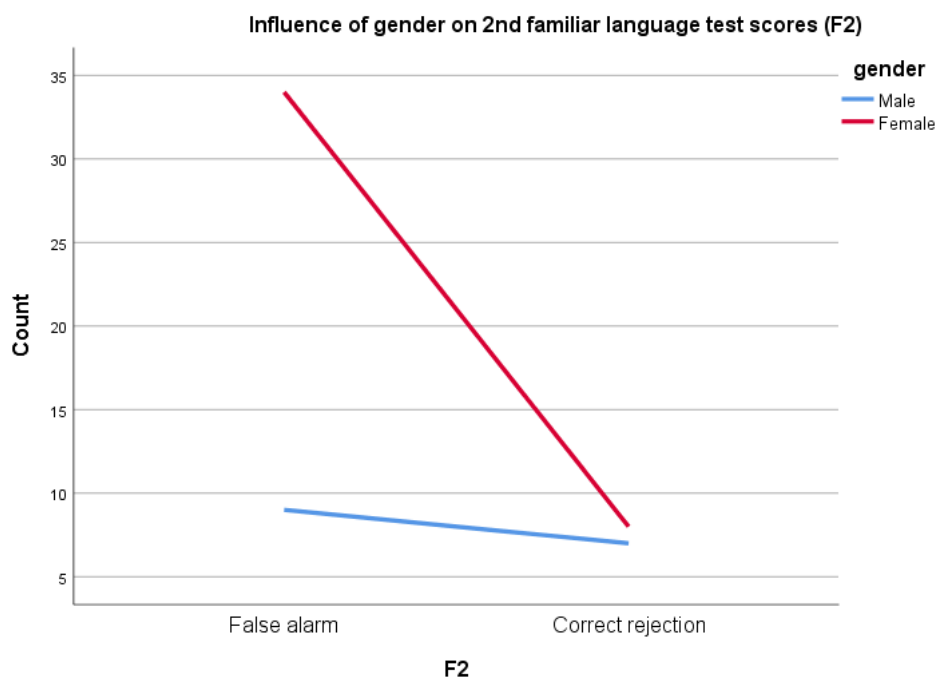


Figure 31. Multiple line graph of gender's influence upon F2's scores in the Spanish group (identification tests).

The second familiar language test (F2) displays a similar pattern to L1's, even when sharing *gender* as the predictor for test score results. In this example, and as shown in the graph above, females in the target population appear more prone to false alarms than males, which sharply contrasts with the latter response type (correct rejection and false alarm) distribution. Albeit non-significant, the decreasing slope on the males' side contrasts with L1's ascending slope, thus reflecting lower and higher chances of ending up with correct rejections and hits, respectively.

As for the remaining language tests (L2 and U2), there is no statistically significant relationship asserted between *age* and/or *gender* as predictors on language test scores.

4.4.1.3. British and Spanish group

An initial exploration of both British and Spanish jurors' distribution of age and gender is shown below:

Age	Gender		Total	
	Male	Female		
18-22	14	56	70	107
+22	12	25	37	
Total	26	81	107	

Table 68. Distribution of age and gender across British and Spanish jurors.

In line with the separate accounts of each group of jurors, distances between the typical sub-categories only increase when grouping participants together, as table 68 shows. Just as in the previous cases, females aged 18-22 are the most numerous sub-group surveyed (56), whereas males over 22 are the least represented (12). It is also important to note that, despite increasing the number of respondents, figures increase proportionally maintaining roughly the same ratio between the variables and their respective sub-categories. This is indicative of the sample obtained being representative of the intended surveyed groups, and thus the same rationale as the one exposed above applies in this respect, too.

The interaction term *gender*age* could not predict overall scores effectively, nor could it bring out significant correlations within the sub-set of language tests separately (F1, L1,

U1, and F2, L2, U2). It is therefore excluded, and only *age* and *gender* main effects are subsequently considered in the model.

Neither age nor gender are not found to be significant predictors for overall scores in the first assortment of language-tests (F1, L1, U1). However, this situation is changed when moving to the second experimental condition. In the first case, *gender*'s correlation with F2 test scores is found relevant.

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.123180	.076453	104	14.691	.000	.971571	1.274790
[age=1]	.107543	.087062	104	1.235	.220	-.065105	.280191
[age=2]	0 ^b	0
[gender=1]	.203527	.096550	104	2.108	.037	.012065	.394989
[gender=2]	0 ^b	0

a. Dependent Variable: F2.

b. This parameter is set to zero because it is redundant.

Table 69. Estimates of age and gender in F2's scores (British and Spanish group, identification tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

The reported p-value (0.037) in table 69 finds that F2 test scores are influenced the most by *gender1* (male), as the positive estimate (0.20) shows. Also, the standard error (0.09) is smaller than in the previous cases, and the CI appears to be considerably narrow (0.01-0.39), which again reinforces the idea of greater scores being ascribed to *gender1*.

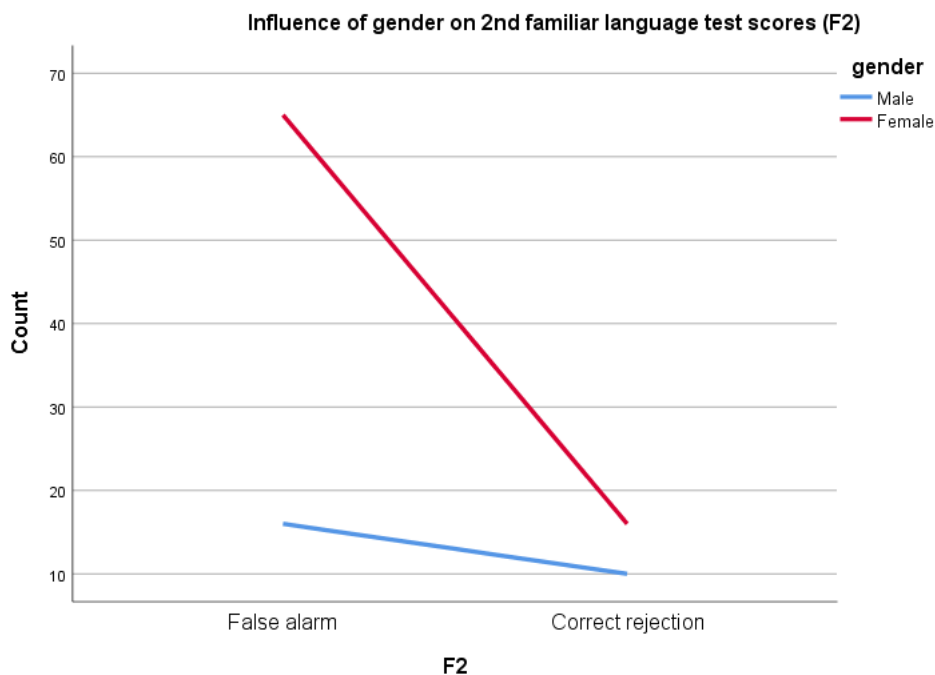


Figure 32. Multiple line graph of gender’s influence upon F2’s scores in the British and Spanish group (identification tests).

In light of the statistic measures shown on estimates of fixed effects’ table, the multiple line graph above shows that, despite obtaining less correct rejections than false alarms, males manage to keep the balance between failing and succeeding the speaker identification test, much in contrast with females’ high scores on false alarms. This very relationship appearing in figure 32 resembles very closely the one reported in the Spanish group’s F2 case (figure 31). As a matter of fact, it could be argued that such existing relationship observed in the Spanish case has influenced the outcome in this grouped account. Nevertheless, it must be noted that the British group could not find any significant predictor for F2 scores. The resulting correlation is, therefore, strong enough to be considered regardless of the cultural group surveyed.

On another note, the target-absent learned language test (L2) found *age* as the relevant predictor for its scores, instead.

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.422172	.075662	104	18.796	.000	1.272133	1.572212
[age=1]	-.271357	.086161	104	-3.149	.002	-.442217	-.100496
[age=2]	0 ^b	0
[gender=1]	.031635	.095550	104	.331	.741	-.157845	.221115
[gender=2]	0 ^b	0

a. Dependent Variable: L2.

b. This parameter is set to zero because it is redundant.

Table 70. Estimates of age and gender in L2's scores (British and Spanish group, identification tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

As the negative estimate of *age1* (-0.27) reflects, it seems that the younger age group (18-22) are less likely to score better results. As table 70 suggests, this relationship appears fairly reinforced with a tiny p-value (0.002), and with a small standard error (0.08). Furthermore, the CI interval of values appear to be narrower (-0.4 to -0.1) than in the previous cases, which seem to reinforce the certainty of this correlation.

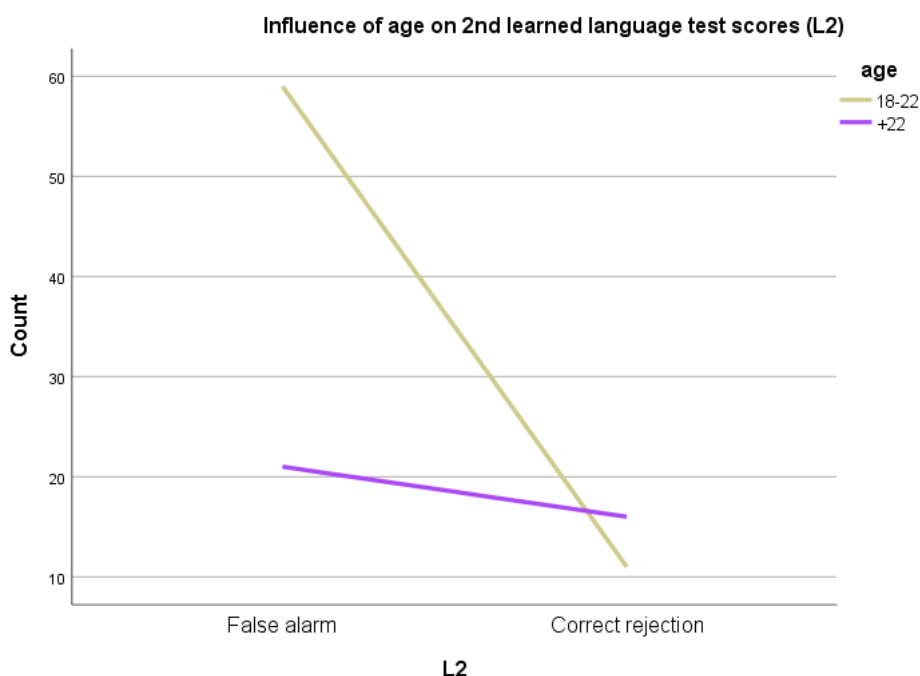


Figure 33. Multiple line graph of age's influence upon L2's scores in the British and Spanish group (identification tests).

These premises are confirmed in the visual representation of the data shown in figure 33 above. In this graph, false alarms appear prominently higher in the first younger group (18-22) than in the second age group (+22). In terms of correct rejections, the opposite trend can be observed: they are facilitated in jurors over 22 years old, while the younger group cannot reach similar figures.

Just as British and Spanish' F2 scores were influenced by *gender* in the Spanish group-only analysis, L2's relationship with *age* appears correlated in a similar vein. Specifically, it is now the British group the one that contributes to bringing out *age*'s significance as a predictor. However, and as in the previous scenario, the Spanish group did not find any relevant sociolinguistic predictor for L2, and thus its influence on the grouped account is somewhat limited. Not only this, but bear in mind that the resulting p-value (0.002) in this particular case (British and Spanish) is remarkably smaller in comparison with the one retrieved from the British group's analysis (0.029), which translates into a stronger correlation in the former. As suggested in F2's test scores being influenced by *gender*, it seems that the finding of L2's results varying across age groups is not affected by adding or removing cultural groups.

As for the second unknown language test (U2), no significant relationships were found between sociolinguistic predictors and test scores.

4.4.2. Discrimination

The second part of this analytical procedure attempts to investigate discrimination tests, only with the target-present condition (F1.Dis, L1.Dis, U1.Dis) with the purpose of finding significant predictors to the scores obtained in the aforementioned aural-perception tests. As explained before, a linear fixed effects model takes *age* and *gender* to predict overall scores first, followed by individualised accounts of each test in isolation thereafter. The analysis begins with the British group, Spanish group, and ends with the British and Spanish grouped account.

4.4.2.1. British group

After running a linear fixed effects model, *age* and *gender* (and their interaction term) were not considered relevant predictors neither for overall scores nor for any of the discrimination tasks in the target-present condition. Here follows an account of the mean test scores obtained in each test amongst the 49 British participants:

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
F1.Dis	49	1.00	2.00	1.92	.277
L1.Dis	49	2.00	2.00	2.00	.000
U1.Dis	49	1.00	2.00	1.98	.143
Valid N (listwise)	49				

Table 71. Descriptive statistics of British discrimination tests, 1st exp. condition.

As observed initially in table 71, mean test scores are incredibly high, since the maximum score is 2.00, which is the assigned code for correct rejections/hits. In the table above, it remains clear that not only most of the respondents were successful in the discrimination task, but also that the learned language's case (L1) rendered no errors in any of the 49 respondents. The other two tests contemplated a few errors, but mean test scores still surpass 1.90 with significantly low standard deviations. As it stands right now, this scenario displaying a lack of significant variation in the data set does not allow for this research to find relevant predictors.

4.4.2.2. Spanish group

The intended linear fixed effects model devoted to the Spanish group did not deem *age* and *gender* as statistically significant predictors for overall scores in discrimination tasks. The same conclusion was reached when predicting each individual language test score with the above-mentioned independent variables.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
F1.Dis	58	1.00	2.00	1.97	.184
L1.Dis	58	1.00	2.00	1.98	.131
U1.Dis	58	1.00	2.00	1.98	.131
Valid N (listwise)	58				

Table 72. Descriptive statistics of Spanish discrimination tests, 1st exp. condition.

From table 72 above on descriptive statistics, it can be surmised that jurors excel at discrimination tasks irrespective of the type of language test employed, as noticed by looking at mean test scores higher than 1.90 with reduced standard deviations. As a matter of fact, finding significant predictors that explain the variance of test scores appears unattainable when said variance is nearly non-existent.

4.4.2.3. British and Spanish group

After conducting a linear fixed effects model, the resulting model comprising *age* and *gender* (as well as their interaction term) as predictors for the target-present language test scores in discrimination tasks (overall scores and separate tests) found no statistically significant correlations among the variables already mentioned.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
F1.Dis	107	1.00	2.00	1.9439	.23115
L1.Dis	107	1.00	2.00	1.9907	.09667
U1.Dis	107	1.00	2.00	1.9813	.13607
Valid N (listwise)	107				

Table 73. Descriptive statistics of British and Spanish discrimination tests, 1st exp. condition.

As shown in table 73 above, discrimination test scores' variance is not high enough to determine whether any sociolinguistic predictor is influential in its outcome, as inferred by the few instances of false alarms (with mean scores nearly reaching 2.00), and low standard deviations (hence indicating values close to the mean of the sample). As explained before, discrimination tests' variables are also coded dichotomously, and thus lesser scores (1.00) stand for false alarms, whereas a higher number (2.00) reflects a correct rejection. Having clarified this point, there seems to be no feasible method to

predict discrimination scores even when putting together the scores generated by the two groups of jurors.

4.4.3. Summary of results

As a final remark, the fourth hypothesis (the efficiency at speaker recognition is conditioned by age and gender) can be accepted in some cases, and discarded on others. If the statement refers to both age and gender (both separately and their interaction term) impinging on every language test score, the null hypothesis must be retained. In spite of this, some patterns were discerned in identification tests:

- British L2 correlated with *age* ($p= 0.029$).
- Spanish L1 correlated with *gender* ($p= 0.022$).
- Spanish F2 correlated with *gender* ($p= 0.048$).
- British and Spanish F2 correlated with *gender* ($p= 0.037$).
- British and Spanish L2 correlated with *age* ($p= 0.002$).

The list above draws two tendencies that surfaced during this experiment: the first one being the stability of language tests' correlations (L2 with *age*, and F2 with *gender*) across cultural groups. What is more, their relationships appear to be strengthened with a reduced p-value in the grouped account (British and Spanish) even if one of the groups did not find significant predictors in said language tests (no correlations found in British F2, nor in Spanish L2). In such cases, correlations in the grouped scenario would be expected to diminish given the reason exposed above. However, it seems that this set up enriches the results for both groups rather than canceling each other out. The second trend observed is the consistency of sub-categories relationships with the language test scores mentioned above. It seems that *gender1* (male) and *age2* (over 22) are the most reliable strata in the target population, which exhibit less false alarm rates than *gender2* (female) and *age1* (18-22).

It could be argued that this conclusion emerged as a result of an unbalanced sample which, coincidentally, includes less participants over 22 years old and less males. Even so, it must be reminded that success rates equal, and at times surpass, those scored in the

majority group (females aged 18-22), which would have, in principle, more chances to score hits/correct rejections. In this sense, the differences lie in the proportional distance between false alarms and hit/correct rejections for each group, and so it is proven that said distance is much shorter in males (Spanish L1, F2, and British and Spanish F2) and in jurors over 22 years old (British L2, and British and Spanish L2). Despite this, replicating the current research design with a more balanced sample is preferable and encouraged for future research.

As for discrimination tasks, the null hypothesis is undoubtedly retained given the lack of statistically significant results in this sub-section.

4.5. CULTURAL GROUPS AND LINGUISTIC ENVIRONMENT

With the aim of standardising the auditory abilities impacting on speaker recognition processes, the fifth hypothesis (H₅): is formulated below, along with the null hypothesis (H₀):

- H₀: Speaker recognition capabilities are influenced by cultural groups (Spanish or British) and linguistic environment (monolingual or bilingual).
- H₅: Speaker recognition capabilities are not influenced by cultural groups (Spanish or British) nor by linguistic environment (monolingual or bilingual).

To address this research question, each set of language test is grouped together with the categories mentioned above, namely cultural group (British and Spanish), and linguistic environment (monolingual or bilingual). This will in turn give rise to specific variables when analysing cultural groups: those concerned with identification tasks in the target-present (F1.All, L1.All, U1.All) and target-absent (F2.All, L2.All, U2.All) condition, and those in discrimination tests examined on the target-present (F1.Dis.All, L1.Dis.All, U1.Dis.All) experimental condition only.

Variable	Options	Code
Cultural group	British	2
	Spanish	1
Linguistic environment	Monolingual	1
	Bilingual	2
Test scores	False alarm/Miss	1
	Hit/Correct rejection	2

Table 74. Dummy codes assigned for each variable considered in hypothesis 5.

When it comes to spot differences amongst linguistic environments, the *All* label is removed from each abbreviation. Hence the names assigned in table 74 for each test: F1, L1, U1, F2, L2, U2, F1.Dis, L1.Dis, and U1.Dis. The resulting composite arrangement of test scores is analysed through a set of Mann-Whitney U tests, which, unlike Wilcoxon signed.rank tests, take into account unrelated samples. This statistical test suits the interests of this specific research question, as test scores have been mixed from the two groups (British and Spanish) and sub-groups (monolingual and bilingual). After a conclusion is reached on discerning whether there are distinctions between the aforementioned groups, a measure of effect size (r) is incorporated to the analysis which, similar to Phi's coefficient and Cramer's V, report the magnitude of the existing association between the selected variables.

Divergences amongst cultural groups are examined first, and linguistic environments are analysed separately thereafter. Identification tests take priority, and discrimination tasks shall follow shortly after. As customarily done throughout this analytical chapter, the order shall inspect the British jurors' group and proceed to the Spanish case afterwards. Even though the compiled table on statistical significances shall display all language tests, comments are made on target-present aural-perception tests, followed by the target-absent experimental condition.

4.5.1. Cultural groups

As a starting point, individual language tests results are juxtaposed and compared amongst British (n=49) and Spanish (n=58) jurors for hypothesis testing. Since every

participant provides one answer per language test, a total of 107 responses are registered according to success (2) or failure (1) in speaker recognition tasks.

4.5.1.1. Identification

An initial step looks at cultural groups’ influence on identification tasks. Specifically, the table below summarises the extant relationships found:

Test Statistics ^a						
	F1.All	F2.All	L1.All	L2.All	U1.All	U2.All
Mann-Whitney U	1270.500	1372.500	1137.000	1280.000	1270.500	1223.000
Wilcoxon W	2495.500	2597.500	2848.000	2991.000	2495.500	2448.000
Z	-1.087	-.408	-2.091	-1.172	-1.087	-1.870
Asymp. Sig. (2-tailed)	.277	.683	.037	.241	.277	.061

a. Grouping Variable: country

Table 75. Mann-Whitney U test on the influence of country (cultural groups) upon identification scores. Statistically significant values are marked in bold ($\alpha = 0.05$).

Through the familiar (F1.All, F2.All), unknown (U1.All, U2.All) and second learned language tests (L2.All), there was no significant statistical difference when using the country of origin (Spanish or British) as the grouping variable, as table 75 asserts. However, a Mann-Whitney U test indicated that scores in the 1st test for the learned language (target-present) was greater in the British group (M= 59.80) than in the Spanish group (M= 49.10, U = 1137, p= 0.037, r = 0.20).

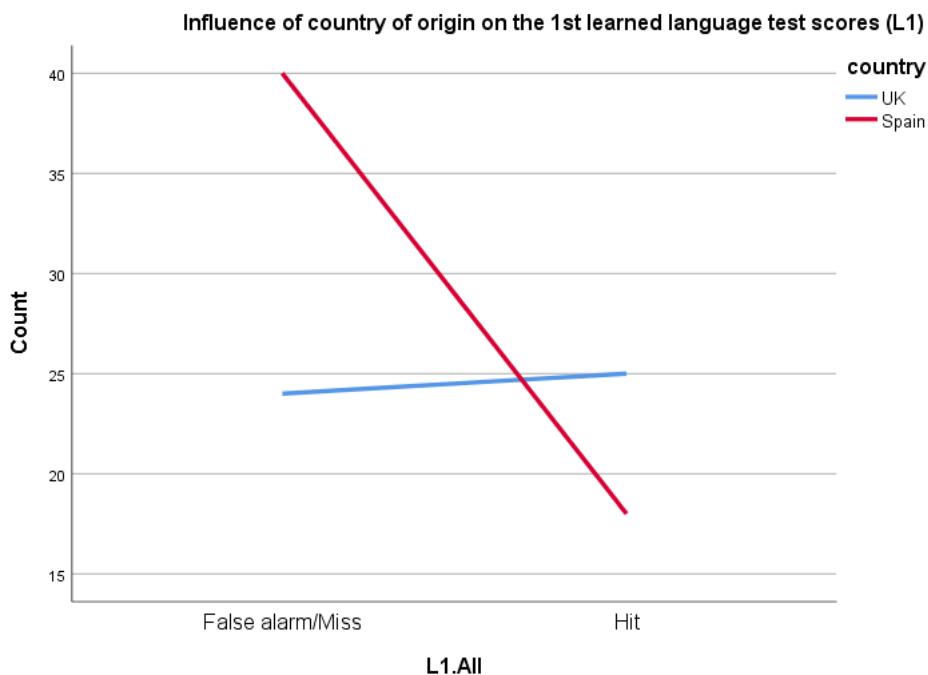


Figure 34. Multiple line graph of country of origin’s influence upon L1’s scores in the British and Spanish group (identification test).

As the absolute value of the Pearson product-moment coefficient implies ($r = 0.20$), the preponderance of British scores over the Spanish jurors is near a medium effect size. As figure 34 illustrates, Spanish jurors are exceedingly prone to false alarms, while hit rates are significantly inferior in comparison. Much in contrast with this group, British jurors keep the balance between the two outcomes, and even manage to slightly raise hit rates over false alarm/miss.

4.5.1.2. Discrimination

Secondly, the thesis proceeds to explore whether discrimination test scores display significant differences across cultural groups (whose grouping variable is labelled as *country*).

Test Statistics^a			
	F1.Dis.All	L1.Dis.All	U1.Dis.All
Mann-Whitney U	1354.000	1396.500	1416.500
Wilcoxon W	2579.000	3107.500	2641.500
Z	-1.051	-.919	-.120
Asymp. Sig. (2-tailed)	.293	.358	.905

a. Grouping Variable: country

Table 76. Mann-Whitney U test on the influence of country (cultural groups) upon discrimination scores.

As noticed in table 76, a Mann-Whitney U test has proven that no perception test is influenced by the grouping variable *country* (Spanish or British group) within the target-present without background noises experimental condition. As for the second post-tests (target-absent with background noises), they do not need to undergo the same statistical process, since all responses will invariably yield correct rejections due to the suspect being absent from the voice line-up itself.

4.5.2. Linguistic environment

Once the analysis of cultural groups is concluded, the next layer of sub-groups comprised in said cultural groups is examined here. The mentioned levels of analysis are, in order of appearance, British monolingual (28) and bilingual (21), and Spanish monolingual (33) and bilingual (25) linguistic environments. The following sub-sections shall deal with identification tests (4.5.2.1.) and discrimination tasks (4.5.2.2.), respectively.

4.5.2.1. Identification

Starting with the British group, it is reminded that the data coming from monolingual linguistic environment refers to universities located at the South East England areas (Winchester, Southampton, and Roehampton), whereas the bilingual side represents the Welsh-speaking community (Swansea, Bangor, and Cardiff). The table below reports the statistical measures needed to spot significant relationships:

Test Statistics^a

	F1	F2	L1	L2	U1	U2
Mann-Whitney U	245.000	252.000	227.500	255.500	245.000	290.500
Wilcoxon W	476.000	483.000	458.500	661.500	476.000	521.500
Z	-1.155	-1.174	-1.551	-.974	-1.155	-.135
Asymp. Sig. (2-tailed)	.248	.240	.121	.330	.248	.893

a. Grouping Variable: ling.environment

Table 77. Mann-Whitney U test on the influence of linguistic environment upon identification scores in the British group.

As noted in table 77 above, none of the language tests completed by the British jurors exhibits significant differences across monolingual and bilingual speech communities. It is worth noting that this situation applies in identification tests regardless of the type of language (familiar, learned, an unknown) and experimental condition (target-present and target-absent).

As for the Spanish group of respondents, the monolingual group belongs to the Andalusian universities of Seville and Granada, whereas the bilingual community is centered on the Catalan-speaking population (València, Barcelona, and Girona). Similar to British group’s analysis, a table reporting the needed statistical correlations is displayed hereby:

Test Statistics^a

	F1	F2	L1	L2	U1	U2
Mann-Whitney U	344.000	399.000	376.500	407.500	402.000	384.500
Wilcoxon W	669.000	724.000	937.500	732.500	727.000	945.500
Z	-1.245	-.279	-.705	-.112	-.191	-.593
Asymp. Sig. (2-tailed)	.213	.780	.481	.911	.849	.553

a. Grouping Variable: ling.environment

Table 78. Mann-Whitney U test on the influence of linguistic environment upon identification scores in the Spanish group.

In a similar vein, Spanish identification test scores’ variance does not seem to be altered excessively when entering *linguistic environment* as the grouping variable, according to the reported values in table 78 above. Just as in the previous case, this tendency is reflected throughout every language test, be it with or without suspect/background noises.

4.5.2.2. Discrimination

Moving to discrimination tests, it is sought here to find out whether the monolingual/bilingual distinction is relevant within the British and Spanish target population surveyed. Firstly, results from the British jurors are drawn here:

	F1.Dis	L1.Dis	U1.Dis
Mann-Whitney U	262.500	294.000	280.000
Wilcoxon W	493.500	525.000	511.000
Z	-1.342	.000	-1.155
Asymp. Sig. (2-tailed)	.180	1.000	.248

a. Grouping Variable: ling.environment

Table 79. Mann-Whitney U test on the influence of linguistic environment upon discrimination scores in the British group.

As inferred above and, as commented previously, the target-absent tests were removed due to the lack of variation in their values. Target-present tests, however, do not reveal significant results either, as table 79 demonstrates. Not only this, but test scores appear quite similar across both sub-groups, let alone L1 test scores, which reflect an identical distribution of values across the studied linguistic environments.

Secondly, the test scores provided by the Spanish community are consulted to test if the same scenario is repeated.

	F1.Dis	L1.Dis	U1.Dis
Mann-Whitney U	408.500	396.000	400.000
Wilcoxon W	733.500	721.000	961.000
Z	-.199	-1.149	-.870
Asymp. Sig. (2-tailed)	.842	.251	.384

a. Grouping Variable: ling.environment

Table 80. Mann-Whitney U test on the influence of linguistic environment upon discrimination scores in the Spanish group.

Notably, no statistically significant differences are found for F1.Dis, L1.Dis, and U1.Dis when considering *linguistic environment* as the grouping variable, as shown in table 80.

This finding is in line with the British case, where none of the language tests were reported to be dissimilar enough in terms of succeeding or failing the speaker discrimination task.

4.5.3. Summary of results

In this specific hypothesis, no concerns around the validity of the results obtained should arise, since the existing proportions of sub-categories within the main variables reflected a balanced sample, both in cultural groups and linguistic environments. To answer the proposed hypothesis (speaker recognition capabilities are not influenced by cultural groups nor by linguistic environment), it should be split into two halves according to its two constituents (cultural groups and linguistic environment).

In this regard, the null hypothesis is rejected only in the identification learned language (target-present) case, where British jurors performed significantly better than Spanish respondents. It could be hypothesised that the former group is endowed with better auditory capacities in comparison with Spanish participants, or perhaps that the voice samples employed for Spanish L1's test (English input) were perceptually more similar to the suspect's than British L1's (Spanish input) test. This question shall be addressed in chapter 6 (*Discussion*) after putting together the results from the acoustic-phonetic analysis (3.7.3.) on inter- and intra-speaker variability.

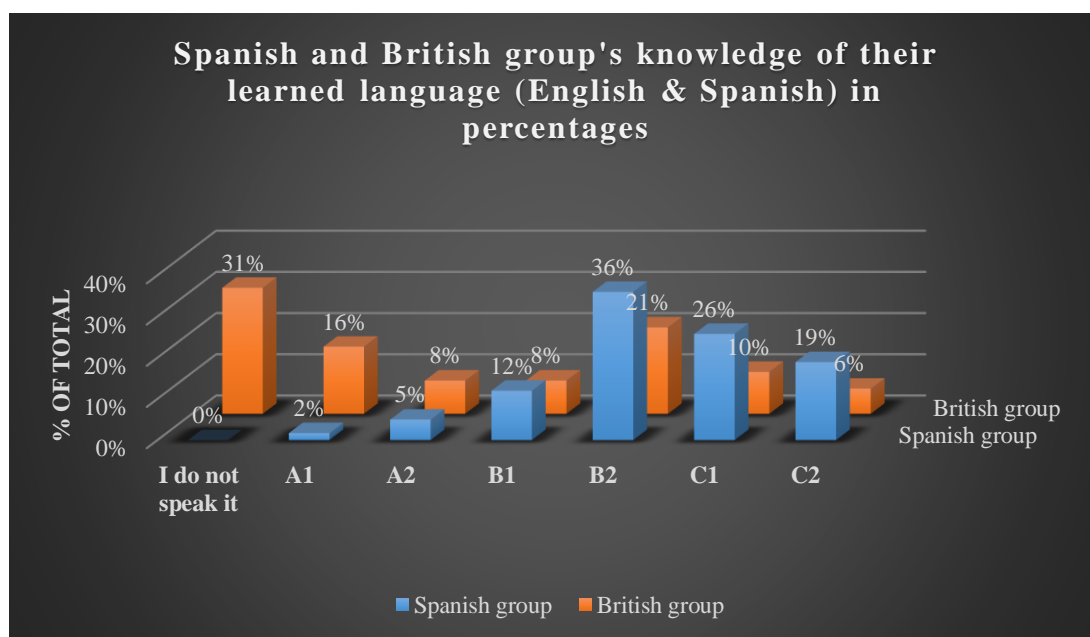


Figure 35. Bar graph on British and Spanish jurors' knowledge of their learned language.

On another note, the distribution of linguistic proficiencies for each target population could have had an effect on the found differences amongst British (n=49) and Spanish (n=58) L1 scores. Figure 35 offers an overview about each cultural group's knowledge of their respective learned language. In the Spanish case, higher proportions of the target population remain on upper-intermediate and advanced tiers. British jurors, however, are centered mainly on not knowing Spanish (31%), even if a decent number of participants (21%) possess upper-intermediate (B2) linguistic skills. If a linear positive relationship between language knowledge and recognition of speakers of said language is assumed, the Spanish group would be in an advantageous position given the percentages shown above. Nevertheless, the case seems reversed, as British participants demonstrated their efficiency at this task through L1 scores. It should not be overlooked that this differentiation between British and Spanish jurors does not occur in the target-absent condition for the learned language (L2), and thus other factors besides linguistic skills seem to be at play.

In contrast with the above, the remaining identification language tests displayed no distinctive test scores across cultural groups and linguistic environments. The same principle is true for discrimination scores, which did not vary significantly through British and Spanish groups, nor through monolingual and bilingual linguistic environments. Therefore, the null hypothesis is retained in the above-mentioned cases.

4.6. BACKGROUND NOISES AND FALSE ALARMS

In this section, the scores deriving from the two experimental conditions employed in this study shall be compared for the sake of discovering whether background noises affect human aural-perceptual skills. Consequently, hypothesis 6 is formulated alongside its null hypothesis:

- H_0 : Background noises do not hinder voice recognition, and its correlations with false alarms occur by chance.
- H_6 : Background noises hinder voice recognition, thus resulting in a higher frequency of false alarms.

It should be reminded that the target-present condition (1st) displays the voice samples without noise disturbances, while target-absent tests (2nd) add background noises for an increased difficulty. The possible combinations of language tests and experimental conditions is illustrated in the table below:

Experimental condition	Language test	Options	Code
Target-present without background noises	Identification F1, L1, U1	False alarm	1
		Miss	1
		Hit	2
	Discrimination F1.Dis, L1.Dis, U1.Dis	False alarm	1
Correct rejection		2	
Target-absent with background noises	Identification F2, L2, U2	False alarm	1
		Correct rejection	2
	Discrimination F2.Dis, L2.Dis, U2.Dis	Correct rejection	2

Table 81. List of variables (and their assigned codes) contemplated for hypothesis 6.

To refresh the basic concepts around the proposed perception tests, these are broken down into two distinct categories depending on the main purpose, to either identify the suspect in the voice line-up (or detect that the suspect is absent in such cases), and to point at the most dissimilar voice from the suspect's (discrimination). Succeeding in the latter set of tests means selecting any voice sample in the line-up but the suspect's, which is why target-absent discrimination tests only contemplate correct rejections (the juror is not able to select a speaker who is not present there). On identification target-absent tests, however, the option *none of the above* leads to correct rejections. As noticed in the first row in table 81, identification tests in the target-present condition offer three outcomes. Nevertheless, the underlying criterion established in this research decides to ascribe codes on the basis of success (2) or failure (1) in speaker recognition tasks.

The focal point in this sub-section is indeed the comparison between target-present and target-absent conditions. Thus, identification tests are explored first, and discrimination tasks afterwards. Both sections shall initiate their respective analyses with the British group, and proceed with the Spanish group thereafter.

4.6.1. Identification

The content covered in this sub-section includes the comparison between identification target-present (F1, L1, U1) and target-absent (F2, L2, U2) test scores in the British (4.6.1.1.) and Spanish (4.6.1.2.) group.

4.6.1.1. British group

To begin with, the comparison of experimental conditions involving British jurors shall proceed first. The table below displays the relevant statistic measures employed to that end.

	Test scores		
	<u>F1-F2</u>	<u>L1-L2</u>	<u>U1-U2</u>
Z-score	-2.887	-1.768	-3.266
Asymp Sig. (2-tailed)	0.004	0.077	0.001
a. Wilcoxon Signed Ranks Test			

Table 82. Pairwise comparisons on identification language tests' scores across two experimental conditions in the British group. Statistically significant values are marked in bold ($\alpha = 0.05$).

The pairwise comparisons exhibited through the Wilcoxon signed-rank tests in table 82 above show a statistically significant difference between familiar (F1-F2) and unknown language (U1-U2) pairs. The learned language is close to statistical significance ($p=0.07$), although without reaching it.

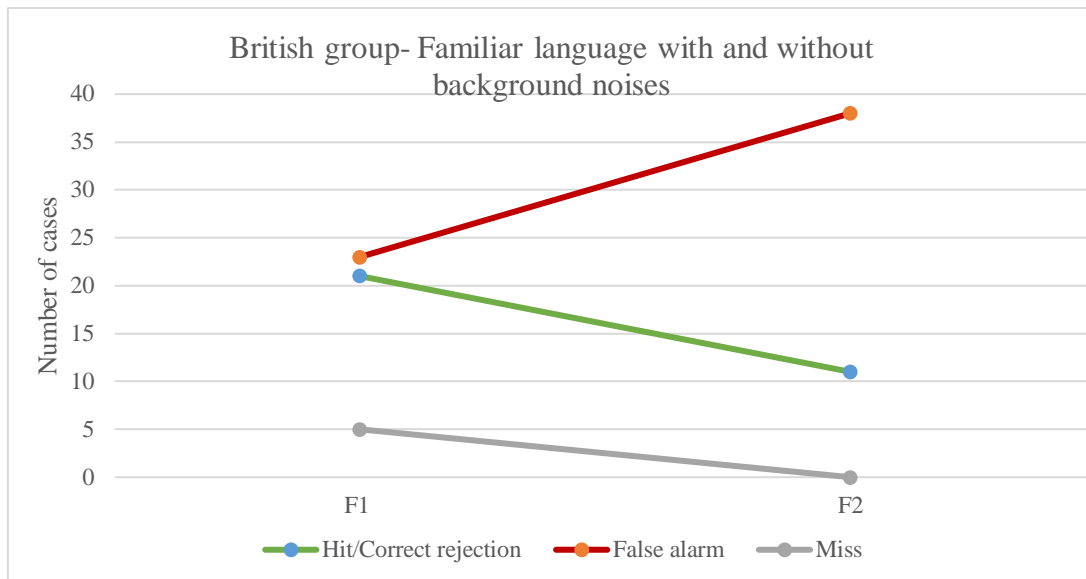


Figure 36. Multiple line graph on British F1-F2 identification tests' comparison.

The negative values resulting from Z-scores are indicative of a negative correlation between F1 and F2, as shown in figure 36 above. For the familiar ($Z=-2.887$) test comparison, it is seen that F2 counts on hit rates are much smaller than the target-present condition (with no noise disturbances). Additionally, false alarms increase dramatically while exposed to background noises (F2).

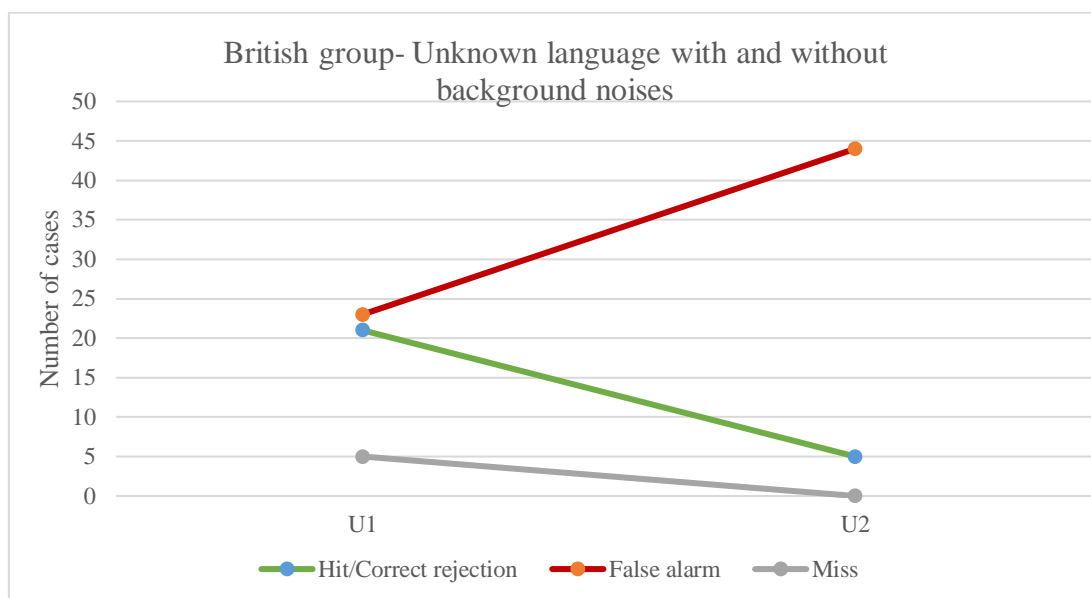


Figure 37. Multiple line graph on British U1-U2 identification tests' comparison.

In a similar fashion, figure 37 shows that the unknown language tests' negative correlation ($Z=-3.266$) reflects a worsened scenario in the second language test (U2),

whereas U1’s test scores fare relatively better. It is worth mentioning that, despite target-present tests being more efficient overall, their hit rates do not surpass the count on false alarms.

4.6.1.2. Spanish group

Secondly, the Spanish group’s pairwise comparison between target-present (without background noise) and target-absent (with background noises) conditions is explored hereby:

		Test scores		
		F1-F2	L1-L2	U1-U2
Z-score		-3.024	-1.279	-3.053
Asymp Sig. (2-tailed)		0.002	0.201	0.002
a. Wilcoxon Signed Ranks Test				

Table 83. Pairwise comparisons on identification language tests’ scores across two experimental conditions in the Spanish group. Statistically significant values are marked in bold ($\alpha = 0.05$).

As in the British’ case, the Spanish group follows the same pattern whereby familiar (F1-F2) and unknown (U1-U2) language test pairs differ substantially. As the Wilcoxon signed-rank tests have proven in table 83, the p-value obtained from the aforementioned perception surveys’ pairs is statistically significant ($p < 0.05$).

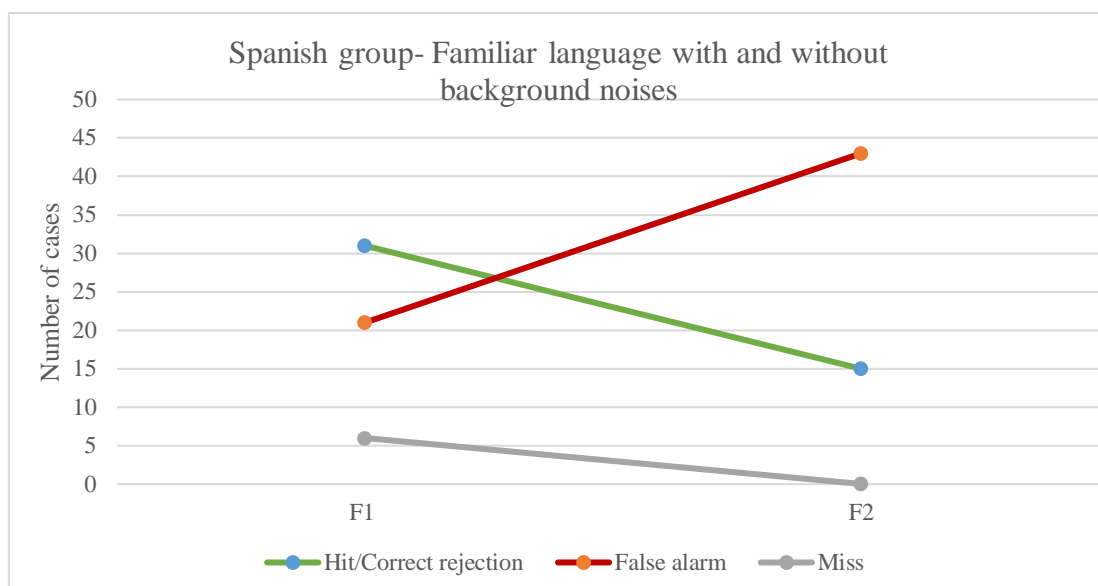


Figure 38. Multiple line graph on Spanish F1-F2 identification tests' comparison.

First of all, figure 38 shows that background noises do affect familiar language test scores negatively, as F2 demonstrates with a low chance of success and higher odds at false alarms. As for the target-present noiseless language test, not only hit rates are exponentially higher than those in F2, but they surpass the occurrences of false alarms.

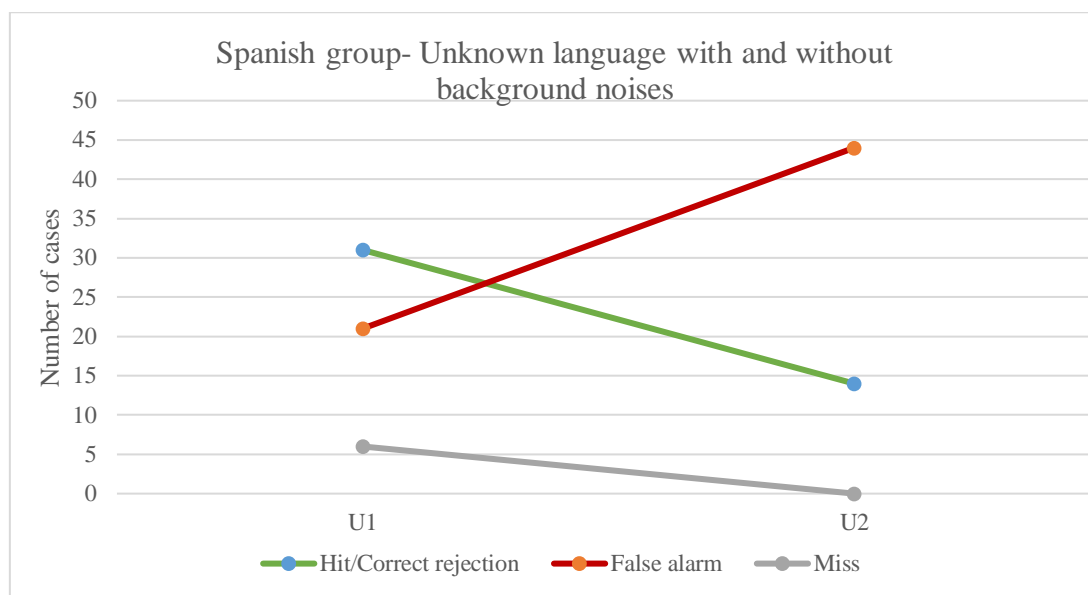


Figure 39. Multiple line graph on Spanish U1-U2 identification tests' comparison.

On a similar note, the unknown language test pair (U1-U2) bears a close resemblance to the extant relationship observed in familiar language tests. As a matter of fact, figure 38

and 39 exhibit near-identical proportions in their distribution of hits/correct rejections, missing the target, and false alarms.

It is therefore surmised in both cultural groups that noise disturbances do hinder the hearer’s ability to recognise unfamiliar voices in familiar and unknown languages.

4.6.2. Discrimination

As a complementary analysis to the first sub-section, this one tackles discrimination test scores’ variance across target-present (F1.Dis, L1.Dis, U1.Dis) and target-absent (F2.Dis, L2.Dis, U2.Dis) experimental conditions. Similarly, British and Spanish groups are investigated in this order.

4.6.2.1. British group

To start this sub-section, the British group inspects the scenario whereby jurors are asked to select the most dissimilar voice to the targeted speaker in the voice line-up, discrimination tasks. To this end, the table below gathers the required statistical measures for detecting differences amongst groups of language tests.

	Test scores		
	<u>F1.Dis-F2.Dis</u>	<u>L1.Dis-L2.Dis</u>	<u>U1.Dis-U2.Dis</u>
Z-score	-2.000	0.000	-1.000
Asymp Sig. (2-tailed)	0.046	1.000	0.317
a. Wilcoxon Signed Ranks Test			

Table 84. Pairwise comparisons on discrimination language tests’ scores across two experimental conditions in the British group. Statistically significant values are marked in bold ($\alpha = 0.05$).

In this particular setting, it seems that only discrimination tests in the familiar language (F1.Dis-F2.Dis) domain report statistical significances between the absence or presence of background noises. As the Wilcoxon signed-rank test indicates in table 84, learned language test scores (L1.Dis-L2.Dis) do not undergo variations in their values in either of

the experimental conditions. As for the unknown perception test, its p-value is too low to be accounted as statistically significant.

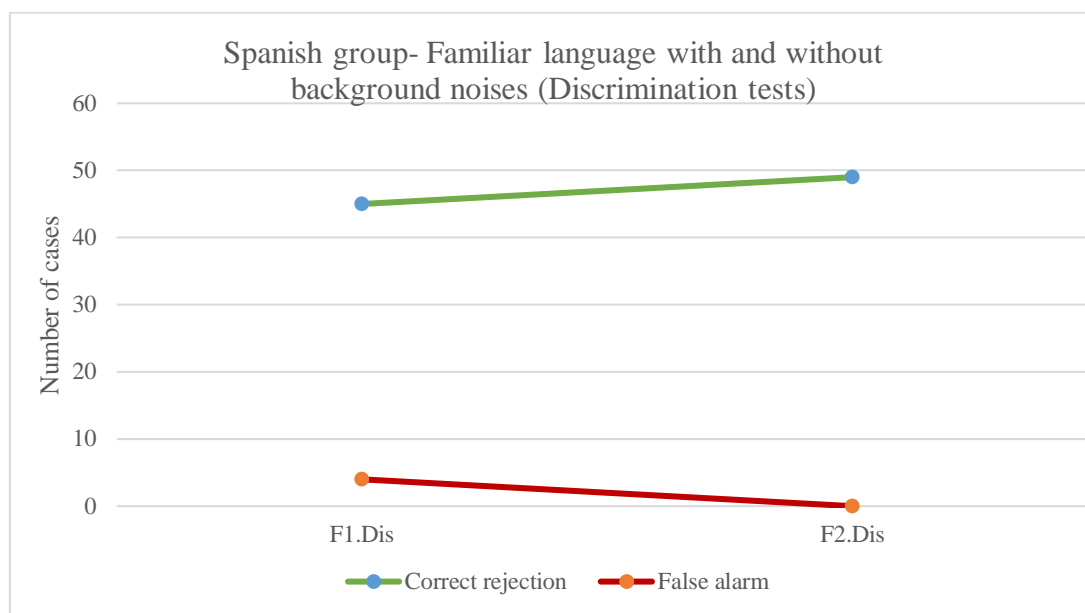


Figure 40. Multiple line graph on British F1.Dis-F2.Dis (discrimination) tests' comparison.

Upon a closer inspection on the relationship between F1.Dis and F2.Dis through the multiple line graph shown in figure 40 above, it is seen that results become more promising for the test with background noises. Nevertheless, it should be clarified that, given the perfect scores obtained in F2.Dis owing to the current research design, any deviation from that is perceived as a sharp contrast. Such is the case of F1.Dis, which only scored four false alarms out of 49 responses. In this respect, this relationship could hardly be deemed as statistically significant.

4.6.2.2. Spanish group

Once discrimination tests have been analysed in the British group, the same procedure is applied to the Spanish jurors. Expectedly, an initial perspective on the data set is offered by means of a table on Wilcoxon signed-rank test's results:

	Test scores		
	<u>F1.Dis-F2.Dis</u>	<u>L1.Dis-L2.Dis</u>	<u>U1.Dis-U2.Dis</u>
Z-score	-1.414	-1.000	-1.000
Asymp Sig. (2-tailed)	0.157	0.317	0.317
a. Wilcoxon Signed Ranks Test			

Table 85. Pairwise comparisons on discrimination language tests' scores across two experimental conditions in the Spanish group.

From table 85 above, it could be inferred that no relevant correlations were drawn due to the scarce variance of the values listed in each of the language test pairs, which renders non-significant p-values. This finding is in line with the one found in the British population, as long as the alleged F1.Dis-F2.Dis correlation from said cultural group is overlooked.

4.6.3. Summary of results

After conducting a series of Wilcoxon signed-ranks tests, statistically significant negative correlations were found for the identification familiar (F1-F2) and unknown (U1-U2) language test pairs in both groups. Results reflect decreased numbers in the second test (target-absent with background noises), hence mirroring less advantageous conditions for successful aural-perceptual speaker recognition. As for the learned language test comparison (L1-L2), it does follow the same trend with a negative correlation, albeit without reaching statistical significance.

Concerning discrimination tests, a series of Wilcoxon Signed Ranks tests spotted only one negative statistically significant correlation in for the familiar (F1-F2) language test pair for the British group ($p < 0.05$). Despite that, F1's mean test score is nearly perfect (1.92), and also shows a small standard deviation (0.277). The statistical significance emerging from F1-F2 comparison may derive from the fact that F2's discrimination tests (and the second experimental condition within discrimination tests, in general) do not contemplate the production of false alarms (since the target is absent, and therefore the juror is bound to make a right judgement regardless of the voice selected). Hence the higher influence that a few mistakes (4/49 in this case) can make to the mean test score.

As a result, discrimination tests undergo negative scores in the first experimental condition (target-present without background noises), whilst always reaching the 100% of success rates in the second experimental condition (2.0).

4.7. EPILOGUE: LEVEL OF STUDIES

The upcoming epilogue is conceived as a follow-up study of hypothesis 4 (*age and gender*) which adds jurors' level of studies to the equation. Consequently, the formulated null hypothesis (H_0) and the original alternative hypothesis (H_4) are slightly modified as follows:

- H_0 : The efficiency at speaker recognition is not conditioned by age, gender and level of studies.
- $H_{4.1}$: The efficiency at speaker recognition is conditioned by age, gender and level of studies.

Since it is not a newly postulated hypothesis, but a more specified version of the original, it has been decided to label it as hypothesis 4.1. The studied variables are exactly the same as the ones exposed previously, which includes overall scores, and individual language test scores, both with identification/discrimination tasks and target-present/target-absent conditions. As for the relevant predictors attempting to predict the dependent variable's values, the introduced variable *studies* is coded as 1 for education up until the undergraduate level (up to BA), and with a 2 being ascribed to postgraduate education (MA/PhD). Due to its irrelevance in the previous study (4.4. *Age and gender*), the interaction term *age*gender* is not considered for the linear fixed effects models conducted in this section. As will be explained in the following sections, this epilogue is written separately from its original version due to the target population being too stratified, which could compromise or call into question, at the very least, the validity of the findings that ensue. Further explanations are offered on the basis of the typical distribution of variables (*age, gender, and studies*) within the target participants and, as such idiosyncrasies may emerge, profiles are bound to differ as well.

As for the order of the elements being analysed, it also follows the same structure as its original counterpart, with a first exploration on identification tests being followed by discrimination tasks. Cultural groups are, too, presented in the same order (British, Spanish, and British and Spanish). Concerning the types of language tests and experimental conditions, overall scores are examined first, and the sub-sections proceed to individual aural-perception tests (target-present and target-absent) thereafter.

4.7.1. Identification

In the upcoming section, identification test scores (both overall scores and individual language tests) are arranged through different groups (British, Spanish, and British and Spanish jurors) to investigate the power of the studied predictors: *age*, *gender*, *studies*, and their interaction terms (*age*studies*, and *gender*studies*).

4.7.1.1. British group

Before exploring the correlations between predictors and dependent variables, a first look at the population’s distribution of said features is offered hereby:

		Gender				Total	
		Male		Female			
		Age	18-22	+22	18-22	+22	
Studies	BA	7	0	23	10	40	49
	MA/PhD	0	3	1	5	9	
Total		7	3	24	15		
		10		39			
		49					

Table 86. Distribution of age, gender, and studies across British jurors.

The stratification displayed in table 86 appears to show an unbalanced sample with some underrepresented sub-categories or, at times, a few which are not represented altogether (male aged 18-22 with postgraduate studies, and male over 22 with undergraduate studies). This observation matches the typicality of the target population, as the instances of students within the two specific cases mentioned above are significantly scarce.

Nevertheless, caution is advised when considering the relationships between two sub-types which appear proportionately different.

The main effects *gender* and *age* were considered along with *studies* and their interaction terms (*age*studies*, and *gender*studies*) to predict *overall.scores* and individual test scores, thus discarding *gender*age* due to the negligible effect it had on 4.4. (*Age and gender*). However, such interaction terms could not reach the established level of significance ($\alpha = 0.05$), and thus were removed from the model. Consequently, only main effects have been considered. One notable exception, however, can be found in the interaction terms enhancing F2's model, with *age* being classified as a significant predictor ($p= 0.020$), apart from *studies* ($p= 0.076$) and *age*studies*' ($p= 0.089$) p-values being close to statistical significance. This appears to suggest a close link between *age* and *studies*.

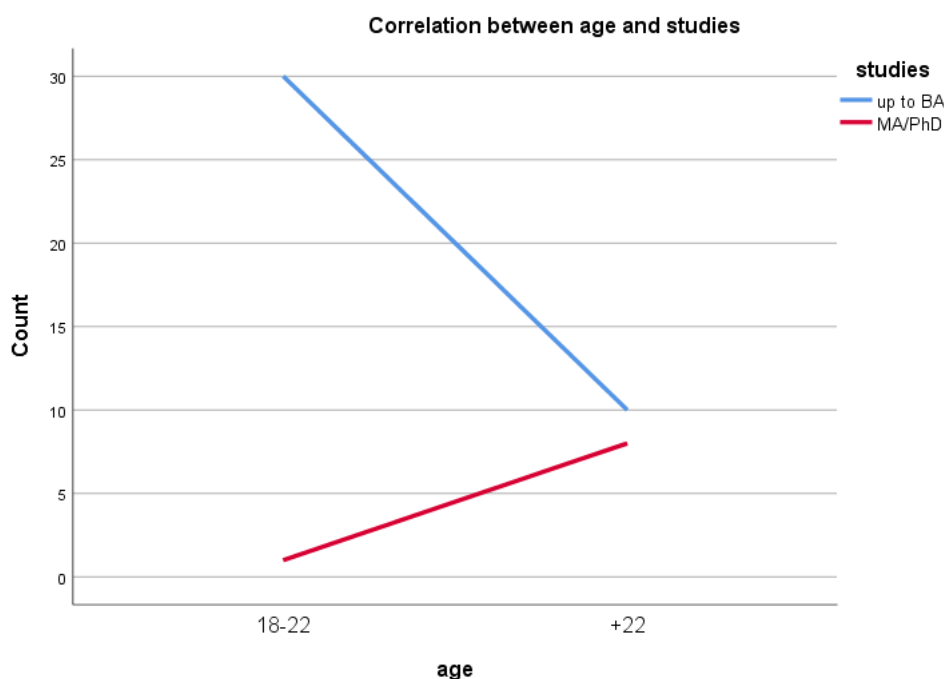


Figure 41. Multiple line graph on the existing correlation between age and studies in the F2 (British group, identification tests).

To test said claim, a Kendall's tau test has been conducted to determine the strength of association (and its orientation) existing between the ordinal variables *age* and *studies*. The results reflect a strong association significant at the 0.01 level ($\tau_b = 0.513$, $p < 0.01$), which is illustrated in figure 41. As seen in the plot above, and as would be expected from

the target population, older (+22) university students are more likely to be gathered around postgraduate studies, whilst younger students (18-22) shall typically be undertaking undergraduate degrees or other forms of vocational training. Nevertheless, their interaction terms do not amount to statistical significance, and so it is decided to remove them from each and every language test. As a result, F2 test scores cannot be effectively predicted by *age*, nor by other main effects factors.

Without interaction terms, L2 results can be predicted by *studies* with a near significant p-value ($p= 0.053$), as shown in table 87 below:

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.741190	.154088	45	11.300	.000	1.430840	2.051539
[age=1]	-.138940	.152409	45	-.912	.367	-.445908	.168028
[age=2]	0 ^b	0
[gender=1]	-.177256	.158312	45	-1.120	.269	-.496113	.141600
[gender=2]	0 ^b	0
[studies=1]	-.380965	.191502	45	-1.989	.053	-.766669	.004739
[studies=2]	0 ^b	0

a. Dependent Variable: L2.

b. This parameter is set to zero because it is redundant.

Table 87. Estimates of age, gender, and studies in L2's scores (British group, identification tests).

Going back to fixed effects estimates, *studies1* (up to BA) seems to be the least efficient stratum in identifying an input from a learned language with a negative estimate (-3.8). The almost significant p-value (0.053) asserts this claim, and its CI bounds appear a bit too far apart (-0.7 to 0.004).

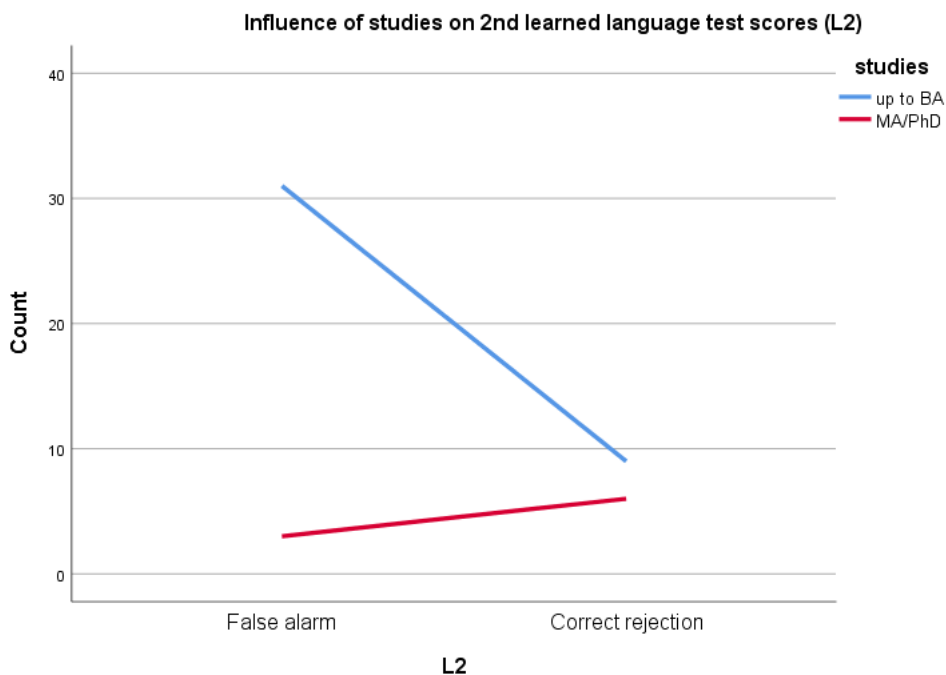


Figure 42. Multiple line graph of studies’ influence upon L2’s scores in the British group (identification tests).

In a similar vein to the first linear fixed effects model (including *gender* and *age* only), L2’s scores in the British group tend to remain stable as the juror’s experience increases with advanced studies (MA/PhD), whereas false alarms are prone to increase dramatically with less academic formation (up to BA), as it is readily observable in figure 42. Despite F2’s case shown above, the results in L2 do not seem linked with the association of *age* and *studies* explained previously. In this sense, this model’s *studies* variable could not be interpreted interchangeably with the variable of *age*. However, it is noticeable that *age*’s influence on L2 test scores in the first study (4.4. *age and gender*) resembles closely the relationship obtained in this epilogue with *studies*.

4.7.1.2. Spanish group

Once the British group’s analysis is cleared, the Spanish participants shall follow. Their target population surveyed is stratified in the following manner:

		Gender				Total	
		Male		Female			
Age		18-22	+22	18-22	+22		
Studies	BA	7	4	31	5	47	58
	MA/PhD	0	5	1	5	11	
Total		7	9	32	10		
		16		42			
		58					

Table 88. Distribution of age, gender, and studies across Spanish jurors.

Just as in the British case, some strata appear either underrepresented or not missing from the chart (male aged 18-22 with postgraduate studies) appearing in table 88. With the exception of females aged 18-22 with undergraduate studies, the other cells values do not seem so distant amongst themselves. Even with the typicality of the target population’s profiles that this sample reflects, interpretation of results is exercised with caution.

In the Spanish group, *overall.scores* did not find significant predictors amongst the variables entered (*age*, *gender*, and *studies*) and their interaction terms (*gender*studies*, *age*studies*). *Age*gender* has been discarded due to its irrelevance on 4.4. (*Age and gender*). When entering the interaction terms, neither *gender*studies* nor *age*studies* were deemed as significant predictors for identification tests (F1, F2, L1, L2, U1, U2). For this reason, the resulting model omits such variables and considers the main effects only.

In the first learned language test (L1), a significant predictor was found: gender.

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.225354	.151749	54	8.075	.000	.921115	1.529593
[age=1]	-.111888	.162086	54	-.690	.493	-.436851	.213075
[age=2]	0 ^b	0
[gender=1]	.326599	.138861	54	2.352	.022	.048198	.604999
[gender=2]	0 ^b	0
[studies=1]	.086543	.188108	54	.460	.647	-.290591	.463677
[studies=2]	0 ^b	0

a. Dependent Variable: L1.

b. This parameter is set to zero because it is redundant.

Table 89. Estimates of age, gender, and studies in L1’s scores (Spanish group, identification tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

As could be inferred by looking at the positive estimates from *gender1* (male) in table 89, higher chances of success are ascribed to males with a significant p-value ($p= 0.022$). The CI bounds, too, are not spread out excessively (0.048 to 0.60).

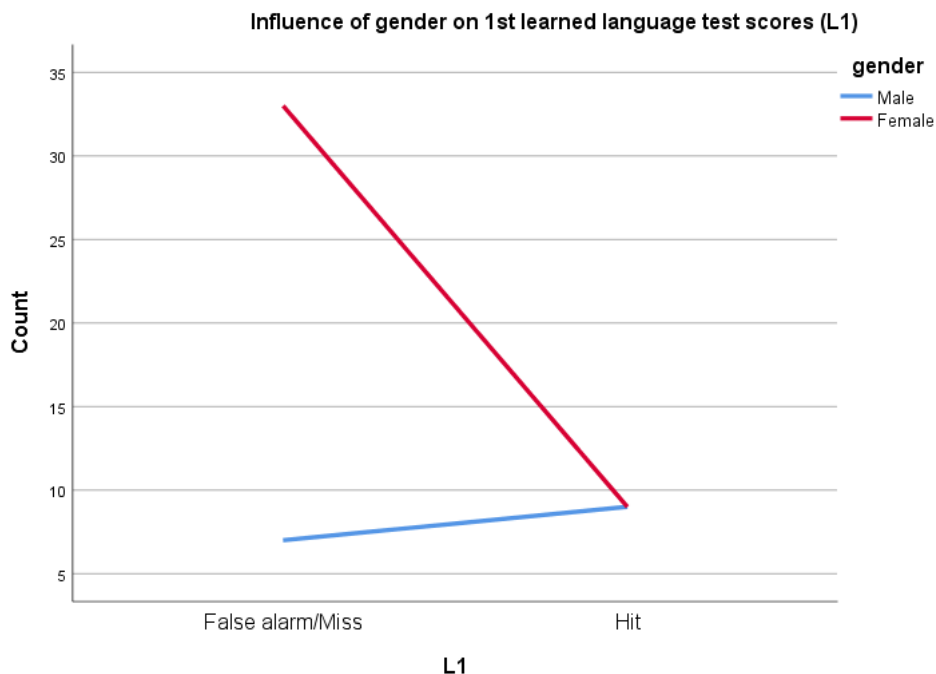


Figure 43. Multiple line graph of gender’s influence upon L1’s scores in the Spanish group (identification tests).

Just as the estimates on the fixed effects ascribed for gender predicted, males appear to fare better in speaker identification tests in the target-present condition for the learned language (L1) in the Spanish group, as figure 43 shows. This scenario does not seem to have varied through the previous study (4.4. *age and gender*) and the current epilogue, since the resulting correlation observed here is fairly similar the one highlighted previously.

In the second familiar language (F2), the same predictor (gender) is found close to significance levels, as the table below demonstrates:

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.142419	.147799	54	7.730	.000	.846099	1.438739
[age=1]	.080680	.157867	54	.511	.611	-.235825	.397185
[age=2]	0 ^b	0
[gender=1]	.270542	.135247	54	2.000	.051	-.000612	.541696
[gender=2]	0 ^b	0
[studies=1]	-.015649	.183212	54	-.085	.932	-.382967	.351669
[studies=2]	0 ^b	0

a. Dependent Variable: F2.

b. This parameter is set to zero because it is redundant.

Table 90. Estimates of age, gender, and studies in F2's scores (Spanish group, identification tests).

From the positive estimate ascribed to *gender1* (0.27) in table 90, F2 is seemingly bearing the same correlation as the one witnessed in L1, although this time gender's p-value is nearly significant (p= 0.051). As a consequence, the values displayed in the CI interval (-0.0006 to 0.54) appear contrastively more far apart than in L1's case.

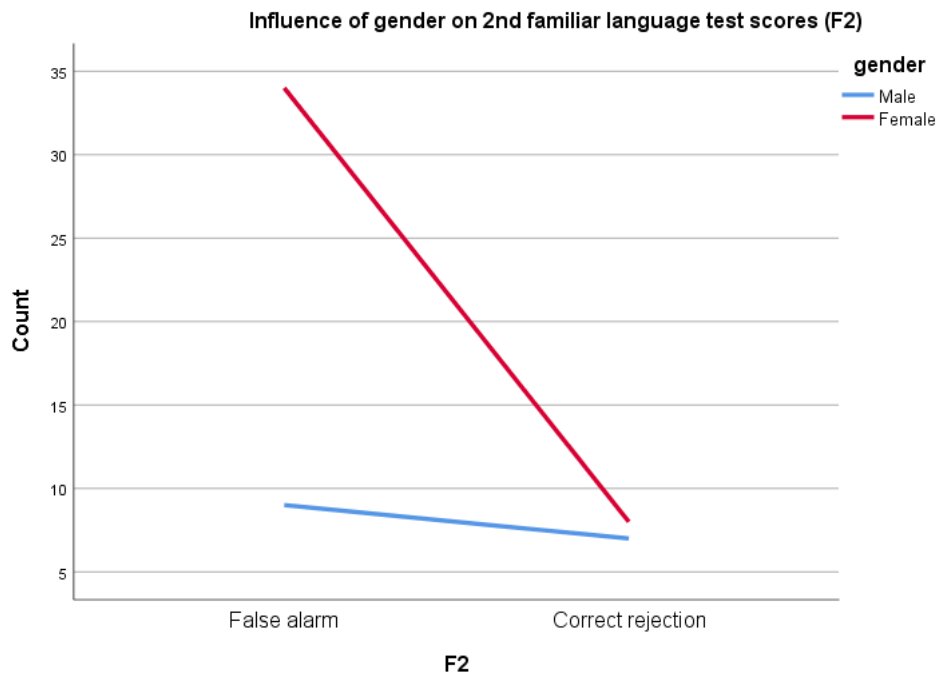


Figure 44. Multiple line graph of gender’s influence upon F2’s scores in the Spanish group (identification tests).

Again, the multiple line-graph in figure 44 is nearly identical to the one drawn in the previous study which did not consider *studies* within the model. This version, however, seems to put correct rejections made by both males and females almost on equal footing, despite both being represented with a descending line (less correct rejections than false alarms).

4.7.1.3. British and Spanish group

Following the traditional approach employed for analytical purposes, this last section gathers the test scores produced by both group of jurors. The table below displays the counts on the sub-categories found in the target population:

	Age	Gender				Total	
		Male		Female			
		18-22	+22	18-22	+22		
Studies	BA	14	4	54	15	87	107
	MA/PhD	0	8	2	10	20	
Total		14	12	56	25		
		26		81			
		107					

Table 91. Distribution of age, gender, and studies across Spanish jurors.

As discussed in the two previous sub-sections, the balance and representativeness of specific sub-categories are far from the ideal in experiments of this kind. However, and despite the roughly 80%/20% distribution in some of them, the number of cases registered in table 91 has increased and thus results appear less prone to error.

Moving to the statistical analysis itself, the studied interaction terms (*studies*age* and *studies*gender*) were not significant to predict *overall.scores*, nor each individual language test scores. For this reason, they were removed from the conclusive model and therefore only main effects were considered.

Specifically, the second language for the familiar language (F2) found gender as an effective predictor:

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.157322	.103629	103	11.168	.000	.951799	1.362846
[age=1]	.135616	.104476	103	1.298	.197	-.071587	.342819
[age=2]	0 ^b	0
[gender=1]	.197790	.097609	103	2.026	.045	.004206	.391374
[gender=2]	0 ^b	0
[studies=1]	-.062864	.128239	103	-.490	.625	-.317195	.191467
[studies=2]	0 ^b	0

a. Dependent Variable: F2.

b. This parameter is set to zero because it is redundant.

Table 92. Estimates of age, gender, and studies in F2's scores (British and Spanish group, identification tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

Similar to the Spanish group-only group analysis, *gender1*'s positive estimate (0.19) shown in table 92 reflects better results related to male participants with a significant p-value (0.045) and a low standard error (0.09). In addition, the CI bounds (0.004-0.3) are narrower than in the Spanish case commented above, which increases the certainty of the claim in this particular scenario.

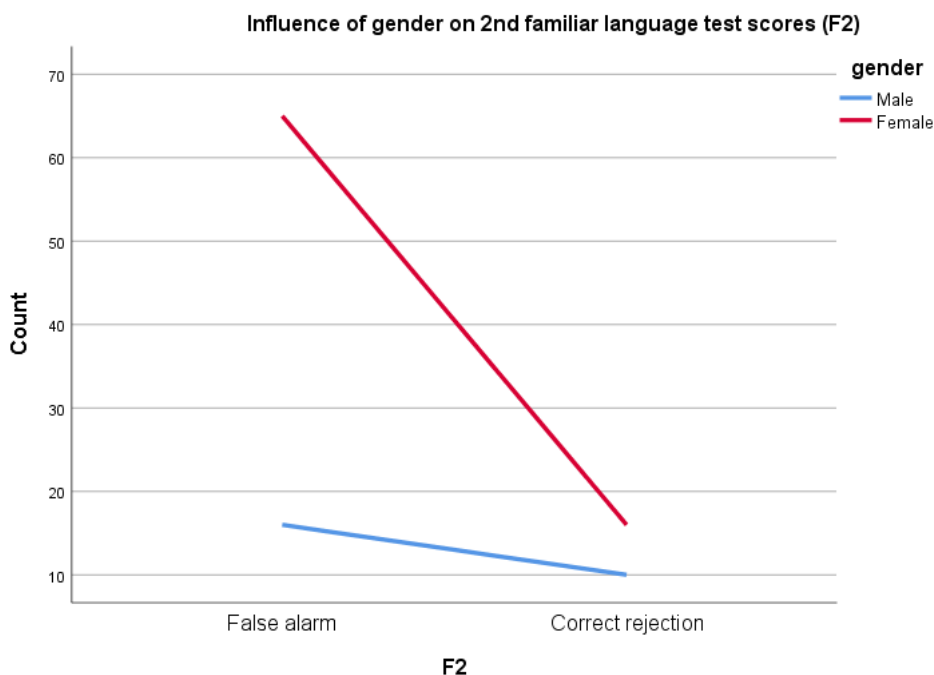


Figure 45. Multiple line graph of gender's influence upon F2's scores in the British and Spanish group (identification tests).

As noticed in figure 45, better results do not necessarily entail a higher production of correct rejections, but rather refer to the proportions found between false alarms and successful responses in this case. In fact, males do score less correct rejections than females, although the ratio between the latter's correct rejections and false alarms is larger than the former's.

On the other hand, *studies* seems an efficient predictor for L2 test scores, as table 93 shows below:

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.562378	.100597	103	15.531	.000	1.362867	1.761888
[age=1]	-.156075	.101419	103	-1.539	.127	-.357216	.045067
[age=2]	0 ^b	0
[gender=1]	.008075	.094753	103	.085	.932	-.179846	.195995
[gender=2]	0 ^b	0
[studies=1]	-.258151	.124487	103	-2.074	.041	-.505041	-.011260
[studies=2]	0 ^b	0

a. Dependent Variable: L2.

b. This parameter is set to zero because it is redundant.

Table 93. Estimates of age, gender, and studies in L2's scores (British and Spanish group, identification tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

In this case, the predictor's (*studies*) influence on L2 test scores resembles the one found in the British jurors' analysis. Interestingly, this grouped account reports a significant p-value ($p = 0.041$), whereas the case previously mentioned yielded a p-value slightly above the significance level ($p = 0.053$). This is all the more intriguing given the fact that the Spanish jurors did not find any predictor for L2 test scores, and thus a less significant p-value should be expected in this account, given the addition of a group (Spanish) whose values are non-significant in this particular domain.

As for the values observed in table 93, the negative estimate assigned for *studies1* matches the findings described in the British group, which implies lesser scores overall for those students at undergraduate level. The CI bounds (-0.5 to -0.01) reflect, in turn, a closer range of numbers, and a higher confidence in the correlation found.

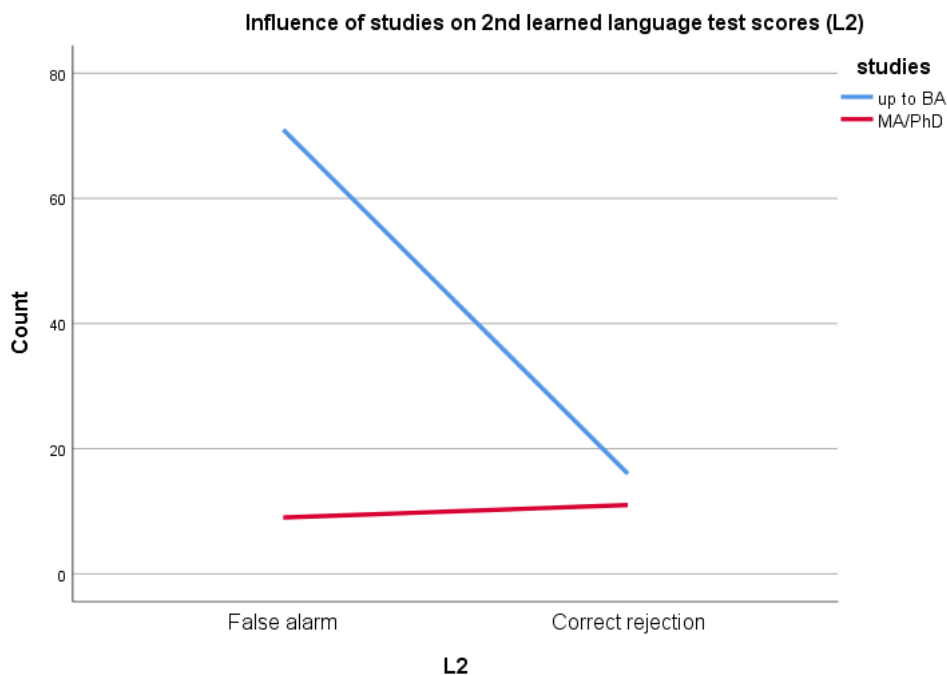


Figure 46. Multiple line graph of studies’ influence upon L2’s scores in the British and Spanish group (identification tests).

Lastly, the predictor’s influence on British and Spanish learned language test scores (L2) is illustrated in figure 46. Just as in the British-only case, *studies* relationship with L2 remains practically the same, even though this time it does reach statistical significance. In such a situation, it would appear that participants with postgraduate studies are not as prone to false alarms as those at undergraduate level.

4.7.2. Discrimination

Once the main effects’ (*age*, *gender*, and *studies*) influence on identification test scores has been studied as well as their interaction terms (*age*studies*, *gender*studies*), this epilogue proceeds to scrutinise discrimination tasks by conducting a series of linear fixed effects models for British, Spanish, and British and Spanish groups.

4.7.2.1. British group

On the discrimination side, the little variation that occurs in said aural-perception tests could not be explained by *age*, *gender*, *studies*, nor by their interaction terms

(*age*studies*, and *gender*studies*) in the British group. What is more, in L1's case, the linear fixed effects model cannot even be computed due to its lack of variance (all jurors scored 2 points, therefore signalling a correct rejection).

4.7.2.2. Spanish group

Similarly, discrimination tests yielded no statistically significant results when inputting interaction terms (*age*studies*, and *gender*studies*) of the main variables involved (*age*, *gender*, and *studies*). Once considering the three predictors separately, only the first familiar language discrimination test (F1.Dis) contained a predictor (*gender*) accounting for the variance of the data.

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	2.055290	.059778	54	34.382	.000	1.935443	2.175137
[age=1]	.038288	.063849	54	.600	.551	-.089722	.166299
[age=2]	0 ^b	0
[gender=1]	-.129296	.054701	54	-2.364	.022	-.238964	-.019627
[gender=2]	0 ^b	0
[studies=1]	-.098539	.074100	54	-1.330	.189	-.247101	.050023
[studies=2]	0 ^b	0

a. Dependent Variable: F1.Dis.

b. This parameter is set to zero because it is redundant.

Table 94. Estimates of age, gender, and studies in F1.Dis' scores (Spanish group, discrimination tests). Statistically significant values are marked in bold ($\alpha = 0.05$).

As noticed in the table of estimates of fixed effects (table 94), *gender1* (male) bears a negative figure (-1.29), which entails worse results for this stratum. The resulting p-value is significant at the 0.05 level ($p= 0.022$), and the 95% CI remains on negative values as well (-0.23 to -0.019).

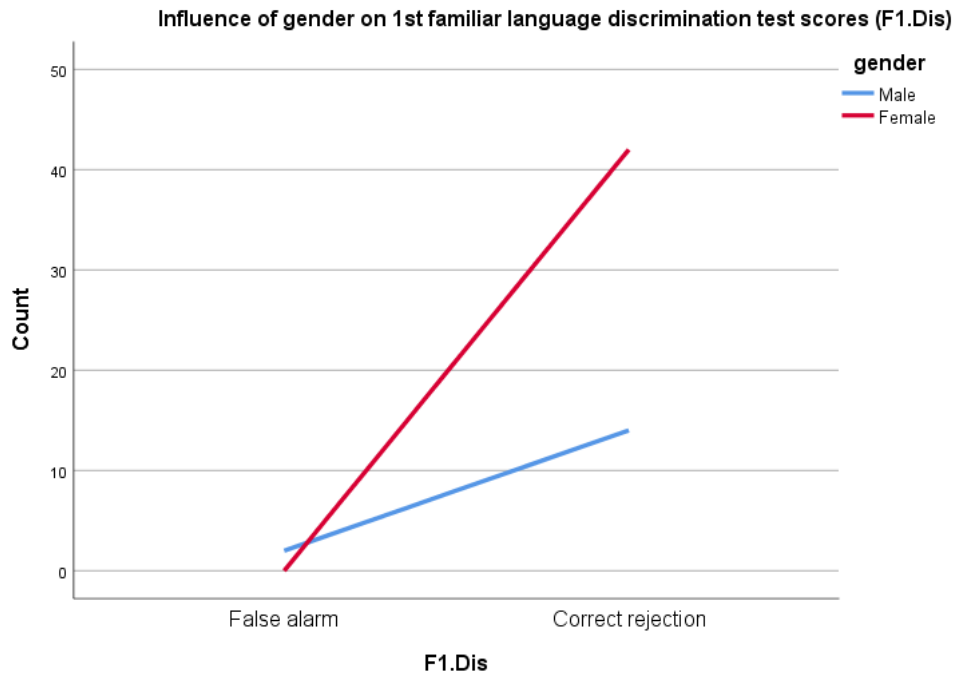


Figure 47. Multiple line graph of gender’s influence on F1.Dis’ scores in the Spanish group (discrimination tests).

To illustrate this correlation, a multiple line graph is plotted in figure 47. It is confirmed that, indeed females’ responses appear more efficient than males’, with a higher number of correct rejections and a count of false alarms on par with males. However, it must be reminded that F1.Dis’ scores do not undergo significant variation ($M=1.97$), and thus results in this section should be interpreted with caution.

4.7.2.3. British and Spanish group

After entering the interaction terms *studies*gender* and *studies*age* along with their individual main effects (*age*, *gender*, and *studies*), no significant predictors were found for the target-present aural-perception tests for British and Spanish groups. Notice that the target-absent condition and the interaction term *age*gender* were removed for their redundancy, as no significance can be extracted from either factors. Furthermore, main effects alone could not predict discrimination tests effectively (F1.Dis, L1.Dis, U1.Dis).

4.7.3. Summary of results

To get a more comprehensive view on how the findings obtained in the epilogue compare to the ones retrieved from the original study (including *age* and *gender* only), the following table (table 95) covers the significant predictors found in both cases, as far as identification tests are concerned:

4.4. Age and gender		4.7. Age, gender, and studies	
Relationship	p-value	Relationship	p-value
British L2 with age	0.029	British L2 with studies	0.053
Spanish L1 with gender	0.022	Spanish L1 with gender	0.022
Spanish F2 with gender	0.048	Spanish F2 with gender	0.051
British and Spanish F2 with gender	0.037	British and Spanish F2 with gender	0.045
British and Spanish L2 with age	0.002	British and Spanish L2 with studies	0.041

Table 95. Findings in the original study and its extended version (epilogue).

With the exception of the identical relationship found between *gender* and Spanish L1 test scores, the rest of aural-perception tests report higher p-values in the epilogue, which is indicative of less certain statements. Not only this, but the predictor *studies* appears to replace *age*'s influence to British and British and Spanish test scores when it is added to the model, which could refer to the close association between these two variables: undergraduate students are likely to be younger than their peers studying at postgraduate level. In this sense, *age* and *studies* could be used interchangeably to account for the variance of L2 scores, since the orientation of the data is practically the same (younger participants with less education levels are more prone to false alarms), although *studies* is less effective as a predictor (higher p-values).

As for the other language tests and predictors, the same relationships are spotted in both studies: Spanish L1 and F2, and British and Spanish F2 correlated with *gender*, where males display less tendencies towards false alarms than females. One particularity of these tests which applies across both the first study and the epilogue emerges with the Spanish L1's case, where males surpass their hit rates over the count of false alarms, in contrast with Spanish, and British and Spanish F2. The conclusions drawn from this epilogue are, therefore, much in line with the ones exposed in 4.4. (*Age and gender*), only that this time

the most reliable predictor within *age* (over 22) is replaced by its equivalent in *studies* (MA/PhD).

As for discrimination tests, the retrieved significance found on F1.Dis in the Spanish group should be limited, given the fact that its scores are not varied enough. In fact, only 2 out of 58 responses accounted for false alarms, which were made by males. This, together with the higher number of female participants (42) in contrast with male jurors (16), renders questionable results, whose interpretation should be exercised with care. For the purposes of this research, it is concluded that discrimination perception tests do not allow for significant correlations to be encountered, given the low variation in their values.

CHAPTER 5

RESULTS:

ACOUSTIC-PHONETIC ANALYSIS

Once the jurors' impressions have been registered and conclusions have been drawn (see *chapter 7. Conclusions* for more details), the fifth chapter turns its focus towards inspecting the stimuli employed in said perception surveys, namely the informants' voice samples. Despite being separated in different chapters, this analytical section on acoustic-phonetics also includes statistical analyses for the sake of conducting a proper forensic voice comparison¹⁴. However, this section is centered on the acoustic properties of the voice, rather than on the perception thereof. In this line of thought, the present thesis proceeds to address the formulated hypotheses in this domain. Firstly, hypothesis 7 is tested (*5.1. Intravariability of suspects*), subsequently followed by hypothesis 8 dealing with distractors' voice comparison (*5.2. Intervariability of foil speakers*), and a

¹⁴ Forensic voice comparison could also tap into biometry, which alludes to physiological and/or behavioural features (highly speaker-dependent) involved in speech production (Farrús 2011: 42). Methodologies including biometric features are typically related to automatic speaker recognition software (Jiménez et al. 2014:37), and thus greater overall results are expected (leading to an increase in its use). Despite seemingly unrelated, this thesis also considers some of the variables which yield more promising results in said discipline, especially at the segmental level.

comparative analysis on the results gathered from chapter 4 and 5 (5.3. *Acoustic-phonetic analysis or jurors' verdict?*) provides the final remarks on both analytical sections. It is reminded that the voice samples analysed throughout the three sub-sections are identical to the ones employed in the aural-perception tests, with all the drawbacks associated with it (such as the audio material's short duration or the intended researcher's adjustment of some samples to equate them with the rest of voices in the line-up). This scenario is indeed far from the ideal conditions assumed for an efficient acoustic-phonetic analysis. Nevertheless, this methodological change is needed in order both to enable a fair comparison with the jurors' responses (chapter 4's findings) and to obtain a closer resemblance to the possible materials gathered in real-life contexts.

Two distinct analytical measures have been selected according to the types of data processed in suprasegmental and segmental features. The former refers to individual points of data (i.e. with no expected variance within their values), whereas the latter gathers several instantiations of each segmental variable (e.g. the sound [s] being represented through lexical items such as *stop*, *son*, and *safe*). In this regard, a dissimilarity matrix fulfils the role of calculating Euclidean distances amongst each speaker's suprasegmental variable, whose values are transformed into standardised Z-scores for the sake of comparison (Barrett 2005: 12). The suprasegmental variables of choice match those found useful in previous studies (Rose 2002: 150), especially the ones concerned with speech tempo (Künzel 1997, Lindh 2009).

When it comes to analyse segmental features, an ANOVA is computed (with the required assumption testing procedure) based on the premise that a set of sounds uttered by different speakers is bound to yield some variation, both in within-subjects and between-subjects' experiments. Even if it is not always the case, it is normally assumed that intra-speaker variation exhibits less variation than inter-speaker variation (Fernández Planas 1998: 157). Contrary to suprasegmental variables, the variance observed in segmental variables is accounted for through measures like mean values and standard deviations. Again, the segmental units of analysis covered in the following sections were chosen according to the results exposed by previous research, ranging from the variables within voiced/voiceless plosives (Clegg & Fails 2017, Whiteside et al. 2004) to the units related to voiced/voiceless alveolar sibilants (Gordon et al. 2002, Koenig et al. 2013, and Univaso

et al. 2014). In this respect, it is sought to decipher the independent variables' (suprasegmental and segmental features) influence upon the dependent (speakers) types.

5.1. INTRAVARIABILITY OF SUSPECTS

Hypothesis 7 (H₇) poses that the existing within-speaker variation related to suspects' voice samples (semi-spontaneous data without control of the segmental units produced) with differing intonation contours (rising and falling intonation patterns) is not statistically significant, and thus a successful identification is plausible despite such differences. Hence, the current hypothesis is formally formulated as follows:

- H₀: Intravariability of the suspects' voice samples with differing intonation contour (rising and falling intonation) and uncontrolled segmental phenomena is statistically significant.
- H₇: Intravariability of the suspects' voice samples with differing intonation contour (rising and falling intonation) and uncontrolled segmental phenomena is not statistically significant.

The upcoming table spells out the exact variables studied in hypothesis 7:

Independent variable		Dependent variable		
		Speakers		Code
Suprasegmental		English informants	Simon T. Elliott	1
Pitch	Mean pitch ($P\bar{x}$)			
	25% Pitch			
	50% Pitch			
	75% Pitch			
	Min. intensity ($I\downarrow$)			
	Max. intensity ($I\uparrow$)			
Mean intensity ($I\bar{x}$)	SUSPECT (Simon T. Elliott)	2		
Pauses	DurPaus	Spanish informants	M12_020	1
	N_paus/min			
	Pause_%			
	N_paus			
	Speech rate			
	Articulation rate			
	ASD (Average Syllable Duration)	SUSPECT (M12_020)	2	
Segmental		Dutch informants	DVA8-F20L	1
[b, d, g]	VOT (Voice Onset Time)			
and [k, p, t]	Release burst intensity			
[s] and [z]	Spectral peak location			
	COG (Center of Gravity)			
	Noise duration			
	Noise amplitude			
	F1	SUSPECT (DVA8-F20L)	2	
F2				
F3				

Table 96. Selected variables for hypothesis 7 testing, displaying the sub-types of both independent (left column) and dependent (right column) variables¹⁵.

As observed in table 96, both suprasegmental and segmental features are covered in the following sub-sections, with individualised observations for each corpus (English,

¹⁵ Independent variables are described in their respective sub-sections depending on their sub-type: suprasegmental (5.1.1.) and segmental (5.1.2.) features.

Spanish, and Dutch voice samples). In this fashion, such independent variables are broken down further into measurements related to pitch and pauses (suprasegmental), and segmental units such as voiced [b, d, g] and voiceless plosives [k, p, t], and the voiceless alveolar sibilant [s]¹⁶. As for the dependent variables, the analysis considers only the suspects' voice samples appearing in the voice line-up (with instantiations of uptalk or rising intonation), and the recordings (with falling intonation) which were used to introduce the suspect to identify at the beginning of each stage (English, Spanish, and Dutch perception tests). Also, the segmental variation is implied within the experimental conditions of this research, since the nature of the data (semi-spontaneous exchanges) does not allow for a controlled account of the segmental units uttered. The codes assigned to speakers are treated as categorical numerical variables, which are needed in some of the statistical tests run in the upcoming sections, like in ANOVAs and Welch's tests.

As hinted in the table above, the order of elements in this analysis goes as follows: suprasegmental features (5.1.1.) come first alongside each group of informants studied (5.1.1.1. *English voice samples*, 5.1.1.2. *Spanish voice samples*, and 5.1.1.3. *Dutch voice samples*), followed by segmental features (5.1.2.) with its respective sub-sections (5.1.2.1. *English voice samples*, 5.1.2.2. *Spanish voice samples*, and 5.1.2.3. *Dutch voice samples*).

5.1.1. Suprasegmental features

In the suprasegmental domain, features concerned with speakers' pitch (mean pitch, 25% quantile, 50% quantile, 75% quantile, minimum/maximum intensity, and mean intensity) and pauses (pauses duration per minute, number of pauses per minute, percentage of pauses per sample, number of pauses in the sample, speech rate, articulation rate, and ASD) are measured in this within-speaker variation section (see 3.7.3.1. *Suprasegmental features* for more details). Since the values extracted offer only one point per case (thus removing measures of data dispersion and variation like mean values and standard deviation) through selecting the whole excerpt (4-14 sec. of duration) as a whole, the option of calculating Euclidean distances with standardised Z-scores seems the best fit for this analytical section. This method enables a comparison of distances between the

¹⁶ Please note here that the voiced alveolar sibilant [z] is only added to the analysis concerned with English informants due to the shortage of its voiceless counterpart [s] in some of the subjects examined.

points of data gathered through suprasegmental variables across the selected speakers. Pitch-related variables were extracted through a Praat script (Lennes 2013). The same is true for pausing measures (De Jong & Wempe 2008). However, the variables concerned with intensity are extracted manually instead, since unwanted background noises can be removed in this manner, if necessary.

The analysis covers each group of informants' recordings in the following order: English (5.1.1.1.), Spanish (5.1.1.2.), and Dutch (5.1.1.3.) voice samples.

5.1.1.1. English voice samples

To start with, English voice samples target Simon T. Elliott and his recording as a suspect. For the sake of a fair comparison between such audio material, the analysis proceeds to calculate Euclidean distances amongst the variables of interest with standardised Z-scores, since the latter can be converted easily to p-values (a significant z-score at the 0.05 significance level is roughly 1.645). The following table summarises the z-scores and p-values obtained the suprasegmental variables related to pitch measurements:

Pitch		
Variable	Z-score	Sig. (p-value)
Mean pitch ($P\bar{x}$)	0.967	0.166
25% Pitch	0.936	0.174
50% Pitch	0.818	0.206
75% Pitch	1.068	0.142
Min. intensity ($I\downarrow$)	0.164	0.434
Max. intensity ($I\uparrow$)	0.023	0.490
Mean intensity ($I\bar{x}$)	0.066	0.473

Table 97. Within-speaker variation of English voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott) in terms of pitch-related measurements.

As inferred from table 97 above, there is no statistically significant difference between differing intonation contours in Simon T. Elliott (rising intonation) and his recording as a suspect (falling intonation), as far as pitch-related features is concerned. As for the

already commented measures on pausing, the following table illustrates their relevance in differentiating the target speakers:

Pauses		
Variable	Z-score	Sig. (p-value)
DurPaus	0.095	0.462
N_paus/min	0.686	0.246
Pause_%	0.091	0.464
N_paus	0.871	0.192
Speech rate	0.332	0.370
Articulation rate	0.557	0.289
ASD	0.708	0.240

Table 98. Within-speaker variation of English voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott) in terms of pauses.

Similarly, table 98 demonstrates that pausing features do not display statistically significant results, which seems indicative of a robust parameter for within-speaker variation, at least when inspecting the selected English voice samples. In this sense, the null hypothesis can be rejected, and thus it is concluded in this sub-section that intravariability of English suspects' voice samples with differing intonation contour (rising and falling intonation) is not statistically significant within the specifications of this particular research.

5.1.1.2. Spanish voice samples

Secondly, Spanish informants' suprasegmental features are extracted and summarised in table 99 below:

Pitch		
Variable	Z-score	Sig. (p-value)
Mean pitch ($P\bar{x}$)	1.207	0.113
25% Pitch	0.711	0.238
50% Pitch	1.003	0.157
75% Pitch	1.793	0.036
Min. intensity ($I\downarrow$)	0.101	0.459
Max. intensity ($I\uparrow$)	0.277	0.390
Mean intensity ($I\bar{x}$)	0.222	0.412

Table 99. Within-speaker variation of Spanish voice samples (M12_020- SUSPECT M12_020) in terms of pitch-related measurements. Statistically significant values are marked in bold ($\alpha = 0.05$).

As far as pitch-related measurements is concerned, it seems that only the 75% quantile on pitch (75% Pitch) is able to differentiate between M12_020's voice sample and her recording as a suspect.

Pauses		
Variable	Z-score	Sig. (p-value)
DurPaus	1.683	0.046
N_paus/min	0.015	0.494
Pause_%	1.668	0.048
N_paus	1.328	0.092
Speech rate	1.688	0.046
Articulation rate	2.358	0.009
ASD	2.397	0.008

Table 100. Within-speaker variation of Spanish voice samples (M12_020- SUSPECT M12_020) in terms of pauses. Statistically significant values are marked in bold ($\alpha = 0.05$).

As for pausing measurements, table 100 suggests that duration of pauses, percentage of pauses, speech rate, articulation rate, and ASD are distinguishable enough between the samples with rising (M12_020) and falling (SUSPECT M12_020) intonation patterns. A visual representation of such differences can be consulted in figure 48 below:

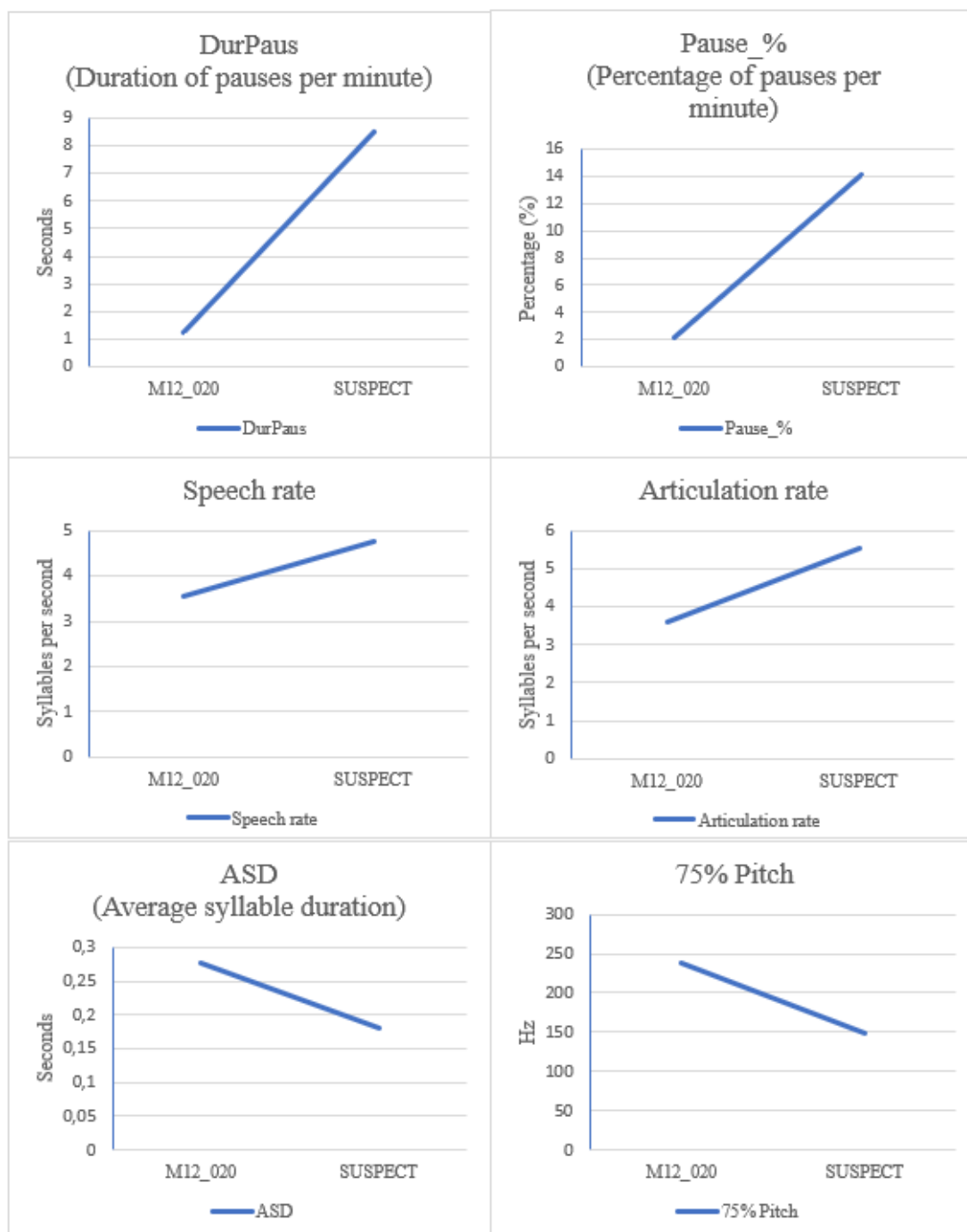


Figure 48. Statistically significant differences in suprasegmental features across the Spanish suspect’s voice samples (M12_020- SUSPECT M12_020).

According to Gros et al. (1999), ‘articulation rate increases with longer words as average syllable duration tends to decrease with more syllables in a word’ (p. 3). By looking at figure 48, a discernable pattern of this kind is perceived (higher values in speech/articulation rate in the suspect’s sample in correlation with a decrease in her ASD

values), and thus it is inferred that the recording used as the voice line-up's suspect contains longer words than the one appearing in the body of distractors (M12_020).

In this regard, this particular experiment concludes that the null hypothesis (the intravariability of suspects' voice samples with differing intonation contour is statistically significant) is retained in the case of Spanish speakers, in the suprasegmental domain at the very least.

5.1.1.3. Dutch voice samples

This third sub-section examines the significant suprasegmental variables emerging from Dutch recordings to gauge degrees of within-speaker variation. The table below:

Pitch		
Variable	Z-score	Sig. (p-value)
Mean pitch ($P\bar{x}$)	0.648	0.258
25% Pitch	0.385	0.350
50% Pitch	0.616	0.268
75% Pitch	0.705	0.240
Min. intensity ($I\downarrow$)	2.657	0.003
Max. intensity ($I\uparrow$)	2.584	0.004
Mean intensity ($I\bar{x}$)	2.395	0.008

Table 101. Within-speaker variation of Dutch voice samples (DVA8-F20L- SUSPECT DVA8-F20L) in terms of pitch-related measurements. Statistically significant values are marked in bold ($\alpha = 0.05$).

As seen in table 101, only the variables related to intensity (min./max. and mean intensity) are statistically different between the DVA8-F20L's voice sample and her recording as a suspect. Such differences can be observed in the following graphics:

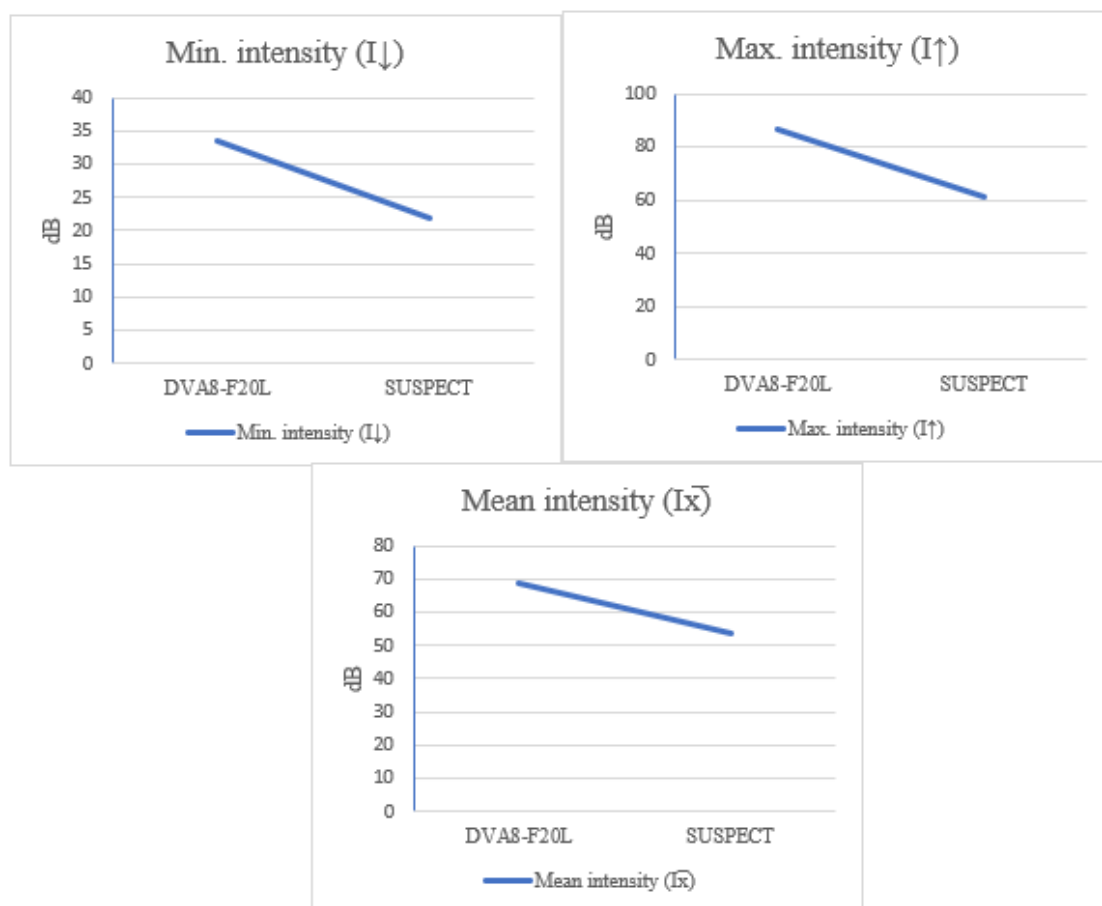


Figure 49. Statistically significant differences in pitch-related measures across the Dutch suspect’s voice samples (DVA8-F20L- SUSPECT DVA8-F20L).

It can be observed from figure 49 above that DVA8-F20L’s minimum intensity (33.37 dB), maximum intensity (86.93 dB) and mean intensity (68.72 dB) are significantly higher than her recording as a suspect, which gathers lower minimum intensity (21.82 dB) maximum intensity (61.43 dB) and mean intensity (53.39 dB). Leaving intensity aside, the following table covers pausing measurements:

Pauses		
Variable	Z-score	Sig. (p-value)
DurPaus	0.698	0.242
N_paus/min	1.141	0.127
Pause_%	0.698	0.242
N_paus	2.601	0.005
Speech rate	0.686	0.246
Articulation rate	0.321	0.374
ASD	0.338	0.368

Table 102. Within-speaker variation of Dutch voice samples (DVA8-F20L- SUSPECT DVA8-F20L) in terms of pauses. Statistically significant values are marked in bold ($\alpha = 0.05$).

As seen in table 102, the number of pauses per excerpt (N_paus) are statistically different between the speakers DVA8-F20L and SUSPECT DVA8-F20L. Such discrepancies are illustrated in the graph below:

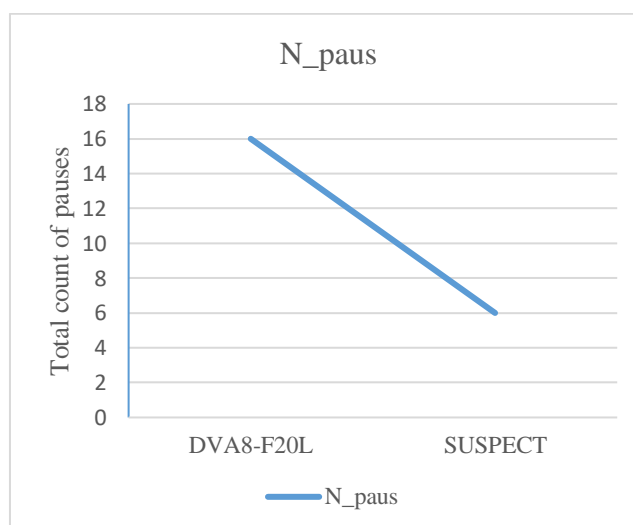


Figure 50. Statistically significant differences in the number of pauses across the Dutch suspect's voice samples (DVA8-F20L- SUSPECT DVA8-F20L).

As figure 50 illustrates, the recording appearing amongst the body of distractors (DVA8-F20L) registers a higher number of pauses per excerpt than her recording as a suspect. Nevertheless, such dissimilarity does not seem to affect other pausing measures, since the rest of variables do not report statistically significant differences, as seen in table 102 above.

Just as in the Spanish' case, the conclusion for Dutch voice samples' variation in relation to suprasegmental features seems to lean towards the null hypothesis (the intravariability of suspects' voice samples with differing intonation contour is statistically significant). Likewise, measures stemming from both pitch-related measurements and pausing measures appear to reinforce said hypothesis within the boundaries of this research.

5.1.2. Segmental features

As explained in 5.1. (*Intravariability of suspects*), the analysis centered on segmental features include units of measurement concerned with voiced ([b, d, g]) and voiceless plosives ([k, p, t]), and the voiceless alveolar sibilant [s] (and the voiced alveolar sibilant [z] in the English informants' case). Variables such as VOT and the release burst intensity are gathered in the former, whereas the latter measures spectral peak location, COG, noise duration, noise amplitude, and F1-F3 values¹⁷. VOT and release burst intensity are measured manually, whereas fricatives' COG, noise duration, and noise amplitude are gathered through a Praat script (Elvira-García 2014), as well as their F1-F3 values (Kawahara 2010). As for spectral peak location, a specific procedure using spectral slices in Praat (Jongman et al. 2000: 1255) is followed to extract said values (see 3.7.3.2. *Segmental features* for more details).

In contrast with the analysis on suprasegmental features, the current inspection on segmental units does take into account measures such as mean values and standard deviations, since such variables are expected to contain more than one observation per unit of measurement. Nevertheless, the statistical measures adopted in this sub-section are more diverse depending on the nature of the data being analysed, based on whether the sample comes from a normal or a non-normal distribution (hence using a Shapiro-Wilk test to this end). A Mann-Whitney U test (since only two recordings are compared here) assesses the on-going relationships amongst segmental units (i.e. noise duration between [s] and [z]) and whether their variances differ substantially across speakers, should the data contain a non-normal distribution of values. Otherwise, an ANOVA would fulfil said role, with a Games-Howell (not assuming equal variances or sample

¹⁷ As discussed previously, using the concept *formant* in fricative consonants seems counterintuitive. However, this thesis takes Univaso et al.'s (2014) notion of concentrations of energy, instead.

sizes) post-hoc test which enables pairwise comparisons. In case the assumption of homogeneity of variances is violated with a significant value on Levene test (set at the 0.05 level), a more robust alternative is chosen (Welch’s test), unless otherwise specified.

As a reminder, the upcoming sub-sections deal with within-speaker variation in English (5.1.2.1.), Spanish (5.1.2.2.), and Dutch (5.1.2.3.) voice samples, accordingly.

5.1.2.1. English voice samples

To begin with, the analysis offers a first glimpse at the variables included within the voiced [b, d, g] and voiceless [k, p, t] plosives, as table 103 shows below:

[b, d, g] and [k, p, t]			
Variable	Sound differences	Test stats.	Sig. (p-value)
VOT	NO	Mann Whitney U	0.310
Release burst intensity	NO	Levene test	0.912
		ANOVA	0.867

Table 103. Within-speaker variation of variables within voiced/voiceless plosives in English voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).

First of all, a Shapiro-Wilk test is conducted in order to assess whether the combination of variables being tested exhibit normally distributed values. After checking for data normality, the appropriate statistical tests (Mann-Whitney U test and ANOVA in this case) are undertaken to find out whether the variables on the left column are influenced by the variable *Sound* ([b, d, g, k, p, t]). As noticed in the second column, the values seen in the variables *VOT* and *release burst intensity* do not differ significantly across sounds.

Secondly, the acoustic-phonetic variables (*VOT* and *release burst intensity*) are tested against the dependent variable *speaker*, following the same analytical procedure of checking for data normality as in the previous step. A Mann-Whitney U test indicated that the mean values of *VOT* in Simon T. Elliott (Mean rank= 16.71) are not statistically different from his suspect’s recording (Mean rank= 13.40, U = 81, p= 0.310). Similarly,

an ANOVA could not detect statistically significant differences in *release burst intensity* values amongst the selected speakers [$F(1, 27) = 0.029, p = 0.867$].

[s] and [z]			
Variable	Sound differences	Test stats.	Sig. (p-value)
Spectral peak location	NO	Levene test	0.737
		ANOVA	0.018
COG	NO	Levene test	0.435
		ANOVA	0.008
Noise duration	NO	Levene test	0.041
		Welch's test	0.156
Noise amplitude	NO	Levene test	0.906
		ANOVA	0.577
F1	NO	Levene test	0.164
		ANOVA	0.067
F2	NO	Levene test	0.077
		ANOVA	0.231
F3	NO	Levene test	0.785
		ANOVA	0.025

Table 104. Within-speaker variation of variables within voiced/voiceless alveolar sibilants in English voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott). Statistically significant values are marked in bold ($\alpha = 0.05$).

Once voiced/voiceless plosives have been proven to be fairly consistent through voice samples containing differing intonation contours, the analysis turns towards the variables gathered through voiced/voiceless alveolar sibilants, as seen in table 104 above. For the sake of brevity, this second sub-section shall only cover those variables which do display significant differences across the recordings of choice. Before proceeding further, it is worth noting that none of the variables are statistically different across the sounds [s] and [z].

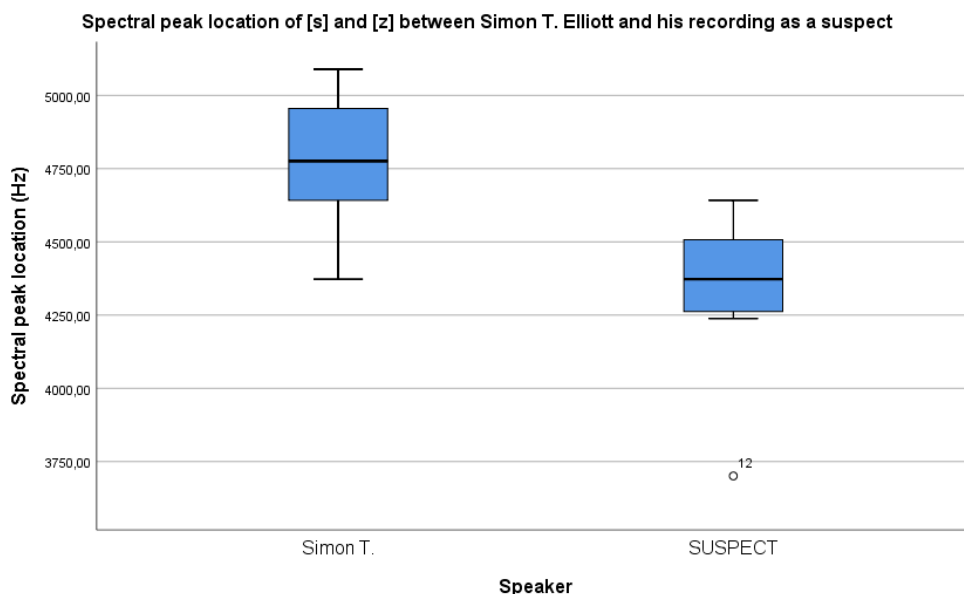


Figure 51. Statistically significant differences in spectral peak location values of [s] and [z] across the English suspect’s voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).

As the boxplots illustrate in figure 51 above, the interquartile range of values seen in the distractor recording (Simon T. Elliott) is significantly higher than the excerpt representing the suspect to identify (SUSPECT Simon T. Elliott), in terms of [s] and [z] spectral peak location [$F(1, 11) = 7.748, p= 0.018$].

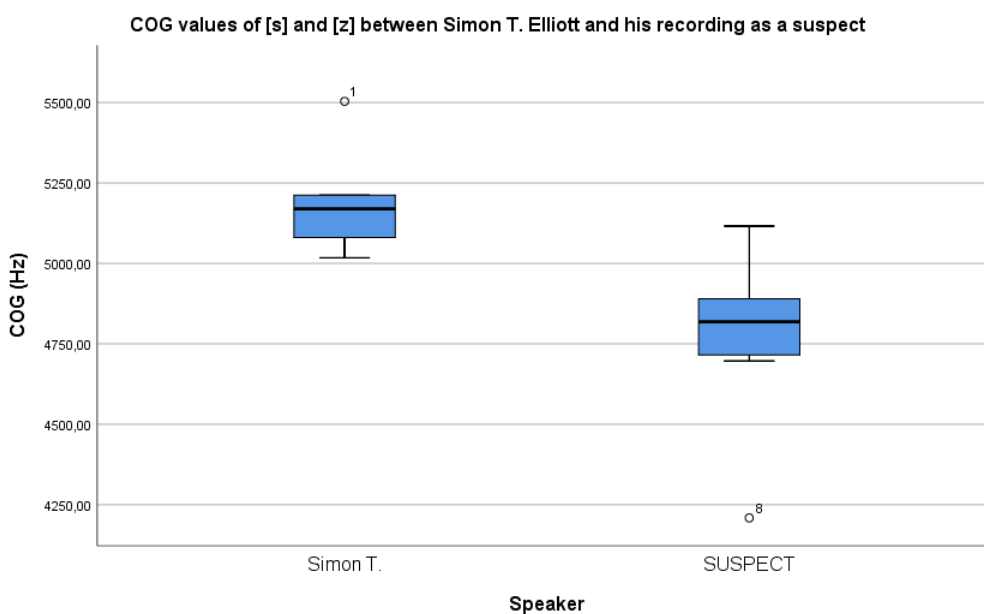


Figure 52. Statistically significant differences in COG values of [s] and [z] across the English suspect’s voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).

In a similar vein, figure 52 shows that [s] and [z] COG values in Simon T. Elliott are statistically different (with higher values) from his suspect's recording [$F(1, 11) = 10.466$, $p = 0.008$].

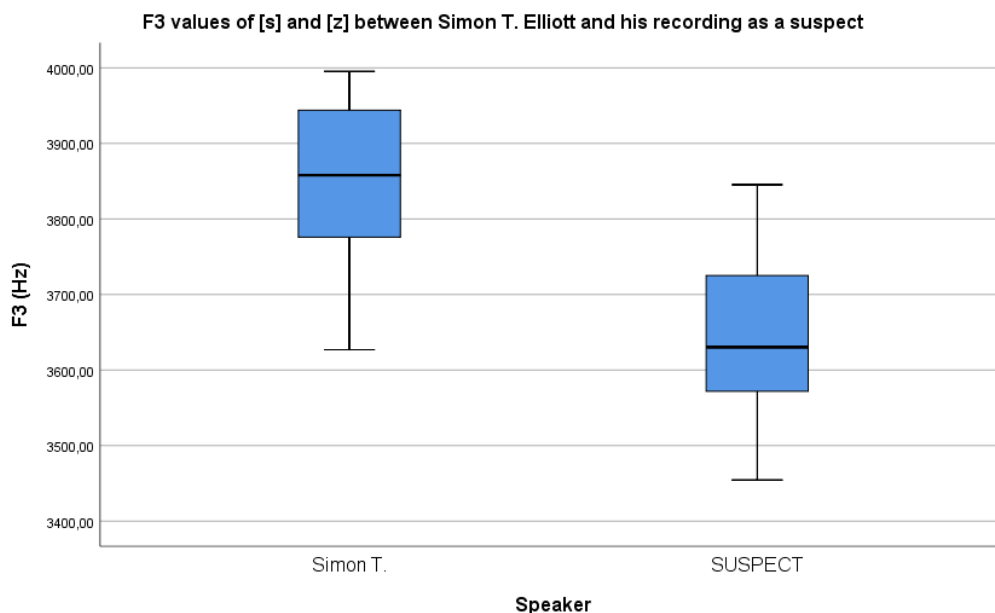


Figure 53. Statistically significant differences in F3 values of [s] and [z] across the English suspect's voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).

Lastly, figure 53 proves that the values for [s] and [z] observed under the label of F3 follow the same trend as in the previous acoustic-phonetic variables, as the suspect's recording values are significantly lower than Simon T. Elliott's [$F(1, 11) = 6.696$, $p = 0.025$].

Notwithstanding the significant differences seen in spectral peak location, COG, and F3 values between [s] and [z], it could be argued that the alternative hypothesis (the intravariability of suspects' voice samples with differing intonation contour is not statistically significant) applies to the rest of variables, which displayed non-significant results and are thus compliant with the formulation of said research hypothesis.

5.1.2.2. Spanish voice samples

This sub-section tackles the segmental analysis for the Spanish group of recordings. The following table displays the required statistical tests and their outcome for the variables concerned with the voiced [b, d, g] and voiceless [k, p, t] plosives:

[b, d, g] and [k, p, t]				
Variable	Sound differences	Test stats.		Sig. (p-value)
VOT	YES	[b, d, g]	Mann Whitney U	0.154
		[k, p, t]	Mann Whitney U	0.672
Release burst intensity	NO	Levene test		0.346
		ANOVA		0.006

Table 105. Within-speaker variation of variables within voiced/voiceless plosives in Spanish voice samples (M12_020- SUSPECT M12_020). Statistically significant values are marked in bold ($\alpha = 0.05$).

In this group of informants, table 105 suggests that VOT values do differ across the segmental features examined, as the conducted Welch's test confirms [$F(5, 10.046) = 20.984, p < 0.05$]. Figure 54 illustrates such differences below:

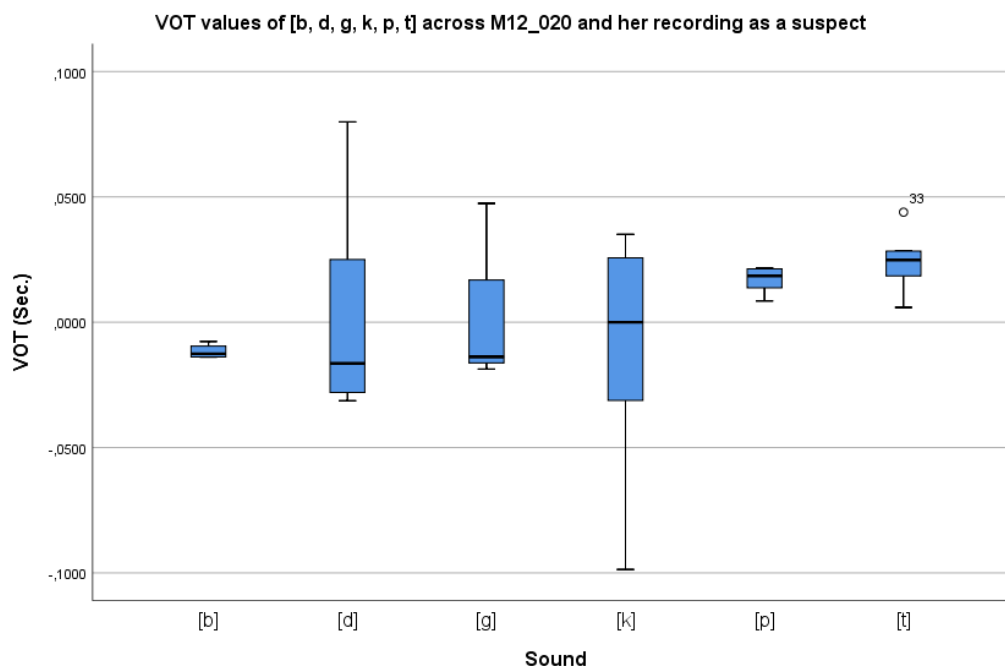


Figure 54. Statistically significant differences in VOT values across segmental units in the Spanish suspect’s voice samples (M12_020- SUSPECT M12_020).

As discerned in the boxplots above, voiced plosives [b, d, g] have noticeably lower values than voiceless plosives [k, p, t]. However, the post-hoc Games-Howell test reveals that there are only significant differences between [b]-[p, t], and [d]-[p, t]. This is to say that [g] and [k] fall somewhere in between said variables, as noticed by their overlapping interquartile range. As a result, VOT values across speakers have been segregated into voiced [b, d, g] and voiceless [k, p, t] plosives. By doing this, a Mann Whitney U test indicated that [b, d, g] mean VOT values extracted from M12_020 (Mean rank= 11.00) are not statistically different from her suspect’s recording (Mean rank= 6.27, U = 3, p= 0.154). Likewise, another Mann Whitney U test demonstrated that the mean VOT values for [k, p, t] in speaker M12_020 (Mean rank= 9.40) are not significantly different from the Spanish suspect’s recording, in statistical terms (Mean rank= 10.87, U = 32, p= 0.672).

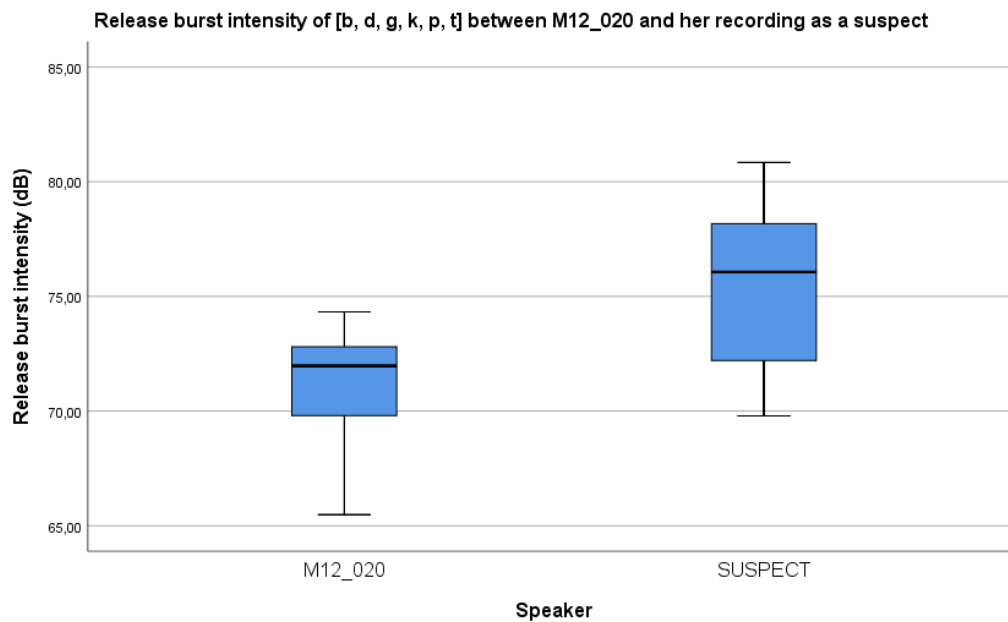


Figure 55. Statistically significant differences in all segmental units' release burst intensity across the Spanish suspect's voice samples (M12_020- SUSPECT M12_020).

Additionally, an ANOVA revealed that release intensity burst is significantly different between M12_020 and the suspect's tape [$F(1, 31) = 8.751, p = 0.006$], as observed in figure 55. Since the statistical tests did not find statistical differences amongst the stop consonants studied, it is surmised here that the suspect's stop consonants are not only louder, but the range of their values are also wider.

[s]		
Variable	Test stats.	Sig. (p-value)
Spectral peak location	Levene test	0.424
	ANOVA	0.357
COG	Levene test	0.094
	ANOVA	0.769
Noise duration	Levene test	0.179
	ANOVA	0.650
Noise amplitude	Levene test	0.559
	ANOVA	0.385
F1	Levene test	0.102
	ANOVA	0.565
F2	Levene test	0.348
	ANOVA	0.607
F3	Levene test	0.937
	ANOVA	0.991

Table 106. Within-speaker variation of variables within voiceless alveolar sibilants in Spanish voice samples (M12_020- SUSPECT M12_020).

A quick glance at table 106 seems enough to note that none of the variables under the sound [s] yield significant p-values when computing them alongside the dependent variable *speaker*. In this respect, this finding seems to indicate that such variables do not exhibit great within speaker variability and are thus fruitful for the purposes of this research hypothesis. With the exception of voiced/voiceless plosives release burst intensity, it can be surmised in this sub-section that the alternative hypothesis (the intravariability of suspects' voice samples with differing intonation contour is not statistically significant) is true.

5.1.2.3. Dutch voice samples

The third group of informants' voice samples is inspected here. To follow the same analytical procedure, this analysis starts with an examination on the variables attached to

the voiced [b, d, g] and voiceless [k, p, t] plosives. Table 107 summarises the outcome of the statistical tests undertaken:

[b, d, g] and [k, p, t]				
Variable	Sound differences	Test stats.		Sig. (p-value)
VOT	YES	[b, d, g]	Mann Whitney U	0.606
		[k, p, t]	Levene test	0.603
			ANOVA	0.910
Release burst intensity	NO	Levene test		0.611
		ANOVA		0.085

Table 107. Within-speaker variation of variables within voiced/voiceless plosives in Dutch voice samples (DVA8-F20L- SUSPECT DVA8-F20L).

Just as in the Spanish case, table 107 suggests that Dutch informants display relevant differences amongst their segmental features in relation to their VOT values. A Kruskal-Wallis test reported significant results in this respect ($H= 24.310$, $p= 0.000$, $df= 4$). The following graph illustrates said relationships:

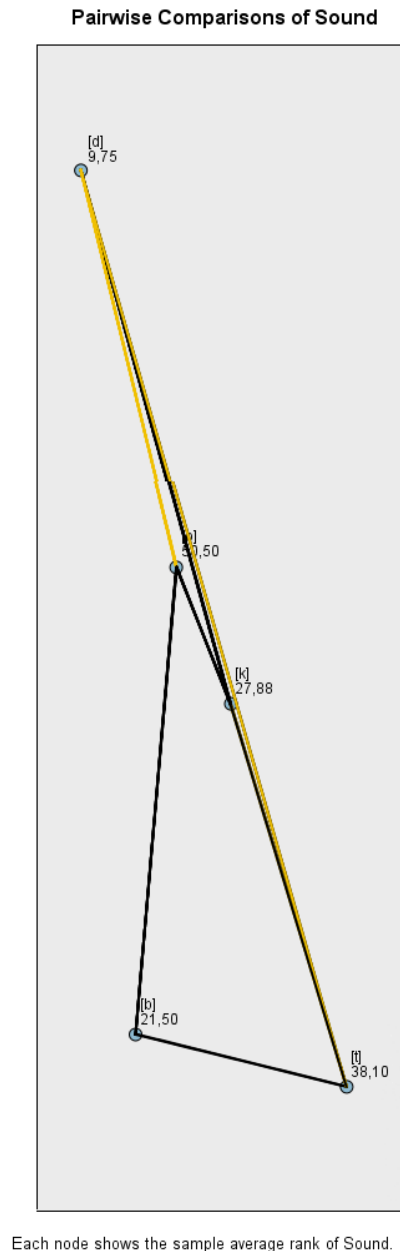


Figure 56. Statistically significant differences in [b, d, g] and [k, p, t] VOT values across the Dutch suspect's voice samples (DVA8-F20L- SUSPECT DVA8-F20L).

As appreciated in figure 56, the non-parametric test converted the values seen on each segmental unit to average ranks, rather than calculating mean values per variable. The resulting pairwise comparison (with adjusted p-values based on applying Bonferroni corrections) highlights a clear difference between one of the voiced plosives [b] and the rest of its voiceless counterparts [k, p, t]. Given the circumstances, voiced and voiceless plosives are calculated separately, as far as VOT values is concerned. Nevertheless, a Mann Whitney U test indicated that [b, d, g] VOT values do not seem to vary significantly

between DVA8-F20L (Mean rank= 7.00) and her suspect's voice sample (Mean rank= 8.40, $U = 27$, $p = 0.606$). Similarly, an ANOVA discerned no statistically significant differences amongst the aforementioned speakers and their [k, p, t] VOT values [$F(1, 42) = 0.013$, $p = 0.910$].

[s]		
Variable	Test stats.	Sig. (p-value)
Spectral peak location	Levene test	0.005
	Welch's test	0.117
COG	Levene test	0.012
	Welch's test	0.132
Noise duration	Levene test	0.155
	ANOVA	0.020
Noise amplitude	Levene test	0.795
	ANOVA	0.813
F1	Levene test	0.181
	ANOVA	0.423
F2	Levene test	0.368
	ANOVA	0.063
F3	Levene test	0.568
	ANOVA	0.213

Table 108. Within-speaker variation of variables within voiceless alveolar sibilants in Dutch voice samples (DVA8-F20L- SUSPECT DVA8-F20L). Statistically significant values are marked in bold ($\alpha = 0.05$).

Moving to the acoustic-phonetic measures concerned with the sibilant [s], most of the variables do not reflect significant differences between the two recordings chosen, as table 108 shows. However, an ANOVA found that the values registered in noise duration are statistically different between DVA8-F02L and SUSPECT DVA8-F20L [$F(1, 8) = 8.444$, $p = 0.020$].

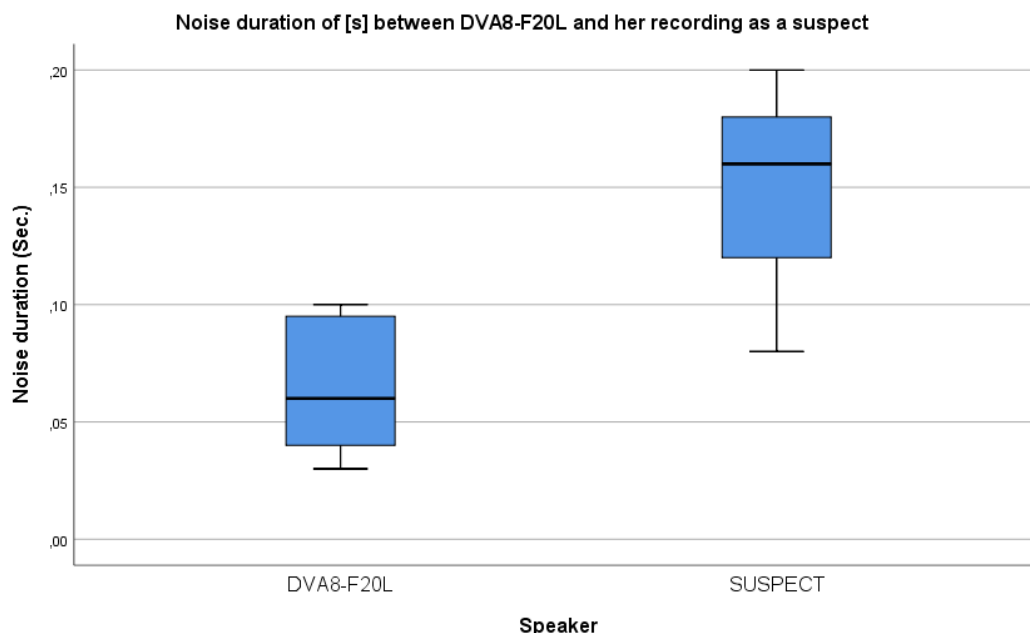


Figure 57. Statistically significant differences in [s]’ noise duration across the Dutch suspect’s voice samples (DVA8-F20L- SUSPECT DVA8-F20L).

As figure 57 shows, the boxplot drawn on the left side reports a median placed roughly at the 0.06 seconds mark for the distractor’s recording, whereas the voice sample used as a suspect exhibits a median which is close to 0.16 seconds. Additionally, it can be noticed that the whiskers in DVA8-F20L’s boxplot do not extend themselves too much from the interquartile range. The suspect’s frication noise duration, however, can reach up to 0.20 seconds, as the whisker at the top seems to indicate. This differentiation in frication noise duration could lead in theory to perceptually different voices, as the distractor’s (left side) frication noise is shorter in duration than the suspect’s (right side).

5.1.3. Summary of results

After examining every possible variable related to suprasegmental and segmental features in intra-speaker variation across English, Spanish, and Dutch voice samples, this subsection shall offer an overview of the most important findings. As a reminder, this hypothesis (n° 7) is formulated as follows: Intravariability of the suspects’ voice samples with differing intonation contour (rising and falling intonation) and uncontrolled segmental phenomena is not statistically significant. For the purposes of hypothesis testing, the results for suprasegmental-related variables are displayed in the following table:

Variables \ Informants	English	Spanish	Dutch
Mean pitch ($P\bar{x}$)	-	-	-
25% Pitch	-	-	-
50% Pitch	-	-	-
75% Pitch	-	X	-
Min. intensity ($I\downarrow$)	-	-	X
Max. intensity ($I\uparrow$)	-	-	X
Mean intensity ($I\bar{x}$)	-	-	X
DurPaus	-	X	-
N_paus/min	-	-	-
Pause_%	-	X	-
N_paus	-	-	X
Speech rate	-	X	-
Articulation rate	-	X	-
ASD	-	X	-

Table 109. Within-speaker variation of suprasegmental parameters across English, Spanish, and Dutch informants. The variables that yielded significant differences (at the 0.05 level) are marked with a cross.

As indicated in table 109 above, those variables whose cells are marked with a cross ended up spotting statistically significant differences between the voice samples with rising (recordings acting as distractors) and falling (recordings acting as suspects) intonation patterns. This is to say that the current research hypothesis is applicable to the variables whose cells are left unmarked. Quite noticeably, only English voice samples were similar enough in terms of suprasegmental features so as not to yield statistically different results. As for the Spanish and Dutch informants, there seems to be a mismatch as for what features caused a difference between the two types of speakers studied. Even if pausing measures like speech rate, articulation rate and ASD (Average Syllable Duration) do not seem to be strictly related to intonation patterns, they do report differences within the Spanish group of speakers (along with *DurPaus* and *Pause_%*). This observation seems to suggest that the aforementioned variables do not fare well in accounting for within-speaker variation in the experimental conditions of this study,

specifically. Likewise, measures of intensity exhibit too much variation, but said statement is only applicable to Dutch voice samples.

Variables \ Informants	English	Spanish	Dutch
VOT	-	-	-
Release burst intensity	-	X	-
Spectral peak location	X	-	-
COG	X	-	-
Noise duration	-	-	X
Noise amplitude	-	-	-
F1	-	-	-
F2	-	-	-
F3	X	-	-

Table 110. Within-speaker variation of segmental parameters across English, Spanish, and Dutch informants. The variables that yielded significant differences (at the 0.05 level) are marked with a cross.

As for the segmental features of interest, table 110 does not seem to report a clear pattern highlighting less efficient parameters in within-speaker variation. Rather, it seems that each group’s recordings display unequal values in very specific areas. Such is the case in Spanish and Dutch informants, since their audio material could only be differentiated in terms of release burst intensity and frication noise duration, accordingly. As for the English case, spectral peak location, COG, and F3 are the variables which detected differences amongst the research subjects. However, this perceived instability of values in the English group could be due to the inclusion of both [s] and [z] in the data set, as opposed to the other two groups of informants.

To conclude, differences in acoustic-phonetic parameters are inherently expected to emerge amongst voice samples, even if the audio material being analysed targets the same speaker under different circumstances. Nevertheless, this research hypothesis attempts to find the most robust segmental and suprasegmental features to intra-speaker variation. As a result, the alternative hypothesis is accepted in those features which did not find significant differences across the three groups of informants examined, namely *Mean pitch (P \bar{x})*, *25% Pitch*, *50% Pitch*, *N_paus/min*, *VOT*, and *sibilant’s noise amplitude*, *F1*,

and *F2* values. It should be noted, however, that longer voice samples could have changed the results to some extent, since the analysis would gather more representative data for each variable.

5.2. INTERVARIABILITY OF FOIL SPEAKERS

On the flip side, hypothesis 8 (H_8) asserts that intervariability of foil speakers' voices adopting similar intonation patterns (rising intonation) and uncontrolled segmental phenomena (semi-spontaneous data) is statistically significant. This hypothesis advocates for a successful discrimination of speakers by means of acoustic-phonetic analysis despite their similarities in terms of intonation contours. Therefore, the following null hypothesis (H_0) is formulated alongside its alternative hypothesis (H_8):

- H_0 : Intervariability of the foil speakers' voice samples with similar intonation patterns (rising intonation) and uncontrolled segmental phenomena is not statistically significant.
- H_8 : Intervariability of the foil speakers' voice samples with similar intonation patterns (rising intonation) and uncontrolled segmental phenomena is statistically significant.

Much in line with the previous analysis (H_7), the eighth hypothesis also considers suprasegmental and segmental features as independent variables. However, dependent variables (speakers) have changed in the current scenario, as the table below illustrates:

Independent variable		Dependent variable		
		Speakers		Code
Suprasegmental		English informants	Alan McElligott	1
Pitch	Mean pitch ($P\bar{x}$)		Alan Burbidge	2
	25% Pitch		Jez Riley	3
	50% Pitch		Peter Toll	4
	75% Pitch		Richard Beard	5
	Min. intensity ($I\downarrow$)		Richard Youell	6
	Max. intensity ($I\uparrow$)		Simon K. Bearder	7
	Mean intensity ($I\bar{x}$)		Simon T. Elliott	8
Pauses	DurPaus	Spanish informants	M12_020	1
	N_paus/min		M12_030	2
	Pause_%		M12_036	3
	N_paus		M13_008	4
	Speech rate		M13_010	5
	Articulation rate		M13_016	6
	ASD (Average Syllable Duration)		M13_016_hab2	7
Segmental		Dutch informants	DVA8-F20K	1
[b, d, g]	VOT (Voice Onset Time)		DVA8-F20L	2
[k, p, t]	Release burst intensity		DVA9-F21M	3
[s], ([z])	Spectral peak location		DVA9-F21N	4
	COG (Center of Gravity)		DVA10-F18O	5
	Noise duration		DVA10-F19P	6
	Noise amplitude		DVA11-F28Q	7
	F1		DVA11-F28R	8
	F2			
	F3			

Table 111. Selected variables for hypothesis 8 testing, displaying the sub-types of both independent (left column) and dependent (right column) variables¹⁸.

As disclosed in table 111, each corpus contains either 7 (Spanish informants) or 8 (English and Dutch informants) voice samples. Rather than being classified as ordinal variables, the codes assigned to each subject are merely categorical. In other words, the numbers do

¹⁸ Independent variables are described in their respective sub-sections depending on their sub-type: suprasegmental (5.2.1.) and segmental (5.2.2.) features.

not reflect a pre-established hierarchy, but rather sets each case apart from each other. In this setting, it should be noted that all audio material analysed here has a common feature, which is the intonation contour (falling intonation), since these samples were used in the construction of the voice line-ups appearing in perception surveys. Conversely, segmental units could not be controlled due to the nature of the data set (semi-spontaneous recordings), which removes the researcher's agency upon the actual speech produced.

Regarding the independent variables of choice, table 111 refers to the same exact variables studied as in the previous point 5.1. (*Intravariability of suspects*), namely suprasegmental features including measures related to pitch and pausing, and the segmental parameters related to voiced/voiceless plosives and voiced/voiceless alveolar sibilants. Likewise, the established order of elements of analysis has not been modified, and thus follows the same line as in the previous analysis: suprasegmental (5.2.1.) and segmental (5.2.2.) features with their respective groups of informants (English, Spanish, and Dutch voice samples).

5.2.1. Suprasegmental features

As explained in point 5.2. (*Intervariability of foil speakers*), the suprasegmental section covers pitch-related measurements (mean pitch, 25% quantile, 50% quantile, 75% quantile, minimum/maximum intensity, and mean intensity) and those concerned with pauses (pauses duration per minute, number of pauses per minute, percentage of pauses per sample, number of pauses in the sample, speech rate, articulation rate, and ASD). In this case, between-speaker variation is measured on the basis of the aforementioned variables, and thus two types of observations may derive from said analysis: the efficiency of acoustic-phonetic variables and the individual differences found amongst foil speakers. The upcoming sub-sections deal with the former, whereas the latter can be consulted in the respective appendixes made for each group (English, Spanish, and Dutch). The statistical method of choice is a dissimilarity matrix which computes the Euclidean distances between independent variables (with standardised Z-scores) across speakers, since said independent variables lack variation in their values (a single point of data per variable).

The extraction of said data points entailed the selection of the whole excerpt (4-14 sec. of duration) to calculate said variables's values through Praat scripts. In this vein, pitch-related measures were extracted by using the *draw_pitch_histogram_from_sound* Praat script (Lennes 2013), whereas pausing variables tapped into *Syllable Nuclei v2* Praat script (De Jong & Wempe 2008) for analytical purposes. As for those variables concerned with intensity, a manual extraction is deemed more fitting due to the fact that the researcher is able to remove unwanted noise disturbances, should an intervention of this kind be needed (see 3.7.3.1. *Suprasegmental features* for more details).

Again, it should be underlined that the upcoming sub-sections provide a summary of the total amount of cases (through independent and dependent variables) which displayed significant differences. Appendixes 7-9 contain a full description of the exact significant pairwise comparison within each group of informants. In this regard, this analysis is broken down into: English (5.2.1.1.), Spanish (5.2.1.2.), and Dutch (5.2.1.3.) voice samples.

5.2.1.1. English voice samples

To discern how effective suprasegmental features are in discriminating different speakers with the same intonation pattern (rising), the upcoming table displays the number of cases where significant differences were found:

Informants Variables	Alan McElligott	Alan Burbidge	Jez Riley	Peter Toll	Richard Beard	Richard Youell	Simon K. Bearder	Simon T. Elliott
Mean pitch (\bar{P})	3	4	1	1	0	1	1	1
25% Pitch	3	3	1	1	0	0	1	1
50% Pitch	1	4	1	1	0	1	0	0
75% Pitch	1	4	1	3	1	1	0	1
Min. intensity (I_{\downarrow})	2	1	1	1	6	1	1	1
Max. intensity (I_{\uparrow})	2	3	1	5	1	2	0	2
Mean intensity (\bar{I})	1	1	1	6	0	1	1	1
DurPaus	1	3	3	5	1	2	2	1
N_paus/min	1	2	2	2	2	0	5	4
Pause_%	1	3	3	5	1	2	2	1
N_paus	1	2	1	1	1	1	1	6
Speech rate	0	3	2	2	2	5	1	1
Articulation rate	1	1	1	2	1	4	4	2
ASD	1	1	1	2	1	4	4	2

Table 112. Between-speaker variation of suprasegmental parameters across English voice samples. The total count of cases signalling significant pairwise differences (at the 0.05 level) are noted in each cell.

By looking at table 112, some cells appear to point at certain speakers whose speech is significantly different from other foil speakers'. For instance, Alan McElligott and Alan Burbidge reported higher cases of dissimilar values in measures related to pitch measurements, while Peter Toll and Richard Beard's intensity values stand out from the rest in certain categories (maximum and mean intensity for Peter Toll, and minimum intensity for Richard Beard).

As for the second array of variables belonging to measures of pausing, the reported cases of dissimilarities appear to be more evenly distributed amongst English subjects. Most notably, Simon T. Elliott's number of pauses stand out from the rest (n=6), whilst Simon K. Bearder's number of pauses per minute are close to contain unique values in the group (n=5). On another note, the suprasegmental variables which reported more cases of dissimilarities amongst speakers are max. intensity (n= 8) and min. intensity (n= 7) in the pitch-related area. Concerning pausing measurements, N_paus/min (n =9), DurPaus (n= 9), Pause_% (n= 9), speech rate (n=8), articulation rate (n=8), and ASD (n= 8) are the variables which report more auspicious results. See *Appendix 7* for a more detailed breakdown of individual differences between speakers across suprasegmental units of measurement.

5.2.1.2. Spanish voice samples

Secondly, Spanish recordings are examined with the aim of discovering which suprasegmental features are relevant in discrimination of speakers. The following table accounts for the statistically significant number of cases of dissimilarities found:

Informants Variables	M12_020	M12_030	M12_036	M13_008	M13_010	M13_016	M13_016_ hab2
Mean pitch (\bar{P})	1	1	6	1	1	1	1
25% Pitch	1	1	6	1	1	1	1
50% Pitch	1	1	6	1	1	1	1
75% Pitch	1	1	6	1	1	1	1
Min. intensity (I_{\downarrow})	1	2	2	0	2	3	4
Max. intensity (I_{\uparrow})	1	1	0	1	2	0	3
Mean intensity (\bar{I})	1	0	1	1	1	0	4
DurPaus	2	2	3	0	2	0	3
N_paus/min	0	2	0	2	2	0	2
Pause_%	2	2	3	0	2	0	3
N_paus	1	2	1	5	2	1	2
Speech rate	2	2	1	2	1	5	1
Articulation rate	1	0	0	1	1	3	0
ASD	3	2	1	2	1	3	0

Table 113. Between-speaker variation of suprasegmental parameters across Spanish voice samples. The total count of cases signalling significant pairwise differences (at the 0.05 level) are noted in each cell.

The first striking aspect of table 113 is the fact that speaker M12_036's pitch-related values are dissimilar enough from the rest of foil speakers ($n= 6$). When looking at variables concerned with intensity, however, it appears that M13_016_hab2 gathers a greater number of cases of dissimilarities than the rest of speakers.

As for pausing measures, Spanish recordings face a similar scenario as the one perceived in English informants. In other words, there is no clear-cut distinction as to what speaker

stands out the most in this domain. Instead, certain speakers display higher discrepancies in very specific categories. Such is the case of M13_008 (N_paus, n= 5) and M13_016 (speech rate, n= 5). Since it is evident that each variable contains at least 1 relationship of dissimilarities between speakers, it could be said that all of them are worthy of examination in between-speaker variation studies. Nevertheless, speech rate (n= 7), N_paus (n=7), and min. intensity (n= 7) report greater results in this group of subjects (consult *Appendix 8* for a more detailed view on between-speaker variation in suprasegmental variables).

5.2.1.3. Dutch voice samples

This third sub-section covers the importance of suprasegmental features in between-speaker variation. In this occasion, Dutch speakers are the subjects of scrutiny, as the following table shows:

Informants Variables	DVA8- F20K	DVA8- F20L	DVA9- F21M	DVA9- F21N	DVA10- F180	DVA10- F19P	DVA11- F28Q	DVA11- F28R
Mean pitch (\bar{P}_x)	2	1	2	3	5	0	2	1
25% Pitch	1	1	3	1	4	0	1	1
50% Pitch	1	0	1	0	5	1	1	1
75% Pitch	2	1	2	3	5	0	2	1
Min. intensity (I_{\downarrow})	1	3	4	1	2	0	1	2
Max. intensity (I_{\uparrow})	5	5	2	3	2	2	3	2
Mean intensity (\bar{I}_x)	4	4	0	2	2	2	2	0
DurPaus	1	1	1	3	5	3	2	2
N_paus/min	1	2	4	2	2	0	1	0
Pause_%	1	1	1	3	5	3	2	2
N_paus	1	5	1	3	1	1	1	1
Speech rate	1	1	1	1	5	3	1	1
Articulation rate	1	1	1	1	6	2	1	1
ASD	1	1	1	1	6	2	1	1

Table 114. Between-speaker variation of suprasegmental parameters across Dutch voice samples. The total count of cases signalling significant pairwise differences (at the 0.05 level) are noted in each cell.

Much in line with the previous analyses, the Dutch group also contains certain speakers whose speech stand out from the rest in specific areas, as table 114 illustrates. For example, DVA10-F180 exhibits more dissimilar relationships than the rest, as far as pitch-related measures are concerned. In terms of intensity, both DVA8-F20K and DVA8-F20L's samples register a higher number of cases of dissimilarities (in max. and mean intensity).

Once again, pausing measures display a more balanced chart. For instance, DVA9-F21M's values on N_paus/min (n= 5) are significantly different in more cases than in other speakers in the group, while the same applies to DVA8-F20L in N_paus (n= 5). Additionally, the values seen in DurPaus, Pause_%, speech rate, articulation rate, and ASD are outstanding in DVA10-F180 and DVA10-F19P's recordings. Aside from that, the variables which report more discrepancies and are thus useful in measuring between-speaker variation are: max. intensity (n= 12), DurPaus (n= 9), Pause_% (n= 9), mean intensity (n= 8) and mean pitch (n= 8). See *Appendix 9* for a full review of the individual differences amongst Dutch speakers in terms of suprasegmental phenomena.

5.2.2. Segmental features

As previously explained (see 5.2. *Intervariability of foil speakers*), the segmental features of choice for this analysis are variables concerned with voiced ([b, d, g]) and voiceless ([k, p, t]) plosives, on the one hand, and voiceless alveolar sibilants [s] (while the voiced [z] is included in English voice samples), on the other. The first group of segmental units registers variables like VOT and release burst intensity across speakers, which are calculated manually. Moreover, the second group employs a wider variety of variables, namely spectral peak location, COG, noise duration, noise amplitude, and F1-F3 values¹⁹. With the exception of spectral peak location (see 3.7.3.2. *Segmental features* for specific details), the rest of variables concerned with [s] (and [z] in English informants) are gathered through Praat scripts. This includes both COG, noise duration, and noise amplitude (Elvira-García 2014), and F1-F3 values (Kawahara 2010).

Since more than one observation is registered per segmental variable (i.e. several realisations for [s]), the statistical tests employed account for said variation of values (including measures like mean values and standard deviations). Before the analytical procedure begins, it is necessary to figure out the best statistical measure for the data set, either with parametric or non-parametric tests. This is accomplished by running a Shapiro-Wilk test. If the values do not come from a normal distribution, A Kruskal-Wallis test is used (since the model contemplates all foil speakers, whose number is higher than

¹⁹ As pointed out already, the notion of *formant* seems inaccurate when talking about fricative consonants. Despite that, this PhD project conceives them as concentrations of energy, as Univaso et al.'s (2014) note.

2 individuals). Otherwise, an ANOVA test (with its respective Games-Howell post-hoc test) shall assess the variance exhibited by segmental variables and determine whether such changes are significant across speakers (dependent variable). A slight variation includes Welch's test, which is a more robust version of ANOVA that is consulted whenever the assumption of homogeneity of variances (calculated through a Levene test) is breached.

Given the main goal of this section (measure between-speaker variation), an additional piece of information is added to the tables with the purpose of underlining significant pairwise comparisons in English (5.2.2.1.), Spanish (5.2.2.2.), and Dutch (5.2.2.3.) voice samples.

5.2.2.1. English voice samples

The first group containing English voice samples is analysed in this sub-section dealing with intervariability of foil speakers. The table below spells out the extant relationships between suprasegmental variables (voiced and voiceless plosives) and sound differences/speakers:

[b, d, g] and [k, p, t]					
Variable	Sound differences	Test stats.		Sig. (p-value)	Significant pairwise comparisons
VOT	YES	[b, d, g]	Kruskal-Wallis	0.382	-
		[k, p, t]	Kruskal-Wallis	0.081	-
Release burst intensity	YES	[b, d, g]	Levene test	0.009	-
			ANOVA	0.004*	
		[k, p, t]	Levene's test	0.007	1-2 1-4 1-5 1-7
			Welch's test	0.004	4-6 4-8 5-6 5-8

Table 115. Between-speaker variation of variables within voiced/voiceless plosives in English voice samples. Note: an asterisk is marked on ANOVAs which violate the homogeneity of variances assumption (confirmed by a significant Levene test at the 0.05 level).

As noticed in table 115, a Kruskal-Wallis test reported significant differences in the distribution of VOT across the categories of sound ($H= 22.256$, $p= 0.000$, $df= 5$). The specific details on differentiations amongst segmental units are noted in the following table:

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
[b]-[t]	-24.022	7.789	-3.084	0.002	0.043
[b]-[k]	-26.188	8.432	-3.106	0.002	0.040
[d]-[t]	-22.043	7.075	-3.116	0.002	0.039
[d]-[k]	-24.210	7.777	-3.113	0.002	0.039

Table 116. Between-speaker differences amongst English recordings regarding VOT in voiceless and voiced plosives according to a Kurskal-Wallis test.

As discerned in table 116, the differences remain on [b]-[k, t], and [d]-[k, t]. Such differences set voiced apart from voiceless consonants, even though this observation is not applicable through every segmental feature (most likely not reaching significance due to Bonferroni corrections). Even after splitting them up, a Kruskal-Wallis test has determined that [b, d, g] VOT values are not useful in differentiating foil speakers ($H= 7.468, p= 0.382, df= 7$). The same is true for [k, p, t] VOT values ($H= 12.661, p= 0.081, df= 7$).

As for release burst intensity, it remains clear from the table above that significant differences arise amongst sounds, as a Kruskal-Wallis test has confirmed ($H= 15.657, p= 0.008, df= 5$). Such differences are quoted in the following table:

Sample 1- Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
[b]-[t]	29.100	7.789	3.736	0.000	0.004

Table 117. Between-speaker differences amongst English recordings regarding release burst intensity in voiceless and voiced plosives according to a Kurskal-Wallis test.

Despite yielding a low p-value (0.008), only one pairwise comparison turned out to be significant, as table 117 shows. Again, this could be explained on the basis of the resulting adjusted p-values stemming from Bonferroni corrections. Either way, release burst intensity for [b, d, g] could not compute a Welch’s test because at least one group has 0 variance. Said cases are speaker 3 (Jez Riley) and speaker 7 (Simon K. Bearder), who display 1 observation in one of the segmental variables of analysis²⁰. A significant ANOVA ensued [$F(7, 22) = 4.353, p= 0.004$]. However, when looking at pairwise comparisons, no speakers display statistically significant differences (at the 0.05 level), which is indicated in the table above with an asterisk. In the case for [k, p, t], a Welch’s test was consulted [$F(7, 6.879) = 10.116, p= 0.004$]., since an ANOVA would not be reliable owing to significant results in a Levene test (< 0.05). In this specific scenario, a Tukey’s HSD post-hoc test is run instead of the already established Games-Howell, due to the fact that the number of occurrences per subject is not as unbalanced as in other acoustic-phonetic variables. Besides that, Tukey’s HSD test rendered more accurate results in accordance with how the data is shown through plots such as the following:

²⁰ In order to enable a pairwise comparison through post-hoc tests, the only point of data registered in these speakers has been duplicated ($n= 2$). Understandably, their values still display 0 variance.

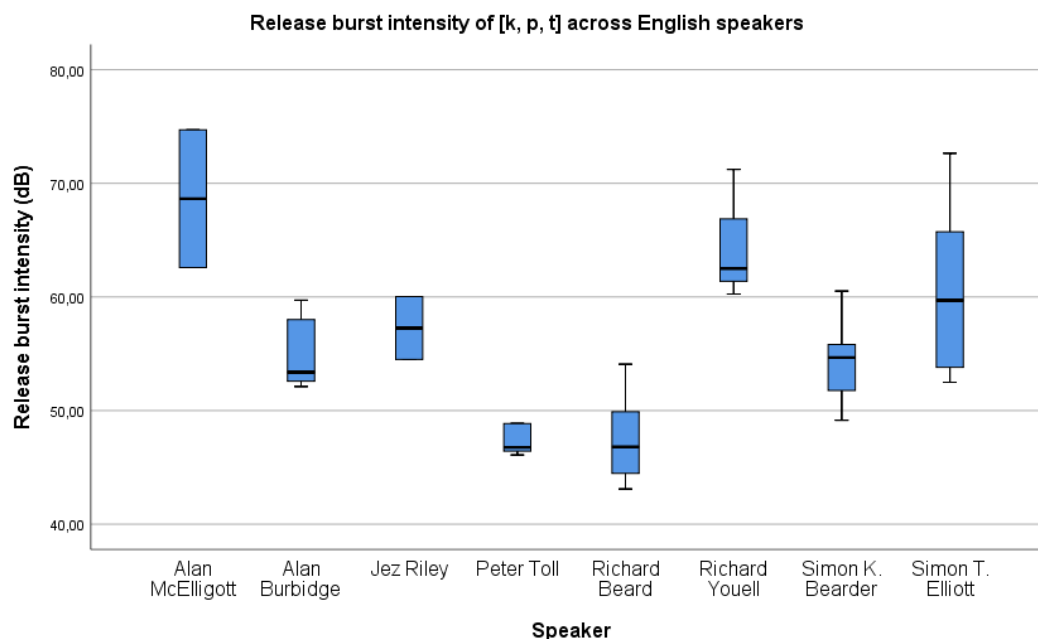


Figure 58. Between-speaker variation of release burst intensity in [k, p, t] across English voice samples.

As the boxplots illustrated in figure 58 show, voiceless plosives are statistically different between speaker 1 (Alan McElligott) and speakers 2 (Alan Burbidge), 4 (Peter Toll), 5 (Richard Beard), and 7 (Simon K. Bearder), on the one hand. Additionally, 4's (Peter Toll) release burst intensity is statistically different from 6 (Richard Youell) and 8's (Simon T. Elliott). As a final remark, Richard Beard's (speaker n° 5) values are dissimilar from 6 (Richard Youell) and 8 (Simon T. Elliott).

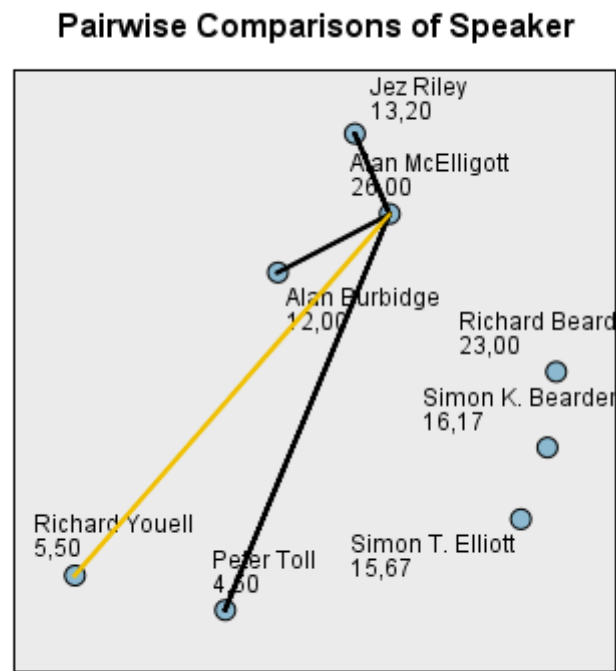
Before proceeding any further, it should be warned that one of the subjects (Richard Beard) displays no instances of [s], which excludes him from pairwise comparisons specifically in this segmental unit's analysis. Throughout the voiced alveolar sibilant [z] and its voiceless counterpart [s], some speakers may register only one point of data per segmental feature. In order to enable pairwise comparisons, such observations are noted twice (n= 2), without affecting their mean values or standard deviations (which are non-existent).

[s] and [z]					
Variable	Sound differences	Test stats.		Sig. (p-value)	Significant pairwise comparisons
Spectral peak location	NO	Kruskal-Wallis		0.018	1-6
COG	YES	[s]	Levene test	0.022	1-7
			ANOVA		0.000*
		[z]	Levene test	0.000	2-7
			ANOVA		0.915*
				3-7	
				3-8	
				4-7	
				4-8	
				6-7	
Noise duration	NO	Kruskal-Wallis		0.360	-
Noise amplitude	NO	Levene test		0.009	1-2
		ANOVA			0.000*
					2-8
					4-5
					4-6
					4-7
					4-8
F1	YES	[s]	Levene test	0.026	1-4
			ANOVA		0.000*
					1-8
					2-4
					2-6
					2-8
					3-8
					6-7
					7-8

		[z]	Levene test	0.000	-
			ANOVA	0.511*	
F2	YES	[s]	Levene test	0.018	1-4
			ANOVA	0.010*	4-7
		[z]	Levene test	0.065	3-5
			ANOVA	0.039	
F3	YES	[s]	Levene test	0.059	1-8
			ANOVA	0.000	2-8
					3-8
				7-8	
		[z]	Levene test	0.220	-
			ANOVA	0.546	

Table 118. Between-speaker variation of variables within voiced/voiceless alveolar sibilants in English voice samples. Note: an asterisk is marked on ANOVAs which violate the homogeneity of variances assumption (confirmed by a significant Levene test at the 0.05 level).

Starting from the first row at the top of table 118, spectral peak location values are significant across speakers, as a Kruskal-Wallis test confirmed ($H= 16.837$, $p= 0.018$, $df= 7$). However, Bonferroni corrections adjusted the resulting p-values, and thus only 1 pairwise comparison is found significant (1. Alan McElligott and 6. Richard Youell), as shown in figure 59 below:



Each node shows the sample average rank of Speaker.

Figure 59. Between-speaker variation of [s] and [z] spectral peak location across English voice samples.

The next variable, COG, displays differences between [s] and [z], as a Kruskal-Wallis test has proven ($H= 11.637$, $p= 0.001$, $df= 1$). COG values for [s] and [z] could not be processed with a Welch's test, because of the above-mentioned issue (one of the speakers lacks variance in relation to this variable). Even so, a Tukey's HSD post-hoc test shows differences across speakers, but only in [s]:

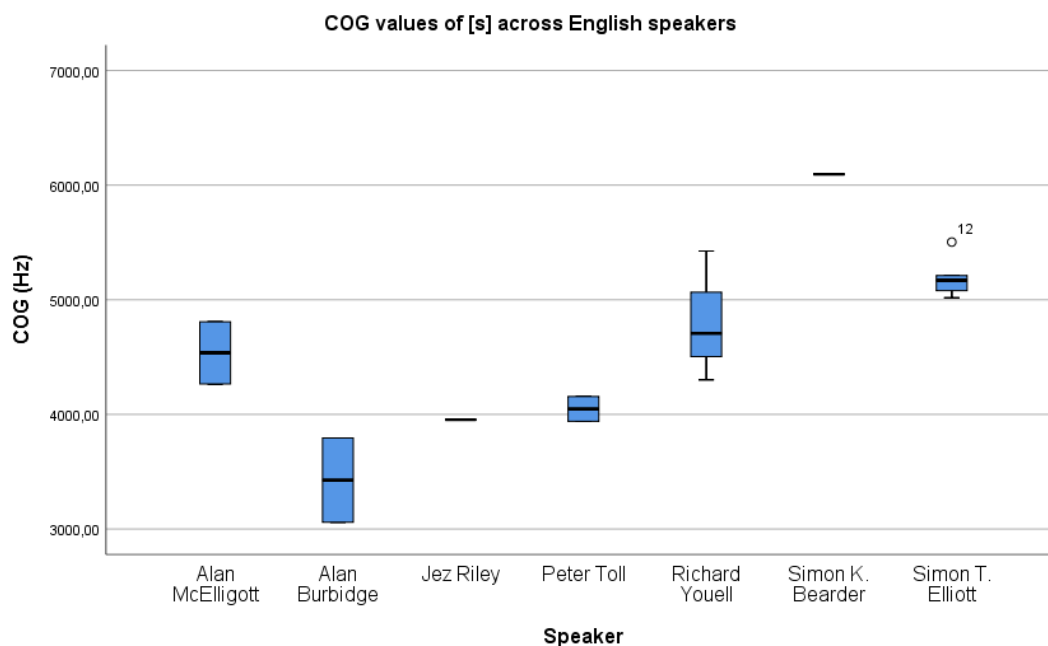


Figure 60. Between-speaker variation of [s] COG across English voice samples.

Despite showing a significant p-value in Levene test (based on mean, but non-significant when based on median and with adjusted df), an ANOVA of COG's [s] showed significant differences across speakers [$F(6, 12) = 17.269, p= 0.000$].

As shown in figure 60, the speakers in the middle (Jez Riley, Peter Toll, and Richard Youell) and Alan McElligott on the left are those who display a central tendency on COG values in [s] realisations. This is to say that discrepancies occur mainly between the aforementioned group of speakers and those with lower (Alan Burbidge) and higher (Simon K. Bearder and Simon T. Elliott) range of values. Needless to state that these two extremes shall also unveil statistically different results.

Noise amplitude is proven non-significant across sounds but appears significant across speakers. Just as in the previous cases, a Welch's test could not be calculated. Despite a positive result in Levene's test (0.009), an ANOVA still displays significant p-values [$F(7, 22) = 7.007, p= 0.000$]. Again, a Tukey's HSD post-hoc test reveals significant pairwise comparisons:

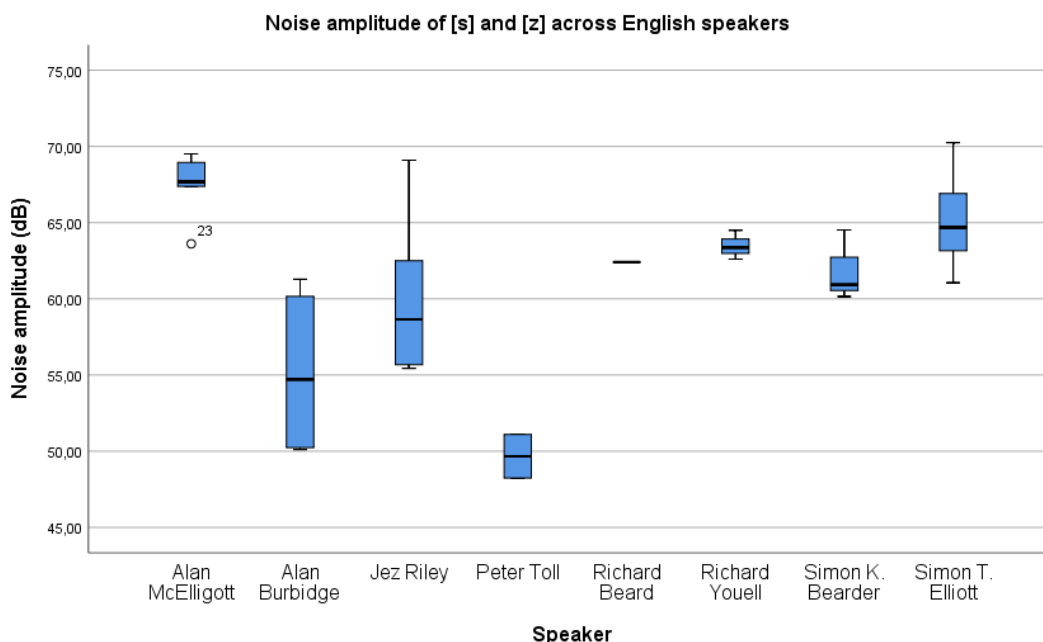


Figure 61. Between-speaker variation of noise amplitude in [s] and [z] across English voice samples.

The first remarkable aspect of figure 61 is that the group to the right (Richard Beard, Richard Youell, Simon K. Bearder, and Simon T. Elliott) share similar noise amplitudes along with the first speaker (Alan McElligott. In this sense, all the significant differences amongst subjects seem to concern those speakers in the middle, especially speaker 2 (Alan Burbidge) and 4 (Peter Toll). Speaker 3 (Jez Riley), however, assumes the most centric tendency of all and thus does not differ substantially from any of the speakers in the graph.

Moreover, F1 values seem to be different across [s] and [z], given the significance of a Kruskal-Wallis test ($H= 8.708, p= 0.003, df= 1$). Due to the same reasons exposed above, a Welch’s test could not be computed, and thus an ANOVA yielded significant results in F1’s [s] only [$F(6, 12) = 17.460, p= 0.000$]. A Tukey’s HSD post-hoc test detects significant pairwise comparisons, which are illustrated in the graph below:

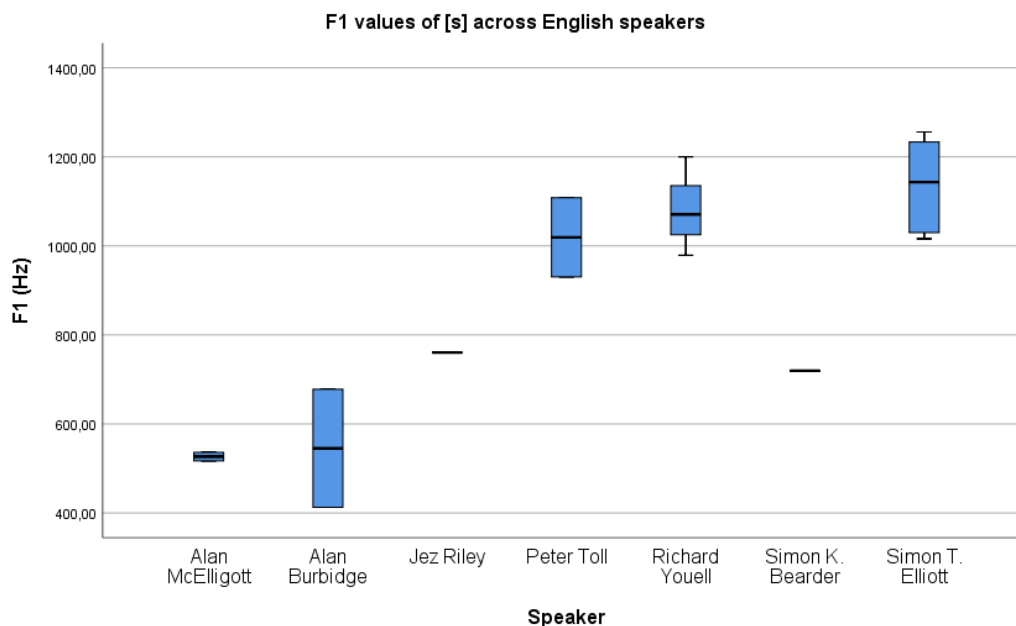


Figure 62. Between-speaker variation of F1 values in [s] across English voice samples.

By looking at figure 62 above, the differences between the boxplots from the first two speakers (Alan McElligott and Alan Burbidge) is discernible in comparison with speakers 4 (Peter Toll), 6 (Richard Youell), and 8 (Simon T. Elliott). On the other hand, speakers 3 (Jez Riley) and 7 (Simon K. Bearder) seem to fit within everyone’s range of values, except from Simon T. Elliott’s.

An ANOVA found differentiated F2 values across segmental units [$F(7, 22) = 7.385, p= 0.000$]. In the case of F2’s [s], a Welch’s test is not available. Hence an ANOVA is consulted, which revealed significant differences arising within the variable *speaker* [$F(6, 12) = 4.780, p= 0.010$].

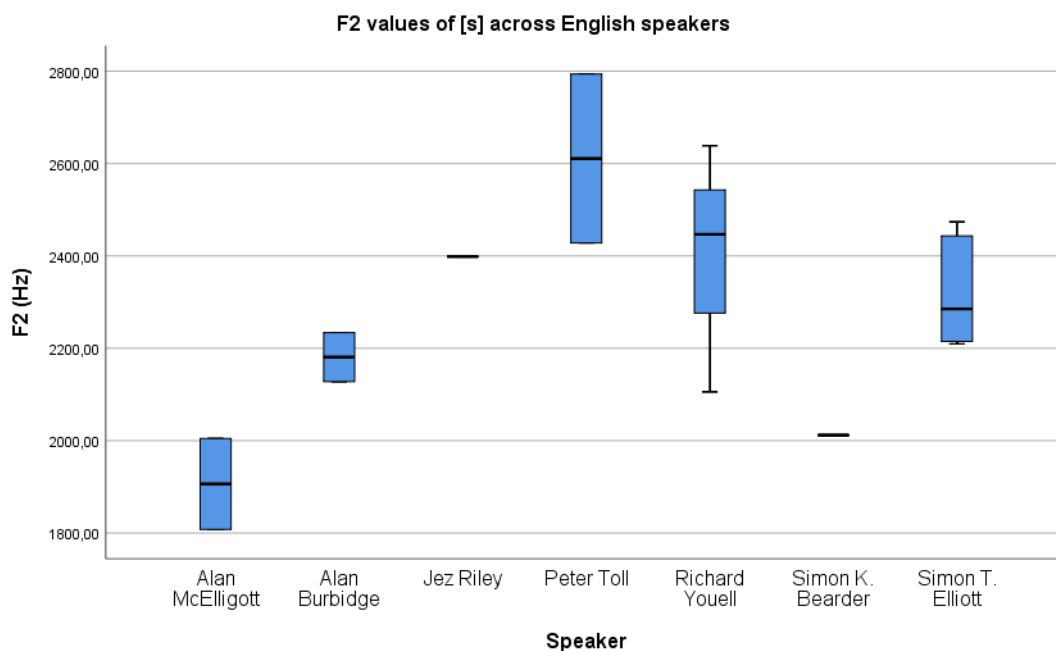


Figure 63. Between-speaker variation of F2 values in [s] across English voice samples.

A Tukey HSD post-hoc appears to signal differences between speakers 1 (Alan McElligott) and 4 (Peter Toll), as shown in figure 63. Furthermore, Peter Toll's F2's values in [s] realisations are in turn dissimilar enough from speaker Simon K. Bearder's (speaker 7).

Also, the voiced alveolar [z] exhibits statistically significant differences across speakers in terms of F2 values, which is corroborated by an ANOVA [$F(4, 8) = 4.244, p = 0.039$].

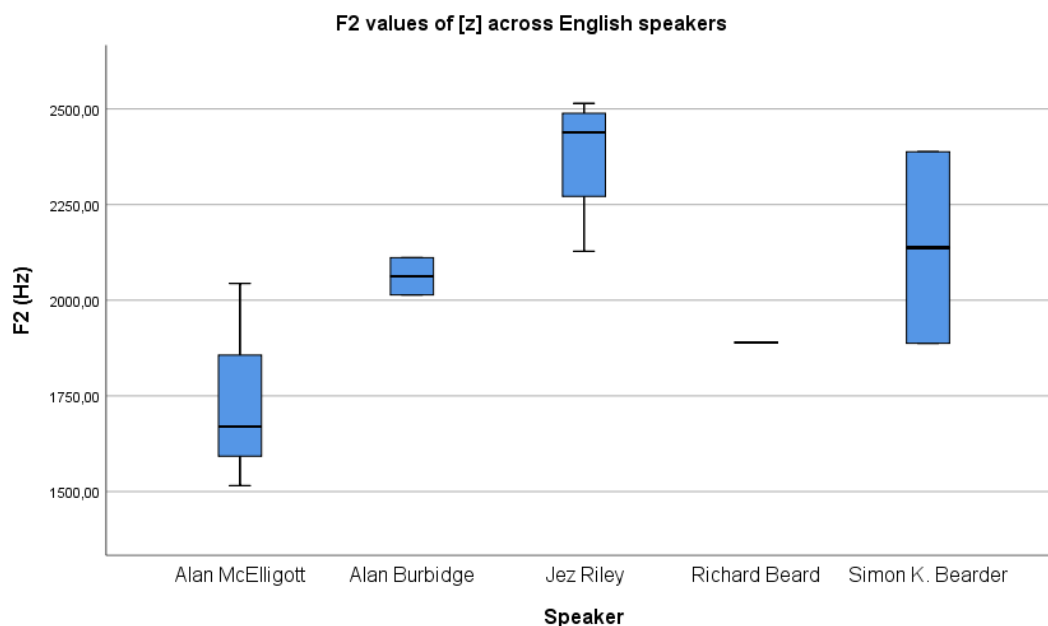


Figure 64. Between-speaker variation of F2 values in [z] across English voice samples.

In this case, a Tukey HSD post-hoc test could not be computed. The alternative, a Games-Howell test, discerns a significant difference between speaker 3 (Jez Riley) and speaker 5 (Richard Beard), as figure 64 illustrates.

Lastly, an ANOVA was run [$F(1, 28) = 7.272, p = 0.012$], which concludes that F3 values also do make a distinction between [s] and [z]. For the latter segmental unit, there are no distinguishable traits amongst speakers. However, F3's [s] appears to signal differences amongst speakers, with a significant ANOVA [$F(6, 12) = 13.066, p = 0.000$].

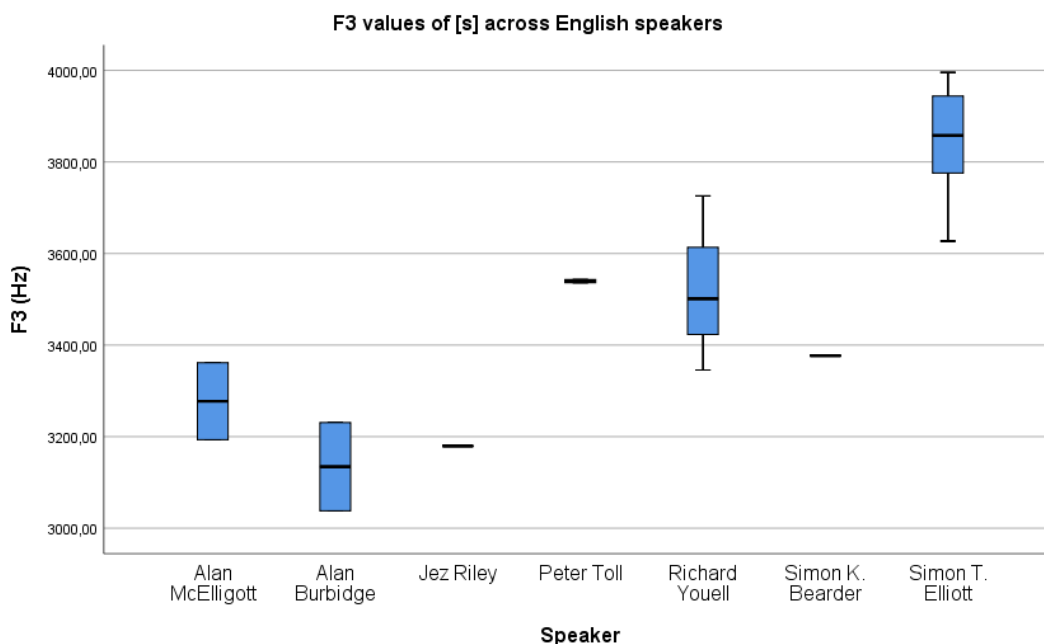


Figure 65. Between-speaker variation of F3 values in [s] across English voice samples.

Figure 65 above reflects the results obtained from a Tukey’s HSD post-hoc test, which asserts differences between speaker 8 (Simon T. Elliott) and those with lower range of values placed at the left side (Alan McElligott, Alan Burbidge, and Jez Riley). Additionally, speaker 7 (Simon K. Bearder) is also dissimilar from Simon T. Elliott, as far as F3 values in [s] realisations are concerned.

5.2.2.2. Spanish voice samples

This second sub-section shall address whether segmental units are relevant features in discriminating speakers. The first step consists of analysing the cases listed for voiced and voiceless plosives. The table below provides a summary of the main findings obtained in this domain:

[b, d, g] and [k, p, t]					
Variable	Sound differences	Test stats.		Sig. (p-value)	Significant pairwise comparisons
VOT	YES	[b, d, g]	Kruskal-Wallis	0.021	2-5
		[k, p, t]	Levene test	0.007	2-7
			Welch's test	0.043	
Release burst intensity	NO	Kruskal- Wallis		0.000	1-6 1-7 2-5 2-7 4-7 5-6 5-7

Table 119. Between-speaker variation of variables within voiced/voiceless plosives in Spanish voice samples. Statistically significant values are marked in bold ($\alpha = 0.05$).

As seen at the top of table 119, a Kruskal-Wallis test detected significant differences across sounds in terms of VOT ($H = 48.244$, $p = 0.000$, $df = 5$). The specific discrepancies are shown below:

Sample 1- Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
[b]-[k]	-41.421	9.319	-4.445	0.000	0.000
[b]-[p]	-43.350	11.221	-3.863	0.000	0.002
[b]-[t]	-44.500	8.742	-5.091	0.000	0.000
[d]-[k]	-36.046	8.820	-4.087	0.000	0.001
[d]-[p]	-37.975	10.811	-3.513	0.000	0.007
[d]-[t]	-39.125	8.208	-4.767	0.000	0.000

Table 120. Between-speaker differences amongst Spanish recordings regarding VOT in voiceless and voiced plosives according to a Kruskal-Wallis test.

From the information displayed above in table 120, it seems that both [b] and [d] are statistically different from the voiceless group of consonants [k, p, t]. The only voiced plosive not appearing in the pairwise comparison is [g], which could be absent due to scarce realisations thereof, or because of Bonferroni corrections (or both).

VOT for [b, d, g] seems significant after applying a Kruskal-Wallis test ($H= 14.964$, $p= 0.021$, $df= 6$), whose results are disclosed in table 121 below:

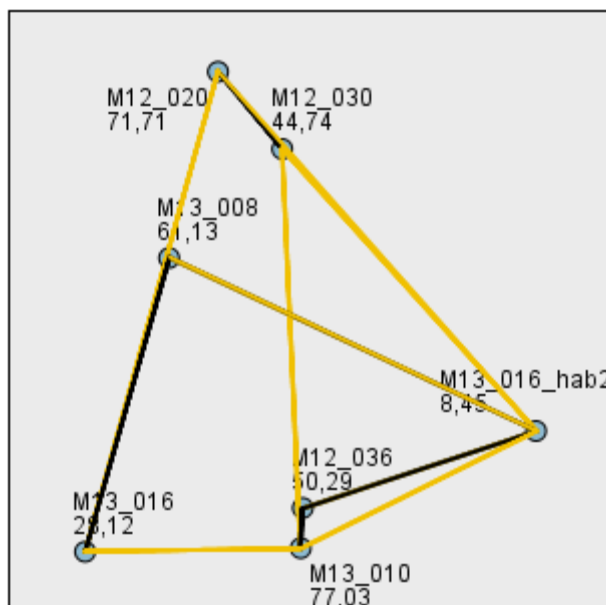
Sample 1- Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
M13_010- M12_030	16.869	5.294	3.186	0.001	0.030

Table 121. Between-speaker differences amongst Spanish recordings in voiced plosives' VOT according to a Kurskal-Wallis test.

VOT for [k, p, t] is significant, as the Welch's test indicates [$F(6, 18.464) = 2.757$, $p= 0.043$]. Since the Welch's test is successfully calculated (unlike in the English group), a Games-Howell post-hoc test discerned a significant pairwise comparison ($p= 0.035$) between speaker 2 (M12_030) and speaker 7 (M13_016_hab2).

Leaving VOT values aside, release burst intensity appears non-significant across sounds, but is significant across speakers, as a Kruskal-Wallis demonstrates ($H= 54.842$, $p= 0.000$, $df= 6$).

Pairwise Comparisons of Speaker



Each node shows the sample average rank of Speaker.

Figure 66. Pairwise comparison of release burst intensity values in voiced/voiceless plosives across Spanish recordings.

A total of seven significant pairwise comparisons highlighted in the graph depicted in figure 66. Specifically, speaker 7 (M13_016_hab2) differs from speakers 1 (M12_020), 2 (M12_030), 4 (M13_008), and 5 (M13_010). Speaker 5's (M13_010) average rank is in turn different from speaker 2's (M12_030). Also, speaker 6 (M13_016) is differentiated from speaker 1 (M12_020) and speaker 5 (M13_010).

Proceeding to the voiceless alveolar sibilant [s], The upcoming analysis shall cover only those variables which display statistically significant results, much in accordance with the results exposed in the table below:

[s]			
Variable	Test stats.	Sig. (p-value)	Significant pairwise comparisons
Spectral peak location	Kruskal- Wallis	0.275	-
COG	Levene test	0.112	-
	ANOVA	0.199	
Noise duration	Kruskal- Wallis	0.336	-
Noise amplitude	Levene test	0.428	1-2 1-7 2-4 3-4
	ANOVA	0.000	4-7 5-7 6-7
F1	Levene test	0.009	-
	Welch's test	0.019	
F2	Levene test	0.331	2-4
	ANOVA	0.001	2-6 4-5
F3	Levene test	0.398	2-4
	ANOVA	0.002	

Table 122. Between-speaker variation of variables within voiceless alveolar sibilants in Spanish voice samples. Statistically significant values are marked in bold ($\alpha = 0.05$).

As discerned in table 122, the first variable that detects significant differences amongst speakers is noise amplitude, which is verified with the calculation of an ANOVA [$F(6, 33) = 14.672, p = 0.000$].

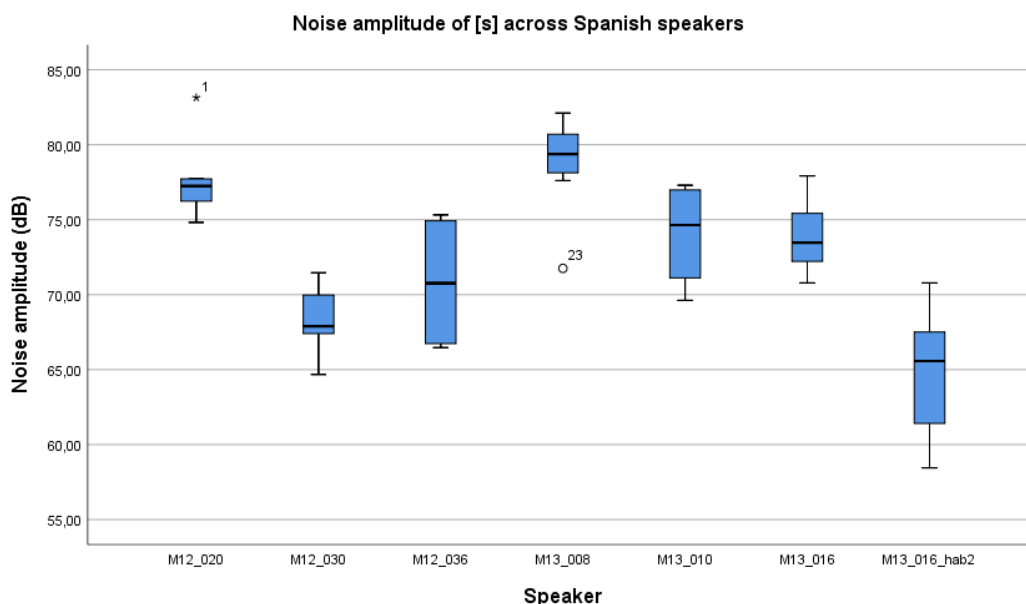


Figure 67. Between-speaker variation of [s] noise amplitude across Spanish voice samples.

A Tukey's HSD post-hoc test spotted differing noise amplitudes of [s] across the individuals appearing in figure 67 above. Speaker 4's (M13_008) high range of values contrasts sharply with those at lower thresholds like speakers 2 (M12_030), 3 (M12_036), and 7 (M13_016_hab2). The latter subject (M13_016_hab2) also displays differing values in contrast with other individuals, such as speakers 1 (M12_020), 5 (M13_010), and 6 (M13_016). As a final remark, the first speaker's (M12_020) noise amplitude is statistically different from speaker 2's (M12_030).

F1 Welch's test appears significant [$F(6, 12.600) = 3.936, p = 0.019$]. However, pairwise comparisons do not reach statistically significant differences. However, an ANOVA did find differences across speakers' [s] F2 [$F(6, 33) = 5.193, p = 0.001$].

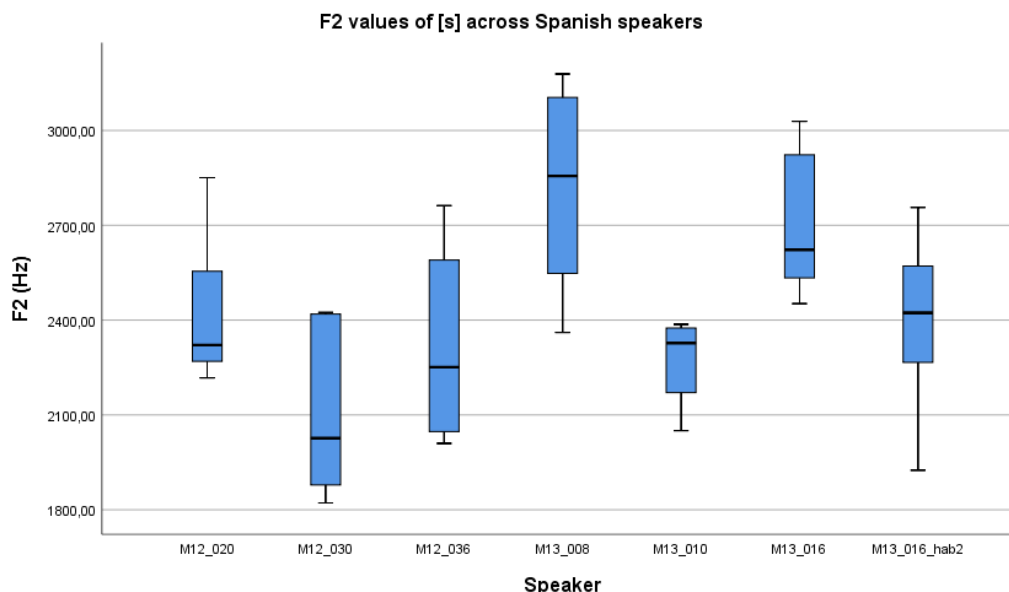


Figure 68. Between-speaker variation of F2 values in [s] across Spanish voice samples.

As figure 68 reveals, a Tukey’s HSD post-hoc test considers the second speaker’s (M12_030) F2 values as statistically different from the fourth (M13_008) and sixth speakers (M13_016). Also, speaker 5’s (M13_010) values are significantly lower than those encountered in speaker 4 (M13_008).

Lastly, a high p-value in Levene test allows for an ANOVA to be computed on [s] F3 values, which is proven to be significantly different across speakers [$F(6, 33) = 4.332, p = 0.002$]. After consulting a Tukey’s HSD post-hoc test, it turns out that only one pairwise comparison reaches significance levels ($p < 0.05$), namely that of speaker 2 (M12_030) and speaker 4 (M13_008).

5.2.2.3. Dutch voice samples

This final sub-section measures whether the variance exhibited by the segmental phenomena studied is statistically different across Dutch speakers (independent variable). The table below breaks down the variables found in voiced and voiceless plosives, along with the statistical test employed, the resulting p-value, and significant pairwise comparisons (if any):

[b, d, g] and [k, p, t]					
Variable	Sound differences	Test stats.		Sig. (p-value)	Significant pairwise comparisons
VOT	YES	[b, d, g]	Kruskal-Wallis	0.044	1-7
		[k, p, t]	Kruskal-Wallis	0.018	-
Release burst intensity	NO	Levene test		0.092	1-2 1-5 1-6 1-7 1-8 2-3 2-7 3-4 3-5
		ANOVA		0.000	3-6 3-7 3-8 4-7 4-8 5-7 6-7 7-8

Table 123. Between-speaker variation of variables within voiced/voiceless plosives in Dutch voice samples. Statistically significant values are marked in bold ($\alpha = 0.05$).

Just as in the two previous groups of informants, VOT values in table 123 can be divided into two groups depending on their categorisation: voiced and voiceless plosives. A Kruskal-Wallis test was run to attest this assertion, which yielded significant results ($H=84.312$, $p=0.000$, $df=4$).

Sample 1- Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
[b]-[k]	-78.361	17.165	-4.565	0.000	0.000
[b]-[p]	-85.551	22.802	-3.752	0.000	0.002
[b]-[t]	-88.122	16.100	-5.473	0.000	0.000
[d]-[k]	-74.175	12.297	-6.032	0.000	0.000
[d]-[p]	-81.365	19.403	-4.193	0.000	0.000
[d]-[t]	-83.936	10.760	-7.801	0.000	0.000

Table 124. Between-speaker differences in voiced/voiceless plosives' VOT values amongst Dutch recordings according to a Kruskal-Wallis test.

Expectedly, table 124 shows that the main differences arise between voiced and voiceless plosives, namely the group [b]-[k, p, t], and [d]-[k, p, t]. In the case of [g], it is either underrepresented in the samples of choice or corrected through the adjustment of p-values through Bonferroni corrections.

Voiced plosives [b, d, g] values scored in VOT were proven statistically different through conducting a Kruskal-Wallis test ($H= 14.407$, $p= 0.044$, $df= 7$). However, only one pairwise comparison reached the established level of significance ($p < 0.05$). This distinction refers to speaker 1 (DVA8.F20K) and speaker 7 (DVA11.F28Q).

As for their voiceless counterparts, [k, p, t] VOT values are statistically different across speakers, as a Kruskal-Wallis asserts ($H= 16.956$, $p= 0.018$, $df= 7$). Even so, no significant pairwise comparisons were drawn due to the adjustment of thresholds related to significance (p-values) by means of applying Bonferroni corrections.

Moving to release burst intensity, types of sound were not deemed statistically different from each other. As a result both [b, d, g] and [k, p, t] were put together in the model. As for the differences amongst speakers, an ANOVA yielded significant results [$F(7, 201) = 18.356$, $p= 0.000$].

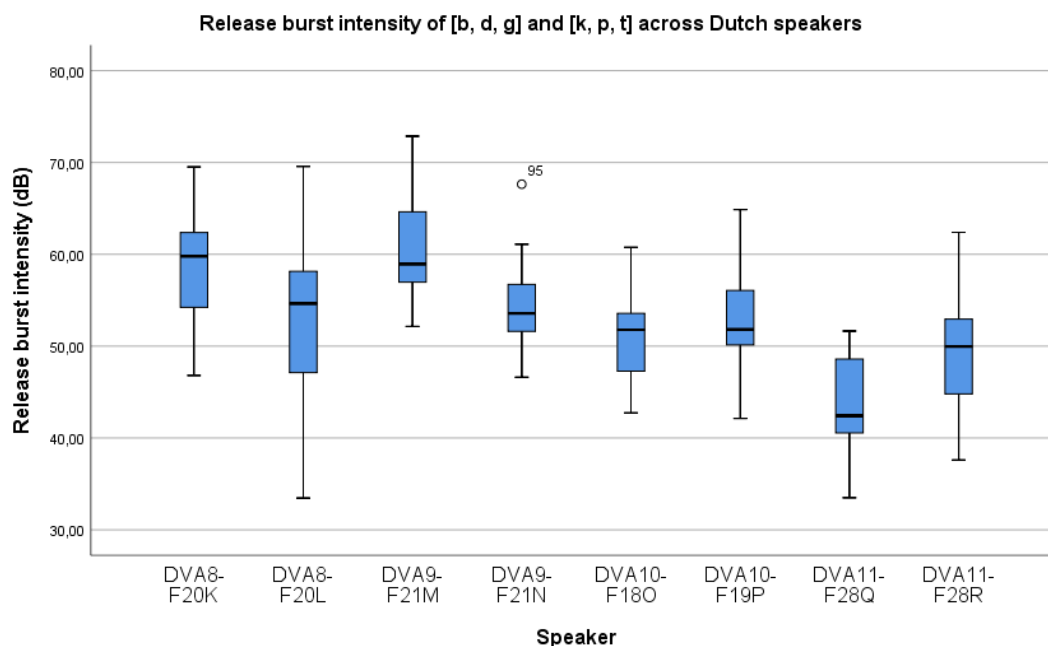


Figure 69. Between-speaker variation of release burst intensity in [b, d, g] and [k, p, t] across Dutch recordings.

By looking at figure 69 above, it can be noted that, according to a Tukey’s HSD post-hoc test, release burst intensity differs between the first speaker (DVA8-20K) and speakers 2 (DVA8-F20L), 5 (DVA10-F18O), 6 (DVA10-F19P), 7 (DVA11-F28Q), and 8 (DVA11-F28R). Also, the second speaker (DVA8-F20L) is seen as statistically different from speakers 3 (DVA9-F21M) and 7(DVA11-F28Q). Besides that, speaker 3’s (DVA9-F21M) values are different from those speakers located on the right side of the plot, namely from speaker 4 (DVA9-F21N) to speaker 8 (DVA11-F28R). Speaker 8 is in turn markedly different from speakers 3 (DVA9-F21M), 4 (DVA9-F21N), and 7 (DVA11-F28Q). The last group of pairwise comparisons also point at speaker 7 as a subject whose release burst intensity values are significantly different from speakers 4 (DVA9-F21N), 5 (DVA10-F18O), and 6 (DVA10-F19P).

Moving to the voiceless alveolar sibilant [s], the following table illustrates each variable’s significance across speakers.

[s]			
Variable	Test stats.	Sig. (p-value)	Significant pairwise comparisons
Spectral peak location	Kruskal- Wallis	0.095	-
COG	Kruskal- Wallis	0.206	-
Noise duration	Kruskal- Wallis	0.016	1-5
Noise amplitude	Levene test	0.002	1-7 2-3 2-7 3-4 3-5
	Welch's test	0.000	3-6 3-7 4-7 5-7
F1	Levene test	0.123	1-2
	ANOVA	0.001	1-4 1-5
F2	Kruskal- Wallis	0.072	-
F3	Levene test	0.360	1-2
	ANOVA	0.194	

Table 125. Between-speaker variation of variables within voiceless alveolar sibilants in Dutch voice samples. Statistically significant values are marked in bold ($\alpha = 0.05$).

As table 125 shows above, noise duration on [s] does make significant distinctions across speakers, according to a Kruskal-Wallis test ($H= 17.173$, $p= 0.016$, $df= 7$).

Sample 1- Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
DVA8.F20K- DVA10.F18O	26.238	7.897	3.323	0.001	0.032

Table 126. Between-speaker variation of noise duration in [s] across Dutch voice samples according to a Kruskal-Wallis test.

As disclosed in table 126, only one pairwise comparison resulted significant amongst all the possible combinations. It should be reminded, that each row in the table tests the null hypothesis (Sample 1 and Sample 2 distributions are the same). In this fashion, asymptotic significances (2-sided tests) are shown at the 0.05 level of significance. Consequently, p-values have been adjusted by Bonferroni corrections applied for multiple tests.

The next significant segmental variable is noise amplitude, whose values have been deemed as statistically different with a Welch's test [$F(7, 14.651) = 26.142, p = 0.000$].

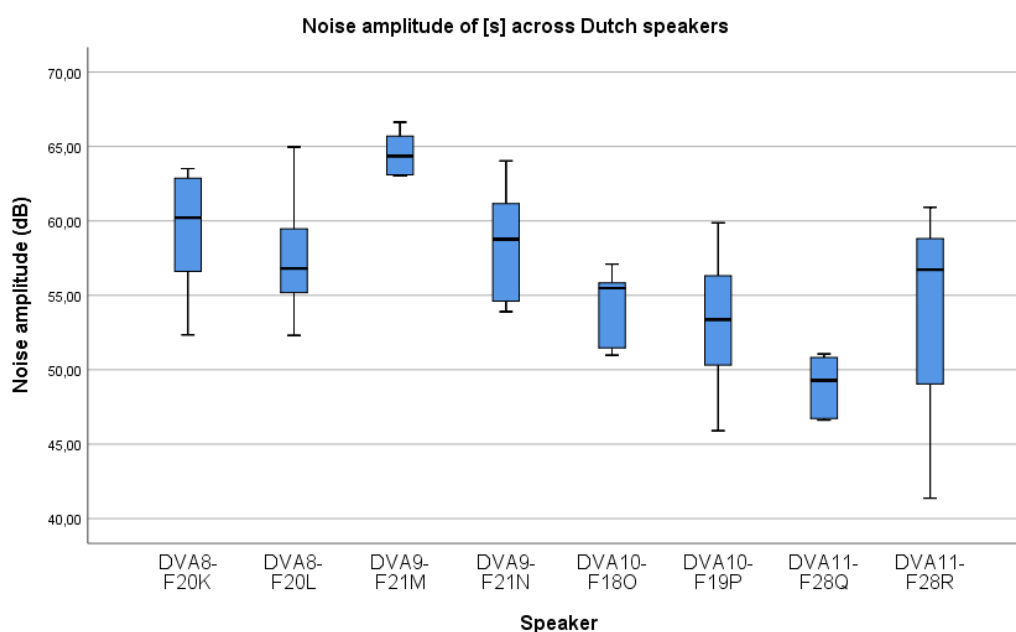


Figure 70. Between-speaker variation of noise amplitude in [s] across Dutch recordings.

After running a Games-Howell post-hoc test, significant differences emerge between speaker 7 (DVA11-F28Q) and speakers 1 (DVA8-F20K), 2 (DVA8-F20L), 3 (DVA9-F21M), 4 (DVA9-F21N), and 5 (DVA10-F18O). As the boxplots drawn in figure 70 above illustrate, a second distinction can be made between speaker 3 (DVA9-F21M) and speakers 2 (DVA8-F20L), 4 (DVA9-F21N), 5 (DVA10-F18O), and 6 (DVA10-F19P).

Another segmental unit which render statistically significant results is F1, whose ANOVA revealed significant differences across speakers [$F(7, 41) = 4.441, p = 0.001$].

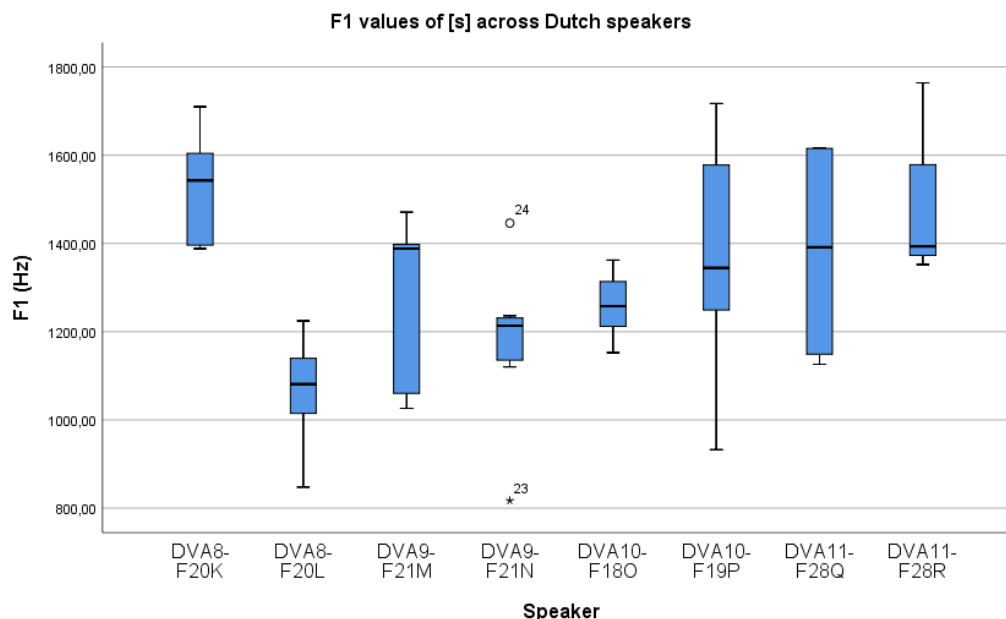


Figure 71. Between-speaker variation of F1 values in [s] across Dutch recordings.

In this scenario, figure 71 displays a rather balanced distribution of values amongst all participants except for the first one (DVA8-F20K). As a matter of fact, the found dissimilarities by a Tukey HSD post-hoc test are focused between the first speaker and those with a lower range of values such as speakers 2 (DVA8-F20L), 4 (DVA9-F21N), and 5 (DVA10-F18O).

Even if the ANOVA undertaken for F3 values was not significant [$F(7, 41) = 1.502, p = 0.194$], a Games-Howell post-hoc test detected one significant difference between the speakers DVA8-F20K and DVA8-F20L ($p = 0.050$), which is illustrated hereby:

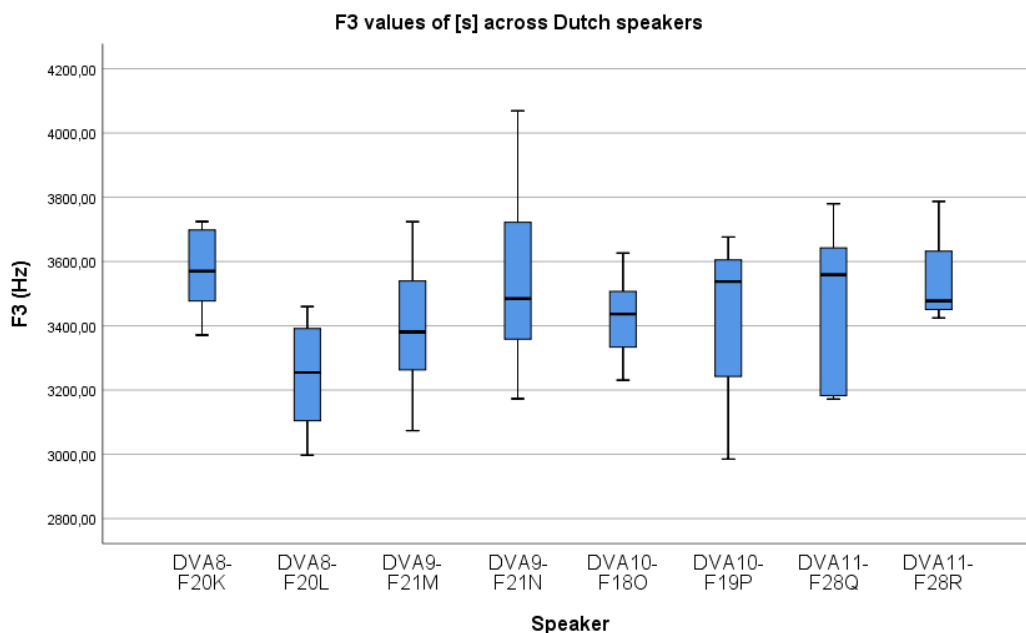


Figure 72. Between-speaker variation of F3 values in [s] across Dutch recordings.

Just like figure 72 illustrates, most of the speakers in the graph share a similar interquartile range, with the exception of the first two speakers. Incidentally, it should be noted that, even when ANOVAs display negative results, a post-hoc test is still required so as not to miss any possible significant pairwise comparison which might otherwise be overlooked.

5.2.3. Summary of results

Before commenting on the results stemming from this analysis, let us formulate hypothesis 8 again: Interspeaker variability of the foil speakers' voice samples with similar intonation patterns (rising intonation) and uncontrolled segmental phenomena is statistically significant. In the suprasegmental domain, it has been proven that every single variable has registered more than one case of discrepancies between speakers. However, it is also noticeable that some speakers report 0 cases of dissimilarities in specific cells (variables). In the light of such findings, it has been decided to choose the most efficient suprasegmental features based on two criteria: those which report a higher number of cases of dissimilarities (criterion 1), and those which register at least one case across every single speaker within each group of informants (criterion 2). The ensuing findings can be consulted in table 127 below:

Criterion 1			Criterion 2		
English	Spanish	Dutch	English	Spanish	Dutch
Max. intensity	Min. intensity	Mean pitch	Min. intensity	Mean pitch	Max. intensity
Min. intensity	Speech rate	Max. intensity	DurPaus	25% pitch	DurPaus
DurPaus	N_paus	Mean intensity	Pause_%	50% pitch	Pause_%
N_paus/min		DurPaus	N_paus	75% pitch	N_paus
Pause_%		Pause_%	Articulation rate	Speech rate	Speech rate
Speech rate			ASD	N_paus	Articulation rate
Articulation rate					ASD
ASD					

Table 127. Efficient suprasegmental features across groups of informants and research criteria (between-speaker variation).

It is interesting to note that, despite showing no standardised parameters across the selected groups of informants (with the exception of *N_paus* being listed simultaneously in every group inside the criterion 2), some patterns can still be appreciated within each of them according to the criteria established: English (min. intensity, DurPaus, Pause_%, Articulation rate, and ASD), Spanish (speech rate and *N_paus*), and Dutch (max. intensity, DurPaus, and Pause_%) group of informants.

Unlike suprasegmental phenomena, each group of voice samples has not registered significant pairwise comparisons through all the contemplated segmental units of analysis. The following table covers all the significant variables found across the three groups of voice samples:

English informants			Spanish informants			Dutch informants		
Variable	N° Comp.	Crit. 2	Variable	N° Comp.	Crit. 2	Variable	N° Comp.	Crit. 2
[k, p, t] Release burst int.	8	NO	[b, d, g] VOT	1	NO	[b, d, g] VOT	1	NO
[s] & [z] Spectral peak location	1	NO	[k, p, t] VOT	1	NO	[b, d, g] & [k, p, t] Release burst int.	17	YES
[s] COG	9	YES	[b, d, g] & [k, p, t] Release burst int.	7	NO	[s] Noise duration	1	NO
[s] & [z] Noise amplitude	7	NO	[s] Noise amplitude	7	YES	[s] Noise amplitude	9	NO
[s] F1	9	YES	[s] F2	3	NO	[s] F1	3	NO
[s] F2	2	NO	[s] F3	1	NO	[s] F3	1	NO
[z] F2	1	NO						
[s] F3	4	NO						

Table 128. Efficient segmental variables, number of significant pairwise comparisons, and compliance with research criterion n° 2 (between-speaker variation).

According to criterion 1 (variables which register higher numbers of dissimilarities), it seems that release burst intensity (both in voiceless plosives only and in voiced/voiceless plosives together) and noise amplitude ([s] and [z] for the English group, and [s] for the Spanish/Dutch recordings) are consistent regardless of the group of informants examined. As table 128 suggests, the third column on the right side of each group reveals whether the second criterion (the variable should signal at least 1 dissimilar case through every subject) is accomplished or not. Differences arise in variables such as the voiceless alveolar sibilant's COG and F1 (English informants), its noise amplitude (Spanish informants), and the release burst intensity of voiced/voiceless plosives (Dutch informants). Although, as noticed, an agreement is not reached amongst the three types of recordings.

As a limitation of the present analysis, it should be reminded that a few variables may have absent values in certain subjects, such as Richard Beard's case, whose excerpt does not include any realisation of the [s] sound (hence the addition of its voiced counterpart [z]). In this sense, complying with criterion 2 entails gathering dissimilar cases according to the subjects who are considered for said variable, thus disregarding those who are not listed due to their lack of data points.

In the light of the above considerations, the eighth hypothesis can be accepted both in segmental and suprasegmental features, given that every single variable has encountered at least one significant distinction amongst speakers. The most crucial difference is that suprasegmental variables are proven as significant throughout each and every group of informants, whereas segmental variables' efficiency is distributed amongst specific groups. Nevertheless, a further distinction could be made according to the researcher's notion of *efficiency*: either related to high performance (criterion 1) or accuracy (criterion 2). When employing said criteria as a method to standardise the results, it seems that the number of pauses per extract (N_paus) remains as a suprasegmental feature which is shared by every group of informants within the second criterion, whereas segmental measures such as *release burst intensity* and *noise amplitude* comply with the first criterion, irrespective of the voice sample's language.

5.3. ACOUSTIC-PHONETIC ANALYSIS OR JURORS' VERDICT?

This last section tackles hypothesis 9 with a contrastive study which draws the contributions from both the previous perception survey-based analysis (chapter 4) and the current acoustic-phonetic analysis (chapter 5) to find out whether the researcher's inspection on voice samples prevails over the immediate juror's judgement. Since successful identifications were already explored throughout chapter 4.1 (*Language familiarity*), this hypothesis is conceived as a follow-up investigation on acoustic-phonetic parameters that may explain the results obtained in 4.2. (*Discrimination or identification?*), where discrimination tasks hinted at a better performance on the juror's side, in contrast with identification tests.

Chapter 4's analysis spotted successful discriminations whenever a juror chose a speaker other than the suspect to identify, but did not make distinctions about the other constituents of the voice-line up, nor about the likelihood of them being wrongly selected as the culprit (due to shared similarities with the suspect's recording). In this regard, this sub-section seeks to determine whether the hearer's accuracy on discriminating speakers is perceptually accurate according to the acoustic properties of the voices examined. This analysis is mainly focused on discrimination tasks, and so results from identification tests are quoted at the summary of this analysis for the sake of comparison, since such findings can be easily inferred by looking at success rates, which equate to the percentages of speakers who rightfully chose one specific speaker over the others. With that in mind, let us formulate the ninth hypothesis along the null hypothesis:

- H_0 : Foreign and native speaker recognition using acoustic-phonetic analysis is not more accurate than the lay listener's (jurors) judgement.
- H_9 : Foreign and native speaker recognition using acoustic-phonetic analysis is more accurate than the lay listener's (jurors) judgement.

To address this matter, perception surveys' answers to the question (*in your opinion, whose voice differs most from the suspect's?*) are consulted so as to find out whether jurors' choices were perceptually accurate. In this sense, the informant (voice sample) who reports the highest number of dissimilarities in all the studied significant variables (both segmental and suprasegmental features) shall be conceived as the most accurate choice (as this would, in theory, point at the most distinguishable subject in the voice line-up). A further observation to the formulated hypothesis is that jurors' responses are expected to be less accurate in the second experimental condition (with noise disturbances) than in the first condition (no background noises).

In order to achieve this, a table is compiled to illustrate the total amount of times a specific speaker has been deemed as statistically different (both at the segmental and suprasegmental level) from another subject, as well as including specific dissimilarities amongst all the other remaining foil speakers. Once an overall number is drawn per speaker, a percentage shall be calculated, and thus a hierarchy of dissimilarities is established. Additionally, a second table shall incorporate the suspects' dissimilarities in relation to the rest of voice recordings (including their own excerpts used as distractors).

The next step juxtaposes these results with the ones obtained from the perception surveys-based analysis, whose numbers shall be converted to percentages as well for the sake of comparison²¹.

It should be reminded that one of the particularities of this research is the intentional enhancement of some recordings. With this adjustment, a portion of the audio material exposed to jurors is enhanced with the purpose of placing each audio file on equal footing in perceptual terms. In this regard, the increased volume of audio files renders them audible, which otherwise would be imperceptible if left unchanged (especially during the second perception tests which include background noises). The speakers whose voice samples have been modified through Camtasia Studio are the following: Richard Beard (English), Peter Toll (English), Jez Riley French (English), Alan Burbidge (English), and Simon K. Bearder (English), and M12_016_hab2 (Spanish). Note that no modifications were required in the Dutch group of informants. This analysis is structured according to the set of recordings employed in voice line-ups, namely English (5.3.1.), Spanish (5.3.2.), and Dutch (5.3.3.) voice samples.

5.3.1. English voice samples

To begin with, a first glimpse at the data unveils the relationships of dissimilarities found amongst specific English speakers, while also calculating the total number of segmental and suprasegmental features which make certain informants stand out from the rest.

²¹ Please note that the upcoming graphs do not display voice line-up's speakers in alphabetical/numerical order. Instead, the original arrangement used in perception surveys is followed (see *Appendix 4*).

Chapter 5- Results: Acoustic-phonetic analysis

	Alan McElligott	Alan Burbidge	Jez Riley	Peter Toll	Richard Beard	Richard Youell	Simon K. Bearder	Simon T. Elliott
Alan McElligott	-	7	0	9	2	1	7	5
Alan Burbidge	7	-	7	9	4	5	5	7
Jez Riley	0	7	-	2	2	5	4	5
Peter Toll	9	9	2	-	3	7	9	11
Richard Beard	2	4	2	3	-	5	2	4
Richard Youell	1	5	5	7	5	-	5	4
Simon K. Bearder	7	5	4	9	2	5	-	0
Simon T. Elliott	5	7	5	11	4	4	0	-
Total	31	44	25	50	22	32	32	36
Total %	11.40%	16.18%	9.19%	18.38%	8.09%	11.76%	11.76%	13.24%

Table 129. Number of dissimilarities found amongst English foil speakers (and total number of differences per speaker) in relation to segmental and suprasegmental features.

As discerned in table 129 above, the most dissimilar recordings encountered amongst foil speakers refer to Peter Toll (n= 50, 18.38%) and Alan Burbidge (n=44, 16.18%), respectively.

When comparing foil speakers to the appointed English suspect (Simon T. Elliott), the findings remain nearly unchanged:

	Alan McElligott	Alan Burbidge	Jez Riley	Peter Toll	Richard Beard	Richard Youell	Simon K. Bearder	Simon T. Elliott
SUSPECT Simon T. Elliott	4	11	2	7	3	3	1	0
Total %	12.90%	35.48%	6.45%	22.58%	9.68%	9.68%	3.23%	0%

Table 130. Number of dissimilarities found between English foil speakers and the selected suspect in relation to segmental and suprasegmental features.

As can be observed from table 130, the most dissimilar speakers in relation to the suspect are still Alan Burbidge and Peter Toll. In this case, however, their positions are reversed, and thus Burbidge (n= 11, 35.48%) displays more differentiated acoustic-phonetic features than Peter Toll (n= 7, 22.58%). It is worth noting that no significant differences were found between the suspect’s recording and his voice sample used in the line-up (0%). For more details on pairwise comparisons between the English suspect and foil speakers at the segmental and suprasegmental level, see *Appendix 10*.

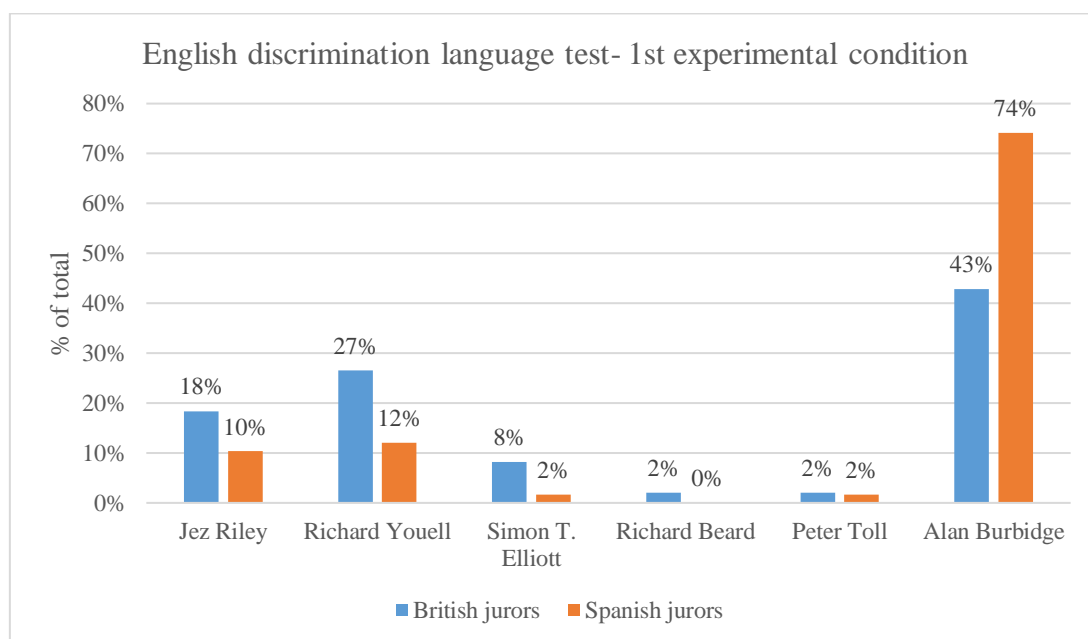


Figure 73. English discrimination language test’s responses in British and Spanish jurors (target-present condition without background noises).

As for the results gathered from the perception surveys themselves, figure 73 above discloses the most voted speakers across British and Spanish jurors in the first

experimental condition (no background noises). As noted in the bar graphs above, Alan Burbidge is perceived as the most differentiated speaker from the suspect by both British (43%) and Spanish (74%) participants, which matches the results discussed above. Expectedly, a low error rate emerged in British (8%) and Spanish (2%) participants, who wrongly pointed at Simon T. Elliott’s recording as being different from the suspect.

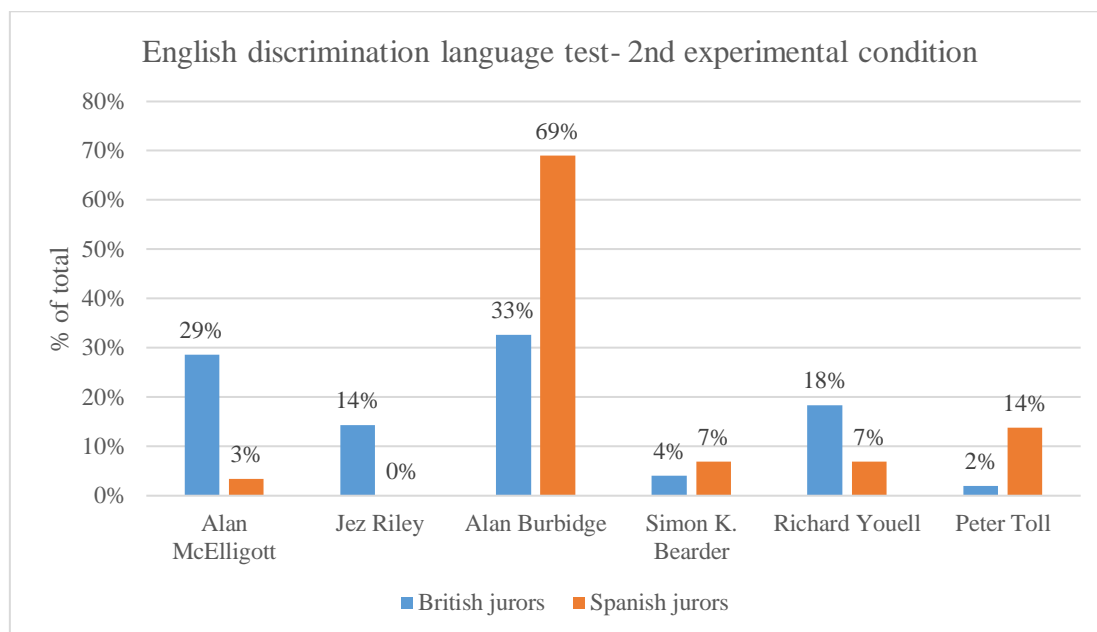


Figure 74. English discrimination language test’s responses in British and Spanish jurors (target-absent condition with background noises).

When noise disturbances are added (second experimental condition), jurors’ responses still refer to Burbidge as the most dissimilar speaker in relation to the suspect’s speech, as noted in figure 74. It must be noted that this tendency applies across groups of jurors and experimental conditions, only that percentages are slightly reduced during noisy conditions: British respondents’ percentage is reduced to 33%, while a similar reduction is applied to Spanish hearers (69%).

5.3.2. Spanish voice samples

Moving to Spanish recordings, an initial overview is seen on the significant pairwise comparisons on the basis of differing segmental and suprasegmental features. The results can be consulted in table 131 below:

	M12_020	M12_030	M12_036	M13_008	M13_010	M13_016	M13_016_hab2
M12_020	-	3	7	1	0	4	7
M12_030	3	-	6	7	3	2	7
M12_036	7	6	-	6	6	6	6
M13_008	1	7	6	-	3	4	4
M13_010	0	3	6	3	-	5	9
M13_016	4	2	6	4	5	-	2
M13_016_hab2	7	7	6	4	9	2	-
Total	22	28	37	25	26	23	35
Total %	11.22%	14.29%	18.88%	12.75%	13.27%	11.73%	17.86%

Table 131. Number of dissimilarities found amongst Spanish foil speakers (and total number of differences per speaker) in relation to segmental and suprasegmental features.

Similar to English voice samples, the percentages found amongst the selected speakers are somewhat balanced. In this particular case, the speakers M12_036 (18.88%) and M13_016_hab2 (17.86%) are perceived as the ones carrying more distinguishable features in their speech, possibly increasing the odds for them to be chosen as the most dissimilar voice in the group of distractors.

	M12_020	M12_030	M12_036	M13_008	M13_010	M13_016	M13_016_hab2
SUSPECT M12_020	6	7	4	4	7	3	6
Total %	16.21%	18.92%	10.81%	10.81%	18.92%	8.12%	16.21%

Table 132. Number of dissimilarities found between Spanish foil speakers and the selected suspect in relation to segmental and suprasegmental features.

As for the specific scenario where distractors' voices are compared with the suspect's, the distribution of percentages vary amongst foil speakers, as shown in table 132. Unlike in the general overview of foil speakers, M12_030 and M13_010 are perceived significantly different (18.92%) from the suspect, while M12_020 and M13_016_hab2 follow shortly after (16.21%). The differences emerging between the Spanish suspect and her voice as a distractor refer to the second highest percentage shown in table 132 above (16.21%). For more information, consult *Appendix 11* for a detailed view on significant

pairwise comparisons between the Spanish suspect and foil speakers at the segmental and suprasegmental level.

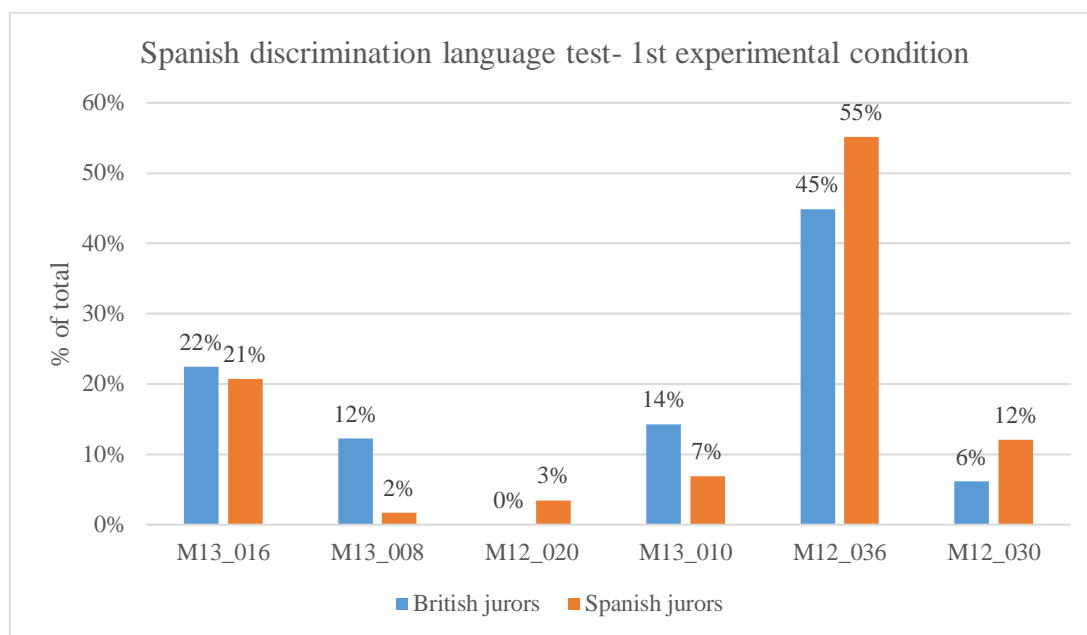


Figure 75. Spanish discrimination language test’s responses in British and Spanish jurors (target-present condition without background noises).

Regarding the first experimental condition, figure 75 illustrates that a great percentage of jurors perceive M12_036 as the most differentiated speaker from the suspect, both in the British (45%) and the Spanish (55%) group. According to the table on multiple comparisons with the suspect (table 132), said speaker would be distinguishable enough only at an intermediate point (10.81%). In the only-distractors pairwise comparison, however, M12_036 is indeed the most outstanding sample in the group (18.88%). On a side note, M12_020 (the suspect’s voice sample used as a distractor) is wrongly assumed to be the culprit only by a few Spanish respondents (3%).

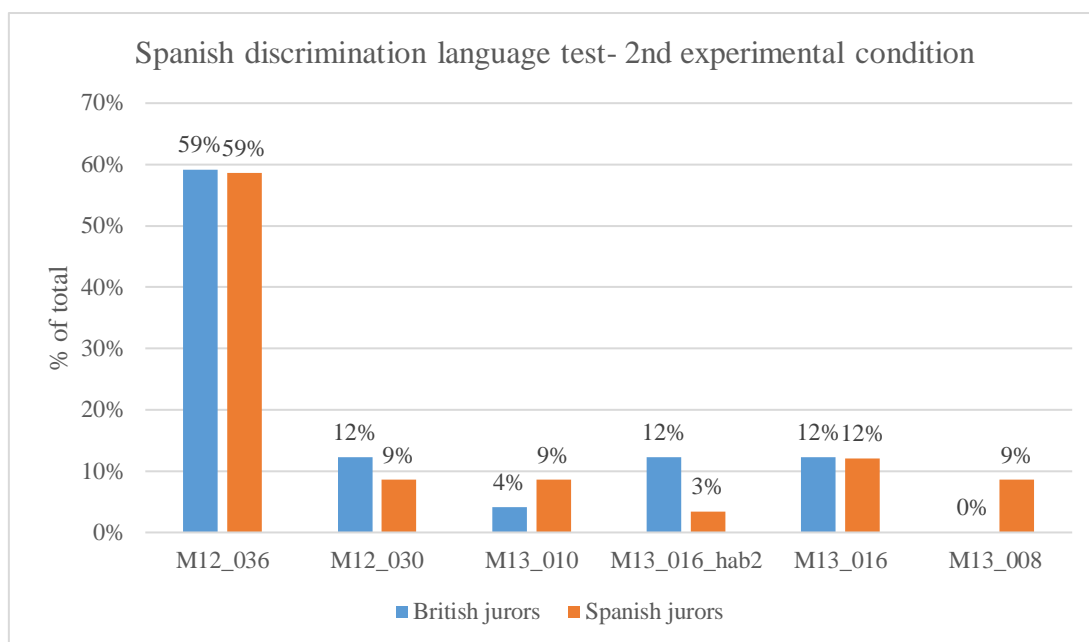


Figure 76. Spanish discrimination language test’s responses in British and Spanish jurors (target-absent condition with background noises).

Following the results in the target-present language test (1st condition), it seems that M12_036 keeps being the most differentiated speaker from the suspect in figure 76. Not only this, but a higher percentage of the surveyed population is more certain of this assertion (59%), regardless of the nationality.

5.3.3. Dutch voice samples

Lastly, Dutch informants’ voice samples are inspected so as to find significant differences amongst foil speakers in terms of segmental and suprasegmental phenomena. The resulting table can be consulted below:

	DVA8-F20K	DVA8-F20L	DVA9-F21M	DVA9-F21N	DVA10-F18O	DVA10-F19P	DVA11-F28Q	DVA11-F28R
DVA8-F20K	-	4	2	5	14	3	5	2
DVA8-F20L	4	-	4	5	10	5	3	1
DVA9-F21M	2	4	-	7	12	2	5	3
DVA9-F21N	5	5	7	-	6	0	5	3
DVA10-F18O	14	10	12	6	-	6	7	5
DVA10-F19P	3	5	2	0	6	-	4	3
DVA11-F28Q	5	3	5	5	7	4	-	1
DVA11-F28R	2	1	3	3	5	3	1	-
Total	35	32	35	31	60	23	30	18
Total %	13.26%	12.12%	13.26%	11.74%	22.73%	8.71%	11.36%	6.82%

Table 133. Number of dissimilarities found amongst Dutch foil speakers (and total number of differences per speaker) in relation to segmental and suprasegmental features.

By looking at table 133 above, it can be surmised that DVA10-F18O exhibits the highest percentage (22.73%) of significant differences shown amongst foil speakers.

	DVA8-F20K	DVA8-F20L	DVA9-F21M	DVA9-F21N	DVA10-F18O	DVA10-F19P	DVA11-F28Q	DVA11-F28R
SUSPECT								
DVA8-F20L	4	4	3	1	7	2	4	3
Total %	14.29%	14.29%	10.71%	3.57%	25%	7.14%	14.29%	10.71%

Table 134. Number of dissimilarities found between Dutch foil speakers and the selected suspect in relation to segmental and suprasegmental features.

On the other hand, an intermediate degree of intra-speaker variation seems to ascribe DVA8-F20L's recording as the most dissimilar from her voice sample used as the suspect (14.29%). Leaving aside this false alarm, the most notable speaker appearing in table 134

is DVA10-F18O (25%), and thus matching the results shown in the comparison amongst foil speakers. As a reminder, the table above gathers all the instances in which the suspect's speech was detected as statistically different from any of the remaining voice samples appearing at the voice line-up. For the sake of brevity, this sub-section reveals only the total number of significant cases obtained through statistical tests, but does not specify every correlation found in terms of pairwise comparisons. Instead, such details can be consulted in *Appendix 12*.

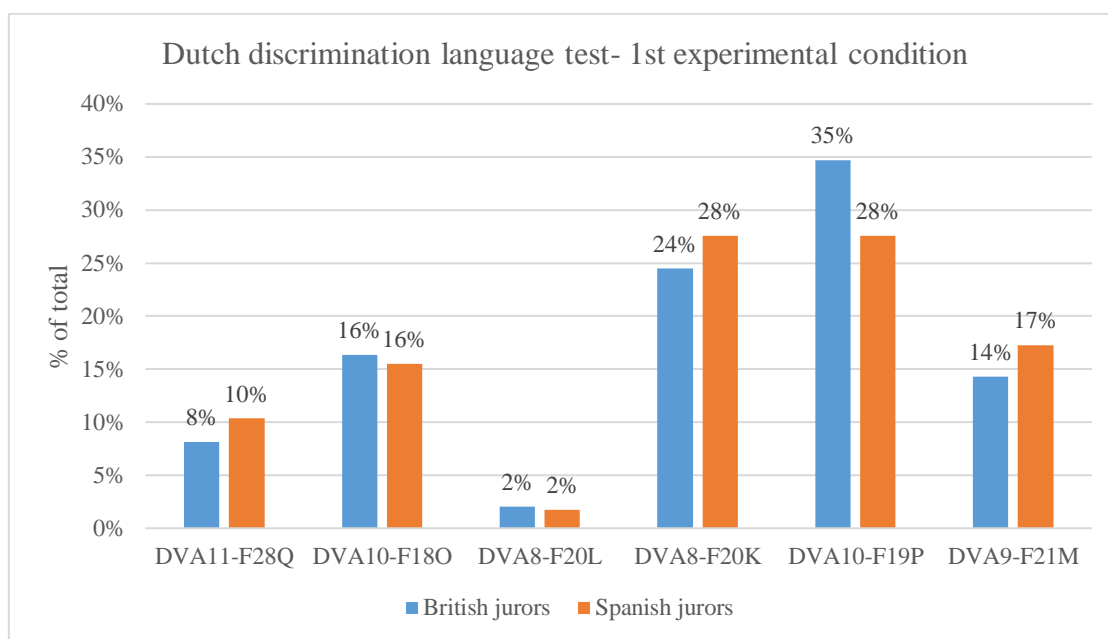


Figure 77. Dutch discrimination language test's responses in British and Spanish jurors (target-present condition without background noises).

Contrary to English and Spanish language tests, the percentages shown in figure 77 appear more evenly distributed across the available speakers. The ones that gather the highest percentages in the first experimental condition in the Dutch language test are DVA8-F20K and DVA10-F19P, the former being a right choice (second highest percentage, 14.29%) based on the multiple pairwise comparison with the suspect's voice sample. Also, the false alarms appearing through analysing segmental and suprasegmental features follow the trends exhibited in English and Spanish discrimination language tests in a similar vein, as proved by their low percentage shown in both groups of jurors (2%).

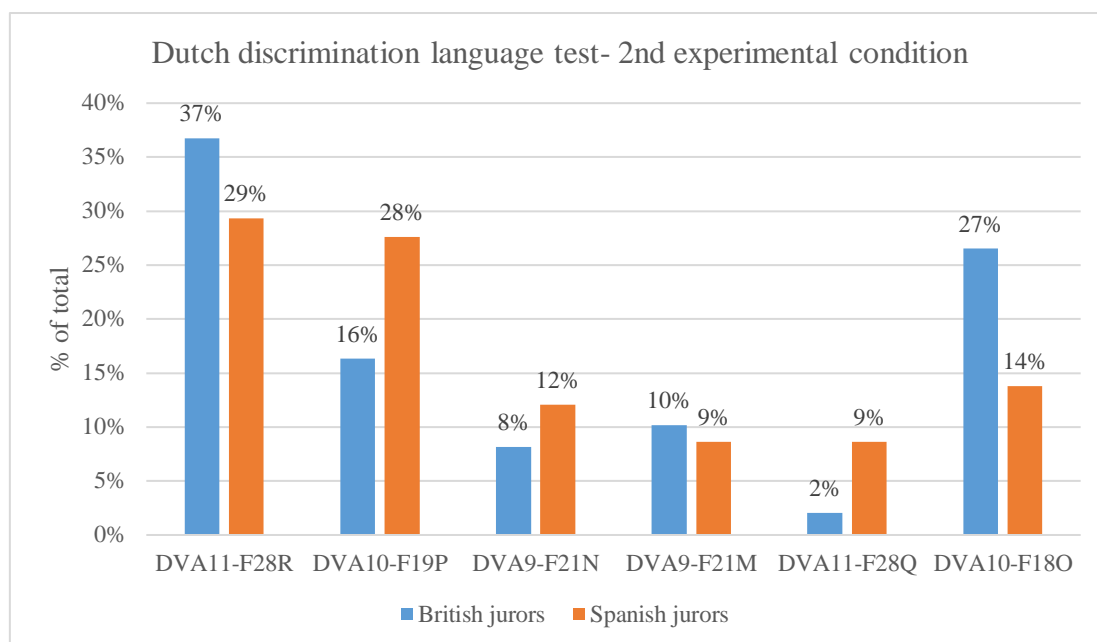


Figure 78. Dutch discrimination language test's responses in British and Spanish jurors (target-absent condition with background noises).

A similar scenario is repeated during the second experimental condition in the Dutch language test, whose distribution of responses appear more balanced than those shown in English and Spanish perception tasks, as figure 78 illustrates. It is only in this particular case that the groups of jurors disagree on a couple of foil speakers. British participants (27%) identify correctly the speaker DVA10-F18O as a dissimilar voice to the suspect's, whereas Spanish respondents focus on DVA10-F19P (28%), who scores the second lowest score on dissimilarities (in relation to the suspect's voice sample). On a side note, a unified inaccurate verdict is provided by both groups when voting DVA11-F28R with the highest percentage of hearers in the British (37%) and the Spanish (29%) group.

5.3.4. Summary of results

Before answering the ninth hypothesis, this summary refers back to point 4.2. (*Discrimination or identification?*) to determine how/if discrimination tests yield better results overall than identification tasks, and to what extent. Unlike in the point already mentioned, discrimination tests' accuracy in this section account for success rates pointing at specific individuals whose speech differs significantly from the suspect's by means of segmental and suprasegmental phenomena. The following table establishes a comparison between identification and discrimination tests' success rates:

		Voice samples					
		English		Spanish		Dutch	
		1 st . cond.	2 nd cond.	1 st cond.	2 nd cond.	1 st cond.	2 nd cond.
Identification	British jurors	42.86%	22.45%	51.02%	30.61%	42.86%	10.20%
	Spanish jurors	31.03%	20.69%	53.45%	25.86%	53.45%	24.14%
Discrimination	British jurors	43%	33%	20%	16%	16%	27%
	Spanish jurors	74%	69%	19%	18%	16%	14%

Table 135. Success rates in identification and discrimination tasks across voice samples and group of jurors.

Before commenting on the results seen on table 135, it should be noted that the percentages shown in discrimination tests stem from the selection of the most differentiated speakers in relation to the suspect, namely Burbidge (35.48%), M12_030 and M13_010 (18.92%), and DVA10-F18O (25%). In this sense, it seems that performance in discrimination tasks is greatly improved in English voice samples across the two groups of jurors. This situation is reversed in the other language tests, with the exception of the 2nd Dutch language tests performed by British jurors, whose success rate is significantly higher (27%) than the one obtained through identification tests (10.20%). A consistent finding is that the first experimental condition (target-present, no background noises) yields better results than the second experimental condition (target-absent with noise disturbances), irrespective of linguistic input or groups of jurors. Besides that, relationships of language familiarity and voice line-up's outcome seem to vary slightly in discrimination tests (see *chapter 6. Discussion*).

Once the follow-up analysis is cleared, the section proceeds to formulate hypothesis n° 9 again: Foreign and native speaker recognition using acoustic-phonetic analysis is more accurate than the lay listener's (jurors) judgement. In this regard, error rates (false alarms) shall be consulted, since the notion of success rates in an acoustic-phonetic analysis would be challenging to pin down. In this vein, the upcoming table illustrates the error rates

found in speaker recognition tests (both identification and discrimination tasks) and in acoustic-phonetic analysis:

		Voice samples		
		1 st experimental condition		
		English	Spanish	Dutch
Identification	British jurors	46.94%	20.41%	46.94%
	Spanish jurors	58.63%	36.21%	36.21%
Discrimination	British jurors	8%	0%	2%
	Spanish jurors	2%	3%	2%
Acoustic-phonetic analysis		0%	16.21%	14.29%

Table 136. Error rates in speaker recognition tests and in acoustic-phonetic analysis across voice samples (1st experimental condition).

As noted in table 136, only the first experimental condition is examined due to the fact that discrimination tasks cannot produce false alarms in this scenario (since the target is absent and hence cannot be selected as a dissimilar speaker). Furthermore, the acoustic-phonetic analysis does not make distinctions in this regard, either (all voice samples were analysed with clear sound). The percentages shown in the identification department were obtained through subtracting the chances for success rates and missing the target. Expectedly, identification tasks are more prone to errors given that the odds are higher than in discrimination tasks (false alarms entail the selection of more than one speaker in identification tasks, whereas discrimination tests' error rates consist of erroneously choosing the suspect's voice recording listed in the voice line-up). As far as identification tests are concerned, the acoustic-phonetic analysis is more efficient than the lay listener's judgement (thus accepting the alternative hypothesis).

On the other hand, only false alarms stemming from acoustic-phonetic analysis and discrimination tests are able to put forward a fair comparison, since such errors occur due to the wrong selection of one definite speaker (suspect's voice sample) in both cases. From such a comparison, the null hypothesis can be rejected only when dealing with

Chapter 5- Results: Acoustic-phonetic analysis

English voice samples. Concerning the Spanish and Dutch recordings, the null hypothesis must be retained, and therefore it is concluded that foreign and native speaker recognition using acoustic-phonetic analysis is not more accurate than the lay listener's judgement. Suffice to say that this finding is only applicable to the current conditions of this research (an acoustic-phonetic analysis done with limited audio material). Had it been a thorough analysis with samples of longer duration, results could have varied significantly.

CHAPTER 6

DISCUSSION

After concluding the perception surveys-based analysis and the one devoted to acoustic-phonetic measurements, this chapter provides a joint discussion of the ensuing findings deriving from said analytical stages. In doing so, each hypothesis' results are consulted so as to discern how they fit with the general aim of this research, and how they relate to the previous literature concerned with the variables of interest, if applicable.

Once such information is contrasted, the study will proceed to highlight the most relevant contributions of this piece of work in chapter 7 (*Conclusions*) with an assessment of the theoretical and methodological aspects surrounding the results obtained. In this regard, said section discusses the extent to which these findings may be reliable in a forensic phonetics context, though not without acknowledging the potential pitfalls and shortcomings that may arise.

As a final step, this empirical investigation will formulate recommendations for future research in chapter 8, dealing both with practicalities at the theoretical (forensic phonetics

in general) and at the practical level (current regulations on the applications of voice line-ups).

With the purpose of reviewing the concepts revolving around the present thesis, it should be reminded that its main objective is to undertake a perceptual study through the conducting of online perception surveys in order to test the assumptions of voice line-up's applications from a theoretical perspective. Accordingly, the two analytical stages contemplated for this research have been split up depending on the agents involved in said voice line-up, namely the jurors (participants whose perception abilities are being tested) and informants (speakers who provide voice samples to build up a voice line-up). Hence, the first analysis (perception surveys-based analysis) investigates whether specific sociolinguistic profiles shape the juror's proficiency at foreign and native speaker recognition tasks in different experimental conditions (hypothesis 1-6). The remaining sub-objectives entail an exhaustive acoustic-phonetic analysis on informants' recordings. In this vein, some segmental and suprasegmental phenomena are registered to gauge levels of intra-speaker variation between the speakers acting as suspects with differing intonation patterns (hypothesis 7), while also measuring the intervariability of foil speakers with a rising intonation (hypothesis 8). Additionally, segmental phenomena remain uncontrolled in both hypotheses due to the specificities of the data set (semi-spontaneous data). As a final note, hypothesis 9 juxtaposes the results of the previous analytical stages with the purpose of explaining the choices made by the jurors on the basis of distinctive acoustic-phonetic properties of the selected informants' speech.

Despite being involved in the speaker recognition procedure, a retention period has not been established in the present research to monitor how the stimuli of short duration is stored in the long-term memory. As a result, the detrimental effects of such delays (Papcun et al. 1989, Yarmey 1995) remain unattested in this experiment. It could be argued that the time spent between the first (target-present) and second (target-absent) language tests could pose an additional difficulty to the hearers, since they are instructed not to rewind the audio file which introduces the suspect to identify. However, previous research shows that the auditory memory does not undergo noticeable losses in such short periods of time (Legge et al. 1984), all the more considering that the exposure to the stimuli is shorter than 30s (Kerstholt et al. 2004).

Chapter 6- Discussion

Besides that, informants' emotional states are expected to yield little variation, as their recordings took place under similar circumstances. However, the limitations remain on simulating the intended emotions at the time of the theoretical incident (Rodríguez Bravo et al. 2003: 33), since such semi-spontaneous exchanges are far from the mood expected in a criminal offence. Likewise, the situational context for the hearers (completing an online survey) contrasts sharply with the one experienced by the actual victims who face a voice line-up in a real-life case, hence the discrepancies in terms of psychological states.

The first hypothesis seeks to investigate if aural-perceptual recognition is exacerbated by the extant relationships of familiarity (or rather the lack thereof) between jurors and the exposed languages (stimuli), as previous research noted (Köster & Schiller 1995, 1997). As discussed previously, the present experiment did not find an evident linear relationship between response types and linguistic input (that familiar languages would facilitate speaker recognition to a greater extent than learned or unknown languages).

Not only this, but the comparison between the results obtained in identification tests through familiar and unknown languages did not yield significant differences, in terms of hits and false alarm rates. This observation applies across the two groups of jurors (British and Spanish). Instead, it was the learned language the one that displayed differentiated results from familiar and unknown language tests, even though its influence is not standardised across cultural groups (British group's learned language produced better results overall than familiar/unknown language tests, whereas the opposite is true for the Spanish group). It is also worth noting that such trend is observed in all identification tests irrespective of the experimental condition (with the exception of the target-absent Spanish test, whose shown differences did not reach statistical significance). As for discrimination tasks, no significant differences were found between voice line-up's outcome and type of language exposed.

As a complementary note, hypothesis 2 concludes that jurors' performance is improved significantly in discrimination tasks, much in contrast with the results obtained in identification tests. Such tendency is consistent across groups of jurors and experimental conditions.

Chapter 6- Discussion

The third hypothesis poses that a heightened self-perceived confidence level should have a positive effect on the jurors' actual performance at speaker recognition tests. The evidence found in this thesis proves, however, that this was the case only in two specific cases in identification tasks: British group's learned language test (target-absent condition with background noises) and Spanish group's unknown language test (target-present condition with no noise disturbances). As for discrimination tests, no significant correlation was found in this respect. Much in line with Kerstholt et al. (2004: 335), it is surmised that confidence levels should be treated with caution, since no clear patterns were discerned across groups of jurors, language familiarities, and experimental conditions. If required, confidence levels should be considered only at the highest peaks on the Likert scale (7-10 points), otherwise this variable may be rendered futile.

Despite an unbalanced sample in relation to the sub-categories *gender* and *age*, the fourth hypothesis spots older (over 22) males as more reliable than younger (18-22) females. Specifically, male participants perform better in identification tasks when exposed to the familiar language (target-absent condition with background noises) in both groups of participants (and in the first learned language test for the Spanish group-only analysis, too). Regarding the influence of *age*, it seems that older students perform better at the learned language test (again during the second experimental condition), regardless of the group surveyed. As usual, discrimination tests found no significant results in this domain, either.

The fifth hypothesis studies whether the chosen cultural groups of jurors (British and Spanish) have an influence on their speaker recognition capabilities, along with their further sub-division (monolingual and bilingual linguistic environments). As it would seem logical to assume after investigating the matter, it was concluded that said sociolinguistic factors do not influence speaker recognition abilities, with the exception of an advantageous position of British jurors over the Spanish group during the first learned language test concerned with identification tasks.

The pairwise comparison of foil speakers offered in the acoustic-phonetics analysis revealed that speaker M12_030 is the one with the second highest number of dissimilarities amongst the constituents of that particular line-up (see *Appendix 4* for full details on the arrangements of voice samples). Coincidentally, she receives the highest

percentage of wrong votes (false alarms) by British participants (10.20%) in identification tasks (1st experimental condition). Besides that, said speaker is remarkably dissimilar from the suspect's introductory recording, even reaching the highest dissimilarity rate in the group (18.92%). On the other hand, the most voted English distractors erroneously chosen in identification tests (1st experimental condition) by Spanish jurors are Peter Toll and Jez Riley (24.14%). The former is the most dissimilar voice amongst foil speakers (18.38%), whereas the latter is the second least dissimilar speaker in the voice line-up (9.19%). When contrasting their acoustic parameters with the suspect's sample (SUSPECT Simon T. Elliott), Peter Toll keeps the same ranking (22.58%), only surpassed by Alan Burbidge (35.48%), while Jez Riley becomes one of the voice samples with less differences in relation to the suspect (6.45%). In light of the above, it could be theorised that jurors target those voices who either stand out the most in the voice line-up, or match more closely the acoustic properties displayed by the suspect to identify (Jez Riley's case) in identification learned language tests. Albeit non-significant, a similar pattern is discerned in the second experimental condition. However, it should be noted that other external factors might be interfering with the results, and hence a more intensive investigation is required.

The exceptional nature exhibited by the learned language test is exhibited again during the sixth hypothesis, which poses that background noises hinder aural-perceptual speaker recognition. As far as identification tasks are concerned, this seems to be the case for familiar and unknown languages. The learned language, however, does not show statistically significant differences between the first (no background noise) and the second (with background noise) experimental condition. After a thorough inspection of the data produced by discrimination tests, it is concluded that no significant negative correlations are found across experimental conditions.

The epilogue adds the variable *studies* to the statistical model which already included *age* and *gender* in hypothesis 4 (hence hypothesis 4.1.). Again, this sub-section should be interpreted with care due to the unbalanced distribution of variables due to an over-stratified population sample. Much in line with its original version, this epilogue spots the same predictors in the same combinations of perception tests and group of jurors. The only remarkable difference is the replacement of the variable *age* for its near-equivalent *studies*. In this respect, the predictor *studies* displays larger p-values in comparison with

age, but still displays significant (or near-significant, as the British group's case reflects in the second learned language test) p-values, nonetheless.

Moving to the acoustic-phonetic department, hypothesis 7 investigates whether informants' intra-speaker variation is different enough at the segmental and suprasegmental level when analysing samples with differing intonation patterns (rising and falling intonation) and uncontrolled segmental units (semi-spontaneous data). The conclusion states that within-speaker variation is inevitable in at least one group of the studied acoustic-phonetic variables (with the exception of English samples, which display no differentiations in terms of suprasegmental phenomena). In spite of that, the following variables were reported as robust measures owing to their consistency throughout the three groups of voice samples (English, Spanish, and Dutch) examined: *Mean pitch ($P\bar{x}$)*, *25%-50% Pitch*, *N_paus/min*, *VOT*, and *alveolar sibilant's noise amplitude, F1-F2 values*.

Hypothesis 8 seeks to prove that intervariability of foil speakers' voice samples is significant enough, even when sharing similar intonation contours (rising intonation), aside from segmental units remaining uncontrolled (semi-spontaneous recordings). According to the findings, every single segmental and suprasegmental unit of measurement has found at least one significant differentiated pair of speakers in one of the groups of informants (English, Spanish, and Dutch). However, two criteria have been put forward to refine the results even further: to spot the variables which report the highest number of found cases of dissimilarities (criterion 1), and those which register at least one relationship of dissimilarity across the established groups of voice samples (criterion 2). As a result, those variables that comply with the aforementioned conditions are *N_paus* (number of pauses per extract) *release burst intensity* (either voiceless plosives only or both voiced and voiceless plosives) and *noise amplitude* (including the voiced alveolar sibilant [z] in English voice samples).

Lastly, the ninth hypothesis compares the results obtained from the perception surveys-based analysis and the one focused on acoustic-phonetic measurements. The further distinction that is made in this sub-section apropos the definition of discrimination tests (it refers now to the deliberate choice made towards one particular speaker whose speech is clearly differentiated from the suspect's, rather than the selection of any speaker in the

voice line-up but the suspect, as contemplated in the previous sections) allows for a complementary analysis on previous hypotheses. According to the first hypothesis, discrimination tests did not find significant correlations amongst levels of language familiarities. However, in this scenario, Spanish jurors' discrimination abilities are far more efficient in the learned language (74%) than in discriminating Spanish (19%) and Dutch (16%) speakers. On the other hand, British jurors' results contrast sharply with the ones obtained in identification tests, since this time their ability to discriminate speakers in their learned language is significantly reduced (20%) in comparison with English (43%) speakers. Additionally, success rates in discriminating Dutch speakers is diminished even further (16%). Following the trends seen in the first hypothesis, discrimination tests do not show an obvious pattern, either. The only distinguishable trait is that the jurors' performance associated with learned language tests is reversed across identification and discrimination tasks (Spanish participants show higher success rates in discriminating English speakers, and yet identifying them proves to be troublesome. Likewise, British jurors successfully identify Spanish speakers with higher odds than familiar and unknown language tests, whilst said success rates drop drastically when it comes to discriminate said group of informants).

As described previously in hypothesis 2, discrimination tests fare undeniably better than identification tasks when the concept of *discrimination* entails the selection of any speaker other than the intended suspect. Nevertheless, when adopting the notion used in hypothesis 9 (deliberate choice of the most dissimilar speaker), this assertion only applies in English voice samples (both cultural groups and experimental conditions) and in the second voice line-up for Dutch recordings (applicable only for British jurors), whereas identification success rates surpass those obtained in discrimination tasks in the rest of cases. Besides that, the sixth hypothesis is reinforced in this scenario, since background noises diminish success rates in discrimination tasks as well (with the notable exception of the Dutch language test undertaken by British jurors).

As a final remark, the contrast shown between speaker recognition tests (juror's choice) and the acoustic-phonetic analysis revealed that the latter displays less error rates than identification tasks. In spite of this, jurors' discrimination abilities present less false alarms than the percentages gathered during the acoustic-phonetic analysis (with the exception of the analysis on English voice samples). It should be reminded, however, that

Chapter 6- Discussion

this statement only applies within the specificities of this particular research, given the limitations of the audio material analysed.

CHAPTER 7

CONCLUSIONS

Once all the areas of interest have been explored, this thesis proceeds to elaborate the main conclusions deriving from the findings of this empirical research. Through this retrospective review, it is sought to highlight the most prominent aspects related to methodology, results, variables, and their implications to the theoretical framework.

As commented already, this study has explored the intricacies of aural-perceptual speaker recognition (identification and discrimination of speakers) by offering a set of sociolinguistic and acoustic-phonetic variables with the purpose of accounting for the varying responses obtained from British and Spanish jurors through online perception surveys. Nevertheless, such phenomenon is acknowledged as an interdisciplinary work, and thus calls for a multi-layered approach which includes other disciplines than what is commonly referred to as forensic phonetics, such as sociology, psychology, cognitive neuroscience, etc. In the grand scheme of things, the analytical work provided in this piece of work offers only a one-dimensional perspective of the issue through forensic phonetics lenses, which renders a partially unexplained account on the otherwise complex

interplay of factors involved in aural-perceptual speaker recognition. For instance, despite the efforts made to simulate a realistic scenario which would get closer to a real-life case, listeners' psychological states are not induced nor controlled by the main researcher for obvious ethical reasons. Hence the limitations in the scope of the present investigation, since emotional responses are one of the main key factors regulating auditory memory's encoding (and its possible retrieval).

One of the aims of this study is to conduct a series of voice line-ups which would replicate more realistic conditions (semi-spontaneous exchanges, audio of short duration, and the addition of noise disturbances) than what the traditional literature has shown hitherto (controlled laboratory settings), with all the consequences that ensue. To achieve this, short excerpts from each group of informants' voice samples were analysed to gauge levels of inter- and intra-speaker variation, which is far from the ideal procedure to follow in acoustic-phonetics analysis, since the shortage of voice parameters may not reveal a representative sample of the target speakers' speech. On the other hand, this methodological decision enables a fair comparison between the results obtained through online perception tests and the subsequent acoustic-phonetic analysis, since the stimuli presented to British and Spanish jurors is the same as the one the researcher has analysed. Besides that, this thesis aims at forensic speaker comparison of pre-selected audio material (speakers) rather than extracting the differences between two voice samples based on the idiosyncrasies of a larger population (Nolan 2001: 14), which would require a reference corpus and specialised means (Fernández Planas 1998: 165) to this end.

Even with such disadvantageous conditions, intra-speaker variation seems to be low when comparing only the two excerpts from the suspect identify, and said variation still prevails after adding the rest of distractors to the model (as in the case of English informants, whose suspect exhibits 0% dissimilarities between his introductory recording and his voice sample used as a distractor). In spite of this, it must be noted that the two remaining groups of informants display higher percentages of dissimilarities in terms of within-speaker variation, both the Spanish (16.21%) and the Dutch (14.29%) group. On a side note, the features that exhibit some intra-speaker variation do not follow an established pattern, but rather each pair of voices differs in very specific areas. Curiously enough, those variables which do not show within-speaker variation across the three major group of informants (English, Spanish, and Dutch) are those concerned with pitch (mean pitch,

and 25%-50% quantiles), pausing (number of pauses per minute), voiced/voiceless plosives (VOT) and alveolar sibilants (noise amplitude, and F1-F2 values). Since intra-speaker variation is inherently expected, this suggests that forensic speaker comparison is still feasible even in adverse conditions (the shortage of audio material and also the differing intonation patterns selected for the excerpts representing the suspects and for those used as foil speakers).

Concerning inter-speaker variation, all the selected variables for the analysis proved to be fruitful in distinguishing at least one pair of speakers across the three groups of voice samples. Two criteria have been put forward to classify the efficiency of such variables, which relate to their superior numbers in registered cases of dissimilarities (criterion 1) and the fact that they could spot differences in every speaker within the group (criterion 2). In this sense, those units of measurement endowed with higher discriminatory power (and consistent across groups of informants) according to the number of reported differentiations are *release burst intensity* (in voiceless plosives and/or in both voiced and voiceless plosives) and *noise amplitude* ([s] in all excerpts, and [z] & [s] in English voice samples). Similarly, the variable *N_paus* (number of pauses per excerpt) complies with the second criterion, which is applicable through the three groups of voice samples. Again, the above-mentioned findings suggest that differentiating speakers by means of forensic speaker comparison procedures is still possible even with the controversial methodological change.

As for the sociolinguistic variables of study, the influence of age, gender, and studies on speaker recognition performance could be questioned due to the unbalanced distribution of certain strata within the group of respondents (especially males and those participants over 22 years old). Nevertheless, it could be argued that, despite such stratification, the sample reflects the idiosyncrasies shown in the target population (higher proportion of 18-22-year-old female university students in language-related degrees). Besides that, correlations are calculated proportionally according to the number of subjects in each sub-category. Having said that, male jurors and those participants over 22 years old with higher studies seem to be the most reliable jurors amongst the selected participants.

Also, discrimination tasks are split into two according to their interpretation of *success rates*: either avoiding the suspect to identify, or to actively select a speaker who presents

more dissimilarities to the suspect than the rest of constituents of the voice line-up. This distinction remains crucial in the interpretation of some of the accepted hypotheses. For instance, when studying the effect of language familiarity upon perception scores, it was found that the familiar and unknown languages' scores did not differ significantly in identification tests (and in the British discrimination test), as it would be expected based on the literature on the topic. If the first meaning of *success rates* is assumed in this research question, it would be concluded that all discrimination tasks (in both cultural groups) succeed with a percentage exceeding the 90% through any experimental condition. At any rate, learned language tests appear to report disparate results within their scores (either higher than average or below average). So much so that hypothesis 5 detects higher L1's perception scores from British jurors in relation to Spanish participants.

Moreover, whether discrimination tasks fare better than identification tests is subject to interpretation, too. When discrimination implies the conscious selection of a dissimilar speaker, this assertion appears true only in English voice samples (and in Dutch recordings with background noises, only applicable for British jurors), whilst identification tests exhibit greater success rates in the rest of cases (Spanish voice line-ups and most of Dutch perception tests). Otherwise, it would be assumed that discrimination tests' rates of correct rejections reach nearly 100%. Contrastively, the negative influence that background noises exert upon the hearer's speaker recognition capabilities is attested in all experimental conditions (with the already commented exception of British jurors completing the 2nd Dutch language test), irrespective of cultural groups. Likewise, the use of confidence levels to predict perception scores does not seem to be accurate overall, but only when it reaches the highest points (7-10) in very specific cases within identification tests (British L2 and Spanish U1).

To conclude this chapter, the contrastive study of the perception surveys-based analysis and the one driven by acoustic-phonetic units has proved that, albeit limited, the proposed analysis on voice parameters contains less error rates than jurors' responses registered through identification tests. Notwithstanding this difference, hearers seem less prone to false alarms than the acoustic-phonetic method in discrimination tasks. However, as mentioned already, a more thorough acoustic-phonetic analysis may have rendered better

Chapter 7- Conclusions

results, and thus the appropriateness of analysing short voice samples under the circumstances of this research is subject to debate.

It should be underlined that, even if the present thesis does not provide a clear-cut answer to the formulated hypotheses with sound and standardised statistically significant results, it still provides valuable insight on semi-spontaneous corpora, which tend to be more challenging to analyse than the ones recorded under laboratory conditions. In this regard, the efforts made to replicate a scenario closer to an actual forensic phonetics' case could be worthy of consideration, given that the average phonetician may lack access to sophisticated instruments of automatic speaker recognition such as the ones employed by law enforcement officers (Morrison 2009: 304) and other relevant institutions (Jiménez et al. 2014: 37).

CHAPTER 8

RECOMMENDATIONS FOR FUTURE RESEARCH

After drawing the main conclusions of this study, it remains clear that some limitations require further work in order for the formulated hypotheses to be validated. First of all, it would be interesting to conduct studies whereby aural-perceptual recognition is measured at three levels, namely jurors' perception, the limited acoustic-phonetic analysis on the same stimuli used for voice line-ups, and an exhaustive account on voice parameters that contemplates larger voices samples. By doing this, a progression could be seen amongst the selected levels in terms of speaker recognition's accuracy.

Research on forensic voice comparison should not be exclusively limited to comparing a narrow selection of speech samples, but said comparison should also point at the probability of a certain speaker matching the voice parameters of another one, given the characteristics of the target population of interest (Nolan 2001: 14). In this line of thought, it is suggested that future research could tackle this issue by compiling acoustic-phonetic data towards the goal of building speech databases including dialectal information

(Fernández Planas 1998: 165), which could be used as reference corpora in forensic phonetics' research.

Concerning sociolinguistic predictors, a more representative sample should be gathered with a higher number of participants so as to clarify the alleged influence that age, gender, and level of studies have upon the jurors' identification/discrimination abilities. As Manzanero & Barón (2017) suggest, an additional step would entail controlling whether false alarms increase or decrease when the target's gender matches that of the juror's.

Due to the unpredictable nature of learned language tests observed in speaker recognition tasks, it is encouraged to conduct more voice line-ups with a multilingual data set that includes several degrees of familiarity in relation to the hearer's linguistic habits and experiences. In this regard, it could be tested whether acquiring a language (or not) makes a difference in relation to learning a language in terms of successfully identifying an intended suspect.

As far as discrimination tasks are concerned, further research should compare them with identification tasks with the purpose of proving if discriminating a specific speaker as the most dissimilar voice to the suspect's is easier for a hearer than actually identifying the suspect himself/herself, and in what experimental conditions. This comparison could even be complemented with an acoustic-phonetic analysis that establishes relationships of perceptual similarities between distractors and suspects to test whether false alarms refer to those voices which bear more resemblances to the suspect's. In this manner, trends on dissimilar/similar voices could be discerned, which could in turn point at potential aural-perceptual cognitive biases.

On the practical side, some suggestions could be made to further improve voice line-ups' guidelines (Broeders & van Amelsvoort 1999, 2001; De Jong-Lendle et al. 2015, and Hollien 2012). Firstly, segregating speakers according to their exposure to certain linguistic input (along with sociolinguistic background) could put them on equal footing, since the current thesis suggests that an intermediate exposure (learned language) yields unpredictable results.

Chapter 8- Recommendations for future research

Despite the fact that voice parades display clear audio tapes, an assessment of the hearer's adaptability to noisy conditions could be crucial to evaluate the witness' accuracy in his/her decision (since background noises are proven to reduce the percentages of success rates), provided that the offence took place in a space with such characteristics.

The procedure could benefit from including a discrimination task whose wording directs the speaker to identify the least similar speaker to the target. According to the error rates registered in the present thesis, listeners appear less prone to false alarms in discrimination tests than in identification tasks.

CHAPTER 9

BIBLIOGRAPHIC REFERENCES

BIBLIOGRAPHIC REFERENCES

- Abdi, H. (2007). The Kendall rank correlation coefficient. In: Salkind, N. (Ed.) *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage, pp. 1-7.
- Acosta, S. A. (2009). La psicología del testimonio en el ámbito psicosocial. La veracidad o la mentira, aspectos con los que se enfrenta el psicólogo jurídico. *Revista Electrónica de Psicología Social* 17(1), pp. 1-10.
- Alexander, A., Botti, F., Dessimoz, D., & Drygajlo, A. (2004). The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Science International* 146(1), pp. 95-99.
- Anderson, J. & Bower, G., H. (1974). A propositional theory of recognition memory. *Memory and Cognition* 2(1), 406-412.
- Anderson, J. R. (2014). *Cognitive psychology and its implications* (8th ed.). New York: Worth Publishers. p. 124-180.
- Arce, R. & Papillon, M. (2002). Desarrollo y evaluación de un procedimiento empírico para la detección de la simulación de enajenación mental en el contexto legal. *Revista Anuario de Psicología* 3(33), pp. 385-408.
- Arce, R. & Fariña, F. (2006). Psicología del testimonio: Evaluación de la credibilidad y de la huella psíquica en el contexto penal. In: Consejo General del Poder Judicial (Ed.). *Psicología del testimonio y prueba pericial* (pp. 39-103). Madrid: Consejo General de Poder Judicial.
- British Association for Applied Linguistics (BAAL). (2016). *Recommendations on Good Practice in Applied Linguistics* (3rd ed.). Retrieved from: https://www.baal.org.uk/wp-content/uploads/2016/10/goodpractice_full_2016.pdf [10/01/2017].
- Baldwin, J., & French, P. (1990). *Forensic Phonetics*. New York: Pinter Publishers.
- Barrett, P. T. (2005). *Euclidean Distance: Raw, normalised, and double-scaled coefficients*. Retrieved from: <https://www.pbarrett.net/techpapers/euclid.pdf> [22/05/2018].
- Bell, Allan (1984). Language style as audience design. *Language in Society* 13(1), pp. 145-204.
- Boersma, P. (2019). *Intro 4.2. Configuring the pitch contour*. Retrieved from: http://www.fon.hum.uva.nl/praat/manual/Intro_4_2_Configuring_the_pitch_contour.html [11/08/2018].
- Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer (v. 6.0.25) [Computer Program]. Retrieved from: www.praat.org [02/03/2017].

Chapter 9- Bibliographic references

- Braun, A. (1995). Fundamental Frequency- How Speaker-specific is it. In Braun, A., & Köster, J.P. (Eds.). *Studies in Forensic Phonetics*. Trier: Wissenschaftlicher Verlag, pp. 9-23.
- Braun, A. (2016). *The speaker identification ability of blind and sighted listeners: An empirical investigation*. SRINGER: Wiesbaden, pp. 63-66.
- Britain, D. (2004). Space and spatial diffusion. In Chambers, J. K., Trudgill, P., and Schilling-estes, N. (Eds.) *The Handbook of Language Variation and Change*. UK: Blackwell publishing. pp. 603-637.
- British Library Sound Archive (2016). *Interviews with Wildlife Sound Recordists*. Retrieved from: <http://sounds.bl.uk/Environment/Interviews-with-wildlife-sound-recordists> [07/01/2017].
- Broeders, A.P.A., & van Amelsvoort, A. G. (1999) *Line-up construction for forensic earwitness identification: A practical approach*. Paper presented at the 14th International Congress of Phonetic Sciences. San Francisco: IPA, pp. 1373–1376.
- Broeders, A.P.A., & van Amelsvoort, A.G. (2001). A practical approach to forensic earwitness identification: Constructing a voice line-up. *Problems of Forensic Sciences* 47(1), pp. 237-245.
- Broeders, A.P.A., Cambier-Langeveld, T., & Vermeulen J. (2002) Case Report: Arranging a voice lineup in a foreign language. *The International Journal of Speech, Language and the Law* 9(1), pp. 104-112.
- Butcher, A. (1996) Getting the voice line-up right: Analysis of a multiple auditory confrontation. In: McCormack, P., Russell, A. (Eds.) *Proceedings of the 6th Australian International Conference on Speech Science and Technology*. Canberra: Australian Speech Science and Technology Association. pp. 97-102.
- Cabin, R., J. & Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: When and how are the questions. *Bulletin of the Ecological Society of America* 81(3), pp. 246-248.
- Cerdà-Massó, R. (2008). Sobre alguns aspectes contraposats en fonètica forense. *Estudios de Fonética Experimental XVII* 17(1), pp. 46-64.
- Cerdà-Massó, R. (2011). Creus que la teua veu es única? *Llengua, societat i comunicació: revista de sociolingüística de la Universitat de Barcelona* 9(1), pp. 33-41.
- Chambers, J. K. (2004). Patterns of variation including change. In Chambers, J. K., Trudgill, P., and Schilling-estes, N. (Eds.) *The Handbook of Language Variation and Change*. UK: Blackwell publishing. pp. 349-372.
- Cicres, J. (2007). *Aplicació de l'Anàlisi de l'Entonació i de l'Alineació Tonal a la Identificació de Parlants en Fonètica Forense* (Unpublished doctoral thesis). Universitat Pompeu Fabra, Barcelona.

Chapter 9- Bibliographic references

- Clegg, J. H., & Fails, W. C. (2017). Los fonemas oclusivos. In: Clegg, J. H., & Fails, W. C. (Eds.) *Manual de fonética y fonología españolas*. London: Routledge (Taylor and Francis), pp. 247-284.
- Clifford, B. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior* 4(4), pp. 373-394.
- Clifford, B., Rathborn, H., & Bull, R. (1981). The effects of delay on voice recognition accuracy. *Law and Human Behavior* 5(1), pp. 201-208.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 75-107.
- Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the Face Overshadowing Effect. *British Journal of Psychology* 92(1), pp. 617-629.
- Coulmas, F. (1997). Introduction. In Coulmas, F. (Ed.) *The Handbook of Sociolinguistics*. Padstow, Cornwall: Blackwell Publishing. pp. 1-7.
- Coulthard, M. (2010). Experts and opinions: In my opinion. In Coulthard, M., and Johnson, A. (Eds.) *The Routledge Handbook of Forensic Linguistics* (pp. 473-486). London: Routledge.
- Delgado, C. (2014). La pericia de identificación del habla: El papel fundamental del experto. In: Garayzábal, E., Jiménez, M., Reigosa, M. (Eds.) *Lingüística forense: La lingüística en el ámbito legal y policial* (2nd ed.). Madrid: Euphonía Ediciones, pp. 199-212.
- De Jong, N. H., & Wempe, T. (2008). Syllable Nuclei v2 [Praat script]. Retrieved from: <https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2> [04/05/2018]. Modified by Quené, H., Persoon, I., & De Jong, N., 2010.
- De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods* 41(2), pp. 385-390.
- De Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015, August). *Voice lineups: A practical guide*. Paper presented at the 18th International Congress of Phonetic Sciences. Glasgow: IPA, pp.1-5.
- Dittmar, N. (1996). Explorations in 'idiolects'. In: Sackmann, R. and Budde, M. (Eds.) *Theoretical Linguistics and Grammatical Description: Papers in Honour of Hans-Heinrich Lieb*. Amsterdam: Benjamins, pp. 111-115.
- Dobson, E. J., (1968). *English Pronunciation 1500–1700* (2nd ed.). Oxford: Clarendon Press.
- Dong, J. (2014). Study on gender differences in language under sociolinguistics. *Canadian Social Science* 10(3), pp. 92-96.

Chapter 9- Bibliographic references

- Dumas, B. K. (1990). Voice identification in a criminal law context. *American Speech* 65(4), pp. 341-348.
- Elvira-García, W. (2014). Zero crossing and spectral moments v. 1.3. [Praat script]. Retrieved from: <http://stel.ub.edu/labfon/sites/default/files/zero-crossing-and-spectral-moments13.praat> [10/06/2019].
- Farrús, M. (2011). La prosòdia com a identificador biomètric. *Llengua, societat i comunicació: revista de sociolingüística de la Universitat de Barcelona* 9(1), pp. 42-48.
- Fernández Planas, A. M. (1998). Fonètica forense. L'anàlisi pericial de la veu com una aplicació de la fonètica. In: Pradilla, M. A. (Ed.) *El món dels sons*. Benicarló: Alambor, pp. 153-166.
- Fernández Planas, A. M. (2007). ¿Para qué sirve la fonética? *Onomázein* 15 (1), pp. 39-51.
- Field, A. P., & Hole, G. J. (2003). *How to design and report experiments*. London: Sage Publications, pp. 235-239.
- French, J. P., & Harrison, P. (2007). Position statement concerning the use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law* 14(1), pp. 137-144.
- Fought, C. (2004). Ethnicity. In Chambers, J. K., Trudgill, P., and Schilling-estes, N. (Eds.) *The Handbook of Language Variation and Change*. UK: Blackwell publishing. pp. 444-472.
- Foulkes, P. (2005). Sociophonetics. In: Brown, K. (Ed.) *Encyclopedia of Language and Linguistics* (2nd ed.). Amsterdam: Elsevier, pp. 495-500.
- Giancarlo, M. (2001). The rise and fall of the Great Vowel Shift? The changing ideological intersections of philology, historical linguistics, and literary history. *Representations* 76(1), pp. 27-60.
- Gil, J., & San Segundo, E. (2014). La cualidad de voz en fonética judicial. In: Garayzábal, E., Jiménez, M., Reigosa, M. (Eds.) *Lingüística forense: La lingüística en el ámbito legal y policial* (2nd ed.). Madrid: Euphonía Ediciones, pp. 153-198.
- Goh, W. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(1), pp. 40-53.
- Goldstein, A., Knight, P., Bailis, K., & Conover, J. (1981). Recognition Memory for Accented and Unaccented Voices. *Bull. Psychonomic Soc.*, 17(1), pp. 217-220.
- González-Rodríguez, J. (2014). Evaluation automatic speaker recognition systems: An overview of the NIST speaker recognition evaluations (1996-2014). *Loquens* 1(1), pp. 1-15.

Chapter 9- Bibliographic references

- Google (2019). Google Drive [Computer Program]. Retrieved from: https://www.google.com/intl/es_ALL/drive/using-drive [02/01/2017].
- Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association* 32(2), pp. 141-174.
- Gros, J., Mihelic, F., & Pavesic, N. (1999). Slovenian speech timing at different speaking rates. In: Ohala, J. J. (Ed.) *Proceedings of the ICPHS99*. San Francisco: University of California, Berkeley. pp. 261-264.
- Hazen, K. (2011). Labov: Language variation and change. In Wodak, R., Johnstone, B., Kerswill, P. E. (Eds.) *The SAGE Handbook of Sociolinguistics*. London: Sage publications. pp. 24-39.
- Hellín, L. E. (2014). Peritaje 2.0: Usos de la telefonía móvil. In: Garayzábal, E., Jiménez, M., Reigosa, M. (Eds.) *Lingüística forense: La lingüística en el ámbito legal y policial* (2nd ed.). Madrid: Euphonía Ediciones, pp. 357-374.
- Henry, A. (2004). Non-standard dialects and linguistic data. *Lingua* 115(1), pp. 1599-1617.
- Hickey, R. (2014). Language variation and change [PowerPoint presentation]. Retrieved from: https://www.uni-due.de/ELE/Language_Variation_and_Change_Introduction.pdf [16/07/2018].
- Hollien, H. (2002). *Forensic voice identification*. London: Academic Press.
- Hollien, H. (2012). On earwitness lineups. *Investigative Sciences Journal* 4(1), pp. 1-17.
- Hollien, H. Didla, G., Harnsberger, J., & Hollien K. (2016). The case for aural perceptual speaker identification. *Forensic Science International* 269(1), pp. 8-20.
- Hulme, C., Maughan, S., & Brown, G. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language* 30(1), pp. 685-701.
- Ibáñez, T. (1979). Factores sociales de la percepción: hacia una psicología del significado. *Quaderns de Psicologia* 1(1), pp. 71-81.
- IBM Corp. (2017). IBM SPSS Statistics for Windows (v. 25.0) [Computer Program]. Retrieved from: <https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-25> [04/02/2017].
- IDESCAT. (2001). *Coneixement del català. Catalunya*. Retrieved from: <https://www.idescat.cat/indicadors/?id=anuals&n=10363&col=1> [05/06/2018].

- IFADV. (2007). *IFA dialog video corpus*. Retrieved from: <http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/> [25/10/2017].
- IVE. (2001). *Coneixement i ús del valencià. Dades comparades dels censos de 1986 a 2011*. Retrieved from: http://www.ceice.gva.es/documents/161863154/162993303/Cens_2011_cvalencia.pdf/f/975646ba-6bb9-4807-a630-000f9cdd61d7 [14/11/2018].
- IWA. (2001). *Wales factfile*. Retrieved from: http://www.iwa.wales/click/wp-content/uploads/5_Factfile_Language.pdf [23/07/2017].
- Jessen, M. (2010). The forensic phonetician. In Coulthard, M., and Johnson, A. (Eds.) *The Routledge Handbook of Forensic Linguistics* (pp. 378-394). London: Routledge.
- Jiménez, M., Reigosa M., & Garayzábal, E. (2014). La lingüística forense: Licencia para investigar la lengua. In: Garayzábal, E., Jiménez, M., Reigosa, M. (Eds.) *Lingüística forense: La lingüística en el ámbito legal y policial* (2nd ed.). Madrid: Euphonía Ediciones, pp. 27-48.
- Johnson, M. K. & Raye, C. L. (1981). Reality monitoring. *Psychological Review* 88(1), pp. 67-85.
- Johnson, A., and Coulthard, R. M. (2010). Introduction: Current debates in Forensic Linguistics. In: Coulthard, R. M. and Johnson, A. (Eds.) *The Routledge Handbook of Forensic Linguistics*. London: Routledge. pp. 1-17.
- Jongman, A., Wayland, R., Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America* 108(1), pp. 1252-1263.
- Kawahara, S. (2010). Get F1, F2, F3 (averages) [Praat script]. Retrieved from: http://user.keio.ac.jp/~kawahara/scripts/get_formants.praat [21/07/2019].
- Kerstholt, J., Jansen, N., Van Amelsvoort, A., & Broeders, A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology* 18(1), pp. 327-336.
- Kerstholt, J., Jansen, N., Van Amelsvoort, A., & Broeders, A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology* 20(1), pp. 187-197.
- Kiparsky, P. (2015). New perspectives in historical linguistics. In Claire Bower (Ed.) *The Routledge Handbook of Historical Linguistics*. London: Routledge. pp. 64-102.
- Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward improved spectral measures of /s/: Results from adolescents. *J Speech Lang Hear Res.* 56(4), pp. 1175-1189.
- Köhnken, G., Manzanero, A., & Scott. M. T. (2015). Análisis de la validez de las declaraciones: Mitos y limitaciones. *Anuario de Psicología Jurídica* 25(1), pp. 13-19.

Chapter 9- Bibliographic references

- Köster, O., Schiller, N.O. & Kühnel, H.J. (1995). *The Influence of Native-language Background on Speaker Recognition*. In: Proc. 13th International Congress in Phonetic Sei., Stockholm. pp. 306-309.
- Köster, O. & Schiller, N. (1997). Different Influences of the Native Language of a Listener on Speaker Recognition. *Forensic Linguistics* 4(1), pp. 18-28.
- Kühnel, H.J. (1995). Field procedures in forensic speaker recognition. In: Windsor Lewis, J. (Ed.), *Studies in General and English Phonetics, Essays in Honour of Professor J.D. O'Connor*. London: Routledge. pp. 68–84
- Kühnel, H. (1997) Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics* 4(1), pp. 48-83.
- Labov, W. (1982). Building on empirical foundations. In: Lehmann, W.P. and Malkiel, Y. (Eds.). *Perspectives on historical linguistics*. Amsterdam and Philadelphia: John Benjamins, pp. 72-92.
- Leemann, A., Kolly, M. J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International* 238(1), pp. 59-67.
- Legge, G., Grosman, C., & Pieper, C. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10(1), pp. 298-303.
- Lenes, M. (2013). Draw_pitch_histogram_from_sound [Praat script]. Retrieved from: https://github.com/FieldDB/Praat/Scripts/blob/master/draw_pitch_histogram_from_sound.praat [16/06/2018].
- Lindh, J. (2009). *Perception of voice similarity and the results of a voice line-up*. In: Proceedings of the 22nd Swedish Phonetics Conference. Stockholm: FONETIK. pp. 186-189.
- Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3), pp. 384-422.
- Loakes, D. (2003). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. Paper presented at the Proceedings of the 15th ICPHS, Barcelona, Spain. Retrieved from: <https://pdfs.semanticscholar.org/0f4b/00acd49b6b30cbfe464bd884932eee0a7a12.pdf> [14/05/2018].
- Loakes, D. (2006). Variation in Long-term Fundamental Frequency: Measurements from Vocalic Segments in Twins' Speech. In: Warren, P., Watson, C. I. (Eds.). *Proceedings of the 11th Australian International Conference on Speech Science & Technology*. Auckland, New Zealand: Australian Speech Science & Technology Association Inc., pp. 205-210.

Chapter 9- Bibliographic references

- Macaulay, R. (2004). Discourse variation. In Chambers, J. K., Trudgill, P., and Schilling-estés, N. (Eds.) *The Handbook of Language Variation and Change*. UK: Blackwell publishing. pp. 283-306.
- Manzanero, A. (2006). Procesos automáticos y controlados de memoria: Modelo Asociativo (HAM) vs. Sistema de Procesamiento General Abstracto. *Revista de Psicología General y Aplicada* 59(3), pp. 373-412.
- Manzanero, A., & Recio, M. (2012). El recuerdo de hechos traumáticos: Exactitud, tipos y características. *Cuad Med. Forense* 18(1), pp. 19-25.
- Manzanero, A., & González, J., (2015). Modelo holístico de evaluación de la prueba testifical (HELPT). *Papeles del Psicólogo* 36(2), pp. 125-138.
- Manzanero, A. & Barón, S. (2017). Recognition and discrimination of unfamiliar male and female voices. *Behavior and Law Journal* 3(1), pp. 52-60.
- Menzer, M. J. (2000). What is the Great Vowel Shift? [Image]. Retrieved from: <http://facweb.furman.edu/~mmenzer/gvs/what.htm> [16/04/2018].
- Martínez, E. (2007). *Análisis espectrográfico de los sonidos del habla* (2nd ed.). Barcelona: Ariel, pp. 29-72.
- Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. (2011). Typicality effects on memory for voice: Implications for earwitness testimony. *Appl. Cognit. Psychol* 25(1), pp. 29-34.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press, pp. 223-293.
- Nolan, F.J. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.
- Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. In: *Proceedings of the conference 'Law and Language: Prospect and Retrospect'*, December 12-15, 2001, Levi (Finnish Lapland). Retrieved from: <http://www.ling.cam.ac.uk/francis/LawLang.doc> [23/01/2018].
- Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law* 12(1), pp. 143-73.
- Ohman, L., Eriksson, A., & Granhag, P. (2013). Angry voices from the past and present: Effects on adults' and childrens' earwitness memory. *Journal of Investigative Psychology and Offender Profiling* 10(1), pp. 57-70.
- Olsson, J. (2008). *Forensic Linguistics* (2nd ed.) London: Continuum International Publishing Group.

Chapter 9- Bibliographic references

- Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America* 85(1), pp. 913-925.
- Patrick, P. L. (2004). The speech community. In Chambers, J. K., Trudgill, P., and Schilling-estes, N. (Eds.) *The Handbook of Language Variation and Change*. UK: Blackwell publishing. pp. 573-598.
- Pereira, D. G., Alfonso, A., & Melo, F. (2015). Overview of Friedman's test and post-hoc analysis. *Communication in Statistics- Simulation and Computation* 44(10), pp. 2636-2653.
- PRESEEA (2014). *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá. Retrieved from: <http://preseea.linguas.net> [07/03/2018].
- Prieto, P. (2002). *Entonació. Models, teoria, mètodes*. Barcelona: Ariel.
- Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory and Cognition* 21(1), pp. 89-102.
- Rasinger, S. M. (2013). *Quantitative Research in Linguistics: An introduction* (2nd ed.). London: Bloomsbury, pp. 41-43.
- Rea, L. M., & Parker, R. A. (1992). *Designing and conducting survey research*. San Francisco: Jossey-Boss.
- Rodríguez Bravo, A., Lázaro Pernias, P., Montoya Vilar, N., Blanco, J. M., Bernadas Suñé, D., Tena Parera, D., Longhi, L., & Oliver Comes, J. M. (2003). *Identificación perceptiva de locutores para la acústica forense: Las RRV*. In: II Congreso de la Sociedad Española de Acústica Forense, Barcelona: SEAF. pp. 23-34.
- Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and simple length on identification of a speaker in a line-up. *Applied Cognitive Psychology* 7(1), pp. 475-481.
- Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.
- San Segundo, E. (2014). El entrenamiento musical y otros factores que pueden influir en el reconocimiento perceptivo de hablantes. In Congosto, Y. (Ed.) *Fonética Experimental, Educación Superior e Investigación*. Madrid: Arco Libros. pp. 571 - 588.
- Sankoff, G. (2004). Linguistic outcomes of language contact. In Chambers, J. K., Trudgill, P., and Schilling-estes, N. (Eds.) *The Handbook of Language Variation and Change*. UK: Blackwell publishing. pp. 638-668.
- Schiller, N.O., & Köster, O. (1996). Evaluation of a foreign speaker in forensic phonetics: A report. *Forensic Linguistics* 3(1), pp. 176-185.

Chapter 9- Bibliographic references

- Schiller, N.O.; Köster, O., & Duckworth, M. (1997). The Effect of Removing Linguistic Information upon Identifying Speakers of a Foreign Language. *Forensic Linguistics* 4(1), pp. 1350-1771.
- Schilling-estes, N. (2004). Investigating stylistic variation. In Chambers, J. K., Trudgill, P., and Schilling-estes, N. (Eds.) *The Handbook of Language Variation and Change*. UK: Blackwell publishing. pp. 375-401.
- Schultz, T. (2007). Speaker characteristics. In Müller, C. (Ed.) *Speaker Classification I: Fundamentals, Features, and Methods*. Berlin: Springer. pp. 47-74.
- Sebastian, S., Suresh, A., K. Sunny, G., & Balraj, A. (2013). An investigation into the voice of identical twins. *Otolaryngology online journal* 3(2), pp. 1-7.
- Squire, L. R. (1987). *Memory and brain*. New York: Oxford University Press.
- Smith, H. M. J., & Baguley, T. (2014). Unfamiliar voice identification: Effect of post-event information on accuracy and voice ratings. *Journal of European Psychology Students* 5(1), pp. 59-68.
- Soto-Barba, J. (1999). Caracterización fonético-acústica de la serie de consonantes /p-t-k/ vs. /b-d-g/. *Onomázein* 4(1), pp. 125-133.
- Stevens, K.N. (1971). *Sources of Inter- and Intra-Speaker Variability in the Acoustic Properties of Speech Sounds*. In: Proceedings of the Seventh International Congress of Phonetic Sciences. Montreal, pp. 206-232.
- Styler, W. (2017). *Using Praat for linguistic research- 1.8.1*. Retrieved from: <http://wstyler.ucsd.edu/praat/UsingPraatforLinguisticResearchLatest.pdf> [12/01/2019].
- TechSmith (2013). Camtasia Studio (v. 8.1.2) [Computer Program]. Retrieved from: <https://www.techsmith.com> [01/01/2017].
- Thompson, C. (1987). A Language Effect in Voice Identification. *Appl. Cogn. Psychol.* 25(1), pp.121-131.
- Tompkinson, J. & Watt, D. (2018). Assessing the abilities of phonetically untrained listeners. *Language and Law/ Linguagem e Direito* 5(1), pp. 19-37.
- Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon Press.
- Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie canadienne* 26(1), pp. 1-12.
- UCL. (2003). *Acoustics of Speech and Hearing- Lecture 2-6: Plosives and Nasals*. Retrieved from: <http://www.phon.ucl.ac.uk/courses/spsci/acoustics/week2-6.pdf> [03/05/2018].

Chapter 9- Bibliographic references

- Univaso, P., Martínez, M., & Gurlekian, J. A. (2014). Variabilidad intra- e inter-hablante de la fricativa sibilante /s/ en el español de Argentina. *Estudios de Fonética Experimental* 23(1), pp. 96-124.
- Van Son, R., Wesseling, W., Sanders, E., & Van den Heuvel, H. (2008). *The IFADV corpus: A free dialog video corpus*. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC). Marrakech, Morocco: ELRA, pp. 1-8.
- Vázquez, V. (2014). ESLORA: Diseño, codificación y explotación de un corpus oral de español de Galicia. In: *II Workshop de Procesamiento Automatizado de Texto y Corpus* (WAPOTEC-2014), November 13-14 2014, Viña del Mar: Pontificia Universidad Católica de Valparaíso. Retrieved from: https://gramatica.usc.es/~vvazq/pdf_publico/eslora_pres.pdf [21/01/2018].
- Warren, P. (2017). The interpretation of prosodic variability in the context of accompanying sociophonetic cues. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1), pp. 1-21.
- Weinreich, U., Labov, W., & Herzog, M. (1968). Empirical Foundations for a Theory of Language Change. In: Lehmann, W., & Malkiel, Y. (Eds.). *Directions for Historical Linguistics*. Austin: University of Texas Press.
- Wells, J.C. (1997). *SAMPA- Computer Readable Phonetic Alphabet*. Retrieved from: www.phon.ucl.ac.uk/home/sampa [24/09/2017].
- Whiteside, S. P., Henry, L., Dobbin, R. (2004). Sex differences in voice onset time: A developmental study of phonetic context effects in British English. *The Journal of the Acoustical Society of America* 116(2), pp. 1179-1183.
- Willis, S. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science and Justice* 49(1), pp. 161-164.
- Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications*. arXiv: 1308.5499. Retrieved from: <https://arxiv.org/ftp/arxiv/papers/1308/1308.5499.pdf> [07/06/2018].
- Yarmey, D., (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law* 1(4), pp. 792-816.
- Yao, Y. (2007). Closure Duration and VOT of Word-initial Voiceless Plosives in English in Spontaneous Connected Speech. *UC Berkeley Phonology Lab Annual Report*. Berkeley: CA. pp. 183- 225.
- Yates, D., Moore, D., & McCabe, G. (1999). *The practice of statistics*. New York: Freeman.
- Zhang, C., & Tan, T. (2008). Voice disguise and automatic speaker recognition. *Forensic Science International* 175(2), pp. 118-122.

CHAPTER 10

APPENDIXES

10.1. APPENDIX 1: ENGLISH INFORMANTS' (WILDLIFE SOUND RECORDISTS) PROFILES (BRITISH LIBRARY SOUND ARCHIVE 2016).

Name	Gender	Age	Place of origin
Beard Richard	M	64	London
Simon K. Bearder	M	71	Oxford
Alan Burbidge	M	53	Long Eaton
David J. Chivers	M	73	Cambridge
Simon T. Elliott	M	62	London
Jez riley French	M	52	London
Dorothy Ireland	F	68	Toton
Alan McElligott	M	47	London
Derek McGinn	M	75	Inverness
John Paterson	M	82	Winchester
Geoff Sample	M	63	Berwick-upon-Tweed
Patrick Sellar	M	88	London
Peter Toll	M	50	Norwich
David Tombs	M	82	Bristol
Nigel Tucker	M	69	Bristol
Dave Williams	M	77	Surrey
Richard Youel	M	51	Waterbeach

The speakers selected for the current experiment are marked in bold above.

10.2. APPENDIX 2: SPANISH CORPUS (ESLORA) INFORMANTS' PROFILES
(VÁZQUEZ 2014).

Sociolect	Speaker	Gender	Studies	Age	Place of birth	1 st lang	2 nd lang
1	M13_008	Fem.	University	29	Santiago de Compostela	Spanish	-
	M13_010	Fem.	University	26	Santiago de Compostela	Galician and Spanish	-
	M13_016	Fem.	University	20	Santiago de Compostela	Spanish	Galician
	M13_016_hab2	Fem.	University	20	Ferrol	Spanish	Galician
2	M12_020	Fem.	Medium	24	Santiago de Compostela	Galician	Spanish
	M12_030	Fem.	Medium	30	Ourense	Spanish	-
	M12_036	Fem.	Medium	32	Santiago de Compostela	Spanish	Galician

10.2.1. Appendix 2.1. ESLORA's signed agreement.

Declaración firmada

Yo, José Vicente Benavent Cháfer, solicito el acceso al material auditivo recogido en el corpus ESLORA de la Universidad de Santiago de Compostela para fines de investigación. Asimismo, me comprometo a cumplir con las dos condiciones siguientes:

- I) el uso de los materiales para fines exclusivos de una investigación concreta (en este caso, para la realización de una rueda de reconocimiento de voz y su posterior análisis fonético acústico), y
- II) el compromiso de no cederlos a otras personas, ni difundirlos más allá del trabajo para el que van destinados.

Firma:



31, de Agosto, del 2017

10.3. APPENDIX 3: DUTCH INFORMANTS' PROFILES (IFADV 2007).

File n°	Speaker	Gender	Education	Age	Place of birth	1st lang	2nd lang
DV_8	F20K	Fem.	Higher	20	Amsterdam	Dutch	-
	F20L	Fem.	Higher	20	Amsterdam	Dutch	Russian
DV_9	F21M	Fem.	Higher	21	Amsterdam	Dutch	-
	F21N	Fem.	Higher	21	Leiderdorp	Dutch	-
DV_10	F18O	Fem.	Higher	18	Amsterdam	Dutch	-
	F19P	Fem.	Higher	19	Naarden	Dutch	-
DV_11	F28Q	Fem.	Higher	28	Roosendaal	Dutch	-
	F28R	Fem.	Higher	28	Alkmaar	Dutch	-

10.4. APPENDIX 4: VOICE LINE-UPS: VOICE SAMPLES ARRANGEMENT.

English voice parade

- Test 1 (without background noise)

English suspect-- Simon T. Elliott 8 sec.

Speaker 1- Jez Riley French	7 sec. (enhanced audio)
Speaker 2- Richard Youell	4 sec.
Speaker 3- Simon T. Elliott	8 sec.
Speaker 4- Richard Beard	9 sec. (enhanced audio)
Speaker 5- Peter Toll	8 sec. (enhanced audio)
Speaker 6- Alan Burbidge	8 sec. (enhanced audio)

- Test 2 (with background noise)

English suspect-- Simon T. Elliott (Absent suspect)

Speaker 1- Alan McElligott	7 sec.
Speaker 2- Jez Riley French	7 sec. (enhanced audio)
Speaker 3- Alan Burbidge	8 sec. (enhanced audio)
Speaker 4- Simon K. Bearder	5 sec. (enhanced audio)
Speaker 5- Richard Youell	4 sec.
Speaker 6- Peter Toll	8 sec. (enhanced audio)

Spanish voice parade

- Test 1 (without background noise)

Spanish suspect-- M12_020 11sec

Speaker 1- M13_016	4 sec.
Speaker 2- M13_008	7 sec.
Speaker 3- M12_020	4 sec.
Speaker 4- M13_010	8 sec.
Speaker 5- M12_036	5 sec.
Speaker 6- M12_030	7 sec.

Chapter 10- Appendixes

- Test 2 (with background noise)

Spanish suspect-- M12_020 (Absent suspect)

Speaker 1- M12_036	5 sec.
Speaker 2- M12_030	7 sec.
Speaker 3- M13_010	8 sec.
Speaker 4- M13_016_hab2	6 sec. (enhanced audio)
Speaker 5- M13_016	4 sec.
Speaker 6- M13_008	7 sec.

Dutch voice parade

- Test 1 (without background noise)

Dutch suspect-- DVA8- F20L 8 sec.

Speaker 1- DVA11-F28Q	12 sec.
Speaker 2- DVA10-F18O	13 sec.
Speaker 3- DVA8-F20L	14 sec.
Speaker 4- DVA8-F20K	14 sec.
Speaker 5- DVA10-F19P	7 sec.
Speaker 6- DVA9-F21M	8 sec.

- Test 2 (with background noise)

Dutch suspect-- DVA8- F20L (Absent suspect)

Speaker 1- DVA11-F28R	10 sec.
Speaker 2- DVA10-F19P	7 sec.
Speaker 3- DVA9-F21N	6 sec.
Speaker 4- DVA9-F21M	8 sec.
Speaker 5- DVA11-F28Q	12 sec.
Speaker 6- DVA10-F18O	13 sec.

The audio's volume of 'Rain&Thunder' appearing in the Dutch test has been capped at 85% to ensure fairness with the rest of slightly less noisy audios (which are played at 100% of their original volume).

10.5. APPENDIX 5: PERCEPTION SURVEYS' STRUCTURE AND DESIGN.

Here are attached some screenshots that illustrate the perception survey's structure and design. Note that only the perception survey aimed at British jurors is displayed here for the sake of simplicity. Similarly, only the voice line-up concerned with the familiar language is shown, since the remaining tests follow the same structure.

- Presentation

Can you recognise the voice of a foreign suspect?

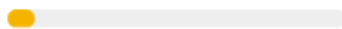
The upcoming survey partakes in the PhD project on auditory perception of foreign languages developed at the University of Barcelona (UB).

In the next page, you are going to hear a series of recordings of the suspects you need to identify. After that, recordings from 6 mock speakers (i.e. distractors) will be displayed. You should guess which recording in the list belongs to the voice you heard beforehand. The voices you hear will be in English, Spanish, and Dutch, accordingly. This test should take 10 min. to complete or less.

Please, note that:

- You are not expected to understand the Spanish or Dutch speakers, but just to distinguish between their voices.
- There is a chance that the suspect is not listed in the voice line-up, so be cautious with your choice.
- Your responses will be kept anonymous.

NEXT

 Page 1 of 12

Never submit passwords through Google Forms.

- Profile

Can you recognise the voice of a foreign suspect?

*Required

Profile

Before starting the test, we need some basic information about you

What is your gender? *

- Male
- Female
- Other: _____

How old are you? *

- 18-22
- 23-27
- 28-32
- over 33 years old

Level of education achieved/ current education level: *

- No academic certificates
- Primary School
- Secondary School
- Vocational training courses
- Undergraduate degree (BA)
- Postgraduate studies (MA, MSc...)
- Postgraduate studies (PhD)

Are you familiar with phonetics and/or linguistics in general? *

- With linguistics
- With phonetics
- With linguistics and phonetics
- No

Have you had any musical training or are you a musician? *

- Yes
- No

What languages do you speak and what is your proficiency level? *

	I do not speak it	A1- Beginner	A2- Elementary	B1- Intermediate	B2- Upper Intermediate	C1- Advanced	C2- Native near-native proficiency
German	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
French	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mandarin	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Spanish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Arabic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Italian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dutch	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Welsh	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
English	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What languages do you understand and what is your proficiency level? *

	I do not understand it	A1- Beginner	A2- Elementary	B1- Intermediate	B2- Upper Intermediate	C1- Advanced	C2- Native near-native proficiency
German	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
French	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mandarin	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Spanish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Arabic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Italian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dutch	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Welsh	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
English	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- If you speak/understand another language besides the ones mentioned above, please list it here (also include your proficiency level):

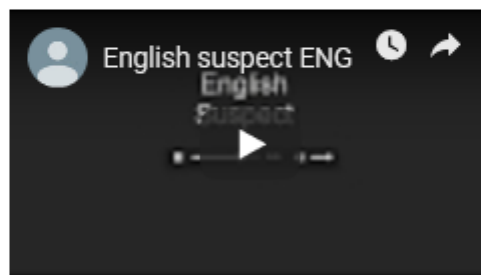
Your answer _____

- Voice line-up 1: British suspect's presentation (familiar language).

Can you recognise the voice of a foreign suspect?

Test 1- The English suspect

Here you have the voice of the suspect you need to identify:
Listen carefully to the audio file, since you are not allowed to go back once you move to the next page.



BACK

NEXT

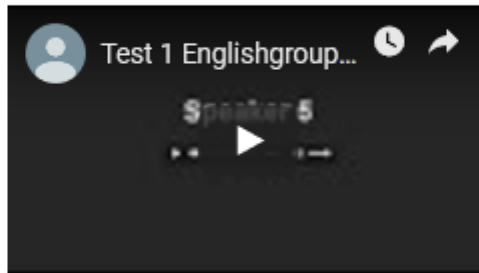
Page 3 of 12

Never submit passwords through Google Forms.

- Voice line-up 1: British suspect's identification (familiar language).

Test 1- Voice line-up 1

Here you can listen to the recordings made for this test:



Whose voice belongs to the suspect? *

1 point

- Speaker 1
- Speaker 2
- Speaker 3
- Speaker 4
- Speaker 5
- Speaker 6
- None of the above

How confident are you in your decision? (1= Not at all, 10= Completely sure) *

1	2	3	4	5	6	7	8	9	10
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- Voice line-up 1: British suspect's discrimination (familiar language).

In your opinion, whose voice differs most from the suspect's? *

- Speaker 1
- Speaker 2
- Speaker 3
- Speaker 4
- Speaker 5
- Speaker 6

How confident are you in your decision? (1= Not at all, 10= Completely sure) *

1	2	3	4	5	6	7	8	9	10
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

BACK

NEXT

Page 4 of 12

Never submit passwords through Google Forms.

- Thank you page.

End of the survey

Click on 'SUBMIT' to finish this survey.

Thank you very much for your time to fill in this survey! Your responses are greatly appreciated!!

- Disclaimer: by clicking on 'SUBMIT', you consent to the usage of the data provided for research purposes only. Said data will be kept anonymous at all times.

BACK

SUBMIT

Page 12 of 12

Never submit passwords through Google Forms.

10.6. APPENDIX 6: PRAAT SCRIPTS

Praat script	Syllable Nuclei v2.
Author/s	Nivia de Jong and Ton Wempe (modified by Hugo Quené, Ingrid Persoon, and Nivia de Jong).
Description	Detects syllable nuclei within an audio file to measure speech rate, articulation rate, ASD (Average Syllable Duration), and measures related to pauses.
<pre> # Praat Script Syllable Nuclei # Copyright (C) 2008 Nivja de Jong and Ton Wempe # # This program is free software: you can redistribute it and/or modify # it under the terms of the GNU General Public License as published by # the Free Software Foundation, either version 3 of the License, or # (at your option) any later version. # # This program is distributed in the hope that it will be useful, # but WITHOUT ANY WARRANTY; without even the implied warranty of # MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the # GNU General Public License for more details. # # You should have received a copy of the GNU General Public License # along with this program. If not, see http://www.gnu.org/licenses/ # # modified 2010.09.17 by Hugo Quené, Ingrid Persoon, & Nivja de Jong # Overview of changes: # + change threshold-calculator: rather than using median, use the almost maximum # minus 25dB. (25 dB is in line with the standard setting to detect silence # in the "To TextGrid (silences)" function. # Almost maximum (.99 quantile) is used rather than maximum to avoid using # irrelevant non-speech sound-bursts. # + add silence-information to calculate articulation rate and ASD (average syllable # duration. # NB: speech rate = number of syllables / total time # articulation rate = number of syllables / phonation time # + remove max number of syllable nuclei # + refer to objects by unique identifier, not by name # + keep track of all created intermediate objects, select these explicitly, # then Remove # + provide summary output in Info window # + do not save TextGrid-file but leave it in Object-window for inspection # (if requested in startup-form) # + allow Sound to have starting time different from zero # for Sound objects created with Extract (preserve times) # + programming of checking loop for mindip adjusted # in the orig version, precedingtime was not modified if the peak was rejected !! # var precedingtime and precedingint renamed to currenttime and currentint # # + bug fixed concerning summing total pause, feb 28th 2011 # counts syllables of all sound utterances in a directory # NB unstressed syllables are sometimes overlooked # NB filter sounds that are quite noisy beforehand # NB use Silence threshold (dB) = -25 (or -20?) # NB use Minimum dip between peaks (dB) = between 2-4 (you can first try; # For clean and filtered: 4) </pre>	

```

form Counting Syllables in Sound Utterances
  real Silence_threshold_(dB) -25
  real Minimum_dip_between_peaks_(dB) 2
  real Minimum_pause_duration_(s) 0.3
  boolean Keep_Soundfiles_and_Textgrids yes
  sentence directory /directory
endform

# shorten variables
silencedb = 'silence_threshold'
mindip = 'minimum_dip_between_peaks'
showtext = 'keep_Soundfiles_and_Textgrids'
minpause = 'minimum_pause_duration'

# print a single header line with column names and units
printline soundname, nsyll, npause, dur (s), phonationtime (s), speechrate (nsyll/dur), articulation rate (nsyll /
phonationtime), ASD (speakingtime/nsyll)

# read files
Create Strings as file list... list 'directory$'/*.wav
numberOfFiles = Get number of strings
for ifile to numberOfFiles
  select Strings list
  fileName$ = Get string... ifile
  Read from file... 'directory$/'/fileName$'

# use object ID
soundname$ = selected$("Sound")
soundid = selected("Sound")

originaldur = Get total duration
# allow non-zero starting time
bt = Get starting time

# Use intensity to get threshold
To Intensity... 50 0 yes
intid = selected("Intensity")
start = Get time from frame number... 1
nframes = Get number of frames
end = Get time from frame number... 'nframes'

# estimate noise floor
minint = Get minimum... 0 0 Parabolic
# estimate noise max
maxint = Get maximum... 0 0 Parabolic
#get .99 quantile to get maximum (without influence of non-speech sound bursts)
max99int = Get quantile... 0 0 0.99

# estimate Intensity threshold
threshold = max99int + silencedb
threshold2 = maxint - max99int
threshold3 = silencedb - threshold2
if threshold < minint
  threshold = minint
endif

# get pauses (silences) and speakingtime
To TextGrid (silences)... threshold3 minpause 0.1 silent sounding
textgridid = selected("TextGrid")
silencetierid = Extract tier... 1
silencetableid = Down to TableOfReal... sounding
nsounding = Get number of rows
npauses = 'nsounding'
speakingtot = 0
for ipause from 1 to npauses
  beginsound = Get value... 'ipause' 1
  endsound = Get value... 'ipause' 2

```

```

speakingdur = 'endsound' - 'beginsound'
speakingtot = 'speakingdur' + 'speakingtot'
endfor

select 'intid'
Down to Matrix
matid = selected("Matrix")
# Convert intensity to sound
To Sound (slice)... 1
sndintid = selected("Sound")

# use total duration, not end time, to find out duration of intdur
# in order to allow nonzero starting times.
intdur = Get total duration
intmax = Get maximum... 0 0 Parabolic

# estimate peak positions (all peaks)
To PointProcess (extrema)... Left yes no Sinc70
ppid = selected("PointProcess")

numpeaks = Get number of points

# fill array with time points
for i from 1 to numpeaks
  t'i' = Get time from index... 'i'
endfor

# fill array with intensity values
select 'sndintid'
peakcount = 0
for i from 1 to numpeaks
  value = Get value at time... t'i' Cubic
  if value > threshold
    peakcount += 1
    int'peakcount' = value
    timepeaks'peakcount' = t'i'
  endif
endfor

# fill array with valid peaks: only intensity values if preceding
# dip in intensity is greater than mindip
select 'intid'
validpeakcount = 0
currenttime = timepeaks1
currentint = int1

for p to peakcount-1
  following = p + 1
  followingtime = timepeaks'following'
  dip = Get minimum... 'currenttime' 'followingtime' None
  diffint = abs(currentint - dip)

  if diffint > mindip
    validpeakcount += 1
    validtime'validpeakcount' = timepeaks'p'
  endif
  currenttime = timepeaks'following'
  currentint = Get value at time... timepeaks'following' Cubic
endfor

# Look for only voiced parts
select 'soundid'
To Pitch (ac)... 0.02 30 4 no 0.03 0.25 0.01 0.35 0.25 450
# keep track of id of Pitch
pitchid = selected("Pitch")

```

Chapter 10- Appendixes

```
voicedcount = 0
for i from 1 to validpeakcount
  querytime = validtime'i'

  select 'textgridid'
  whichinterval = Get interval at time... 1 'querytime'
  whichlabel$ = Get label of interval... 1 'whichinterval'

  select 'pitchid'
  value = Get value at time... 'querytime' Hertz Linear

  if value <> undefined
    if whichlabel$ = "sounding"
      voicedcount = voicedcount + 1
      voicedpeak'voicedcount' = validtime'i'
    endif
  endif
endfor

# calculate time correction due to shift in time for Sound object versus
# intensity object
timecorrection = originaldur/intdur

# Insert voiced peaks in TextGrid
if showtext > 0
  select 'textgridid'
  Insert point tier... 1 syllables

  for i from 1 to voicedcount
    position = voicedpeak'i' * timecorrection
    Insert point... 1 position 'i'
  endfor
endif

# clean up before next sound file is opened
select 'intid'
plus 'matid'
plus 'sndintid'
plus 'ppid'
plus 'pitchid'
plus 'silencetierid'
plus 'silencetableid'

Remove
if showtext < 1
  select 'soundid'
  plus 'textgridid'
  Remove
endif

# summarize results in Info window
speakingrate = 'voicedcount'/originaldur'
articulationrate = 'voicedcount'/speakingtot'
npause = 'npauses'-1
asd = 'speakingtot'/voicedcount'

printline 'soundname$', 'voicedcount', 'npause', 'originaldur:2', 'speakingtot:2', 'speakingrate:2',
'articulationrate:2', 'asd:3'

endfor
```

Praat script	draw_pitch_histogram_from_sound.
Author/s	Mietta Lennes
Description	Calculates and extracts F0 basic statistics such as minimum/maximum/mean pitch, and its quantiles. Also, it draws a histogram based on the pitch points found within the audio file. It saves the registered pitch points in a separate plain text file.
<pre> # This script calculates a Pitch object from a Sound object, # displays basic F0 statistics, draws a histogram according to the distribution # of the calculated pitch points, and saves all the original pitch values to a plain text file. # Exactly one Sound object must be selected in the object window. # This script is distributed under the GNU General Public License. # Copyright Mieta Lennes 30.9.2013 form Draw F0 histogram from Sound object comment Give the F0 analysis parameters: positive Minimum_pitch_(Hz) 80 positive Maximum_pitch_(Hz) 400 positive Time_step_(s) 0.01 comment Save F0 point data to a text file in the directory: text directory comment (Empty directory = the same directory where this script file is.) comment Number of "bars" in the histogram: integer Number_of_bins 30 choice Pitch_scale_for_drawing 1 button Hertz button mel button semitones re 100 Hz button ERB endform Erase all # Define the name of the text file: soundname\$ = selected\$ ("Sound") filename\$ = directory\$ + "f0points_'soundname\$'.txt" # Delete the old file if it exists: if fileReadable(filename\$) pause Do you want to overwrite the old file 'filename\$'? filedelete 'filename\$' endif # Calculate F0 values To Pitch... time_step minimum_pitch maximum_pitch numberOfFrames = Get number of frames # Loop through all frames in the Pitch object: select Pitch 'soundname\$' unit\$ = "Hertz" min_Hz = Get minimum... 0 0 Hertz Parabolic min\$ = "min_Hz" max_Hz = Get maximum... 0 0 Hertz Parabolic max\$ = "max_Hz" mean_Hz = Get mean... 0 0 Hertz mean\$ = "mean_Hz" stdev_Hz = Get standard deviation... 0 0 Hertz stdev\$ = "stdev_Hz" median_Hz = Get quantile... 0 0 0.50 Hertz median\$ = "median_Hz" quantile25_Hz = Get quantile... 0 0 0.25 Hertz quantile25\$ = "quantile25_Hz" </pre>	

```

quantile75_Hz = Get quantile... 0 0 0.75 Hertz
quantile75$ = "quantile75_Hz"
if pitch_scale_for_drawing > 1
    unit$ = unit$ + " 'pitch_scale_for_drawing'"
    min = Get minimum... 0 0 "pitch_scale_for_drawing$" Parabolic
    min$ = min$ + " 'min'"
    max = Get maximum... 0 0 "pitch_scale_for_drawing$" Parabolic
    max$ = max$ + " 'max'"
    mean = Get mean... 0 0 'pitch_scale_for_drawing$'
    mean$ = mean$ + " 'mean'"
    if pitch_scale_for_drawing <> 3
        pitch_scale_short$ = pitch_scale_for_drawing$
    else
        pitch_scale_short$ = "semitones"
    endif
    stdev = Get standard deviation... 0 0 'pitch_scale_short$'
    stdev$ = stdev$ + " 'stdev'"
    median = Get quantile... 0 0 0.50 'pitch_scale_for_drawing$'
    median$ = median$ + " 'median'"
    quantile25 = Get quantile... 0 0 0.25 'pitch_scale_for_drawing$'
    quantile25$ = quantile25$ + " 'quantile25'"
    quantile75 = Get quantile... 0 0 0.75 'pitch_scale_for_drawing$'
    quantile75$ = quantile75$ + " 'quantile75'"
endif

# Print the statistics to the Info window:
echo F0 statistics from 'soundname$'
printline
printline 'unit$'
printline Min 'min$'
printline Max 'max$'
printline Median 'median$'
printline 25% quantile 'quantile25$'
printline 75% quantile 'quantile75$'
printline Mean 'mean$'
printline Stdev 'stdev$'
printline
printline ---
printline Selected options
printline Minimum pitch: 'minimum_pitch' Hz
printline Maximum pitch: 'maximum_pitch' Hz
printline Time step: 'time_step' s
printline Number of bins in the histogram: 'number_of_bins'
# Collect and save the pitch values from the individual frames to the text file:
for iframe to numberOfFrames
    timepoint = Get time from frame... iframe
    f0 = Get value in frame... iframe 'pitch_scale_for_drawing$'
    if f0 <> undefined
        fileappend 'filename$' 'f0'newline$'
    endif
endifor

# Convert the original minimum and maximum parameters in order to define the x scale of the
# picture, if required:
if pitch_scale_for_drawing = 2
    minimum_pitch = hertzToMel(minimum_pitch)
    maximum_pitch = hertzToMel(maximum_pitch)
elseif pitch_scale_for_drawing = 3
    minimum_pitch = hertzToSemitones(minimum_pitch)
    maximum_pitch = hertzToSemitones(maximum_pitch)

elseif pitch_scale_for_drawing = 4
    minimum_pitch = hertzToErb(minimum_pitch)
    maximum_pitch = hertzToErb(maximum_pitch)
endif

# Read the saved pitch points as a Matrix object:

```

Chapter 10- Appendixes

```
Read Matrix from raw text file... 'filename$'  
# Draw the Histogram  
Draw distribution... 0 0 0 0 minimum_pitch maximum_pitch number_of_bins 0 0 yes  
Text bottom... yes 'pitch_scale_for_drawing$'  
printline  
printline The defined pitch values from all frames were saved to the file  
printline 'filename$'.
```



```

##### PREDEFINIDAS #####
#txtName$ = "spectrum-analysis"
select all
numberOfSelectedObjects = numberOfSelected ()
if numberOfSelectedObjects < 0
    pause You have objects in the list. Do you want me to remove them?
    Remove
endif

if praatVersion < 5364
    exit Download Praat version 53.6.4 or later
endif

#####
form Spectral analysis
    comment Write the name of the txt file where data will be store
    comment The file will be created in the same folder where wavs are.
    sentence txtName spectrum-analysis
    comment ¿Do you have the speaker's name in the code? ¿How many characters has it?
    integer speaker_digits 0
    boolean filter 1
    comment Analyse intervals where text equals:
    sentence label nonempty
endform

folder$ = chooseDirectory$ ("Choose the Sound and TextGrid folder:")
txtName$ = folder$ + "/" + txtName$
txtNameExtension$ = txtName$ + ".txt"

##### encabezado #####
if fileReadable (txtNameExtension$)
    pause There is already a file with that name. It will be deleted.
    deleteFile: txtNameExtension$
endif

writeFileLine ("txtName$.txt", "Speaker   ", "File   ", "Interval label   ", "Interval start [ms]   ",
"Interval end [ms]   ", "Interval duration [ms]   ", "Zero crossings 30 ms   ", "Zero crossings
interval   ", "Zero crossings* 10 / interval duration [ms]   ", "Max frequency   ", "Min intensity   ",
"Max intensity   ", "Mean intensity   ", "Center of gravity [Hz]   ", "Skewness   ", "Kurtosis   ",
"Standard deviation [Hz]   ", "Central moment[Hz to power]   ", newline$)

Create Strings as file list.. list 'folder$'/*.wav
numberOfFiles = Get number of strings

#empieza el bucle
for ifile to numberOfFiles
    ##### ACCIONES PARA TODOS LOS INTERVALOS #####
    select Strings list
    fileName$ = Get string: ifile
    base$ = fileName$ - ".wav"

    # Lee el Sonido
    Read from file... 'folder$'/'base$.wav
    Open long sound file: folder$ + "/" + base$ + ".wav"
    # Lee el TextGrid
    Read from file... 'folder$'/'base$.TextGrid

    # Consigue el nombre del informante
    # left$ (a$, n)
    speakersId$ = left$ (base$, speaker_digits)
    # lo escribe

    ##### BUCLE DE INTERVALOS #####
    #Consigue el nombre de cada intervalo
    select TextGrid 'base$'

```

```

numberOfIntervals = Get number of intervals: 1
for n to numberOfIntervals
    select TextGrid 'base$'
    intervalLabel$ = Get label of interval: 1, n
    #lo escribe
    #aquí le digo cuando quiero que analice el intervalo
    if label$ = "nonempty"
        if intervalLabel$ <> ""
            #analiza
            @fric_analysis
        endif
    else
        if intervalLabel$ = label$
            #analiza
            @fric_analysis
        endif
    endif
endif
endfor
#fin del bucle

##### LIMPIEZA FINAL E INFO #####
select all
minus Strings list
Remove
endfor

echo The file has been created.
printline You can find it here 'folder$'.

##### ANÁLISIS #####

procedure fric_analysis
    appendFile ("txtName$.txt", "speakersId$ " , "base$ " , "intervalLabel$ ")
    #saca donde empieza el intervalo
    .intervalStart = Get start point: 1, n
    .intervalEnd = Get end point: 1, n
    .intervalDur = .intervalEnd - .intervalStart
    .intervalStartms = .intervalStart*1000
    .intervalEndms = .intervalEnd*1000
    .intervalDurms = .intervalDur*1000
    .intervalStartms$ = fixed$ (.intervalStartms, 0)
    .intervalEndms$ = fixed$ (.intervalEndms, 0)
    .intervalDurms$ = fixed$ (.intervalDurms, 0)

    select LongSound 'base$'
    #si el intervalo es menor de 0-030 el valor 2 = intervalEnd
    .targetEnd = .intervalStart + 0.030
    if .targetEnd > .intervalEnd
        .targetEnd = .intervalEnd
    endif
    printline '.intervalStart' - '.intervalEnd' targetEnd: '.targetEnd'

    select LongSound 'base$'
    Extract part: .intervalStart, .targetEnd, "yes"
    To PointProcess (zeroes): 1, "yes", "yes"
    .numeroDePuntos = Get number of points
    Remove

    select LongSound 'base$'
    Extract part: .intervalStart, .intervalEnd, "yes"
    Rename: "fricative"
    To PointProcess (zeroes): 1, "yes", "yes"
    .numeroPuntosIntervalo = Get number of points
    .zCrossing = (.numeroPuntosIntervalo*10) / .intervalDurms
    .zCrossing$ = fixed$ (.zCrossing, 2)

```

```

#appendFile ("txtName$.txt", "'intervalStart' ", "'intervalEnd' ", "'intervalDur'",
"numeroDePuntos' ", 'newline$')
appendFile: txtNameExtension$, .intervalStartms$, tab$, .intervalEndms$, tab$, .intervalDurms$,
tab$, .numeroDePuntos, tab$, .numeroPuntosIntervalo, tab$, .zCrossing$, tab$

# MOMENTOS ESPECTRALES
select Sound fricative
# Using a filter is a suggestion by Ricard Herrero and Daniel Recasens
if filter = 1
    Filter (pass Hann band): 1000, 11000, 100
endif

To Ltas: 150
.max_freq = Get frequency of maximum: 0, 0, "Cubic"
.max_freq$ = fixed$ (.max_freq, 0)
appendFile: txtNameExtension$, .max_freq$, tab$

select Sound fricative
To Intensity: 500, 0, "yes"
.min_intensity = Get minimum: 0, 0, "Parabolic"
.max_intensity = Get maximum: 0, 0, "Parabolic"
.mean_intensity = Get mean: 0, 0, "energy"

.min_intensity$ = fixed$ (.min_intensity, 0)
.max_intensity$ = fixed$ (.max_intensity, 0)
.mean_intensity$ = fixed$ (.mean_intensity, 0)
appendFile: txtNameExtension$, .min_intensity$, tab$, .max_intensity$, tab$, .mean_intensity$, tab$

select Sound fricative
To Spectrum: "yes"
.center_gravity = Get centre of gravity: 2
.skewness = Get skewness: 2
.kurtosis = Get kurtosis: 2
.standard_dev = Get standard deviation: 2
.central_moment = Get central moment: 3, 2

.center_gravity$ = fixed$ (.center_gravity, 4)
.skewness$ = fixed$ (.skewness, 4)
.kurtosis$ = fixed$ (.kurtosis, 4)
.standard_dev$ = fixed$ (.standard_dev, 4)
.central_moment$ = fixed$ (.central_moment, 4)
appendFile: txtNameExtension$, .center_gravity$, tab$, .skewness$, tab$, .kurtosis$, tab$,
.standard_dev$, tab$, .central_moment$, newline$
#limpia de la lista de objetos
select Sound fricative
Remove
endproc

```

Praat script	get F1, F2, F3 (averages).
Author/s	Shigeto Kawahara
Description	Calculates the average values of F1, F2, and F3 in all the specified labels within the audio files in the folder of choice.
<pre> # This Praat script will get average F1, F2, and F3 of all the intervals of the all files in the specified folder. # Version: 3 Feb 2010 # Author: Shigeto Kawahara # To use, you must have sound files and the corresponding text grids with the same name. form Get F1, F2, F3 sentence Directory ./ comment If you want to analyze all the files, leave this blank word Base_file_name comment The name of result file (don't change this) text textfile result.txt endform # Write-out the header fileappend result.txt soundname'tab\$'intervalname'tab\$'F1'tab\$'F2'tab\$'F3'tab\$' fileappend result.txt 'newline\$' #Read all files in a folder Create Strings as file list... wavlist 'directory\$'/base_file_name\$*.wav Create Strings as file list... gridlist 'directory\$'/base_file_name\$*.TextGrid n = Get number of strings for i to n clearinfo #We first extract a formant tier select Strings wavlist filename\$ = Get string... i Read from file... 'directory\$'/filename\$ soundname\$ = selected\$ ("Sound") To Formant (burg)... 0 5 5500 0.025 50 # We now read grid files and extract all intervals in them select Strings gridlist gridname\$ = Get string... i Read from file... 'directory\$'/gridname\$ int=Get number of intervals... 1 # We then calculate F1, F2 and F3 for k from 1 to 'int' select TextGrid 'soundname\$' label\$ = Get label of interval... 1 'k' if label\$ <> "" # calculates the onset and offset vowel_onset = Get starting point... 1 'k' vowel_offset = Get end point... 1 'k' select Formant 'soundname\$' f_one = Get mean... 1 vowel_onset vowel_offset Hertz f_two = Get mean... 2 vowel_onset vowel_offset Hertz f_three = Get mean... 3 vowel_onset vowel_offset Hertz resultline\$ = "'soundname\$'tab\$'label\$'tab\$'f_one'tab\$'f_two'tab\$'f_three'tab\$'" fileappend result.txt 'resultline\$' endif </pre>	

Chapter 10- Appendixes

```
endfor
```

```
fileappend result.txt 'newline$'  
endfor
```

```
# clean up  
select all  
Remove
```

10.7. APPENDIX 7: BETWEEN-SPEAKER VARIATION OF SUPRASEGMENTAL FEATURES IN ENGLISH VOICE SAMPLES.

Pitch			
Variable	Speakers	Z-score	Sig. (p-value)
Mean pitch (\bar{P}_x)	McElligott↔Burbidge	3.053	0.0011
	McElligott↔Simon K. Bearder	1.647	0.0497
	McElligott↔Simon T. Elliott	1.651	0.0493
	Burbidge↔Jez Riley	2.537	0.0055
	Burbidge↔Peter Toll	2.930	0.0016
	Burbidge↔Richard Youell	1.802	0.0357
25% Pitch	McElligott↔Burbidge	2.997	0.0013
	McElligott↔Simon K. Bearder	1.679	0.0465
	McElligott↔Simon T. Elliott	1.687	0.0458
	Burbidge↔Jez Riley	2.642	0.0041
	Burbidge↔Peter Toll	2.781	0.0027
50% Pitch	McElligott↔Burbidge	2.944	0.0016
	Burbidge↔Jez Riley	2.657	0.0039
	Burbidge↔Peter Toll	3.019	0.0012
	Burbidge↔Richard Youell	1.850	0.0321
75% Pitch	McElligott↔Burbidge	2.896	0.0018
	Burbidge↔Jez Riley	2.491	0.0063
	Burbidge↔Peter Toll	3.062	0.0010
	Burbidge↔Richard Youell	1.781	0.0374
	Peter Toll↔Richard Beard	1.670	0.0474
	Peter Toll↔Simon T. Elliott	1.806	0.0354
Min. intensity (I↓)	McElligott↔Peter Toll	2.269	0.0116
	McElligott↔Richard Beard	3.178	0.0007
	Burbidge↔Richard Beard	1.929	0.0268
	Jez Riley↔Richard Beard	2.422	0.0077
	Richard Beard↔Richard Youell	2.466	0.0068
	Richard Beard↔Simon K. Bearder	1.662	0.0482
	Richard Beard↔Simon T. Elliott	2.275	0.0114
Max. intensity (I↑)	McElligott↔Burbidge	1.786	0.0370
	McElligott↔Peter Toll	2.550	0.0053

	Burbidge↔ Richard Youell	1.838	0.0330
	Burbidge↔ Simon T. Elliott	2.109	0.0174
	Jez Riley↔ Peter Toll	2.306	0.0105
	Peter Toll↔ Richard Beard	1.743	0.0406
	Peter Toll↔ Richard Youell	2.601	0.0046
	Peter Toll↔ Simon T. Elliott	2.872	0.0020
Mean intensity (\bar{x})	McElligott↔ Peter Toll	3.035	0.0012
	Burbidge↔ Peter Toll	1.704	0.0441
	Jez Riley↔ Peter Toll	2.200	0.0139
	Peter Toll↔ Richard Youell	2.928	0.0017
	Peter Toll↔ Simon K. Bearder	1.866	0.0310
	Peter Toll↔ Simon T. Elliott	2.878	0.0020

Pauses			
Variable	Speakers	Z-score	Sig. (p-value)
DurPaus	McElligott↔ Peter Toll	1.682	0.0462
	Burbidge↔ Jez Riley	2.001	0.0227
	Burbidge↔ Peter Toll	2.796	0.0026
	Burbidge↔ Richard Beard	1.786	0.0370
	Jez Riley↔ Richard Youell	1.806	0.0354
	Jez Riley↔ Simon K. Bearder	1.792	0.0365
	Peter Toll↔ Richard Youell	2.601	0.0046
	Peter Toll↔ Simon K. Bearder	2.587	0.0048
	Peter Toll↔ Simon T. Elliott	1.858	0.0316
N_paus/min	McElligott↔ Simon K. Bearder	2.072	0.0191
	Burbidge↔ Simon K. Bearder	2.819	0.0024
	Burbidge↔ Simon T. Elliott	2.214	0.0134
	Jez Riley↔ Simon K. Bearder	2.451	0.0071
	Jez Riley↔ Simon T. Elliott	1.847	0.0323
	Peter Toll↔ Simon K. Bearder	2.471	0.0067
	Peter Toll↔ Simon T. Elliott	1.866	0.0310
	Richard Beard↔ Simon K. Bearder	2.306	0.0105
	Richard Beard↔ Simon T. Elliott	1.701	0.0444
Pause_%	McElligott↔ Peter Toll	1.682	0.0462
	Burbidge↔ Jez Riley	2.001	0.0227
	Burbidge↔ Peter Toll	2.796	0.0026
	Burbidge↔ Richard Beard	1.786	0.0370

Chapter 10- Appendixes

	Jez Riley↔ Richard Youell	1.806	0.0354
	Jez Riley↔ Simon K. Bearder	1.791	0.0366
	Peter Toll↔ Richard Youell	2.601	0.0046
	Peter Toll↔ Simon K. Bearder	2.586	0.0048
	Peter Toll↔ Simon T. Elliott	1.858	0.0316
N_paus	McElligott↔ Simon T. Elliott	2.220	0.0132
	Burbidge↔ Simon K. Bearder	1.776	0.0378
	Burbidge↔ Simon T. Elliott	3.108	0.0009
	Jez Riley↔ Simon T. Elliott	2.664	0.0038
	Peter Toll↔ Simon T. Elliott	2.664	0.0038
	Richard Beard↔ Simon T. Elliott	2.220	0.0132
	Richard Youell↔ Simon T. Elliott	2.664	0.0038
Speech rate	Burbidge↔ Jez Riley	2.094	0.0181
	Burbidge↔ Peter Toll	1.817	0.0346
	Burbidge↔ Richard Beard	1.967	0.0246
	Jez Riley↔ Richard Youell	2.749	0.0030
	Peter Toll↔ Richard Youell	2.473	0.0067
	Richard Beard↔ Richard Youell	2.623	0.0043
	Richard Youell↔ Simon K. Bearder	2.151	0.0157
	Richard Youell↔ Simon T. Elliott	2.243	0.0124
Articulation rate	McElligott↔ Simon K. Bearder	1.760	0.0392
	Burbidge↔ Simon K. Bearder	1.787	0.0369
	Jez Riley↔ Richard Youell	2.177	0.0147
	Peter Toll↔ Simon K. Bearder	2.284	0.0112
	Richard Beard↔ Richard Youell	2.177	0.0147
	Richard Youell↔ Simon K. Bearder	2.835	0.0023
	Richard Youell↔ Simon T. Elliott	2.379	0.0087
ASD	McElligott↔ Simon K. Bearder	2.046	0.0204
	Burbidge↔ Simon K. Bearder	2.074	0.0190
	Jez Riley↔ Richard Youell	1.908	0.0282
	Peter Toll↔ Simon K. Bearder	2.489	0.0064
	Peter Toll↔ Simon T. Elliott	1.825	0.0340
	Richard Beard↔ Richard Youell	1.908	0.0282
	Richard Youell↔ Simon K. Bearder	2.848	0.0022
	Richard Youell↔ Simon T. Elliott	2.185	0.0144

10.8. APPENDIX 8: BETWEEN-SPEAKER VARIATION OF SUPRASEGMENTAL FEATURES IN SPANISH VOICE SAMPLES.

Pitch			
Variable	Speakers	Z-score	Sig. (p-value)
Mean pitch (\bar{P}_x)	M12_020↔ M12_036	2.568	0.0051
	M12_030↔ M12_036	3.078	0.0010
	M12_036↔ M13_008	2.561	0.0052
	M12_036↔ M13_010	2.090	0.0183
	M12_036↔ M13_016	2.484	0.0064
	M12_036↔ M13_016_hab2	2.394	0.0083
25% Pitch	M12_020↔ M12_036	2.973	0.0014
	M12_030↔ M12_036	2.584	0.0048
	M12_036↔ M13_008	2.819	0.0024
	M12_036↔ M13_010	1.789	0.0368
	M12_036↔ M13_016	1.987	0.0234
	M12_036↔ M13_016_hab2	1.996	0.0229
50% Pitch	M12_020↔ M12_036	2.776	0.0027
	M12_030↔ M12_036	2.927	0.0017
	M12_036↔ M13_008	2.748	0.0029
	M12_036↔ M13_010	2.052	0.0200
	M12_036↔ M13_016	2.235	0.0127
	M12_036↔ M13_016_hab2	2.238	0.0126
75% Pitch	M12_020↔ M12_036	1.663	0.0481
	M12_030↔ M12_036	3.032	0.0012
	M12_036↔ M13_008	2.412	0.0079
	M12_036↔ M13_010	2.336	0.0097
	M12_036↔ M13_016	2.567	0.0051
	M12_036↔ M13_016_hab2	2.504	0.0061
Min. intensity (I↓)	M12_020↔ M13_016_hab2	1.676	0.0468
	M12_030↔ M13_016	2.257	0.0120
	M12_030↔ M13_016_hab2	2.296	0.0108
	M12_036↔ M13_016	1.687	0.0458
	M12_036↔ M13_016_hab2	1.726	0.0421
	M13_010↔ M13_016	2.424	0.0076

Chapter 10- Appendixes

	M13_010↔ M13_016_hab2	2.463	0.0068
Max. intensity (I↑)	M12_020↔ M13_016_hab2	2.405	0.0080
	M12_030↔ M13_010	2.110	0.0174
	M13_008↔ M13_016_hab2	2.221	0.0131
	M13_010↔ M13_016_hab2	2.890	0.0019
Mean intensity (I \bar{x})	M12_020↔ M13_016_hab2	2.853	0.0021
	M12_036↔ M13_016_hab2	1.903	0.0285
	M13_008↔ M13_016_hab2	2.452	0.0071
	M13_010↔ M13_016_hab2	2.776	0.0027

Pauses			
Variable	Speakers	Z-score	Sig. (p-value)
DurPaus	M12_020↔ M12_036	1.982	0.0237
	M12_020↔ M13_016_hab2	2.060	0.0197
	M12_030↔ M12_036	2.271	0.0116
	M12_030↔ M13_016_hab2	2.349	0.0094
	M12_036↔ M13_010	2.271	0.0116
	M13_010↔ M13_016_hab2	2.349	0.0094
N_paus/min	M12_030↔ M13_008	2.464	0.0068
	M12_030↔ M13_016_hab2	2.411	0.0079
	M13_008↔ M13_010	2.464	0.0068
	M13_010↔ M13_016_hab2	2.411	0.0079
Pause_%	M12_020↔ M12_036	1.980	0.0238
	M12_020↔ M13_016_hab2	2.043	0.0205
	M12_030↔ M12_036	2.282	0.0112
	M12_030↔ M13_016_hab2	2.346	0.0095
	M12_036↔ M13_010	2.282	0.0112
	M13_010↔ M13_016_hab2	2.346	0.0095
N_paus	M12_020↔ M13_008	1.984	0.0236
	M12_030↔ M13_008	2.646	0.0041
	M12_030↔ M13_016_hab2	1.984	0.0236
	M12_036↔ M13_008	1.984	0.0236
	M13_008↔ M13_010	2.646	0.0041
	M13_008↔ M13_016	1.984	0.0236
	M13_010↔ M13_016_hab2	1.984	0.0236
Speech rate	M12_020↔ M12_030	1.750	0.0400
	M12_020↔ M13_016	2.495	0.0063

Chapter 10- Appendixes

	M12_030↔ M13_008	2.012	0.0221
	M12_036↔ M13_016	1.943	0.0260
	M13_008↔ M13_016	2.756	0.0029
	M13_010↔ M13_016	2.122	0.0169
	M13_016↔ M13_016_hab2	2.178	0.0147
Articulation rate	M12_020↔ M13_016	2.743	0.0030
	M13_008↔ M13_016	2.665	0.0038
	M13_010↔ M13_016	2.482	0.0065
ASD	M12_020↔ M12_030	1.788	0.0369
	M12_020↔ M12_036	1.710	0.0436
	M12_020↔ M13_016	2.643	0.0041
	M12_030↔ M13_008	1.658	0.0486
	M13_008↔ M13_016	2.514	0.0059
	M13_010↔ M13_016	2.255	0.0121

10.9. APPENDIX 9: BETWEEN-SPEAKER VARIATION OF SUPRASEGMENTAL FEATURES IN DUTCH VOICE SAMPLES.

Pitch			
Variable	Speakers	Z-score	Sig. (p-value)
Mean pitch (P \bar{x})	DVA8-F20K↔ DVA9-F21N	1.945	0.0258
	DVA8-F20K↔ DVA10-F18O	2.736	0.0031
	DVA8-F20L↔ DVA10-F18O	1.720	0.0427
	DVA9-F21M↔ DVA9-F21N	1.891	0.0293
	DVA9-F21M↔ DVA10-F18O	2.682	0.0036
	DVA9-F21N↔ DVA11-F28Q	1.939	0.0262
	DVA10-18O↔ DVA11-F28Q	2.730	0.0031
	DVA10-18O↔ DVA11-F28R	2.231	0.0128
25% Pitch	DVA8-F20K↔ DVA10-F18O	2.657	0.0039
	DVA8-F20L↔ DVA9-F21M	1.824	0.0340
	DVA9-F21M↔ DVA10-F18O	2.857	0.0021
	DVA9-F21M↔ DVA11-F28R	1.718	0.0428
	DVA9-F21N↔ DVA10-F18O	2.544	0.0054
	DVA10-F18O↔ DVA11-F28Q	2.325	0.0100
50% Pitch	DVA8-F20K↔ DVA10-F18O	3.032	0.0012
	DVA9-F21M↔ DVA10-F18O	2.707	0.0033
	DVA10-F18O↔ DVA10-F19P	2.094	0.0181
	DVA10-F18O↔ DVA11-F28Q	2.974	0.0014
	DVA10-F18O↔ DVA11-F28R	2.337	0.0097
75% Pitch	DVA8-F20K↔ DVA9-F21N	1.660	0.0484
	DVA8-F20K↔ DVA10-F18O	2.627	0.0043
	DVA8-F20L↔ DVA10-F18O	1.797	0.0361
	DVA9-F21M↔ DVA9-F21N	1.694	0.0451
	DVA9-F21M↔ DVA10-F18O	2.662	0.0038
	DVA9-F21N↔ DVA11-F28Q	1.915	0.0277
	DVA10-F18O↔ DVA11-F28Q	2.882	0.0019
	DVA10-F18O↔ DVA11-F28R	2.508	0.0060
Min. intensity (I \downarrow)	DVA8-F20K↔ DVA9-F21M	2.101	0.0178
	DVA8-F20L↔ DVA9-F21M	2.996	0.0013
	DVA8-F20L↔ DVA9-F21N	1.673	0.0471

	DVA8-F20L↔ DVA10-F18O	2.331	0.0098
	DVA9-F21M↔ DVA11-F28Q	2.191	0.0142
	DVA9-F21M↔ DVA11-F28R	2.600	0.0046
	DVA10-F18O↔ DVA11-F28R	1.935	0.0264
Max. intensity (I↑)	DVA8-F20K↔ DVA9-F21N	2.266	0.0117
	DVA8-F20K↔ DVA10-F18O	2.056	0.0198
	DVA8-F20K↔ DVA10-F19P	1.899	0.0287
	DVA8-F20K↔ DVA11-F28Q	2.214	0.0134
	DVA8-F20K↔ DVA11-F28R	1.913	0.0278
	DVA8-F20L↔ DVA9-F21N	2.268	0.0116
	DVA8-F20L↔ DVA10-F18O	2.058	0.0197
	DVA8-F20L↔ DVA10-F19P	1.901	0.0286
	DVA8-F20L↔ DVA11-F28Q	2.216	0.0133
	DVA8-F20L↔ DVA11-F28R	1.915	0.0277
	DVA9-F21M↔ DVA9-F21N	1.721	0.0426
	DVA9-F21M↔ DVA11-F28Q	1.669	0.0475
Mean intensity (I \bar{x})	DVA8-F20K↔ DVA9-F21N	2.476	0.0066
	DVA8-F20K↔ DVA10-F18O	2.838	0.0022
	DVA8-F20K↔ DVA10-F19P	2.074	0.0190
	DVA8-F20K↔ DVA11-F28Q	2.011	0.0221
	DVA8-F20L↔ DVA9-F21N	2.177	0.0147
	DVA8-F20L↔ DVA10-F18O	2.538	0.0055
	DVA8-F20L↔ DVA10-F19P	1.774	0.0380
	DVA8-F20L↔ DVA11-F28Q	1.711	0.0435

Pauses			
Variable	Speakers	Z-score	Sig. (p-value)
DurPaus	DVA8-F20K↔ DVA10-F18O	1.703	0.0044
	DVA8-F20L↔ DVA10-F18O	1.686	0.0459
	DVA9-F21M↔ DVA10-F18O	1.653	0.0491
	DVA9-F21N↔ DVA10-F18O	2.714	0.0033
	DVA9-F21N↔ DVA11-F28Q	2.114	0.0172
	DVA9-F21N↔ DVA11-F28R	2.294	0.0109
	DVA10-F18O↔ DVA10-F19P	2.616	0.0044
	DVA10-F19P↔ DVA11-F28Q	2.016	0.0219
	DVA10-F19P↔ DVA11-F28R	2.197	0.0140
N_paus/min	DVA8-F20K↔ DVA9-F21M	2.470	0.0067

	DVA8-F20L↔ DVA9-F21N	1.728	0.0419
	DVA8-F20L↔ DVA10-F18O	1.795	0.0363
	DVA9-F21M↔ DVA9-F21N	2.717	0.0033
	DVA9-F21M↔ DVA10-F18O	2.784	0.0027
	DVA9-F21M↔ DVA11-F28Q	2.546	0.0054
Pause_%	DVA8-F20K↔ DVA10-F18O	1.708	0.0438
	DVA8-F20L↔ DVA10-F18O	1.692	0.0453
	DVA9-F21M↔ DVA10-F18O	1.659	0.0485
	DVA9-F21N↔ DVA10-F18O	2.718	0.0033
	DVA9-F21N ↔ DVA11-F28Q	2.110	0.0174
	DVA9-F21N ↔ DVA11-F28R	2.290	0.0110
	DVA10-F18O↔ DVA10-F19P	2.621	0.0044
	DVA10-F19P↔ DVA11-F28Q	2.012	0.0221
	DVA10-F19P↔ DVA11-F28R	2.192	0.0142
N_paus	DVA8-F20K↔ DVA8-F20L	1.754	0.0397
	DVA8-F20L↔ DVA9-F21N	3.257	0.0006
	DVA8-F20L↔ DVA10-F18O	2.506	0.0061
	DVA8-F20L↔ DVA10-F19P	2.255	0.0121
	DVA8-F20L↔ DVA11-F28Q	2.255	0.0121
	DVA9-F21M↔ DVA9-F21N	2.255	0.0121
	DVA9-F21N↔ DVA11-F28R	1.754	0.0397
Speech rate	DVA8-F20K↔ DVA10-F18O	2.005	0.0225
	DVA8-F20L↔ DVA10-F18O	1.679	0.0466
	DVA9-F21M↔ DVA10-F18O	2.161	0.0153
	DVA9-F21N↔ DVA10-F18O	2.705	0.0034
	DVA10-F18O↔ DVA10-F19P	3.311	0.0005
	DVA10-F19P↔ DVA11-F28Q	1.990	0.0233
	DVA10-F19P↔ DVA11-F28R	2.005	0.0225
Articulation rate	DVA8-F20K↔ DVA10-F18O	1.943	0.0260
	DVA8-F20L↔ DVA10-F19P	1.973	0.0242
	DVA9-F21M↔ DVA10-F18O	2.398	0.0082
	DVA9-F21N↔ DVA10-F18O	1.761	0.0391
	DVA10-F18O↔ DVA10-F19P	3.187	0.0007
	DVA10-F18O↔ DVA11-F28Q	2.429	0.0075
	DVA10-F18O↔ DVA11-F28R	2.762	0.0029
ASD	DVA8-F20K↔ DVA10-F18O	2.105	0.0177
	DVA8-F20L↔ DVA10-F19P	1.786	0.0370
	DVA9-F21M↔ DVA10-F18O	2.552	0.0053
	DVA9-F21N↔ DVA10-F18O	1.978	0.0240

Chapter 10- Appendixes

	DVA10-F18O↔ DVA10-F19P	3.190	0.0007
	DVA10-F18O↔ DVA11-F28Q	2.552	0.0053
	DVA10-F18O↔ DVA11-F28R	2.807	0.0025

10.10. APPENDIX 10: SIGNIFICANT DIFFERENCES BETWEEN THE ENGLISH SUSPECT (SUSPECT SIMON T. ELLIOTT) AND THE VOICES USED AS DISTRACTORS, BOTH AT THE SUPRASEGMENTAL AND SEGMENTAL LEVEL.

Pitch			
Variable	Speakers	Z-score	Sig. (p-value)
Mean pitch ($P\bar{x}$)	SUSPECT Simon T. Elliott ↔ Burbidge	2.444	0.0072
25% Pitch	SUSPECT Simon T. Elliott ↔ Burbidge	2.319	0.0101
50% Pitch	SUSPECT Simon T. Elliott ↔ Burbidge	2.340	0.0096
75% Pitch	SUSPECT Simon T. Elliott ↔ Burbidge	2.393	0.0083
Min. intensity (I↓)	SUSPECT Simon T. Elliott ↔ Richard Beard	2.546	0.0054
Max. intensity (I↑)	SUSPECT Simon T. Elliott ↔ Burbidge	2.087	0.0184
	SUSPECT Simon T. Elliott ↔ Peter Toll	2.852	0.0021
Mean intensity ($I\bar{x}$)	SUSPECT Simon T. Elliott ↔ Peter Toll	2.897	0.0018

Pauses			
Variable	Speakers	Z-score	Sig. (p-value)
DurPaus	SUSPECT Simon T. Elliott ↔ Peter Toll	1.891	0.0293
N_paus/min	-	-	-
Pause_%	SUSPECT Simon T. Elliott ↔ Peter Toll	1.895	0.0290
N_paus	SUSPECT Simon T. Elliott ↔ Burbidge	2.176	0.0148
	SUSPECT Simon T. Elliott ↔ Jez Riley	1.741	0.0408
	SUSPECT Simon T. Elliott ↔ Peter Toll	1.741	0.0408
	SUSPECT Simon T. Elliott ↔ Richard Youell	1.741	0.0408
Speech rate	SUSPECT Simon T. Elliott ↔ Richard Youell	2.064	0.0195
Articulation rate	SUSPECT Simon T. Elliott ↔ Richard Youell	1.969	0.0245
ASD	-	-	-

[b, d, g] and [k, p, t]					
Variable	Sound differences	Test stats.		Sig. (p-value)	Significant pairwise comparisons
VOT	YES	[b, d, g]	Kruskal-Wallis	0.425	-
		[k, p, t]	Kruskal-Wallis	0.058	-
Release burst intensity	YES	[b, d, g]	Levene test	0.040	No significant differences with SUSPECT
			ANOVA	0.009	
		[k, p, t]	Kruskal-Wallis	0.000	SUSPECT- Peter Toll SUSPECT-Richard Beard

[s] and [z]					
Variable	Sound differences	Test stats.		Sig. (p-value)	Significant pairwise comparisons
Spectral peak location	NO	Kruskal-Wallis		0.004	SUSPECT-Alan McElligott
COG	YES	[s]	Levene test	0.045	SUSPECT-Burbidge
			ANOVA	0.000*	SUSPECT-Simon K. Bearder
		[z]	Levene test	0.000	SUSPECT-McElligott
			Welch's test	Cannot be computed	SUSPECT-Burbidge SUSPECT-Richard Beard
Noise duration	NO	Kruskal-Wallis		0.247	-

Chapter 10- Appendixes

Noise amplitude	NO	Levene test		0.065	SUSPECT-Burbidge SUSPECT-Peter Toll
		ANOVA		0.000	
F1	YES	[s]	Levene test	0.067	SUSPECT-McElligott SUSPECT-Burbidge
			ANOVA	0.000	
		[z]	Levene test	0.000	-
			Welch's test	Cannot be computed	
F2	NO	Levene test		0.069	SUSPECT-McElligott
		ANOVA		0.000	
F3	YES	[s]	Levene test	0.079	SUSPECT-Burbidge SUSPECT-Jez Riley
			ANOVA	0.000	
		[z]	Levene test	0.109	-
			ANOVA	0.257	

10.11. APPENDIX 11: SIGNIFICANT DIFFERENCES BETWEEN THE SPANISH SUSPECT (SUSPECT M12_020) AND THE VOICES USED AS DISTRACTORS, BOTH AT THE SUPRASEGMENTAL AND SEGMENTAL LEVEL.

Pitch			
Variable	Speakers	Z-score	Sig. (p-value)
Mean pitch (\bar{P}_x)	SUSPECT M12_020 ↔ M12_036	3.516	0.0002
25% Pitch	SUSPECT M12_020 ↔ M12_036	3.385	0.0003
	SUSPECT M12_020 ↔ M13_010	1.776	0.0378
50% Pitch	SUSPECT M12_020 ↔ M12_036	3.494	0.0002
	SUSPECT M12_020 ↔ M13_010	1.652	0.0492
75% Pitch	SUSPECT M12_020 ↔ M12_020	1.793	0.0364
	SUSPECT M12_020 ↔ M12_036	3.351	0.0004
Min. intensity (I↓)	SUSPECT M12_020 ↔ M13_016	1.658	0.0486
	SUSPECT M12_020 ↔ M13_016_hab2	1.701	0.0444
Max. intensity (I↑)	SUSPECT M12_020 ↔ M13_016_hab2	2.265	0.0117
Mean intensity (\bar{I}_x)	SUSPECT M12_020 ↔ M13_016_hab2	2.725	0.0032

Pauses			
Variable	Speakers	Z-score	Sig. (p-value)
DurPaus	SUSPECT M12_020 ↔ M12_020	1.683	0.0462
	SUSPECT M12_020 ↔ M12_030	1.977	0.0240
	SUSPECT M12_020 ↔ M13_010	1.977	0.0240
N_paus/min	-	-	-
Pause_%	SUSPECT M12_020 ↔ M12_020	1.668	0.0476
	SUSPECT M12_020 ↔ M12_030	1.975	0.0241
	SUSPECT M12_020 ↔ M13_010	1.975	0.0241
N_paus	SUSPECT M12_020 ↔ M12_030	1.992	0.0232
	SUSPECT M12_020 ↔ M13_010	1.992	0.0232
Speech rate	SUSPECT M12_020 ↔ M12_020	1.688	0.0457
	SUSPECT M12_020 ↔ M13_008	1.955	0.0252
Articulation rate	SUSPECT M12_020 ↔ M12_020	2.358	0.0092

	SUSPECT M12_020 ↔M13_008	2.284	0.0112
	SUSPECT M12_020 ↔M13_010	2.111	0.0174
ASD	SUSPECT M12_020 ↔M12_020	2.397	0.0083
	SUSPECT M12_020 ↔M13_008	2.273	0.0115
	SUSPECT M12_020 ↔M13_010	2.023	0.0215

[b, d, g] and [k, p, t]					
Variable	Sound differences	Test stats.		Sig. (p-value)	Significant pairwise comparisons
VOT	YES	[b, d, g]	Kruskal-Wallis	0.027	No significant differences with SUSPECT
		[k, p, t]	Kruskal-Wallis	0.003	SUSPECT- M12_030
Release burst intensity	YES	[b, d, g]	Levene test	0.413	SUSPECT-M12_030 SUSPECT-M13_008 SUSPECT-M13_016
			ANOVA	0.000	SUSPECT-M13_016_hab2
		[k, p, t]	Kruskal-Wallis	0.000	SUSPECT-M12_030 SUSPECT-M13_016 SUSPECT-M13_016_hab2

[s]			
Variable	Test stats.	Sig. (p-value)	Significant pairwise comparisons
Spectral peak location	Kruskal-Wallis	0.345	-
COG	Levene test	0.095	-
	ANOVA	0.183	

Chapter 10- Appendixes

Noise duration	Kruskal-Wallis	0.414	-
Noise amplitude	Levene test	0.245	SUSPECT-M12_030
	ANOVA	0.000	SUSPECT-M13_016_hab2
F1	Levene test	0.017	No significant differences with SUSPECT
	Welch's test	0.016	
F2	Levene test	0.327	No significant differences with SUSPECT
	ANOVA	0.002	
F3	Levene test	0.449	No significant differences with SUSPECT
	ANOVA	0.009	

10.12. APPENDIX 12: SIGNIFICANT DIFFERENCES BETWEEN THE DUTCH SUSPECT (SUSPECT DVA8-F20L) AND THE VOICES USED AS DISTRACTORS, BOTH AT THE SUPRASEGMENTAL AND SEGMENTAL LEVEL.

Pitch			
Variable	Speakers	Z-score	Sig. (p-value)
Mean pitch (\bar{P})	SUSPECT DVA8-F20L ↔ DVA10-F18O	2.456	0.0070
25% Pitch	-	-	-
50% Pitch	SUSPECT DVA8-F20L ↔ DVA10-F18O	2.265	0.0117
75% Pitch	SUSPECT DVA8-F20L ↔ DVA10-F18O	2.586	0.0048
Min. intensity (I↓)	SUSPECT DVA8-F20L ↔ DVA8-F20K	1.815	0.0347
	SUSPECT DVA8-F20L ↔ DVA8-F20L	2.657	0.0039
	SUSPECT DVA8-F20L ↔ DVA11-F28Q	1.900	0.0287
	SUSPECT DVA8-F20L ↔ DVA11-F28R	2.285	0.0111
Max. intensity (I↑)	SUSPECT DVA8-F20L ↔ DVA8-F20K	2.582	0.0049
	SUSPECT DVA8-F20L ↔ DVA8-F20L	2.584	0.0048
	SUSPECT DVA8-F20L ↔ DVA9-F21M	2.053	0.0200
Mean intensity (\bar{I})	SUSPECT DVA8-F20L ↔ DVA8-F20K	2.692	0.0035
	SUSPECT DVA8-F20L ↔ DVA8-F20L	2.395	0.0083

Pauses			
Variable	Speakers	Z-score	Sig. (p-value)
DurPaus	SUSPECT DVA8-F20L ↔ DVA10-F18O	2.405	0.0081
	SUSPECT DVA8-F20L ↔ DVA11-F28Q	1.798	0.0361
	SUSPECT DVA8-F20L ↔ DVA11-F28R	1.981	0.0238
N_paus/min	SUSPECT DVA8-F20L ↔ DVA9-F21M	2.194	0.0141
Pause_%	SUSPECT DVA8-F20L ↔ DVA10-F18O	2.411	0.0079
	SUSPECT DVA8-F20L ↔ DVA11-F28Q	1.795	0.0363
	SUSPECT DVA8-F20L ↔ DVA11-F28R	1.977	0.0240
N_paus	SUSPECT DVA8-F20L ↔ DVA8-F20L	2.601	0.0046
Speech rate	SUSPECT DVA8-F20L ↔ DVA10-F18O	2.451	0.0071
Articulation rate	SUSPECT DVA8-F20L ↔ DVA10-F19P	1.763	0.0389
ASD	SUSPECT DVA8-F20L ↔ DVA10-F18O	1.827	0.0338

[b, d, g] and [k, p, t]					
Variable	Sound differences	Test stats.		Sig. (p-value)	Significant pairwise comparisons
VOT	YES	[b, d, g]	Kruskal-Wallis	0.074	-
		[k, p, t]	Kruskal-Wallis	0.023	No significant differences with SUSPECT
Release burst intensity	NO	Levene test		0.119	SUSPECT- DVA8-F20K
		ANOVA		0.000	SUSPECT- DVA9-F21M

[s]			
Variable	Test stats.	Sig. (p-value)	Significant pairwise comparisons
Spectral peak location	Kruskal-Wallis	0.110	-
COG	Levene test	0.054	SUSPECT- DVA9-F21N
	ANOVA	0.064	SUSPECT- DVA10-F19P SUSPECT- DVA11-F28Q
Noise duration	Kruskal-Wallis	0.007	No significant differences with SUSPECT
Noise amplitude	Levene test	0.003	No significant differences with SUSPECT
	Welch's test	0.000	
F1	Levene test	0.146	No significant differences with SUSPECT
	ANOVA	0.001	

Chapter 10- Appendixes

F2	Levene test	0.069	-
	ANOVA	0.160	
F3	Levene test	0.380	-
	ANOVA	0.242	

CHAPTER 11

LISTS OF FIGURES AND TABLES

11.1. LIST OF FIGURES:

Figure 1. Basic levels of stylistic variation. Adapted from Schultz (2007: 54)..... 44

Figure 2. Great Vowel Shift explained (Menzer 2000)..... 53

Figure 3. Conservative and progressive pronunciations according to Dobson (1968)..... 54

Figure 4. Main areas explored in forensic linguistics..... 56

Figure 5. Bayesian statistical model on likelihood ratios according to Nolan (2001: 14)..... 60

Figure 6. Possible outcomes of a speaker identification experiment (Braun 2016 :63)..... 63

Figure 7. The sequential method according to Hollien (2002: 63)..... 66

Figure 8. Encoding and storage model of the memory. Adapted from Anderson (2014: 127)..... 71

Figure 9. Varieties of memories. Adapted from Squire (1987: 170)..... 72

Figure 10. Continuum on types of retrieval according to Manzanero (2006: 405)..... 75

Figure 11. The experiment: jurors, language tests, and experimental conditions..... 87

Figure 12. LPC slice (left side) and FFT slice (right side) of [s] in the word *Huis*, Dutch speaker (DVA10-F19P)..... 129

Figure 13. British group’s count of response types across three language familiarities in the first experimental condition (target-present, identification tests)..... 138

Figure 14. British group’s count of response types across three language familiarities in the second experimental condition (target-absent, identification tests)..... 141

Figure 15. Spanish group’s count of response types across three language familiarities in the first experimental condition (target-present, identification tests)..... 145

Figure 16. Spanish group’s count of response types across three language familiarities in the second experimental condition (target-absent, identification tests)..... 148

Figure 17. British and Spanish group’s count of response types across three language familiarities in the first experimental condition (target-present, identification tests)..... 151

Figure 18. British and Spanish group’s count of response types across three language familiarities in the second experimental condition (target-absent, identification tests)..... 154

Figure 19. British group’s count of response types across three language familiarities in the first experimental condition (target-present, discrimination tests)..... 157

Figure 20. Spanish group’s count of response types across three language familiarities in the first experimental condition (target-present, discrimination tests)..... 160

Figure 21. British and Spanish group’s count of response types across three language familiarities in the first experimental condition (target-present, discrimination tests)..... 163

Figure 22. Multiple line graph on the L2-CL. L2 correlation in the British group (identification tests)..... 174

Figure 23. Multiple line graph on the U1-CL. U1 correlation in the Spanish group (identification tests)..... 175

Figure 24. Multiple line graph on the F1-CL. F1 correlation in the Spanish group (identification tests)..... 176

Chapter 11- Lists of figures and tables

Figure 25. Multiple line graph on the L2-CL. L2 correlation in the British and Spanish group (identification tests).....	178
Figure 26. Boxplot on the L2-CL. L2 correlation in the British and Spanish group (identification tests).....	179
Figure 27. Multiple line graph on the U1-CL. U1 correlation in the British and Spanish group (identification tests).....	180
Figure 28. Boxplot on the U1-CL. U1 correlation in the British and Spanish group (identification tests).....	180
Figure 29. Multiple line graph of age's influence upon L2's scores in the British group (identification tests).....	188
Figure 30. Multiple line graph of gender's influence upon L1's scores in the Spanish group (identification tests).....	191
Figure 31. Multiple line graph of gender's influence upon F2's scores in the Spanish group (identification tests).....	192
Figure 32. Multiple line graph of gender's influence upon F2's scores in the British and Spanish group (identification tests).....	195
Figure 33. Multiple line graph of age's influence upon L2's scores in the British and Spanish group (identification tests).....	196
Figure 34. Multiple line graph of country of origin's influence upon L1's scores in the British and Spanish group (identification test).....	204
Figure 35. Bar graph on British and Spanish jurors' knowledge of their learned language.....	208
Figure 36. Multiple line graph on British F1-F2 identification tests' comparison.....	212
Figure 37. Multiple line graph on British U1-U2 identification tests' comparison.....	212
Figure 38. Multiple line graph on Spanish F1-F2 identification tests' comparison.....	214
Figure 39. Multiple line graph on Spanish U1-U2 identification tests' comparison.....	214
Figure 40. Multiple line graph on British F1.Dis-F2.Dis (discrimination) tests' comparison.....	216
Figure 41. Multiple line graph on the existing correlation between age and studies in the F2 (British group, identification tests).....	220
Figure 42. Multiple line graph of studies' influence upon L2's scores in the British group (identification tests).....	222
Figure 43. Multiple line graph of gender's influence upon L1's scores in the Spanish group (identification tests).....	224
Figure 44. Multiple line graph of gender's influence upon F2's scores in the Spanish group (identification tests).....	226
Figure 45. Multiple line graph of gender's influence upon F2's scores in the British and Spanish group (identification tests).....	228
Figure 46. Multiple line graph of studies' influence upon L2's scores in the British and Spanish group (identification tests).....	230
Figure 47. Multiple line graph of gender's influence on F1.Dis' scores in the Spanish group (discrimination tests).....	232

Chapter 11- Lists of figures and tables

Figure 48. Statistically significant differences in suprasegmental features across the Spanish suspect's voice samples (M12_020- SUSPECT M12_020).....	244
Figure 49. Statistically significant differences in pitch-related measures across the Dutch suspect's voice samples (DVA8-F20L- SUSPECT DVA8-F20L).....	246
Figure 50. Statistically significant differences in the number of pauses across the Dutch suspect's voice samples (DVA8-F20L- SUSPECT DVA8-F20L).....	247
Figure 51. Statistically significant differences in spectral peak location values of [s] and [z] across the English suspect's voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).....	251
Figure 52. Statistically significant differences in COG values of [s] and [z] across the English suspect's voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).....	251
Figure 53. Statistically significant differences in F3 values of [s] and [z] across the English suspect's voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).....	252
Figure 54. Statistically significant differences in VOT values across segmental units in the Spanish suspect's voice samples (M12_020- SUSPECT M12_020).....	254
Figure 55. Statistically significant differences in all segmental units' release burst intensity across the Spanish suspect's voice samples (M12_020- SUSPECT M12_020).....	255
Figure 56. Statistically significant differences in [b, d, g] and [k, p, t] VOT values across the Dutch suspect's voice samples (DVA8-F20L- SUSPECT DVA8-F20L).....	258
Figure 57. Statistically significant differences in [s]' noise duration across the Dutch suspect's voice samples (DVA8-F20L- SUSPECT DVA8-F20L).....	260
Figure 58. Between-speaker variation of release burst intensity in [k, p, t] across English voice samples.....	276
Figure 59. Between-speaker variation of [s] and [z] spectral peak location across English voice samples.....	279
Figure 60. Between-speaker variation of [s] COG across English voice samples.....	280
Figure 61. Between-speaker variation of noise amplitude in [s] and [z] across English voice samples.....	281
Figure 62. Between-speaker variation of F1 values in [s] across English voice samples.....	282
Figure 63. Between-speaker variation of F2 values in [s] across English voice samples.....	283
Figure 64. Between-speaker variation of F2 values in [z] across English voice samples.....	284
Figure 65. Between-speaker variation of F3 values in [s] across English voice samples.....	285
Figure 66. Pairwise comparison of release burst intensity values in voiced/voiceless plosives across Spanish recordings.....	287
Figure 67. Between-speaker variation of [s] noise amplitude across Spanish voice samples.....	289
Figure 68. Between-speaker variation of F2 values in [s] across Spanish voice samples.....	290
Figure 69. Between-speaker variation of release burst intensity in [b, d, g] and [k, p, t] across Dutch recordings.....	293
Figure 70. Between-speaker variation of noise amplitude in [s] across Dutch recordings.....	295
Figure 71. Between-speaker variation of F1 values in [s] across Dutch recordings.....	296
Figure 72. Between-speaker variation of F3 values in [s] across Dutch recordings.....	297

Chapter 11- Lists of figures and tables

Figure 73. English discrimination language test's responses in British and Spanish jurors (target-present condition without background noises).....	304
Figure 74. English discrimination language test's responses in British and Spanish jurors (target-absent condition with background noises).....	305
Figure 75. Spanish discrimination language test's responses in British and Spanish jurors (target-present condition without background noises).....	307
Figure 76. Spanish discrimination language test's responses in British and Spanish jurors (target-absent condition with background noises).....	308
Figure 77. Dutch discrimination language test's responses in British and Spanish jurors (target-present condition without background noises).....	310
Figure 78. Dutch discrimination language test's responses in British and Spanish jurors (target-absent condition with background noises).....	311

11.2. LIST OF TABLES:

Table 1. Objects of study.....	29
Table 2. Summary of the planned objectives.....	34
Table 3. Hypotheses considered for each analytical stage.....	38
Table 4. Sociolinguistic issues worth addressing according to Hickey (2014: 20).....	48
Table 5. Scale of opinions in reporting authorship identification results (Coulthard 2010: 480)....	60
Table 6. Types of memory recall according to Manzanero (2006: 405-407).....	74
Table 7. Variables that influence the witness' identification ability. Adapted from Manzanero & González (2015: 132).....	79
Table 8. Perception language tests and experimental conditions tested on groups of jurors, along with the resulting test score per correct answer.....	102
Table 9. Variables provided by online perception surveys and their subsequent coding for statistical processing.....	104
Table 10. Types of association depending on ϕ or Cramer's V values. Adapted from Rea & Parker (1992: 203).....	112
Table 11. Categorical variables and their assigned codes for hypothesis 1.....	134
Table 12. Friedman's two-way analysis on <i>language1</i> and <i>response1</i> for hypothesis 1 (British group, identification tests).....	136
Table 13. Chi-square test for hypothesis 1 (British group, identification tests, 1st exp. condition).....	136
Table 14. Cramer's V and Phi coefficient for hypothesis 1 (British group, identification tests, 1st exp. condition).....	137
Table 15. Contingency table for hypothesis 1 (British group, identification tests, 1st exp. condition).....	137
Table 16. Friedman's two-way analysis on <i>language2</i> and <i>response2</i> for hypothesis 1 (British group, identification tests).....	139
Table 17. Chi-square test for hypothesis 1 (British group, identification tests, 2nd exp. condition).....	139
Table 18. Cramer's V and Phi coefficient for hypothesis 1 (British group, identification tests, 2nd exp. condition).....	140
Table 19. Contingency table for hypothesis 1 (British group, identification tests, 2nd exp. condition).....	140
Table 20. Friedman's two-way analysis on <i>language1</i> and <i>response1</i> for hypothesis 1 (Spanish group, identification tests).....	142
Table 21. Chi-square test for hypothesis 1 (Spanish group, identification tests, 1st exp. condition).....	142
Table 22. Cramer's V and Phi coefficient for hypothesis 1 (Spanish group, identification tests, 1st exp. condition).....	143
Table 23. Contingency table for hypothesis 1 (Spanish group, identification tests, 1st exp. condition).....	144
Table 24. Friedman's two-way analysis on <i>language2</i> and <i>response2</i> for hypothesis 1 (Spanish group, identification tests).....	146

Chapter 11- Lists of figures and tables

Table 25. Chi-square test for hypothesis 1 (Spanish group, identification tests, 2nd exp. condition)....	146
Table 26. Cramer's V and Phi coefficient for hypothesis 1 (Spanish group, identification tests, 2nd exp. condition).....	147
Table 27. Contingency table for hypothesis 1 (Spanish group, identification tests, 2nd exp. condition).....	147
Table 28. Friedman's two-way analysis on <i>All.language1</i> and <i>All.response1</i> for hypothesis 1 (British and Spanish group, identification tests).....	149
Table 29. Chi-square test for hypothesis 1 (British and Spanish group, identification tests, 1st exp. condition).....	149
Table 30. Cramer's V and Phi coefficient for hypothesis 1 (British and Spanish group, identification tests, 1st exp. condition).....	150
Table 31. Contingency table for hypothesis 1 (British and Spanish group, identification tests, 1st exp. condition).....	150
Table 32. Friedman's two-way analysis on <i>All.language2</i> and <i>All.response2</i> for hypothesis 1 (British and Spanish group, identification tests).....	152
Table 33. Chi-square test for hypothesis 1 (British and Spanish group, identification tests, 2nd exp. condition).....	152
Table 34. Cramer's V and Phi coefficient for hypothesis 1 (British and Spanish group, identification tests, 2nd exp. condition).....	153
Table 35. Contingency table for hypothesis 1 (British and Spanish group, identification tests, 2nd exp. condition).....	153
Table 36. Friedman's two-way analysis on <i>Dis.language</i> and <i>Dis.response</i> for hypothesis 1 (British group, discrimination tests).....	155
Table 37. Chi-square test for hypothesis 1 (British group, discrimination tests, 1st exp. condition).....	155
Table 38. Cramer's V and Phi coefficient for hypothesis 1 (British group, discrimination tests, 1st exp. condition).....	156
Table 39. Contingency table for hypothesis 1 (British group, discrimination tests, 1st exp. condition).....	156
Table 40. Friedman's two-way analysis on <i>Dis.language</i> and <i>Dis.response</i> for hypothesis 1 (Spanish group, discrimination tests).....	158
Table 41. Chi-square test for hypothesis 1 (Spanish group, discrimination tests, 1st exp. condition)....	158
Table 42. Cramer's V and Phi coefficient for hypothesis 1 (Spanish group, discrimination tests, 1st exp. condition).....	159
Table 43. Contingency table for hypothesis 1 (Spanish group, discrimination tests, 1st exp. condition).....	159
Table 44. Friedman's two-way analysis on <i>Dis.language</i> and <i>Dis.response</i> for hypothesis 1 (British and Spanish group, discrimination tests).....	161

Chapter 11- Lists of figures and tables

Table 45. Chi-square test for hypothesis 1 (British and Spanish group, discrimination tests, 1st exp. condition).....	161
Table 46. Cramer's V and Phi coefficient for hypothesis 1 (British and Spanish group, discrimination tests, 1st exp. condition).....	162
Table 47. Contingency table for hypothesis 1 (British and Spanish group, discrimination tests, 1st exp. condition).....	162
Table 48. Numerical variables assigned for each language test in identification and discrimination tasks for hypothesis 2.....	166
Table 49. Descriptive statistics of each language test in identification and discrimination tasks for hypothesis 2 (British group, 1st exp. condition).....	167
Table 50. Wilcoxon-signed ranks test for each pair of language test in identification and discrimination tasks for hypothesis 2 (British group, 1st exp. condition).....	167
Table 51. Descriptive statistics of each language test in identification and discrimination tasks for hypothesis 2 (British group, 2nd exp. condition).....	168
Table 52. Wilcoxon-signed ranks test for each pair of language test in identification and discrimination tasks for hypothesis 2 (British group, 2nd exp. condition).....	168
Table 53. Descriptive statistics of each language test in identification and discrimination tasks for hypothesis 2 (Spanish group, 1st exp. condition).....	169
Table 54. Wilcoxon-signed ranks test for each pair of language test in identification and discrimination tasks for hypothesis 2 (Spanish group, 1st exp. condition).....	170
Table 55. Descriptive statistics of each language test in identification and discrimination tasks for hypothesis 2 (Spanish group, 2nd exp. condition).....	170
Table 56. Wilcoxon-signed ranks test for each pair of language test in identification and discrimination tasks for hypothesis 2 (Spanish group, 2nd exp. condition).....	171
Table 57. Numerical variables considered in hypothesis 3: test scores and confidence levels.....	172
Table 58. Mean test scores and confidence levels in the British group (identification tasks).....	176
Table 59. Mean test scores and confidence levels in the Spanish group (identification tasks).....	177
Table 60. Mean test scores and confidence levels in the British group (discrimination tasks).....	182
Table 61. Mean test scores and confidence levels in the Spanish group (discrimination tasks).....	182
Table 62. Dependent and independent variables considered for the fourth hypothesis.....	185
Table 63. Distribution of age and gender across British jurors.....	187
Table 64. Estimates of age and gender in L2's scores (British group, identification tests).....	188
Table 65. Distribution of age and gender across Spanish jurors.....	189
Table 66. Estimates of age and gender in L1's scores (Spanish group, identification tests).....	190
Table 67. Estimates of age and gender in F2's scores (Spanish group, identification tests).....	192
Table 68. Distribution of age and gender across British and Spanish jurors.....	193
Table 69. Estimates of age and gender in F2's scores (British and Spanish group, identification tests).....	194
Table 70. Estimates of age and gender in L2's scores (British and Spanish group, identification tests).....	196

Chapter 11- Lists of figures and tables

Table 71. Descriptive statistics of British discrimination tests, 1st exp. condition.....	198
Table 72. Descriptive statistics of Spanish discrimination tests, 1st exp. condition.....	199
Table 73. Descriptive statistics of British and Spanish discrimination tests, 1st exp. condition....	199
Table 74. Dummy codes assigned for each variable considered in hypothesis 5.....	202
Table 75. Mann-Whitney U test on the influence of country (cultural groups) upon identification scores	203
Table 76. Mann-Whitney U test on the influence of country (cultural groups) upon discrimination scores	205
Table 77. Mann-Whitney U test on the influence of linguistic environment upon identification scores in the British group.....	206
Table 78. Mann-Whitney U test on the influence of linguistic environment upon identification scores in the Spanish group.....	206
Table 79. Mann-Whitney U test on the influence of linguistic environment upon discrimination scores in the British group.....	207
Table 80. Mann-Whitney U test on the influence of linguistic environment upon discrimination scores in the Spanish group.....	207
Table 81. List of variables (and their assigned codes) contemplated for hypothesis 6.....	210
Table 82. Pairwise comparisons on identification language tests' scores across two experimental conditions in the British group.....	211
Table 83. Pairwise comparisons on identification language tests' scores across two experimental conditions in the Spanish group.....	213
Table 84. Pairwise comparisons on discrimination language tests' scores across two experimental conditions in the British group.....	215
Table 85. Pairwise comparisons on discrimination language tests' scores across two experimental conditions in the Spanish group.....	217
Table 86. Distribution of age, gender, and studies across British jurors.....	219
Table 87. Estimates of age, gender, and studies in L2's scores (British group, identification tests).....	221
Table 88. Distribution of age, gender, and studies across Spanish jurors.....	223
Table 89. Estimates of age, gender, and studies in L1's scores (Spanish group, identification tests)....	224
Table 90. Estimates of age, gender, and studies in F2's scores (Spanish group, identification tests).....	225
Table 91. Distribution of age, gender, and studies across Spanish jurors.....	227
Table 92. Estimates of age, gender, and studies in F2's scores (British and Spanish group, identification tests).....	227
Table 93. Estimates of age, gender, and studies in L2's scores (British and Spanish group, identification tests).....	229
Table 94. Estimates of age, gender, and studies in F1.Dis' scores (Spanish group, discrimination tests)	231
Table 95. Findings in the original study and its extended version (epilogue).....	233

Chapter 11- Lists of figures and tables

Table 96. Selected variables for hypothesis 7 testing, displaying the sub-types of both independent (left column) and dependent (right column) variables.....	239
Table 97. Within-speaker variation of English voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott) in terms of pitch-related measurements.....	241
Table 98. Within-speaker variation of English voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott) in terms of pauses.....	242
Table 99. Within-speaker variation of Spanish voice samples (M12_020- SUSPECT M12_020) in terms of pitch-related measurements.....	243
Table 100. Within-speaker variation of Spanish voice samples (M12_020- SUSPECT M12_020) in terms of pauses.....	243
Table 101. Within-speaker variation of Dutch voice samples (DVA8-F20L- SUSPECT DVA8-F20L) in terms of pitch-related measurements.....	245
Table 102. Within-speaker variation of Dutch voice samples (DVA8-F20L- SUSPECT DVA8-F20L) in terms of pauses.....	247
Table 103. Within-speaker variation of variables within voiced/voiceless plosives in English voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).....	249
Table 104. Within-speaker variation of variables within voiced/voiceless alveolar sibilants in English voice samples (Simon T. Elliott- SUSPECT Simon T. Elliott).....	250
Table 105. Within-speaker variation of variables within voiced/voiceless plosives in Spanish voice samples (M12_020- SUSPECT M12_020).....	253
Table 106. Within-speaker variation of variables within voiceless alveolar sibilants in Spanish voice samples (M12_020- SUSPECT M12_020).....	256
Table 107. Within-speaker variation of variables within voiced/voiceless plosives in Dutch voice samples (DVA8-F20L- SUSPECT DVA8-F20L).....	257
Table 108. Within-speaker variation of variables within voiceless alveolar sibilants in Dutch voice samples (DVA8-F20L- SUSPECT DVA8-F20L).....	259
Table 109. Within-speaker variation of suprasegmental parameters across English, Spanish, and Dutch informants.....	261
Table 110. Within-speaker variation of segmental parameters across English, Spanish, and Dutch informants.....	262
Table 111. Selected variables for hypothesis 8 testing, displaying the sub-types of both independent (left column) and dependent (right column) variables.....	264
Table 112. Between-speaker variation of suprasegmental parameters across English voice samples.....	267
Table 113. Between-speaker variation of suprasegmental parameters across Spanish voice samples.....	269
Table 114. Between-speaker variation of suprasegmental parameters across Dutch voice samples.....	271
Table 115. Between-speaker variation of variables within voiced/voiceless plosives in English voice samples.....	274

Chapter 11- Lists of figures and tables

Table 116. Between-speaker differences amongst English recordings regarding VOT in voiceless and voiced plosives according to a Kurskal-Wallis test.....	274
Table 117. Between-speaker differences amongst English recordings regarding release burst intensity in voiceless and voiced plosives according to a Kurskal-Wallis test.....	275
Table 118. Between-speaker variation of variables within voiced/voiceless alveolar sibilants in English voice samples.....	278
Table 119. Between-speaker variation of variables within voiced/voiceless plosives in Spanish voice samples.....	286
Table 120. Between-speaker differences amongst Spanish recordings regarding VOT in voiceless and voiced plosives according to a Kurskal-Wallis test.....	286
Table 121. Between-speaker differences amongst Spanish recordings in voiced plosives' VOT according to a Kurskal-Wallis test.....	287
Table 122. Between-speaker variation of variables within voiceless alveolar sibilants in Spanish voice samples.....	288
Table 123. Between-speaker variation of variables within voiced/voiceless plosives in Dutch voice samples.....	291
Table 124. Between-speaker differences in voiced/voiceless plosives' VOT values amongst Dutch recordings according to a Kurskal-Wallis test.....	292
Table 125. Between-speaker variation of variables within voiceless alveolar sibilants in Dutch voice samples.....	294
Table 126. Between-speaker variation of noise duration in [s] across Dutch voice samples according to a Kruskal-Wallis test.....	294
Table 127. Efficient suprasegmental features across groups of informants and research criteria (between-speaker variation).....	298
Table 128. Efficient segmental variables, number of significant pairwise comparisons, and compliance with research criterion n° 2 (between-speaker variation).....	299
Table 129. Number of dissimilarities found amongst English foil speakers (and total number of differences per speaker) in relation to segmental and suprasegmental features.....	303
Table 130. Number of dissimilarities found between English foil speakers and the selected suspect in relation to segmental and suprasegmental features.....	304
Table 131. Number of dissimilarities found amongst Spanish foil speakers (and total number of differences per speaker) in relation to segmental and suprasegmental features.....	306
Table 132. Number of dissimilarities found between Spanish foil speakers and the selected suspect in relation to segmental and suprasegmental features.....	306
Table 133. Number of dissimilarities found amongst Dutch foil speakers (and total number of differences per speaker) in relation to segmental and suprasegmental features.....	309
Table 134. Number of dissimilarities found between Dutch foil speakers and the selected suspect in relation to segmental and suprasegmental features.....	309
Table 135. Success rates in identification and discrimination tasks across voice samples and group of jurors.....	312

Chapter 11- Lists of figures and tables

Table 136. Error rates in speaker recognition tests and in acoustic-phonetic analysis across voice samples
(1st experimental condition).....313