

ALTERAÇÕES ACÚSTICAS E PERCETIVAS INTRODUZIDAS NAS VOZES DE INDIVÍDUOS GÉMEOS E DEVIDAS AO CANAL TELEFÓNICO – UMA DISCUSSÃO DE IMPACTO NA ANÁLISE FORENSE

Vânia Sofia da Costa Fernandes

D

2018



Vânia Sofia da Costa Fernandes

**ALTERAÇÕES ACÚSTICAS E PERCETIVAS
INTRODUZIDAS NAS VOZES DE INDIVÍDUOS GÉMEOS E
DEVIDAS AO CANAL TELEFÓNICO – UMA DISCUSSÃO DE
IMPACTO NA ANÁLISE FORENSE**

Tese realizada no âmbito do Doutoramento em Ciências da Linguagem – Tecnologias da
Linguagem Humana, orientada pelo Professor Doutor Aníbal João de Sousa Ferreira

Faculdade de Letras da Universidade do Porto

Outubro de 2018

Alterações acústicas e percetivas introduzidas nas vozes de
indivíduos gémeos e devidas ao canal telefónico – uma
discussão de impacto na análise forense

Vânia Sofia da Costa Fernandes

Tese realizada no âmbito do Doutoramento em Ciências da Linguagem – Tecnologias da
Linguagem Humana, orientada pelo Professor Doutor Aníbal João de Sousa Ferreira

Júri

PRESIDENTE:

Doutora Ana Maria Barros de Brito, Professora Catedrática do Departamento de Estudos Portugueses e Estudos Românicos da Faculdade de Letras da Universidade do Porto.

VOGAIS:

Doutor Fernando da Assunção Martins, Professor Auxiliar da Faculdade de Letras da Universidade de Lisboa;

Doutor Fernando Santos Perdigão, Professor Auxiliar do Departamento de Engenharia Eletrotécnica e Computadores da Universidade de Coimbra;

Engenheiro André Ferraz Saraiva, Laboratório de Polícia, Polícia Judiciária de Lisboa;

Doutor Aníbal João de Sousa Ferreira, Professor Associado da Faculdade de Engenharia da Universidade do Porto (Orientador);

Doutora Belinda Mary Harper Sousa Maia, Professora Associada aposentada da Faculdade de Letras da Universidade do Porto.

À minha querida Mãe

Sumário

Declaração de honra	11
Agradecimentos	12
Resumo	13
Abstract	15
Índice de ilustrações	17
Índice de tabelas	20
Lista de abreviaturas e siglas.....	23
Capítulo 1 – Introdução.....	26
1.1 Contexto	26
1.2 Motivação	28
1.3 Objetivos	29
1.4 Contribuições	31
1.5 Estrutura da dissertação	33
Capítulo 2 – Enquadramento estrutural	36
2.1 Introdução.....	36
2.2 Sistema humano de produção de fala.....	36
2.2.1 O aparelho fonador	37
2.2.2 Características físicas e percetivas da voz	44
2.2.3 O aparelho fonador de indivíduos gémeos	46
2.2.3.1 Gémeos dizigóticos e monozigóticos	47
2.2.3.2 Características das vozes de indivíduos gémeos	48
2.3 Identificação do orador	51
2.3.1 Parâmetros percetivos e parâmetros acústicos	53
2.3.2 Análise percetiva	57
2.3.3 Análise acústica	58
2.3.4 Reconhecimento automático do falante.....	59
2.3.5 Condições ideais e condições reais	62
2.4 Fonética forense.....	64
2.4.1 Âmbito e expressão em Portugal	64

2.4.2 A pertinência do estudo de vozes de gémeos.....	66
2.5 Conclusão	67
Capítulo 3 – Planeamento e estruturação da base de dados.....	70
3.1 Introdução	70
3.2 Cenário para recolha de dados	70
3.3 Tipos de canal e modalidade de gravação	74
3.4 Questionários	77
3.5 Participantes	78
3.6 Base de dados das gravações	81
3.6.1 Base de dados inicial	81
3.6.2 Isolamento de regiões de vogais sustentadas.....	83
3.6.3 Remoção do segundo orador e conversão da frequência de amostragem.....	84
3.6.4 Subdivisão das gravações na base de dados	85
3.7 Conclusão	89
Capítulo 4 – Ambientes e ferramentas de processamento.....	91
4.1 Introdução.....	91
4.2 <i>Software Audacity e Adobe Audition</i>	91
4.3 <i>Software Praat</i>	93
4.4 <i>Software iVocalise</i>	95
4.5 <i>Software Biometrics</i>	98
4.6 <i>Software VoiceStudio</i>	103
4.7 Análise estatística.....	105
4.8 Conclusão	107
Capítulo 5 – Estudos com base em parâmetros acústicos	109
5.1 Introdução	109
5.2 Vogais e parâmetros acústicos seleccionados	109
5.2.1 Vogais sustentadas.....	110
5.2.2 Parâmetros acústicos relevantes na comparação de vozes	113
5.3 Análise através de parâmetros acústicos	115
5.3.1 Análise com base na frequência fundamental	115
5.3.2 Análise com base no desvio padrão da frequência fundamental	120
5.3.3 Análise com base no parâmetro <i>jitter</i>	122
5.3.4 Análise com base no parâmetro <i>shimmer</i>	125
5.3.5 Análise com base no parâmetro HNR	127

5.3.6 Análise com base nas frequências das formantes.....	132
5.4 Conclusão	148
Capítulo 6 – Estudos com base no reconhecimento automático de orador	153
6.1 Introdução	153
6.2 Pré-condicionamento dos dados.....	160
6.3 Resultados previamente consolidados.....	162
6.4 Segmentação e recomposição da base de dados.....	166
6.5 Testes 1h-1h + 2h-2h.....	168
6.6 Testes 1h-2h.....	170
6.7 Testes usando diversidade de canal.....	174
6.8 Testes com assimetria na influência de canal	180
6.9 Conclusão	185
Capítulo 7 – Caso de estudo.....	188
7.1 Introdução	188
7.2 Trabalho de peritagem	188
7.3 Conclusão	202
Capítulo 8 - Conclusões	204
8.1 Síntese dos estudos realizados.....	204
8.2 Principais contribuições	206
8.3 Trabalho futuro	207
Referências bibliográficas	209
Anexos.....	226
Anexo 1	227
Anexo 2	228
Anexo 3	229
Anexo 4	232
Anexo 5	235
Anexo 6	239
Anexo 7	240

Declaração de honra

Declaro que a presente dissertação é de minha autoria e não foi utilizado previamente noutro curso ou unidade curricular, desta ou de outra instituição. As referências a outros autores (afirmações, ideias, pensamentos) respeitam escrupulosamente as regras da atribuição, e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas de referência. Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Porto, 4 de outubro de 2018

Vânia Sofia da Costa Fernandes

Agradecimentos

Quero expressar os meus sinceros agradecimentos ao Professor Aníbal Ferreira pela transmissão de conhecimento ao longo de todo este processo, pela prontidão incansável em abrir horizontes de aprendizagem, pela empatia, pelos conselhos, pela motivação e pelas palavras de alento nos momentos menos resplandecentes. Bem haja!

Agradeço, também, ao Professor João Veloso pelo ensinamento dos conceitos basilares de Fonética necessários para a redação da tese e à Professora Ana Maria Brito pela gentileza e palavras simpáticas e motivantes.

Desejo agradecer, da mesma forma, aos meus pais, mas, especialmente, à minha Mãe, a quem dedico esta dissertação, pelo incentivo que me deu, desde que me conheço, a sonhar mais alto com tenacidade, encorajando-me, sempre, a estudar, a aprofundar os meus conhecimentos e, acima de tudo, a ser feliz.

Ao meu namorado António Romualdo Carvalho, pelo amor, carinho, ombro amigo, palavras de motivação, risadas, pelo caminho percorrido, juntos, e pelo auxílio com o desespero que o *Word* pode ser, por vezes.

Muito obrigada a todos os participantes do estudo – gémeos e familiares. Sem a vossa disponibilidade, não seria possível.

Agradeço, ainda, a todos os amigos e colegas de trabalho que acompanharam o desenvolvimento deste trabalho e aos que partilharam comigo a saga que uma dissertação de doutoramento pode ser.

Resumo

A gravação de sinais de voz telefónica, em situações de chamadas de emergência ou de investigação criminal, é de extrema importância, não só para permitir dar resposta adequada a pedidos de socorro, como para a realização de estudos periciais que deem resposta a suspeitas de foro criminal. Entre estas, inclui-se a necessidade frequente de constituir prova da correspondência de vozes entre dois ou mais registos telefónicos. Porém, este tipo de investigação depara-se com dificuldades técnicas relacionadas, principalmente, com a qualidade dos registos telefónicos e com a duração desses registos. Por outro lado, a prova de correspondência, não sendo nunca absoluta e definitiva, pode ser baseada na opinião de um painel de ouvintes treinados em tarefas de linguística e reconhecimento de vozes, ou pode ser fundamentada em análise por computador e medidas acústicas quantificadas, como, por exemplo, os designados *likelihood ratios* (razões de verosimilhança).

É neste âmbito que se pretende aferir o impacto da degradação que afeta os registos de voz em várias configurações de equipamento e canal de comunicação (*e.g.* VoIP e GSM). Este objetivo exigiu a construção de uma base de dados de registos de voz, com características peculiares, designadamente quanto à diversidade de oradores incluindo indivíduos gémeos e familiares próximos, e quanto às condições de captação de registos de voz incluindo duas versões de alta qualidade e duas versões de qualidade telefónica. Foram ensaiadas estratégias não-automáticas e automáticas de análise e comparação de registos de voz que permitiram estudar o impacto da degradação devida ao canal de comunicação em tarefas de comparação de vozes e identificação de orador. As estratégias não-automáticas permitiram sobretudo investigar a relevância de diversos parâmetros acústicos em face das diferentes condições de canal, enquanto que as estratégias de reconhecimento automático de orador permitiram sobretudo investigar o impacto das condições de canal de comunicação na eficácia da construção de modelos de orador e na eficiência da identificação de orador. Por último, é discutido um caso de estudo real em que estivemos envolvidos.

Palavras-chave: Reconhecimento de orador, Comunicação GSM, Comunicação VoIP, Fonética Forense, Análise Acústica da Fala

Abstract

The recording of telephone voice signals in situations of emergency calls or criminal investigation is extremely important, not only in order to provide an adequate response to distress calls, but also to conduct expert studies that respond to criminal suspicions. These include the frequent need to provide evidence of voices matching between two or more telephone recordings. However, this type of investigation encounters technical difficulties related, in particular, to the quality and the duration of such telephone recordings. On the other hand, the proof of correspondence, never being absolute and definitive, can be based on the opinion of a panel of listeners trained in linguistic tasks and voice recognition, or it can be based on computer analysis and quantified acoustic measures like the so-called likelihood ratios.

It is within this scope that the degree of degradation that affects telephone voice registers in various equipment configurations (*e.g.* VoIP and GSM communications) is assessed, and it is mainly intended to study the impact of this degradation on voice comparison tasks, with the intention of identifying records of the same speaker, through the voices of twin individuals and their close relatives, using statistical data processing and cross-checking the results of acoustic tests obtained through software. The main goal is to verify the level of success in voices matching with VoIP and GSM communications and to determine which are the most relevant acoustic parameters in speaker identification.

It is within this scope that we intend to assess the impact caused by the degradation affecting voice records in various equipment and communication channel configurations (*e.g.* VoIP and GSM). This objective required the construction of a database of voice records with peculiar characteristics, namely with regard to the diversity of speakers including twins and relative individuals, as well as the voice recording conditions and including two high quality versions and two phone-quality versions. Non-automatic and automatic voice analysis and comparison strategies were tested to study the impact of degradation due to the communication channel on voice

comparison and speaker identification tasks. The non-automatic strategies allowed us to investigate the relevance of different acoustic parameters with respect to different channel conditions, while the strategies of automatic speaker recognition allowed us to investigate the impact of communication channel conditions on the robustness of speaker models and the efficiency of speaker identification. Finally, we discuss a case study in which we were involved.

Keywords: Speech Recognition, GSM Communication, VoIP Communication, Forensic Phonetics, Speech Acoustic Analysis

Índice de ilustrações

Figura 1 - Comunicação GSM.	30
Figura 2 - Comunicação VoIP.	30
Figura 3 - O aparelho fonador (adaptado de [12]).	38
Figura 4 - Visão anterior, posterior e de topo da laringe [14].	39
Figura 5 - Pregas vocais abduzidas (sons surdos - esquerda) e pregas vocais aduzidas (sons sonoros - direita) (adaptado de [16]).	40
Figura 6 - Corte sagital do trato vocal [17].	41
Figura 7 - Teoria Fonte-Filtro [18].	42
Figura 8 - Diferenças físicas entre os aparelhos fonadores do macaco e do homem [27].	43
Figura 9 - Diferenças anatómicas entre o trato vocal de uma criança (esquerda) e o de um adulto (direita) (adaptado de [28]).	44
Figura 10 - Esquema da formação embrionária gemelar: gémeos dizigóticos (esquerda) e gémeos monozigóticos (direita) [36].	48
Figura 11 - Representação da ligação na recolha de amostras de fala com modalidade de comunicação GSM, em ambos os tipos de gravação: HQ e TL.	75
Figura 12 - Representação da ligação na recolha de amostras de fala com modalidade de comunicação VoIP, em ambos os tipos de gravação: HQ e TL.	76
Figura 13 - Datas de nascimento dos oradores do género feminino, compreendidas entre décadas.	78
Figura 14 - Datas de nascimento dos oradores do género masculino, compreendidas entre décadas.	80
Figura 15 - O painel superior, nesta figura, representa a onda osciloscópica original, na emissão, e o painel inferior representa a onda osciloscópica na receção e denotando a presença da voz do interlocutor.	84
Figura 16 - O painel superior, nesta figura, representa registo espectrográfico original, na emissão, e o painel inferior representa o registo espectrográfico na receção e denotando a presença da voz do interlocutor.	85
Figura 17 - Condições de gravação para cada orador, contemplando os dois tipos de canal (HQ e TL) e as duas modalidades de comunicação (GSM e VoIP).	86
Figura 18 - Operação de divisão dos registos existentes na base de dados inicial.	87
Figura 19 - Esquema dos modelos de orador construídos com base em registos de fala não concomitantes.	87
Figura 20 - Método de construção de seis novas bases de dados.	88
Figura 21 - Interface do Audacity (análise acústica da vogal sustentada /a/ de FEM01, em GSM_HQ).	92
Figura 22 - Interface do Adobe Audition (análise acústica da vogal sustentada /a/ de FEM01, em GSM_HQ).	93
Figura 23 - Interface do Praat (análise acústica da vogal sustentada /a/ de FEM01, em registo GSM_HQ).	94
Figura 24 - Interface do iVocalise [91].	96
Figura 25 - Representação do método de obtenção do Likelihood Ratio [64].	100
Figura 26 - Resultados do iVocalise no LR Plot [90].	101

Figura 27 - Resultados do iVocalise no EEG [90].	102
Figura 28 - Resultados do iVocalise no DET Plot [90].	103
Figura 29 - Interface do VoiceStudio (análise acústica de uma vogal de uma oradora feminina).	104
Figura 30 – Representação na onda osciloscópica e no espectrograma da vogal sustentada /a/ de uma oradora, dos 0 aos 0.30 segundos.	111
Figura 31 - Ilustração da influência das frequências formantes sobre o espectro do sinal de fonte laríngea [18].	114
Figura 32 - Resultados da média estimada de F0, incluindo os intervalos de confiança de 95%.	116
Figura 33 - Média da F0 da vogal /u/ dos 19 oradores, no Praat.	119
Figura 34 - Resultados estimados do desvio padrão da F0, incluindo os intervalos de confiança de 95%.	120
Figura 35 - Resultados estimados do jitter PPQ5, incluindo os intervalos de confiança de 95%.	123
Figura 36 - Resultados estimados do shimmer APQ5, incluindo o intervalo de confiança de 95%.	125
Figura 37 - Médias e intervalos de confiança de 95% do shimmer entre os pares TL e HQ, no Praat e no VoiceStudio.	127
Figura 38 - Resultados estimados do HNR, incluindo os intervalos de confiança de 95%.	128
Figura 39 - Resultados do HNR da vogal /u/ no VoiceStudio.	129
Figura 40 - Média e intervalo de confiança (95%) para a diferença de valores HNR obtidos pelo Praat e VoiceStudio.	130
Figura 41 - Correlação entre o Praat e o VoiceStudio para os diferentes parâmetros acústicos (n=20), a média e o intervalo de confiança de 95%.	131
Figura 42 - Média da F1 e respetivo intervalo de confiança de 95% (oradores femininos).	135
Figura 43 - Média da F1 e respetivo intervalo de confiança de 95% (oradores masculinos).	135
Figura 44 - Médias dos desvios relativos de F1 entre as versões TL e HQ. Os resultados com significância estatística assinalam-se a cor sólida (azul).	136
Figura 45 - Média da F2 e respetivo intervalo de confiança de 95% (oradores femininos).	137
Figura 46 - Média da F2 e respetivo intervalo de confiança de 95% (oradores masculinos).	137
Figura 47 - Médias dos desvios relativo de F2 entre as versões TL e HQ. Os resultados com significância estatística assinalam-se a cor sólida (azul).	138
Figura 48 - Média da F3 e respetivo intervalo de confiança de 95% (oradores femininos).	139
Figura 49 - Média da F3 e respetivo intervalo de confiança de 95% (oradores masculinos).	139
Figura 50 - Médias dos desvios relativos de F3 entre as versões TL e HQ. Os resultados com significância estatística assinalam-se a cor sólida (azul).	140
Figura 51 - Média da F4 e respetivo intervalo de confiança de 95% (oradores femininos).	141
Figura 52 - Média da F4 e respetivo intervalo de confiança de 95% (oradores masculinos).	142
Figura 53 - Médias dos desvios relativos de F4 entre as versões TL e HQ. Os resultados com significância estatística assinalam-se a cor sólida (azul).	142
Figura 54 - Comparação entre as médias de F1 e de F2 (oradores femininos). A linha contínua (azul) exprime os resultados do nosso estudo e a interrompida (castanho) representa resultados de [94].	143
Figura 55 - Comparação entre as médias de F1 e de F2 (oradores masculinos). A linha contínua (azul) exprime os resultados do nosso estudo e a interrompida (castanho) representa resultados de [94].	144

Figura 56 - Resultados dos desvios de F1 e de F2 no mapa dos triângulos acústico das vogais (oradores femininos). (GSM HQ – linha pontilhada redonda; GSM TL – linha sólida redonda; VoIP HQ – linha pontilhada quadrada; VoIP TL – linha sólida quadrada).	145
Figura 57 - Resultados dos desvios de F1 e de F2 no mapa dos triângulos acústico das vogais (oradores masculinos). (GSM HQ – linha pontilhada redonda; GSM TL – linha sólida redonda; VoIP HQ – linha pontilhada quadrada; VoIP TL – linha sólida quadrada).	146
Figura 58 - Ilustração da comparação entre dois oradores em que o sinal de voz do orador de teste é classificado de acordo com um modelo estatístico de um orador conhecido.	156
Figura 59 - Ilustração da comparação entre dois oradores em que são primeiro extraídos os seus modelos estatísticos sendo posteriormente comparados para se obter um score de similaridade.	157
Figura 60 - Diagrama de blocos da estrutura e funcionamento do software iVocalise.	160
Figura 61 - Ilustração da operação do pré-processamento VAD. O painel superior representa um sinal de voz em que as regiões de voz ativa são assinaladas pelo detetor VAD, neste caso através de linhas de delimitação de cor magenta. Por último, o processador VAD remove as regiões de silêncio e ruído de fundo de baixa intensidade, preservando somente as de voz ativa, tal como se ilustra no painel inferior.	162
Figura 62 - Resultados de desempenho na identificação de orador, para cada uma das seis situações de teste, quando se inclui, ou não, pré-processamento VAD, e quando se considera, ou não, a remoção do segundo orador (situação em que os ficheiros são modificados).	164
Figura 63 - Alternativas de combinação de duas metades dos registos de voz que representam tipos de canal diferentes e respeitam o objetivo de mútua exclusividade dos registos voz nos estudos de comparação. Quaisquer duas combinações assinaladas com 'X' são viáveis, exceto as seis associações assinaladas através de grafismo oval.	175
Figura 64 - Scores obtidos entre o registo XX01_GSM_TL e todas as amostras restantes de índole telefónica.	193
Figura 65 - Scores obtidos entre o registo XX02_GSM_TL e todas as amostras restantes de índole telefónica [52].	194
Figura 66 - Curvas de densidade de probabilidade para as hipóteses H0 e H1, tendo em conta o conjunto de 17 oradores da população relevante (esquerda) e a base de dados alargada de 19 oradores (direita) [52].	195
Figura 67 - Valores do desempenho EER tendo em conta o conjunto de 17 oradores da “população relevante” (esquerda) e tendo em conta a base de dados alargada de 19 oradores (direita).	195
Figura 68 - Valores estimados da frequência fundamental em ambos os registos telefónicos do orador conhecido (OS01, OS02) e do orador desconhecido (XX01, XX02), juntamente com a representação do intervalo de confiança a 95% para o valor médio encontrado.	198
Figura 69 - Espectrogramas correspondentes a uma ocorrência da palavra “carro” no registo de voz telefónica OS02_GSM_TL (painel superior) e no registo de voz telefónica XX01_GSM_TL (painel inferior).	199
Figura 70 - Espectrogramas correspondentes a uma só ocorrência da palavra “sete” em cada um dos dois registos telefónicos do orador OS (painéis superiores) e em cada um dos dois registos telefónicos do orador XX (painéis inferiores).	200

Índice de tabelas

Tabela 1 - Dados relativos à idade e ao grau de parentesco dos oradores.	82
Tabela 2 - Duração, em minutos e segundos, dos quatro registos de cada orador [52].	82
Tabela 3 - Idade e graus de parentesco dos oradores cujas vogais sustentadas foram analisadas acusticamente.	112
Tabela 4 - Correlações normalizadas da média da F0 (n=19, todos $p < 0.05$).	117
Tabela 5 - Correlações normalizadas do desvio Padrão da F0 (n=19).	121
Tabela 6 - Correlações normalizadas do jitter PPQ5 (n=19).	124
Tabela 7 - Correlações normalizadas para o shimmer APQ5 (n=19).	126
Tabela 8 - Correlações normalizadas de HNR (n=19).	131
Tabela 9 - Correlações Praat-VoiceStudio consideradas para os 19 oradores (n=20).	132
Tabela 10 - Idade, género e grau de parentesco dos oradores cujas frequências formantes foram extraídas e analisadas a partir das vogais sustentadas.	134
Tabela 11 - Valores médios dos resultados agregados, em percentagem, para os desvios relativos das frequências formantes, em função do tipo de canal.	147
Tabela 12 - Valores médios dos resultados agregados, em percentagem, para os desvios relativos das frequências formantes, em função do género do orador.	147
Tabela 13 - Melhores resultados de desempenho EER alcançados em [52] para cada uma das combinações de teste e para cada cenário de género de oradores: FEM, MAL e FEM & MAL.	165
Tabela 14 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos dos (20) oradores do género FEM. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos com metade da duração.	168
Tabela 15 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos dos (17) oradores do género MAL. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos com metade da duração.	169
Tabela 16 - <i>Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos de todos os (37) oradores dos dois géneros. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos com metade da duração.</i>	169
Tabela 17 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos dos (20) oradores do género FEM. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos mutuamente exclusivos e com metade da duração original.	171
Tabela 18 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos dos (17) oradores do género MAL. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos mutuamente exclusivos e com metade da duração original.	172

Tabela 19 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos de todos os (37) oradores dos dois géneros. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos homólogos de registos de voz mutuamente exclusivos e com metade da duração original.	174
Tabela 20 - Resultados de desempenho EER (em %) obtidos com base na comparação das bases de dados 1 e 2 em que os registos de voz incluem duas condições de canal e são mutuamente exclusivos quanto ao conteúdo da fala. Apresentam-se os resultados obtidos para os oradores do género MAL, FEM e junção dos dois. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita os valores previamente obtidos na comparação das bases de dados com uma só condição de canal.....	177
Tabela 21 - Resultados de desempenho EER (em %) obtidos com base na comparação das bases de dados 3 e 4 em que os registos de voz incluem duas condições de canal e são mutuamente exclusivos quanto ao conteúdo da fala. Apresentam-se os resultados obtidos para os oradores do género MAL, FEM e junção dos dois. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita os valores previamente obtidos na comparação das bases de dados com uma só condição de canal.....	178
Tabela 22 - Resultados de desempenho EER (em %) obtidos com base na comparação das bases de dados 5 e 6 em que os registos de voz incluem duas condições de canal telefónico e são mutuamente exclusivos quanto ao conteúdo da fala. Apresentam-se os resultados obtidos para os oradores do género MAL, FEM e junção dos dois. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita os valores previamente obtidos na comparação das bases de dados com uma só condição de canal telefónico.....	179
Tabela 23 - Resultados de desempenho EER (em %) obtidos com base na comparação da base de dados 1 (GSM_HQ_1h_VOI_TL_2h), em que os registos de voz incluem duas condições de canal, com a base de dados GSM_TL_2h e também VOI_TL_1h, cujo conteúdo de fala é mutuamente exclusivo. Apresentam-se os resultados obtidos para os oradores do género FEM, MAL e FEM & MAL. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita valores relevantes previamente obtidos na comparação de bases de dados com uma só condição de canal.	181
Tabela 24 - Resultados de desempenho EER (em %) obtidos com base na comparação da base de dados 3 (VOI_HQ_1h_GSM_TL_2h), em que os registos de voz incluem duas condições de canal, com a base de dados VOI_TL_2h e também GSM_TL_1h, cujo conteúdo de fala é mutuamente exclusivo. Apresentam-se os resultados obtidos para os oradores do género FEM, MAL e FEM & MAL. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita valores relevantes previamente obtidos na comparação de bases de dados com uma só condição de canal.	183
Tabela 25 - Resultados de desempenho EER (em %) obtidos com base na comparação da base de dados 5 (GSM_TL_1h_VOI_TL_2h), em que os registos de voz incluem duas condições de canal, com a base de dados GSM_TL_2h e também VOI_TL_1h, cujo conteúdo de fala é mutuamente exclusivo. Apresentam-se os resultados obtidos para os oradores do género FEM, MAL e FEM & MAL. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita valores relevantes previamente obtidos na comparação de bases de dados com uma só condição de canal.	184
Tabela 26 - Durações das amostras do OS, com e sem silêncio, em minutos e segundos.	191

Tabela 27 - Durações das amostras do orador desconhecido, com e sem silêncio, em minutos e segundos.	191
Tabela 28 - Scores obtidos a partir da comparação entre amostras do orador OS e entre eles próprios e com os registos das interceções telefónicas do orador desconhecido.....	193
Tabela 29 - Scores da comparação entre os registos de interceções telefónicas e de amostras telefónicas das conversações.....	196
Tabela 30 - Resultados LR derivados da comparação entre as amostras das interceções telefónicas e as amostras TL das conversações telefónicas simuladas com o orador OS.	196
Tabela 31 - Frequências da formante F1 e da formante F2 do fone /a/ em /ka/, em Hz, para as amostras OS2 e XX01, em qualidade telefónica de GSM.	199
Tabela 32 - Valores aproximados das frequências das formantes extraídas da vogal semiaberta /e/, pertencente a uma ocorrência da palavra “sete”, em cada um dos quatro registos de voz.	201

Lista de abreviaturas e siglas

1h – *First half*

2h – *Second half*

ADN – *Ácido Desoxirribonucleico*

AES – *Audio Engineering Society*

APQ – *Amplitude Perturbation Quocient*

CVC – *Consoante Vogal Consoante*

dB – *DeciBel*

DET – *Detection Error Trade-off*

E – *Evidence Score*

EEG – *Equal Error Graph*

EER – *Equal Error Rate*

F0 – *Frequência Fundamental*

F1 – *Frequência/Formante 1*

F2 – *Frequência/Formante 2*

F3 – *Frequência/Formante 3*

F4 – *Frequência/Formante 4*

FAR – *False Acceptance Rate*

FRR – *False Rejection Rate*

FEM – *Female*

GMM – *Gaussian Mixture Model*

GSM – *Global System for Mobile Communications*

H0 – *Hipótese Nula*

H1 – *Hipótese 1*

HNR – *Harmonics to-Noise Ratio*

HQ – *High Quality*

Hz – *Hertz*

IAFL – *International Association of Forensic Linguistics*

i-Vectors – Identity Vectors

kHz - kiloHertz

L1 – Língua materna

LDA – Linear Discriminant Analysis

LR - Likelihood Ratios

MAL – Male

MAP – Maximum a Posteriori

MFCC – Mel-frequency cepstral coefficient

OS – Orador Suspeito

PCA – Principal Component Analysis

PLDA – Probabilistic Linear Discriminant Analysis

PPQ5 – Period Perturbation Quocient

ROC – Receiver Operating Characteristic

TL – Qualidade Telefónica

TV – Total Variability

UBM – Universal Background Model

VAD – Voice Activity Detection

VoIP – Voice over Internet Protocol

XX – Orador Desconhecido

Capítulo 1 – Introdução

1.1 Contexto

A Fonética Forense é um ramo da área da Linguística que se ocupa da investigação e perícia de amostras de fala em contexto judicial. Assim, esta ciência oferece os conhecimentos e métodos de análise acústica e linguística que se focam na voz de determinado(s) orador(es), com o objetivo de identificar, através de perícias, o orador de uma gravação ou chamada telefônica, recorrendo a parâmetros acústicos e a aspetos de fala característicos que possam destacar um indivíduo, tais como a proveniência (dialetos) e as características da fala de cada um (socioletos e idioletos).

Assim, na comunicação verbal, de uma forma abstrata, os aspetos paralinguísticos, como por exemplo, a idade, o género, a profissão e o estado de espírito do interlocutor podem ser detetados através das perícias da Fonética Forense [1]. Outros parâmetros, relacionados com as características físicas da voz e passíveis de serem analisados objetivamente por computador, têm a capacidade de auxiliar na identificação do orador, não obstante os desafios incontornáveis devido à variação que a voz pode sofrer ao longo do tempo, por exemplo.

No que diz respeito ao reconhecimento/identificação de falante, cabe ao foneticista forense assumir várias tarefas primárias relacionadas entre si [2], a saber, estimar a identidade dos falantes a partir de uma gravação, conjecturar sobre a origem geográfica de um orador, através do dialeto, verificar se a gravação foi ou não editada, averiguar enunciados impercetíveis ou menos claros, atestar a autenticidade das gravações, analisar a proficiência do falante e, inclusivamente, recrutar possíveis avaliadores que colaborem em tarefas de avaliação percetiva. Na nossa investigação, focamo-nos no reconhecimento/identificação do orador.

A definição de reconhecimento de orador difere conforme os autores. Loakes, através da literatura de Broeders (2001), refere que é através do reconhecimento de orador que alguém é identificado – individualizado – tendo em conta características da

sua voz [2]. Francis Nolan, por seu turno, fala-nos de uma amostra que é atribuída a uma pessoa com base nas características acústico-fonéticas [3]. Este investigador também considera que existem dois tipos de reconhecimento de falante, a saber, o *naive* (indivíduo não treinado na área) e o técnico (perito) [4].

Rose, por sua vez, recorrendo também a Nolan (1983), destaca o facto de que associamos o termo Identificação Forense de Falante pelo Perito quando a deliberação está relacionada com disciplinas como Linguística, Fonética, Acústica, Processamento de Sinal e Estatística. Rose [5], tal como Loakes [2], indica que o Reconhecimento *Naive* do Falante se verifica quando determinado sujeito tem a capacidade auditiva de discriminar um orador e de reconhecer uma voz particular, o que acontece, a título de exemplo, quando uma testemunha é chamada para fazer reconhecimento de voz.

Morrison sugere, ainda, quatro formas de comparar vozes: i) perceptiveamente, ii) recorrendo à técnica fonético-acústica, iii) automaticamente, e iv) socorrendo-se de espectrogramas. A estas formas adiciona o perceptive-espectrográfico e o perceptive-fonético-acústico [6]. Focar-nos-emos, nesta dissertação, no potencial e robustez de certas características acústicas para o objetivo da identificação de falante, e também no desempenho do reconhecimento automático de falante, na medida em que é um meio recente de verificação de identidade nas aplicações em que se faz uso da voz, no quotidiano, ou mesmo em conversas telefónicas entre amigos e familiares ou registos de negócios [7].

O desafio na nossa dissertação em relação aos gémeos, e sem distinguir, para já, os monozigóticos dos dizigóticos, prende-se pelo facto de serem conhecidos por serem similares na forma anatómica [8], por exemplo, o que faz com que o aparelho fonador seja análogo. Esta circunstância dificulta a sua distinção através de gravações telefónicas, especialmente quando não existem diferenças paralinguísticas, como a exposição a diferentes formas de socialização – o que contribui, muitas vezes, para dicções díspares –, os desiguais estados de espírito no momento da gravação, mesmo que seja feita no mesmo dia, ou perturbações na produção dos sons da fala que apenas um dos gémeos possa apresentar. Contudo, os gémeos monozigóticos têm tendência

para partilhar os padrões de articulação, na medida em que possuem as mesmas influências genéticas no discurso e na fala [9].

Em suma, este trabalho de doutoramento insere-se na área da Fonética Forense com especial incidência na comparação de vozes e no reconhecimento automático de orador, com base em características objetivas extraídas do sinal de voz, através de *software* desenhado para o efeito e recorrendo a amostras de voz de indivíduos gémeos e de parentes próximos, como pais, mães, irmãos ou primos.

1.2 Motivação

A nossa curiosidade em relação à Fonética teve início nas aulas desta disciplina, no último ano da licenciatura em Línguas, Literaturas e Culturas. A tese de mestrado em Ensino do Português e do Alemão (terceiro ciclo e ensino secundário) recaiu nas áreas da Fonética e da Fonologia através de exercícios de conhecimento fonético e fonológico implícito e explícito na língua materna e na língua estrangeira e começámos a fortalecer, aqui, o desejo de continuar a aprofundar a aprendizagem em relação a esta ciência. Assim, no doutoramento, propusemo-nos a reforçar competências de Fonética Acústica e de Fonética Forense, tendo especial interesse, para nós, o campo da identificação de orador, na área da comparação de vozes.

Na investigação criminal e nas chamadas de emergência, a gravação de voz é um ponto fulcral, na medida em que permite ao perito forense, mais tarde, a identificação do suspeito, quando comparada uma amostra da sua voz com outras amostras. Desta forma, atendendo ao progresso das tecnologias que envolvem a utilização de *chat*, vídeos e chamadas, como o *Skype*, decorre uma grande necessidade de se realizarem estudos de comparação de vozes de situações de crime, como ameaças pessoais ou fraudes em contexto bancário. Contudo, este tipo de investigação pode ter vários obstáculos técnicos, particularmente devido ao facto de que a qualidade e a duração das gravações telefónicas podem ser insatisfatórias ou desadequadas.

É nossa aspiração que, com esta dissertação, possamos contribuir para destacar a importância da Fonética Acústica e Forense, em Portugal, e salientar a relevância desta

ciência para a lei, o procedimento judicial, a investigação criminal e para os tribunais, tentando compreender e evidenciar quais são os melhores métodos e os parâmetros preferíveis para este propósito, particularmente no que diz respeito às vozes que são muito semelhantes, como as dos gémeos e as dos familiares próximos.

1.3 Objetivos

Com este trabalho, pretendemos, concretamente, avaliar o nível de degradação que perturba as gravações telefónicas no que diz respeito a alguns canais de comunicação, como o GSM¹ e o VoIP², que serão alvo de estudo no Capítulo 3. Neste contexto, um dos nossos objetivos é estudar a influência da degradação das comunicações nas tarefas de comparação de voz que têm como meta reconhecer amostras do mesmo falante. Além disso, temos a intenção de verificar quais são os parâmetros acústicos que permitem discriminar de forma mais eficiente o falante entre si (com várias amostras do mesmo orador) e o falante com vários oradores (havendo, desta forma, várias amostras de falantes diferentes). Nesta tarefa, também é importante determinar qual é o impacto da duração que uma amostra pode ter.

Assim, podemos afirmar que esta dissertação tem três grandes objetivos:

- 1 – A investigação com base em parâmetros acústicos;
- 2 – A investigação com base no reconhecimento automático de orador;
- 3 – A aplicação dos cenários anteriores a um caso real de peritagem.

Estes objetivos de estudo são abordados usando registos de voz que foram captados em duas sessões de conversação, tal como se ilustra na Figura 1 e na Figura 2, a primeira respeitante a comunicação através de um canal GSM, a segunda respeitante a comunicação através de canal VoIP.

¹ **GSM** – Tecnologia celular e digital aberta utilizada para transmitir voz e serviços de dados via comunicação móvel. Significa *Global System for Mobile Communications*.

² **VoIP** – Sistema de encaminhamento de interlocação humana através de uma rede de computadores baseada no Protocolo da Internet. Significa *Voice over Internet Protocol*.

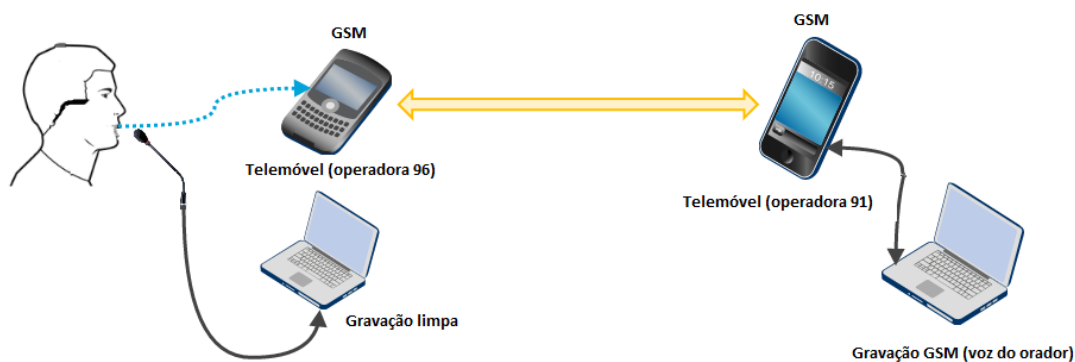


Figura 1 - Comunicação GSM.

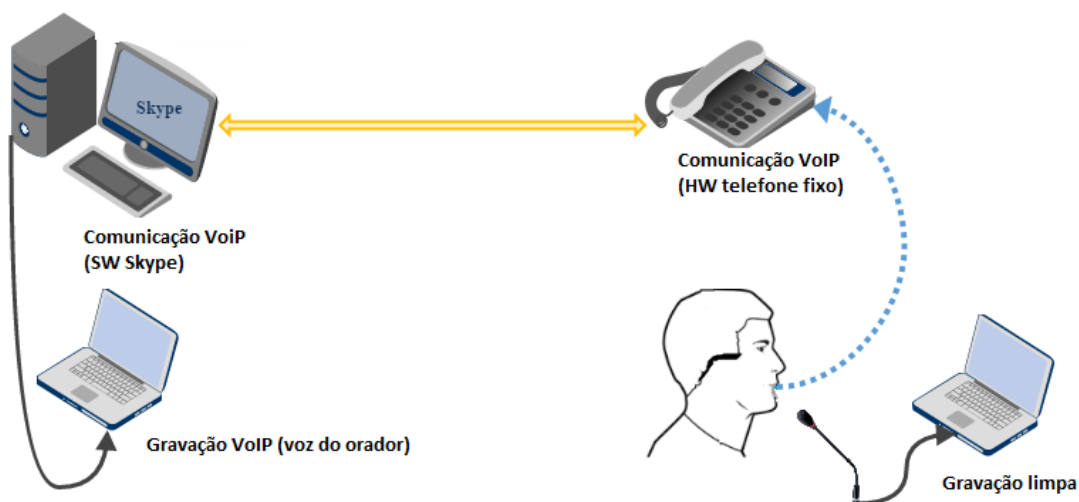


Figura 2 - Comunicação VoIP.

Em ambos os casos, há um registo concomitante (*i.e.* contemporâneo) e local de voz de alta qualidade, para além do registo remoto de qualidade telefónica. Estes cenários e logística de gravação encontram-se ilustrados nas Figuras 1 e 2. Em síntese, tal como será detalhado na Secção 3.3, para cada orador, a nossa base de dados inclui dois registos de qualidade telefónica (um VoIP, outro GSM) e os dois registos correspondentes de alta qualidade.

1.4 Contribuições

Enumeramos, de seguida, as principais apresentações e publicações relacionadas com a dissertação que procuram contribuir para a análise forense de voz e que procuram auxiliar a criar condições para a realização de trabalhos de peritagens, como o que é descrito neste trabalho:

1 – Em outubro de 2012, apresentámos na 3rd *European Conference IAFL (International Association of Forensic Linguistics)* – “*Forensic Linguistics: bridging the gap(s) between language and law*” (Portugal) o resultado de um pequeno trabalho realizado no âmbito de avaliação da unidade curricular de Processamento de Sinais Fisiológicos, lecionada na Faculdade de Engenharia da Universidade do Porto, que se denomina de “Avaliação qualitativa de vogais sustentadas de gémeos: um estudo experimental”. Através da realização desta investigação, verificámos, tendo em conta a amostra disponível, que não existia correspondência exata entre vozes de indivíduos gémeos do mesmo género. Recorremos a parâmetros acústicos e percetivos de avaliação qualitativa, dando início à nossa demanda em relação a esta área;

2 – Em abril de 2017, fomos motivados para colaborar num caso de peritagem, detalhado no capítulo 7, enquanto especialistas em Linguística e análise forense, quer para conduzir e gerar os diálogos simulados com o orador, quer para obter os parâmetros acústicos de fones específicos extraídos dos registos de voz dos oradores desconhecidos, disponíveis numa base de dados previamente organizada, e do suspeito em questão. Este trabalho de peritagem tinha como objetivo emitir um parecer técnico fundamentado sobre as vozes que constavam em interceções telefónicas legais e foi supervisionado pelo Orientador desta dissertação;

3 – Em junho de 2017, na *AES International Conference on Audio Forensics* (Estado Unidos da América), foram apresentados os resultados sobre o trabalho que avaliou a relevância de cinco parâmetros acústicos que expusessem as características idiossincráticas da voz quando o sinal é afetado pelo canal de comunicação. Este artigo tem como título “On the relevance of F0, Jitter, Shimmer and HNR acoustic parameters in forensic voice comparisons using GSM, VOIP and contemporaneous high-quality voice recordings”;

4 – Estudámos o comportamento das frequências formantes³ no que diz respeito à modificação devida ao canal de comunicação (GSM e VoIP) e concluímos que as formantes F1, F3 e F4 apresentam comportamentos diferentes quando submetidas a esse meio de transmissão e que as vozes femininas são, tipicamente, mais afetadas do que as masculinas no que diz respeito à estimativa de erro. Daqui, resultou a apresentação realizada na 13^a Conferência Bienal IAFL (*International Association of Forensic Linguistics*), em Portugal, em junho de 2017, que se intitula de “Voice formants modification due to GSM and VoIP telephonic communication”, em coautoria;

5 – Com o artigo apresentado na 22nd *International Conference on Digital Signal Processing*, em agosto de 2017 (Inglaterra), cujo título é “Consistency of the F0, Jitter, Shimmer and HNR voice parameters in GSM and VOIP communication”, descrevemos os resultados da análise através de duas plataformas concomitantes de gravações telefónicas e de alta qualidade com duas modalidades de comunicação (GSM e VoIP). O objetivo deste trabalho era o de compreender a robustez de cinco parâmetros acústicos, a saber, a média e o desvio padrão da frequência fundamental⁴, o *jitter* PPQ⁵, o *shimmer* APQ⁶ e o HNR⁷, de entender os efeitos do canal de comunicação e de explorar o potencial destes parâmetros acústicos no que diz respeito a um eventual cenário forense que envolva gravação de voz. Todos estes conceitos serão detalhados nos Capítulos 2 e 5.

Deve referir-se ainda, pela sua relevância, que foi desde muito cedo planeada uma publicação de jornal referente aos resultados apresentados no Capítulo 6. Contudo, os planos de desenvolvimento desta investigação foram comprometidos devido a dois fatores essencialmente: 1) um protocolo que foi desenvolvido com uma entidade de Polícia Científica e que visava colaboração entre essa entidade e a Universidade do Porto para recolha de voz telefónica, e 2) a aquisição do *software* comercial usado nos nossos estudos. No primeiro caso, o protocolo evoluiu durante cerca de dois anos de

³ **Formante** – Picos de energia numa região do espectro sonoro. Traduz a ressonância acústica que é produto das diferentes configurações ressoadoras e de filtro do trato vocal. Uma explicação ilustrada é fornecida na Secção 5.2.2.

⁴ **Frequência Fundamental** – É a frequência produzida na glote, em resultado da vibração das pregas vocais.

⁵ **Jitter PPQ5** – Medida que avalia a variabilidade involuntária da frequência fundamental. PPQ significa *Period Perturbation Quocient*.

⁶ **Shimmer APQ5** – Medida que avalia a variabilidade involuntária da amplitude dos ciclos glotais. APQ significa *Amplitude Perturbation Quocient*.

⁷ **HNR** – Estabelece a relação entre a energia dos harmónicos e a energia da componente de ruído do sinal, num sinal de voz. Significa *Harmonics to-Noise Ratio*.

clarificação dos termos de colaboração e envolvendo os Apoios Jurídicos de ambas as Instituições, mas não logrou chegar à fase de aprovação por ambas as partes. Este atraso obrigou ao planeamento e concretização de novas condições técnicas de captação de sinal. Por outro lado, a aquisição de *software* comercial, obrigou a tramitações específicas que introduziram um atraso significativo, designadamente em relação a um período de extensão da licença de utilização de modo a ser viável a conclusão de novos testes que entretanto se revelaram necessários.

Não obstante, está já em curso a preparação de uma submissão de artigo para jornal que será valorizada caso venha ainda a ser viável, através de uma nova mas breve extensão da licença de utilização, a realização de testes adicionais em complemento aos já apresentados na Secção 6.7.

1.5 Estrutura da dissertação

A presente dissertação está dividida em oito capítulos. No primeiro, o da Introdução, figuram o contexto, a motivação, os objetivos do trabalho, as contribuições e a estrutura da tese.

O Capítulo 2 fornece um enquadramento estrutural, focando-se no sistema humano de produção de fala, nomeadamente, a estrutura anatómica do aparelho fonador, as características físicas e percetivas da voz. De seguida, incide sobre o aparelho fonador de indivíduos gémeos, tendo em conta a perspetiva anatómica dos gémeos monozigóticos e dizigóticos e as suas características de voz. A Secção 2.3 foca-se na identificação do orador tendo em conta os parâmetros perceptivos e acústicos, a análise perceptiva, a análise acústica, o reconhecimento automático do falante e as condições ideais *versus* as condições reais. Terminaremos o segundo capítulo com o âmbito e expressão da Fonética Forense em Portugal e a pertinência do estudo de vozes de gémeos.

No Capítulo 3, tecem-se considerações sobre o planeamento e estruturação da base de dados, mostrando o nosso cenário de recolha de amostras e explicando os tipos de modalidade de comunicação (GSM *versus* VoIP) e de canal (alta qualidade em

contraste com qualidade telefônica). Visam-se, ainda, os questionários utilizados, os participantes e a composição da base de dados. Esta subdivide-se em várias fases de condicionamento: i) em base de dados inicial, ii) base de dados após isolamento de regiões de vogais sustentadas, iii) após remoção do segundo orador, iv) após conversão da frequência de amostragem e, por fim, v) após subdivisão e recomposição das gravações na base de dados.

O capítulo 4 aborda os ambientes e ferramentas de processamento, como os ambientes *software* que utilizámos (*Audacity*, *Adobe Audition*, *Praat*, *iVocalise* e *Biometrics*), e a análise estatística.

Os estudos com base em parâmetros acústicos (Capítulo 5) focam-se nas vogais sustentadas que recolhemos, nos parâmetros acústicos que consideramos relevantes para a comparação de vozes e, também, na análise através de parâmetros acústicos (frequência fundamental, desvio padrão da frequência fundamental, *jitter*, *shimmer*, HNR e frequências das formantes).

Por seu turno, o Capítulo 6 discute os estudos com base no reconhecimento automático de orador e de acordo com os seguintes temas principais: pré-condicionamento dos dados, resultados previamente consolidados e segmentação em frações de registo que, depois, são re combinadas de forma diversa para serem possíveis novos arranjos de modelização e identificação do orador; testes de identificação de orador usando variações de duração dos registos de voz, mútua exclusividade na fala, diversidade de condições de canal e assimetria nas condições de canal.

O caso de estudo de peritagem figura no capítulo 7, sendo que a última parte da dissertação. Finalmente, o oitavo capítulo, apresenta as conclusões, resume as principais contribuições e avança com propostas de tarefas futuras que procuram complementar e ampliar o âmbito da investigação.

Capítulo 2 – Enquadramento estrutural

2.1 Introdução

O presente capítulo debruça-se na pesquisa sobre o sistema humano de produção de fala, recorrendo a conceitos como o aparelho fonador e explorando as características físicas e percetivas da voz, na medida em que são os processos biomecânicos que a tornam ímpar. Bastante pertinente para a nossa dissertação é um breve estudo do aparelho fonador dos indivíduos gémeos, tendo em consideração se são monozigóticos ou dizigóticos. O nosso objetivo com este capítulo é, também, o de introduzir conceitos chave como os subjacentes à identificação do orador e envolvendo a perceção e a acústica. Por fim, e porque está relacionada com a nossa investigação, explicaremos, sucintamente, o âmbito e a expressão da Fonética Forense, em Portugal, e qual é a relevância do estudo de vozes de gémeos para esta área científica.

2.2 Sistema humano de produção de fala

O sistema humano de produção de fala é complexo e merece especial atenção. O aparelho fonador é um espantoso mecanismo constituído por órgãos que pertencem aos aparelhos respiratório e digestivo. Assim, a voz humana é o resultado de processos biomecânicos e acústicos complexos que serão explorados nas secções seguintes. A voz não é, apenas, consequência das vibrações das pregas vocais, das caixas de ressonância e da filtragem. Além disso, a voz não é só o resultado do ar que é expulso dos pulmões. Para alguns autores [10], nomeadamente os psicolinguistas, é importante asseverar que o discurso começa no cérebro, depois da criação da mensagem e das estruturas léxico-gramaticais que têm, obrigatoriamente, de obedecer a critérios canónicos pelos quais se rege a gramática de determinada língua. Posteriormente a este processo, um falante humano organiza a sequência correta dos segmentos que irão dar lugar às sílabas, às

palavras e, por sua vez, às frases. Além disso, são necessários comandos motores para dar origem à voz e aos fones.

De facto, a produção de fala, controlada pelo sistema nervoso central, é constituída por três etapas, a saber: i) a respiração, que representa a fonte de energia acústica imprescindível para a produção de voz, ii) a fonação, que é a segunda fase, na qual existe vibração das pregas vocais, devido à pressão do ar que emana dos pulmões, e iii) a articulação, na qual estão envolvidos os órgãos do trato vocal responsáveis pela modelação e filtragem dos sons, como a língua e os lábios, a título de exemplo. De referir que existem sons surdos, que não necessitam da vibração das pregas vocais.

É extremamente importante que o foneticista forense tenha conhecimento de como se dá a produção da voz para aliar esta informação à capacidade percetiva para verificar, por exemplo, se o orador de determinada gravação está a produzir uma africada ou uma fricativa [11]. Assim, os movimentos do aparelho fonador coadjuvam na produção, através da articulação, de determinada consoante ou vogal. Contudo, a entoação, a acentuação, a fluência do falante, a rapidez com que articula o discurso e o timbre devem ser igualmente considerados, porque são aspetos suprasegmentais que podem facilitar ou dificultar a identificação do orador.

2.2.1 O aparelho fonador

O aparelho fonador, como referido, é constituído por órgãos dos aparelhos respiratório e digestivo, sendo uma conceção meramente utilizada nas áreas da Fonética e da Fonologia. Divide-se em três regiões, a saber, a supraglótica, a glótica e a subglótica. As cavidades supraglóticas e a glótica constituem o trato vocal. A imagem ilustrada na Figura 3 ilustra os elementos principais do aparelho fonador.

Os órgãos principais das cavidades supraglóticas são a cavidade bucal, a cavidade nasal e a faringe. É nesta última cavidade que os impulsos do ar se transformam num *input* acústico, na medida em que age como um ressoador acústico que amplifica certas frequências e atenua outras. A laringe é, por excelência, o órgão principal da cavidade glótica. O fenómeno principal que se dá, aqui, é o da fonação, ou seja, a vibração das pregas vocais, que é resultado da interação entre a pressão do ar que

deriva dos pulmões, a elasticidade destas pregas e o Efeito de Bernoulli [76]. A traqueia, os pulmões e o diafragma são os constituintes principais das cavidades infraglóicas e o seu papel fundamental é o de ajustar a corrente do ar do sistema fonador, na medida em que controlam a pressão do ar e transportam-no à cavidade glótica. Nela, não se verificam sons, mas apenas sopros [73].

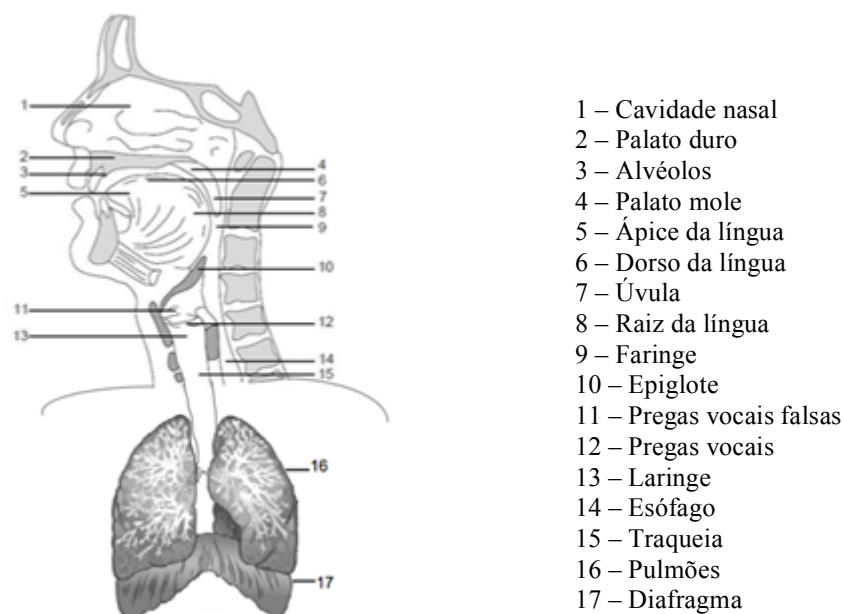


Figura 3 - O aparelho fonador (adaptado de [12]).

Passar-se-á, agora, a descrever a função de alguns elementos do aparelho fonador [18]. O diafragma é um músculo em formato de cúpula incumbido da inspiração e da expiração e separa a caixa torácica das cavidades abdominais. Durante a inspiração, contrai-se e desce, o que provoca a diminuição da pressão intratorácica e faz com que o ar entre nos pulmões. Na expiração, por sua vez, este músculo relaxa e eleva-se, expulsando o ar [73].

A traqueia é um canal de, aproximadamente, 1.5 centímetros de diâmetro e de 10 a 12 centímetros de comprimento, ligando a laringe aos brônquios e conduzindo o ar

dos pulmões durante a respiração. É constituído por um músculo liso, sendo que, externamente, possui anéis de cartilagem [73].

A laringe é constituída por nove cartilagens separadas e é nela que assenta a fonte acústica da voz humana [13]. As cartilagens mais relevantes são a cricoide, a tiroide e as aritenoides. A primeira é o suporte da laringe e funciona como um anel de topo da traqueia. A cartilagem tiroide localiza-se junto à cricoide e à cricotiroide. Possui lâminas que se unem, formando uma proeminência, vulgo “maçã de Adão”, que é mais demarcada nos homens do que nas mulheres, devido ao ângulo das lâminas. As cartilagens aritenoides ligam-se à cricoide via cricoaritnoide, o que faculta o movimento anterior, posterior e lateral das pregas vocais, permitindo o ajuste da pressão. A Figura 4 ilustra a constituição da laringe.

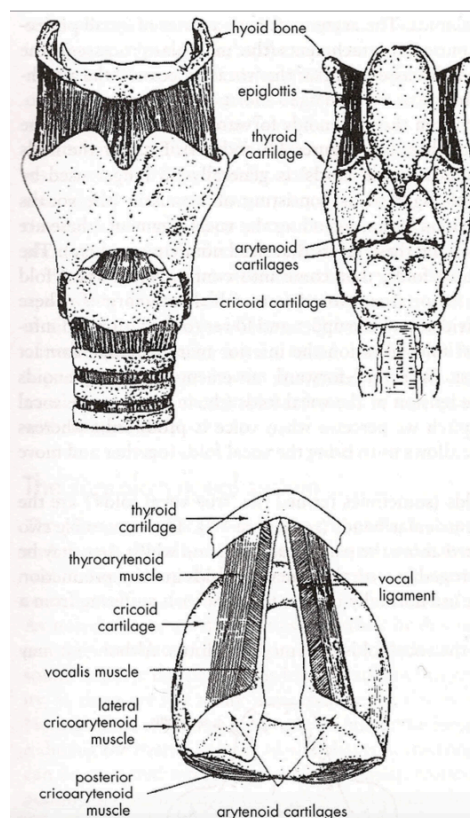


Figura 4 - Visão anterior, posterior e de topo da laringe [14].

É na laringe que estão alojadas as cordas ou pregas vocais, que geram ondas periódicas caracterizadas por uma dada frequência fundamental, que contribui para a distinção entre os oradores, na medida em que está dependente do comprimento, da tensão e da massa deste tecido musculoso [15]. Quando existe vibração das pregas vocais, é produzido um som vozeado ou sonoro. No momento em que o ar passa livremente, é produzido som não vozeado ou surdo [19]. Além disso, as duas pregas podem estar tensas ou relaxadas, produzindo diferentes timbres, mais agudos ou graves

A Figura 5 ilustra a posição das pregas vocais quando se encontram na posição abduzida e na posição aduzida.

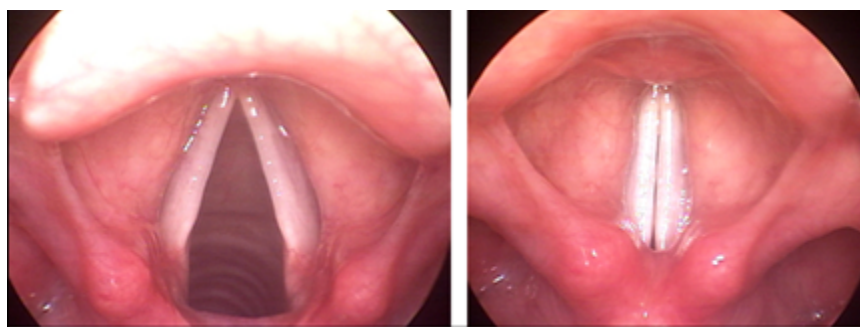


Figura 5 - Pregas vocais abduzidas (sons surdos - esquerda) e pregas vocais aduzidas (sons sonoros - direita) (adaptado de [16]).

As pregas vocais estão sujeitas ao já referido efeito de Bernoulli, que está relacionado com o princípio de conservação da energia, na medida em que a velocidade e a pressão de um fluxo de ar estão inversamente relacionadas, ou seja, se a velocidade aumenta, a pressão diminui [76]. Este fenómeno vai dar origem ao ciclo fonatório que é usualmente medido através do número de vezes que as pregas vocais vibram – a frequência fundamental. A título de exemplo, se este parâmetro for de 120 Hz, representa que as pregas vocais vibram 120 vezes por segundo, o que, tipicamente, ilustra a onda sonora da voz emitida por um orador masculino adulto.

A faringe situa-se entre a laringe e as cavidades orais e nasais, sendo constituída pela nasofaringe, a orofaringe e a hipofaringe, como se pode observar na Figura 6.

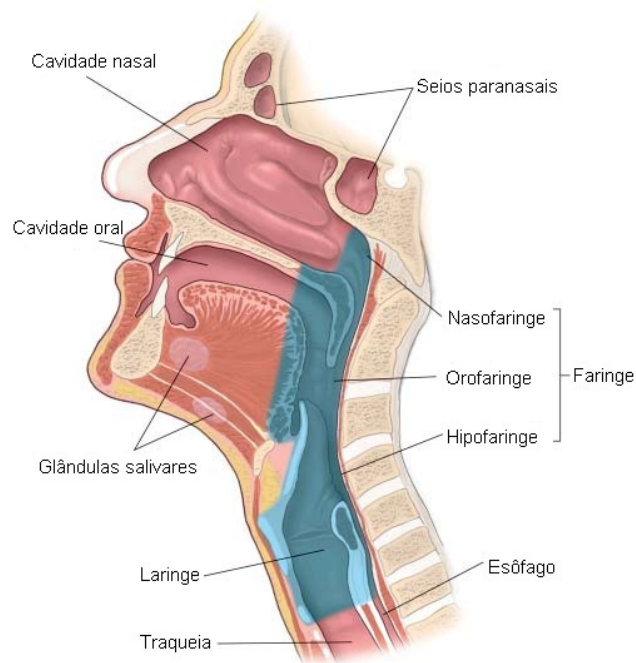


Figura 6 - Corte sagital do trato vocal [17].

As cavidades oral e a nasal também se encontram ilustradas na figura anterior. A primeira, ao contrário da cavidade nasal, pode mudar de forma e de volume, aumentando e diminuindo tendo em conta a projeção dos lábios e o afastamento ou aproximação do maxilar inferior. A língua, alojada na cavidade oral, é um músculo com muita importância para a produção da fala, na medida em que expande ou retrai. É na cavidade oral que a grande maioria das produções vocais se diferencia. Por seu turno, a cavidade nasal é a cavidade ressoadora que atribui a nasalidade aos fones⁸, caso o véu palatino esteja descido e desencostado da parede posterior da faringe. Assim, uma vez que o volume não é variável na cavidade nasal, existe apenas o acréscimo desta

⁸ **Fone** – Unidade de estudo da Fonética. É perceptível num contínuo sonoro, sendo a mais pequena unidade discreta. Constitui a realização física dos sons da fala, ao nível concreto (no nível abstrato, existe o fonema – unidade de estudo da Fonologia).

influência no timbre. A parte superior da cavidade é constituída pelo osso etmoide e pelos nervos olfativos, sendo que a parte inferior é constituída pelo palato duro.

É impossível falar-se do aparelho fonador sem se referir, ainda que brevemente, à teoria fonte-filtro da voz, de Gunnar Fant [18]. Este modelo indica que a voz é resultado da modificação do sinal da fonte, também designado por excitação glótica – gerado na laringe –, através do filtro que exprime a influência combinada da cavidade nasal, da cavidade oral, dos maxilares, dos dentes, lábios, narinas e língua.

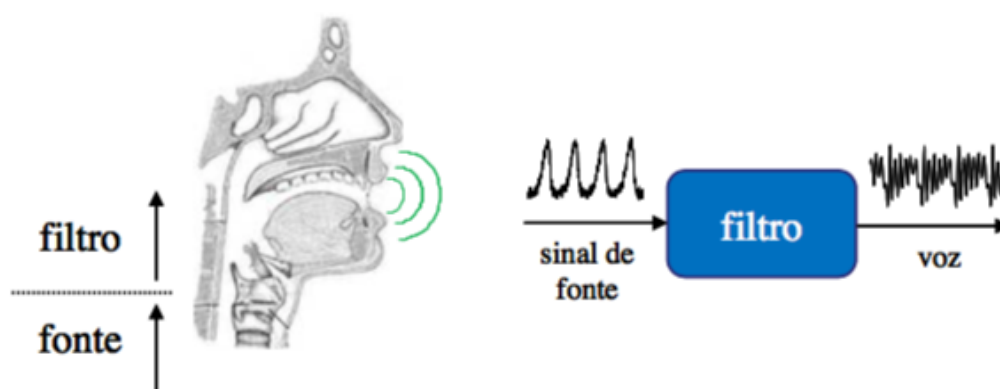


Figura 7 - Teoria Fonte-Filtro [18].

A modificação biológica que se iniciou no Homem com a passagem de quadrúpede a bípede possibilitou-lhe uma reorganização craniana. A área de Broca, localizada no giro frontal inferior, é responsável pelos programas motores que permitem exprimir os conceitos armazenados na área de Wernicke. Consequentemente, esta área está relacionada com o processo de descodificação fonológico, devido à capacidade de conversão dos conceitos numa sequência sonora específica. É a partir daqui que se verifica o planeamento motor complexo e específico da fala [22] [23] [24] [25] [26]. O aparelho fonador permitiu aos humanos pré-históricos alargar a vocalização para outras utilidades comunicativas, como a verbalização de raciocínios e o planeamento de ações em grupo, como a caça. O desenvolvimento da capacidade de monotorização de fala em tempo real permitiu o desenvolvimento do cerebelo. De todas estas transformações, a mais importante foi a descida da laringe [20] [21], consequência da adoção do

bipedismo, na medida em que permitiu a possibilidade de maior autonomia de vocalização e de realização de sequências articuladas de gestos vocais. Na Figura 8, podem-se observar as diferenças físicas entre o aparelho fonador dos humanos e o dos símios, sendo que os primeiros possuem a chamada estrutura em “cotovelo”, facilitando a produção da fala, e os macacos têm um sistema em “L”, o que dificulta. Nos primeiros anos de vida, a criança também possui o aparelho fonador com a estrutura em “L”.

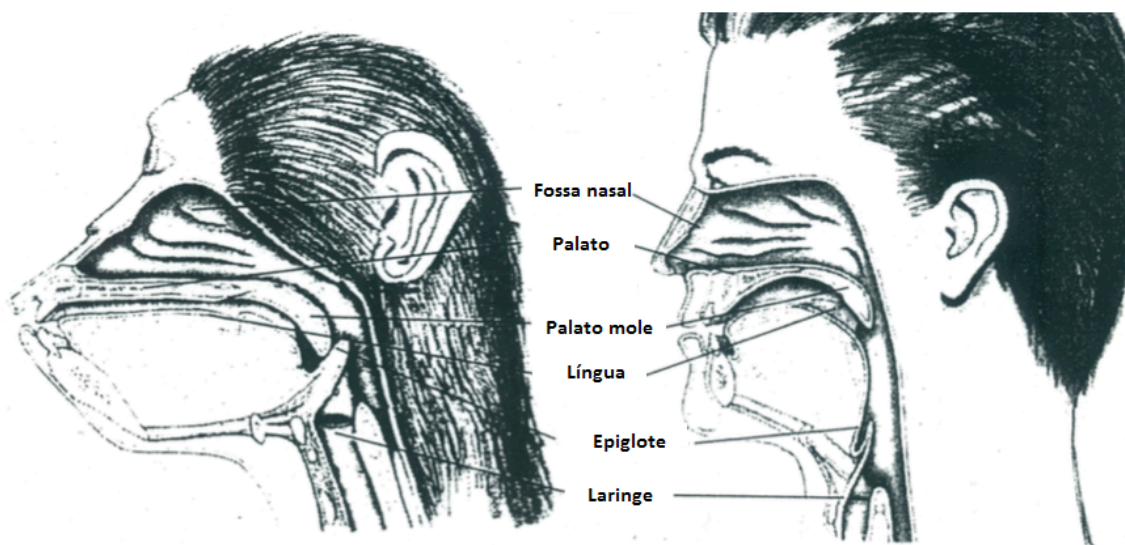


Figura 8 - *Diferenças físicas entre os aparelhos fonadores do macaco e do homem [27].*

De facto, a descida da laringe é consequência anômica da evolução e independente do controlo cortical da vocalização. A postura vertical fez com que a parte frontal da garganta se tenha prolongado, o que é também reforçado pelo deslocamento da cabeça para uma posição posterior em relação à coluna vertical. A laringe cresceu e adquiriu maior mobilidade. A redução da cavidade oral e do rosto, a amplificação da laringe e a maior liberdade de movimentos da língua na cavidade oral contribuíram para este fenómeno, possibilitando a articulação completa da fala. Esta especificidade do Ser Humano permite ao trato vocal assumir as formas necessárias para produzir fones muito diferenciados.

A título de exemplo, o processo da descida da laringe faz com que o trato vocal aumente de tamanho, o que permite diminuir as frequências formantes. Na Figura 9, pode-se observar a diferença do trato vocal de um adulto e de uma criança.

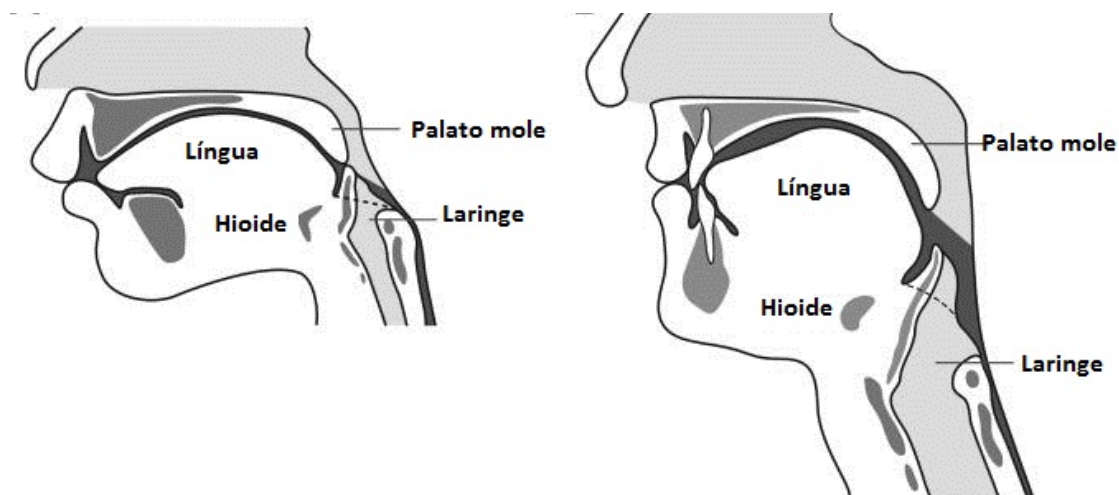


Figura 9 - Diferenças anatômicas entre o trato vocal de uma criança (esquerda) e o de um adulto (direita) (adaptado de [28]).

Assim, podemos observar que o aparelho fonador é resultado de um processo evolutivo de milhares de anos, sendo também modificado, ao longo do desenvolvimento físico do Ser Humano, até atingir a maturação dos órgãos.

2.2.2 Características físicas e perceptivas da voz

Fazendo parte indissociável da socialização humana e da comunicação oral, a voz transporta traços físicos e psicológicos do orador e é produto de características adquiridas ao longo da vida, como as formas e hábitos de articulação. Segundo Sousa [29], o sinal acústico relaciona-se com o fenómeno perceptivo e, desta forma, existe correlação com a qualidade de voz.

Fisicamente, o som caracteriza-se por ser o resultado da estimulação dos átomos de determinado meio – líquido, sólido ou gasoso – e o movimento oscilatório das partículas – energia mecânica. Esse movimento cria estados de compressão e de

rarefação, o que provoca a sensação sonora que se manifesta na sensação auditiva [30] [74]. Esta é a génese da onda acústica que, na voz humana, a nível de fala, é sempre complexa. O canal de comunicação é o ar, por excelência.

De acordo com Sousa [29], os parâmetros acústicos medem grandezas do sinal acústico e que se relacionam com fenómenos físicos. Por seu turno, os parâmetros perceptivos estão relacionados com a sensação auditiva que os estímulos acústicos provocam num ouvinte humano, sendo esta sensação subjetiva, uma vez que vai depender da acuidade auditiva e do treino auditivo, entre outros aspetos. De forma mais lata, um ouvinte humano tem tendência a afirmar se o interlocutor fala “alto” ou “baixo” – relacionando-se com o termo “altura” – e se tem voz mais aguda ou mais grave. De acordo com Ferreira [18], a designação “altura” oferece ambiguidade, na medida em que pode estar relacionada com a intensidade da pressão sonora (medida em deciBel – dB) ou com a organização de frequência de determinado som.

Além disso, sem qualquer tipo de formação ou treino auditivo, um ouvinte humano é capaz de reconhecer, por exemplo, se existe afonia ou disfonia, se existe boa ou má articulação, quais são os aspetos particulares de co-articulação ou, ainda, se o orador apresenta uma voz mais ou menos nasal. Afiguram-se outros parâmetros perceptivos que se aliam aos de qualidade, como, por exemplo, a afinação, a tessitura ou o timbre, sendo que este último é relevante na nossa investigação, na medida em que é ele que permite a um ouvinte humano distinguir as vozes, ou seja, reconhecer sons situados na mesma frequência e gerados por fontes diferentes [31]. Outros dois aspetos primordiais que auxiliam nesta tarefa do ouvido humano são a frequência fundamental e o seu comportamento temporal. A distinção é feita através de uma análise que tem em conta uma organização aproximadamente logarítmica da frequência, através de uma escala, como por exemplo, a escala Mel⁹ e a escala Bark¹⁰ [18].

⁹ **Escala Mel** – É uma escala musical de tons com intervalos avaliados perceptivamente como iguais pelos ouvintes, através de estudos realizados com ondas sinusoidais. Como referência, 1000 Mel corresponde à frequência melódica perceptiva de um som com frequência fundamental de 1000 Hz [102] [103].

¹⁰ **Escala Bark** – É uma escala psicoacústica que abrange 24 bandas críticas na gama audível de frequências. Cada região da membrana basilar pode ser considerada como um filtro passa-banda com uma largura de banda igual a um Bark. Relaciona-se com estudos ligados ao mascaramento de frequências puras por bandas estreitas de ruído e também com a intensidade percebida (*loudness*) [102] [103].

Relacionada com a frequência fundamental, a prosódia é a área da Fonologia que estuda a acentuação e o ritmo dos diversos segmentos fónicos. Assim, é indissociável de três parâmetros, a saber, a intensidade, a variação da altura do tom laríngeo e a duração [32]. Existem três fatores visados pela prosódia: o acento dinâmico, ligado à energia com que o ar é expulso dos pulmões, o acento de entoação, relacionando-se com a elevação da frequência fundamental, e a duração de um fone, interligando-se com a sua sustentação sonora [33]. É através da prosódia que se distinguem as vogais átonas das tónicas. É possível relacionar a prosódia com marcas pessoais de falar, assim como a entoação, que, ao contrário da prosódia, corresponde às variações da frequência fundamental que ocorrem numa sequência maior do que um fone, como, por exemplo, uma palavra ou uma sucessão de palavras. Pode-se, facilmente, reconhecer, através da entoação, se uma frase é afirmativa ou interrogativa. Com efeito, a observação do contorno da frequência fundamental (*rise and fall*) é de grande importância na entoação. Assim, há lugar para a relação da prosódia e da entoação com parâmetros perceptivos que podem, também, ser acústicos, a partir do momento em que são analisados através de *software*, por exemplo.

O campo dos parâmetros perceptivos encontra-se, ainda, em larga medida, inexplorado, na medida em que há mecanismos da audição humana que são desconhecidos numa perspectiva psicoacústica [18]. Existem inclusivamente vários processos de análise acústica, na percepção auditiva, que se afiguram ainda como misteriosos para a Ciência. O efeito da circularidade do som, conhecido como efeito de Shepard, é só um exemplo [75].

2.2.3 O aparelho fonador de indivíduos gémeos

Como referido, o trato vocal é uma parte decisiva na produção dos sons da fala [34]. Devido ao facto de as gravações realizadas no âmbito desta dissertação incluírem amostras de indivíduos gémeos, considera-se pertinente apresentar uma breve secção sobre o aparelho fonador e a perspectiva anatómica dos gémeos monozigóticos e dizigóticos, assim como um resumo das características de voz que apresentam.

2.2.3.1 Gêmeos dizigóticos e monozigóticos

A formação gemelar pode acontecer quando existe a expulsão de dois ovócitos que, posteriormente, são fecundados. Assim, verifica-se a formação de dois zigotos que irão dar origem a gêmeos dizigóticos. Os gêmeos dizigóticos partilham da mesma semelhança genética que irmãos nascidos em períodos diferentes, na medida em que também tiveram origem em ovócitos diferentes [35] [36] [37] [38]. Vulgarmente, são conhecidos por “gêmeos falsos” ou “gêmeos fraternos” e podem ser do mesmo género (dois masculinos ou dois femininos) ou de géneros diferentes (masculino e feminino). Cada embrião possui o seu saco amniótico e o seu saco coriônico. Contudo, também é possível os gêmeos dizigóticos partilharem da mesma placenta, na medida em que é fundida graças à contiguidade dos locais de implantação dos blatocistos que originam este tipo de gêmeos. Esta gravidez gemelar é a mais comum, acontecendo duas em cada três vezes e um fator determinante pode ser a idade tardia da mãe no momento da concepção e a administração de certos medicamentos no tratamento para a possibilidade de engravidar aumentar, como o clomifeno [38].

Os gêmeos monozigóticos, também denominados de “gêmeos verdadeiros” ou “gêmeos idênticos”, prosaicamente, têm origem num só zigoto que se desenvolve entre um a catorze dias depois da fertilização. Por terem início num só zigoto, estes indivíduos partilham da mesma estrutura genética [35] [36] [37] [38]. São, também por essa razão, do mesmo género, excetuando ocorrências atípicas [35]. No caso dos gêmeos monozigóticos, cada embrião tem o seu saco amniótico, mas o saco coriônico rodeia ambos. Assim, existe apenas uma placenta partilhada pelos irmãos [38]. Enquanto que os gêmeos dizigóticos apresentam 50% de correspondência dos genes, os monozigóticos compartilham 100% [37].

A Figura 10 apresenta um esquema simplificado da classificação dos gêmeos [36].

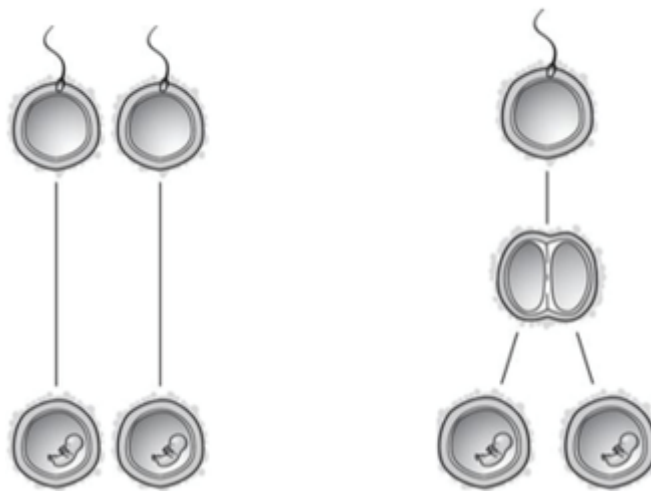


Figura 10 - Esquema da formação embrionária gemelar: gêmeos dizigóticos (esquerda) e gêmeos monozigóticos (direita) [36].

Como referido acima, a forma do trato vocal influencia as características do discurso devido ao filtro e à caixa de ressonância, o que contribui para a diferença entre as frequências formantes dos oradores [39]. Uma vez que os gêmeos monozigóticos partilham 100% do ADN, e como, por exemplo, as características das cartilagens laríngeas são determinadas geneticamente [40], é compreensível que o aparelho fonador seja semelhante, o que contribui para que, muitas vezes, os gêmeos sejam confundidos e não seja possível distinguir, perceptivamente, cada um dos indivíduos. De acordo com Sebastien *et al.* [41], num estudo realizado, mostrou-se que os gêmeos monozigóticos possuem características de voz semelhantes; contudo, alguns intervenientes revelam que, por vezes, conseguem ser distinguidos, o que remete para aspetos que não estão relacionados com os parâmetros acústicos. Lierde *et al.* [40] afirmam, também, que as características perceptivas da voz e as medidas aerodinâmicas laríngeas do tempo máximo de fonação são semelhantes nos gêmeos monozigóticos.

2.2.3.2 Características das vozes de indivíduos gêmeos

Os gémeos monozigóticos são conhecidos por partilharem das maiores semelhanças anatómicas entre os humanos. As variações que podem existir relacionadas com a dentição e o maxilar, como o tamanho ou a abertura da mordida e a inclinação dos dentes, são muito menores nos gémeos monozigóticos do que nos dizigóticos [8]. As características físicas, como o tamanho das pregas vocais, são geneticamente determinadas e estes fatores genéticos influenciam a qualidade vocal [42]. O discurso dos gémeos idênticos é, também, influenciado pelos mesmos problemas de dentição [44]. Além disso, no estudo da aquisição da linguagem, verificou-se, por exemplo, que muitos gémeos desenvolveram códigos apenas compreendidos entre eles. Os erros de pronúncia de certos sons, como substituições, omissões e distorções, são mais comuns nos gémeos monozigóticos do que nos dizigóticos. Observou-se, também, maior semelhança nos padrões da frequência da segunda formante (F2) dos gémeos monozigóticos [8]. Quando existem diferenças entre as vozes de gémeos, podem estar relacionadas com fatores ambientais, o crescimento individual e a adaptação e confraternização na sociedade [43]. Intrínseco a esta tema, surgem também na literatura os conceitos de *nature versus nurture*. O primeiro está relacionado com os fatores biológicos, como a genética, a fisiologia e a biomecânica. *Nurture* refere-se aos parâmetros não biológicos, como o ambiente social, a aprendizagem e os fatores linguísticos a que um falante se encontra exposto ao longo de toda a sua vida [43]. Estes fatores podem contribuir para que os gémeos se expressem de forma diferente e que tenham, até, a intenção de se demarcarem do irmão, por uma questão de identidade [2]. Assim, as diferenças entre os falantes podem ser orgânicas, devido às diferentes estruturas do trato vocal, ou apreendidas, graças aos diferentes hábitos articulatórios [44].

A frequência fundamental das vozes de gémeos é um parâmetro acústico que desperta particular interesse. Segundo Zuo e Mok [44], Debruyne *et al.* [45] e Weinrich e Lancia [50], os valores deste parâmetro aproximam-se mais nos gémeos monozigóticos do que nos dizigóticos e nos falantes que não são parentes. Contudo, alguns autores mediram o tempo máximo de fonação, a frequência mais elevada e a mais baixa, assim como a frequência fundamental e não encontraram diferenças

significativas. Este aspeto pode estar relacionado com os conceitos referidos nesta subsecção, a saber, *nature versus nurture*. Podem existir oradores que, deliberadamente, se tentam demarcar do irmão gémeo e adotam hábitos diferentes de produção de discurso que são significativos e perceptíveis. Estes mecanismos estão excluídos do sistema linguístico e não estão relacionados com o aparelho fonador [9] [43] [45] e podem ser escolhas conscientes ou inconscientes.

Loakes [2], através dos estudos das frequências das formantes do discurso de gémeos idênticos e fraternos, detetou que apenas os monozigóticos não revelaram diferenças, ao contrário dos outros pares, que mostraram diferenças significativas. Whiteside e Rixon [46] investigaram o grau de co-articulação de um par de indivíduos gémeos; concluíram, por exemplo, que a duração das palavras não indicava particular diferença entre os irmãos, na medida em que era semelhante. Os gémeos monozigóticos apresentam diferenças nas características dinâmicas, como os gestos articulatórios. Contudo, o mesmo não se verificou nas características estáticas, como, a título de exemplo, a frequência central das formantes F1, F2 e F3 de 160 palavras monossilábicas com estrutura CVC (consoante-vogal-consoante). Estes resultados suportam a conclusão de que as características estáticas – regiões acústicas estáveis – estão relacionadas com as características apreendidas, enquanto que as dinâmicas decorrem da anatomia [47].

A dinâmica das formantes, ou seja, o comportamento deste parâmetro acústico ao longo do tempo, no espectrograma, foi estudado por Zuo e Mok [44], através de palavras, em indivíduos gémeos bilingues do xangainense e do mandarim. A conclusão foi a de que os gémeos apresentavam diferenças significativas na dinâmica das formantes em ambas as línguas, ao ponto de serem suficientes para distinguirem os irmãos. Contudo, os valores da dinâmica deste parâmetro acústico eram mais semelhantes na língua que dominavam melhor. Este idioma era o usado para os gémeos comunicarem entre si.

Neste estudo, também participaram gémeos que tinham sido separados à nascença; no entanto, os resultados foram tão semelhantes ou ainda mais do que os gémeos que cresceram juntos. Este resultado poderá justificar-se com as características anatómicas do aparelho fonador dos gémeos e à capacidade de escolha dos hábitos

articulatórios [48]. Assim, verifica-se que as escolhas individuais e o fator social, como a identidade, influenciam a produção articulatória.

Um outro estudo, nos âmbitos otorrinológico, auditivo, vocal perceptivo-auditivo e acústico da fonte glótica com gêmeos monozigóticos [49], confirmou que não existem diferenças significativas no que diz respeito ao ataque, articulação, *loudness* e ressonância, na amostra utilizada.

Observa-se que vários estudos acústico-perceptivos comprovam que os gêmeos, especialmente os monozigóticos, se assemelham em questões de valores de parâmetros acústicos, como a frequência fundamental e as frequências das formantes. Além disso, foi estudado que o *jitter*, a frequência formante F1, a frequência formante F2 e o *shimmer* podem ser particularmente evidentes a discriminar os gêmeos [41]. Contudo, não se deve perder de vista que há estudos cujos resultados se distanciam, sendo possível haver a hipótese de estarem relacionados com as escolhas articulatórias pessoais do indivíduo gêmeo, por exemplo.

2.3 Identificação do orador

O Homem possui a capacidade de distinguir vozes, na medida em que armazena informações sobre as características vocais, da mesma maneira que memoriza as particularidades dos atributos físicos, a título de exemplo [51]. Quanto mais familiarizado um ouvinte estiver com uma determinada voz, mais fácil se torna o seu reconhecimento.

Assim, em contexto forense, a identificação do orador está relacionada com tarefas que podem ser executadas, quer através de *software*, quer por profissionais treinados, como os foneticistas. Como a designação indica, esta tarefa consiste no reconhecimento da voz de um indivíduo através de características que possam atestar a verosimilhança quando são comparadas amostras de um indivíduo conhecido com uma ou várias amostras de oradores não identificados.

Segundo Rose [11], Mendes [15] e Pinto [52], existem dois grandes tipos de tarefas de identificação do falante, a saber, a identificação do orador e a verificação do

orador. O objetivo da primeira é identificar uma voz desconhecida do conjunto de amostras de vozes conhecidas ou descartá-la, e este processo é chamado de *closed ou in-set scenario*. Desta forma, aqui, existe a comparação entre a amostra do orador desconhecido e as outras amostras de um grupo de oradores. Por seu turno, a verificação do orador, uma tarefa base no reconhecimento de orador, é o momento em que se verifica se um orador é realmente quem ele afirma ser. Neste procedimento, uma identidade é aceite ou rejeitada, depois de se compararem dois modelos, o do orador e um modelo geral, também designado de *Universal Background Model* (UBM), que pode ser construído com base num conjunto de amostras previamente gravadas de um número alargado de oradores. O reconhecimento de orador implica, ainda, outros métodos, dependentes e independentes de texto, na medida em que as gravações podem ou não corresponder a uma frase singular que foi enunciada por todos os oradores [15].

Alguns dos desafios que dificultam a identificação de orador relacionam-se com fatores que afetam o falante e a voz, como o disfarce, os níveis de ansiedade ou o medo. A doença é, também, um parâmetro que altera os valores dos parâmetros da voz [2]. Outras fontes de variabilidade devem-se ao orador, ao discurso e à tecnologia [52]. Além do que foi acima referido, os níveis de variabilidade de orador também estão relacionados com a variação elementar que existe entre oradores, isto é, a variabilidade inter-orador. A variabilidade relacionada com o discurso tem que ver com a interação com outro orador e o dialeto, por exemplo. A variabilidade relacionada com a tecnologia liga-se à qualidade do áudio e à forma como é capturado, convertido e gravado, havendo articulação com questões eletromecânicas (canal de comunicação), ambientais (ruído de fundo) e de qualidade da amostra (taxa de amostragem, duração, formato de representação e/ou compressão).

A tarefa de discriminação de orador foi originalmente proposta por Doddington, em 1985. Depende de um limiar que pode ser qualitativo – avaliação subjetiva de semelhanças e diferenças – e quantitativo – similaridade expressa numericamente, na medida em que a discriminação está relacionada com a exclusão e a não discriminação tem que ver com a identificação [53].

A identificação forense de orador tem como objetivo contribuir para a identificação de um falante e a exclusão de eventuais suspeitos. Assim, é essencial haver peritos nesta área. Surgem, em particular, nesta temática, sistemas automáticos de reconhecimento de orador em consequência do desenvolvimento da tecnologia e das interações do dia a dia, sejam elas coloquiais ou profissionais [54]. Iremos refletir sobre esta temática na Secção 2.3.4.

2.3.1 Parâmetros perceptivos e parâmetros acústicos

É importante, quando nos referimos a análise acústica forense, selecionar os parâmetros que consideramos serem os mais adequados para a extração de características de voz que possam discriminar o orador de forma eficaz. Assim, nesta subsecção, ilustramos os parâmetros que consideramos mais pertinentes e estudados na literatura, além de serem aqueles com que trabalhamos, na investigação.

Mendes [15] refere-se à informação de alto nível e à informação de baixo-nível, sendo que a primeira está ligada ao estilo de fala e hábitos de dicção e de pronúncia, como o dialeto, e a segunda relaciona-se com as características que resultam da forma do trato vocal, como a frequência fundamental ou a frequência das formantes.

Qualquer ouvinte humano possui a capacidade de reconhecer, implícita e inconscientemente, características de voz através de marcas individuais. Esta tarefa é auxiliada pelo facto de existirem informações que são veiculadas no sinal acústico e que se revelam de âmbito linguístico e sociocultural (contexto) e de índole extralinguístico (características anatómicas do trato vocal) [57].

A caracterização do orador faz-se, portanto, através das marcas dialetais e sociodialetais relacionadas com o género, a formação académica e a localização geográfica, a título de exemplo. Contudo, não se deve perder de vista que as marcas dialetais, por si só, não permitem a imediata identificação do orador, na medida em que são partilhadas por inúmeros falantes. O idioleto é, aqui, um fator fulcral, na medida em que corresponde, num sentido lato, à configuração individual de produção de discurso de cada indivíduo. Todas estas condições colaboram na descrição do perfil de voz – ou assinatura sonora – de um orador.

A qualidade de voz é, também, passível de uma avaliação perceptiva. Laver [59] refere-se à componente orgânica e à componente de ajuste articulatório no que diz respeito a este critério. A componente orgânica não pode ser controlada pelo indivíduo, ao contrário da componente de ajuste articulatório, que é possível dominar deliberadamente. Além disso, a primeira está relacionada com a anatomia do trato vocal de cada orador, que condiciona a qualidade de voz através do volume, massa e fisionomia dos órgãos. O ajuste articulatório diz ainda respeito às configurações articulatorias habituais de cada indivíduo, quando produz discurso, e está relacionado com os *ajustes* supralaríngeo e o fonatório, como o comportamento da laringe e os cinco estilos de vibração: falsete, sussurro, basal, áspero e soproso [59].

No que diz respeito aos parâmetros acústicos, a frequência fundamental é, talvez, um dos mais utilizados na análise acústica. Sucintamente, e como referido na Secção 2.2.1, corresponde à taxa da vibração das pregas vocais, ou seja, o número de vezes que as pregas vocais abrem e fecham, por segundo. De acordo com Kreiman e Sidtis [55], o *pitch* é o correlato perceptivo da frequência fundamental. Além disso, esta frequência está intrinsecamente ligada às características físicas do aparelho fonador do indivíduo e é o atributo que mais realce tem, perceptivamente, para os ouvintes. Este parâmetro possui elevada robustez acústica, mesmo em ambientes poluídos com elevados níveis de ruído de *fundo*, o que o torna muito fiável quando se fala de identificação e de reconhecimento de orador. Não obstante, é importante ressaltar que a frequência fundamental sofre de variações ao longo do dia. A título de exemplo, deve ter-se em consideração as alterações de humor do indivíduo, alguma doença ou a presença de *stress*, como se verifica num estudo levado a cabo por Ghiurcau *et al.* [56]. Aliados à frequência fundamental, existem dois parâmetros que também foram utilizados no estudo, a saber, a média e o desvio padrão [56].

As frequências das formantes, também designadas de padrão acústico das formantes, constituem outro parâmetro acústico relevante para o reconhecimento de orador. As formantes têm especial relevância, aqui, na medida em que representam os resultados dos diferentes volumes das cavidades de ressonância do trato vocal. Segundo Gillier [57], o número das frequências formantes é infinito, mas os dois primeiros, F1 e

F2, são aquelas mais utilizadas, na medida em que se relacionam quer com a abertura do maxilar inferior, quer com os movimentos da língua (avanço e recuo e elevação e abaixamento). São, portanto, formantes responsáveis pela qualidade das vogais e fornecem informação relevante para a identificação do orador, denotando as tendências articulatórias e exprimindo as diferenças de tamanho do trato vocal de cada indivíduo. As duas formantes seguintes de maior frequência, F3 e F4, são, similarmente, fulcrais, uma vez que também estão relacionadas com a anatomia individual do aparelho fonador. Segundo Rose [11], elas refletem o resultado da ressonância das cavidades como a laringe. Frühholz *et al.* [66] referem, ainda, que as formantes F3 e F4 parecem transportar bastante informação sobre o nível de excitação da vocalização afetiva, discriminando várias categorias emocionais. Contudo, beneficiam de pouca informação sobre a distinção das emoções, sendo elas positivas ou negativas.

Na nossa investigação, utilizámos, ainda, três outros parâmetros também relacionados com a perturbação e com a qualidade de voz, concretamente, i) o *jitter*, ii) o *shimmer* e iii) o HNR.

O *jitter* exprime uma “*medida de curto-termo (i.e. ciclo-a-ciclo) de variabilidade involuntária do período fundamental*” [58], e é frequentemente aferida através do quociente de perturbação do período glótico (*Period Perturbation Quocient* – PPQ) envolvendo, por exemplo, 3, 5, 11 ou 55 períodos; apresenta-se em percentagem e, no caso de três períodos, calcula-se através da Equação (1). Nesta equação, $T0_{\text{médio}}$ representa a média dos períodos considerados.

$$PPQ_3 = \frac{\frac{1}{N-2} \sum_{k=2}^{N-1} \left| \frac{T_{k-1} + T_k + T_{k+1}}{3} - T_k \right|}{T0_{\text{médio}}} \quad (1)$$

O *shimmer* exprime uma “*medida de curto-termo (i.e. ciclo-a-ciclo) de variabilidade involuntária da amplitude dos ciclos glóticos*” [58], também se apresenta em percentagem e usualmente é também aferida através do quociente de perturbação da amplitude do ciclo glótico (*Amplitude Perturbation Quocient* – APQ) cujo cálculo é muito similar ao do *jitter*, por exemplo, no caso de se considerarem 3 períodos o cálculo é descrito pela Equação (2) em que $A_{\text{média}}$ representa a amplitude média dos períodos

considerados.

$$APQ_3 = \frac{\frac{1}{N-2} \sum_{k=2}^{N-1} \left| \frac{A_{k-1} + A_k + A_{k+1}}{3} - A_k \right|}{A_{\text{média}}} \quad (2)$$

O HNR (*Harmonics-to-Noise Ratio*) é um parâmetro que exprime a relação entre harmónicos e ruído, em concreto, “(...) relaciona a energia da componente harmónica em relação à energia da componente de ruído do sinal de voz, numa voz saudável (...)”[58]. É calculado através do logaritmo da relação entre a potência média da componente periódica do sinal, representada por P_S , e da potência média da componente ruidosa do sinal, representada por P_R [58], tal como se indica na Equação (3). A unidade é o deciBel (dB).

$$HNR = 10 \log_{10} \left(\frac{P_S}{P_R} \right) \quad (3)$$

As vozes sem patologia, usualmente, possuem um valor *jitter* PPQ5 máximo de 1,5%. Por seu turno, o valor máximo para o *shimmer* APQ5 de vozes saudáveis é de 3,0%. No que diz respeito ao parâmetro HNR, 12dB é o valor mínimo indicativo para uma voz normal, sendo que se o resultado for inferior, pode indicar uma voz soprosa ou disfónica [18]. Normalmente, estes parâmetros avaliam amostras de vogais sustentadas, uma vez que correspondem a um regime fonatório controlado [18]. O comportamento das vogais sustentadas pode ser transposto para o que acontece na voz normal, como as hesitações de discurso [77]. Podemos verificar, então, que o *jitter* e o *shimmer*, enquanto parâmetros de avaliação de perturbação de voz, calculam o comportamento de particularidades temporais do sinal de voz e o HNR avalia os atributos espectrais, sendo um parâmetro de qualidade. Com estes parâmetros, é possível detetar desvios de carácter involuntário face aos valores representativos das vozes normais.

Lindh e Morrison [60] levaram a cabo um estudo em que um *software* e indivíduos voluntários comparavam a performance de análise de voz, através de amostras ouvidas em tempo invertido e na ordem temporal normal. Concluíram que o

ouvido humano consegue distinguir melhor as vozes quando a gravação segue a ordem cronológica natural, uma vez que conseguem extrair informação linguística e fonética, ao contrário do que acontece nas gravações que passaram de trás para a frente, uma vez que essa informação estava mascarada. Por seu turno, o sucesso dos resultados do *software* superou os resultados dos ouvintes, em ambos os tipos das amostras. Neste estudo particular, apresenta-se um processo de conversão das avaliações subjetivas por humanos nas pontuações da razão de verosimilhança, mostrando que o desempenho na comparação de vozes por computadores é melhor do que a comparação de vozes realizada por humanos.

2.3.2 Análise percetiva

A análise percetiva de voz é de natureza subjetiva, na medida em que é realizada por ouvintes humanos. É a mais antiga forma de avaliação qualitativa da voz e é abundantemente recorrente na prática clínica [18]. Compreende a avaliação da voz tendo em conta vários parâmetros como os enunciados na subseção anterior (rouquidão ou aspereza), através da atribuição de uma pontuação entre valores mínimos (quando os valores de voz se aproximam, percetivamente, dos parâmetros de vozes normais) e máximos (quando os valores de uma voz se distanciam, da mesma maneira, dos valores de vozes normais).

Pode acontecer existir variabilidade na análise feita por um indivíduo, mesmo sendo especialista treinado para analisar uma amostra, percetivamente. Este aspeto é denominado de variabilidade intra-avaliador, que se contrapõe à variabilidade inter-avaliador – quando existe variação na avaliação concretizada por distintos profissionais da voz, influenciada pela experiência profissional e pela díspar formação académica [79]. Por seu turno, a variabilidade intra-avaliador está relacionada com a flutuação das referências internas do mesmo avaliador, a que não é alheia a habituação em resultado da prática clínica [78].

A identificação forense de orador começou a beneficiar deste método no Reino Unido entre os anos 60 e os anos 80 do século XX. Foneticistas forenses analisavam o discurso com base em elementos segmentais e prosódicos [61]. Hoje em dia, é

insuficiente para a análise de sinal acústico em contexto de avaliação forense, na medida em que descarta muitos dados que, por exemplo, os parâmetros acústicos podem captar e fornecer. Além disso, o perito pode estar perante dois sinais acústicos diferentes, mas ter a percepção de que são os mesmos. A utilização do Alfabeto Fonético Internacional, um conjunto de símbolos que representam os fones existentes, também faz parte da tradição da análise perceptiva [11].

2.3.3 Análise acústica

Este método de análise de voz apoia-se na avaliação de parâmetros acústicos como a frequência fundamental e o padrão formântico das vogais, estando, por isso, relacionado com a forma, o volume e o tamanho específico do trato vocal de cada orador. Segundo Loakes [2], a análise acústica também se foca, a título de exemplo, na avaliação da co-articulação e na prosódia, que espelham as dimensões distintas, tanto do trato vocal como do comportamento de dicção e dos hábitos articulatórios. De facto, quantos mais parâmetros forem analisados, maior é a probabilidade de haver sucesso na identificação e no reconhecimento do orador.

Dentro do espectro dos parâmetros que podem ser utilizados na análise acústica, Loakes [2] distingue características de curto prazo e características de longo prazo (*short-term/long-term features*). Como ilustração, a avaliação das formantes das vogais faz parte das características de curto prazo e é segmental. Por outro lado, a análise da frequência fundamental e da prosódia faz parte das características de longo prazo.

A análise cepstral é considerada uma forma de análise acústica que faz parte da análise não tradicional, também denominada de automática [62]. Originalmente, foi idealizada para processar os valores da frequência fundamental, mas acabou por auxiliar na identificação e no reconhecimento do orador.

Nos anos 60 do século passado, era comum, nos Estados Unidos, usarem-se *voiceprints* para se proceder à avaliação acústica, tendo sido uma tecnologia desenvolvida nos anos 40 [63]. Este método recorre à interpretação visual dos espectrogramas para analisar enunciados semelhantes e verificar se pertencem ao mesmo orador, o que levanta questões, na medida em que a voz não é única e invariável

tal como as impressões digitais são. Contudo, ainda se utiliza este método em alguns estados dos Estados Unidos da América.

Clinicamente, o objetivo da avaliação acústica é o de detetar conjunturas de patologia através de informação de diagnóstico e de procedimento não-invasivo [18].

2.3.4 Reconhecimento automático do falante

Nolan [4] categorizou o reconhecimento de oradores tendo em conta o reconhecimento técnico e o reconhecimento *naive*. O primeiro associa-se ao reconhecimento automático (utilização de máquinas computadorizadas) e semiautomático (utilização de espectrogramas por peritos), envolvendo técnicas analíticas, humanas e automáticas. O reconhecimento *naive*, como o nome indica, é executado por indivíduos sem formação específica nesta área, recorrendo, apenas, ao reconhecimento percetivo.

A abordagem automática é uma das perspetivas de comparação forense de voz, sendo que as outras são a percetiva, a espectrográfica e a fonético-acústica [63]. Foi desenvolvida por engenheiros da área do processamento de sinal e fundamenta-se na avaliação quantitativa dos atributos acústicos do discurso.

O reconhecimento automático de falante aponta para um processo forense de reconhecimento de voz que opera mecanicamente, na medida em que o processo se baseia na extração de características, na modelagem, na pontuação de comparação e no cálculo de razões de verosimilhança. A deteção de voz pode igualmente estar incorporada neste processo, mas não é uma etapa imperativa. O perito forense tem de estar envolvido em todo o procedimento, quer para o controlar quer para examinar os resultados [64].

De acordo com Pinto [52], a razão de verosimilhança é uma avaliação com base na relação de probabilidades que oferece confiança na identificação de orador. Este conceito baseia-se na definição de duas hipóteses e na especificação das probabilidades associadas. No caso da comparação forense de vozes, testa-se se o orador A produziu ou não determinada amostra e pode-se definir, antes de se efetuar a comparação, que i) H_0 exprime a hipótese de uma amostra ter sido produzida pelo orador e ii) H_1 exprime a

hipótese de uma amostra não ter sido produzida pelo orador. A H_0 está associada uma curva de probabilidades com base em pontuações de comparação (*scores*) de amostras de voz do mesmo orador, e a H_1 está associada uma outra curva de probabilidades com base em *scores* de comparação de amostras de voz de oradores diferentes. Com base nestas curvas e com base no score da comparação entre uma amostra de voz que se sabe pertencer ao orador A e uma outra amostra de um indivíduo suspeito, a razão de verosimilhança consiste simplesmente, com base no score obtido, na relação entre a probabilidade da hipótese H_0 ser verdadeira, e a probabilidade da hipótese H_1 ser verdadeira. Para que H_0 seja verosímil, existe um limiar para a razão de verosimilhança que tipicamente deve exceder, em muito, a unidade. Este tema será aprofundado na Secção 4.5. Uma interpretação verbal para certas gamas do valor da relação de verosimilhança é apresentada em Rose [11].

Nos sistemas automáticos de reconhecimento de orador, as características extraídas do sinal de fala constituem uma das configurações mais importantes desses sistemas, na medida em que são elas que fornecem pistas idiossincráticas sobre os oradores e possibilitam arquitetar os seus modelos ou protótipos matemáticos. Uma vez que a nossa dissertação não se centra especificamente em terminologia da área da engenharia, iremos resumir algumas das propriedades que se encontram organizadas em hierarquias, sendo que as características do trato vocal têm primazia em relação aos hábitos articulatórios do orador. Catalogam-se, assim, como informação de baixo nível e informação de alto nível [52].

Já referidas na Secção 2.3.3, as características de curto termo (*short-term features*) consistem em parâmetros extraídos de segmentos curtos do sinal de fala, tipicamente entre 10 ms e 100 ms, e que se relacionam com os Coeficientes Cepstrais de frequência Mel e com as características fundamentadas na codificação linear preditiva [52]. Por seu turno, as *high-level features* são usadas para a extração de características de alto nível através de *tokens* que retratam aceções linguísticas relevantes [52] [54]. Em relação às características de curto termo, para evitar a sua degradação, existem técnicas de normalização como a *Mean Subtraction* – subtração do vetor médio de características de todos os vetores de particularidades do indivíduo, para cada amostra –

e a *Mean Variance Normalization* – dedução do vetor médio de características de todos os vetores de atributos individuais e divisão pelo desvio padrão [52] [54] [65].

Os classificadores são elementos fulcrais no reconhecimento automático de orador, uma vez que decidem o reconhecimento com base numa modelação na qual são reunidas as propriedades idiossincráticas de um orador. Como exemplo, existe a modelação estatística de características do orador através de uma combinação curvas Gaussianas (*Gaussian Mixture Model* – *GMM*) que, desejavelmente, modela o que há de específico num orador. Por outro lado, existe o *Universal Background Model (UBM)*, também na forma de um *Gaussian Mixture Model*, no qual se representam as características de um conjunto elevado de oradores, ou seja, modela o que há de típico num orador médio, isto é, não diferenciado. A classificação decorre da relação entre a similaridade obtida entre o modelo GMM da amostra de um orador desconhecido e o modelo GMM de um orador suspeito – e, portanto, conhecido –, em relação à similaridade entre o primeiro modelo GMM e o relativo ao UBM. Uma correspondência é assumida (“match”) quando a relação entre a especificidade e a tipicidade ultrapassa um determinado limiar. Mais recentemente, esta mesma abordagem é usada, não com modelos GMM, mas com uma forma de modelação mais elaborada e conhecida por *i-Vectors (Identity Vectors)* [52]. A motivação para os *i-vectors* está relacionada com a modelização robusta do orador, através de *supervectors*, procurando-se reduzir a sua dimensionalidade, o que facilita a classificação, e procurando-se, simultaneamente, minimizar as influências do canal de comunicação [52] [54]. Esta discussão será detalhada no Capítulo 6 quando se abordar uma perspectiva histórica do desenvolvimento nesta área que conduza à compreensão da arquitetura do *software iVocalise*.

A variação no discurso humano faz com que o reconhecimento automático de orador seja um desafio permanente. Deve enfatizar-se que o reconhecimento automático, por *software*, tem o obstáculo de comparar discurso que pode estar afetado pelo disfarce, doença, stress, esforço vocal ou emoção. Além disso, analisar uma conversa entre duas ou mais pessoas é diferente de analisar discurso de um humano proferido para um dispositivo (como o telemóvel) ou mesmo uma amostra com um texto lido. O simples facto de se isolar a voz que se pretende estudar, eliminando outras

que possam constar da gravação, altera a amostra e pode ser um desafio para o reconhecimento automático de falante.

Uma vez que os sistemas de reconhecimento automático se baseiam, portanto, nos parâmetros do aparelho fonador e explicam não só características estáticas¹¹, como também características dinâmicas relacionadas por exemplo com a co-articulação, distinguir gémeos monozigóticos pode tornar-se numa dificuldade. Künzel [8], num estudo levado a cabo com recurso ao *software* *Batvox* 3.1, concluiu que nem os gémeos idênticos podem ser réplicas exatas no que diz respeito às características da voz e ao discurso, na medida em que foi possível distinguir os irmãos uns dos outros. Observou, também, que as vozes femininas são mais difíceis de analisar, porque a frequência fundamental é mais elevada, o que provoca que o espaçamento entre os harmónicos seja menos denso do que no género masculino.

2.3.5 Condições ideais e condições reais

Idealmente, o foneticista forense teria total controlo sobre as amostras, evitando, por exemplo, a variabilidade que fomos discutindo ao longo das subsecções anteriores. Contudo, esse cenário não se verifica na realidade.

De acordo com Rose [11], parte da variação pode ser aleatória e verificar-se em função de diferentes fatores durante o discurso. Contudo, também é possível ser sistemática, estando relacionada com as diferenças de tamanho no trato vocal (variações entre-falantes) e com as emoções (variação intra-falante). Desta forma, dependendo das circunstâncias e do interlocutor, o registo de voz vai oscilar, até porque se adequa o registo à situação conversacional, o que dificulta o reconhecimento automático de orador. Assim, é necessário conhecer em que medida as vozes variam segundo diferentes momentos e registos do dia, aliados a emoções ou até a doenças, como laringites, e em que medida as situações afetam, por exemplo, os valores da frequência fundamental.

¹¹ Tal como referido na Secção 2.2.3.2, presume-se uma escala temporal na ordem das poucas dezenas de milissegundos.

A redução da quantidade de parâmetros de voz é outra questão de grande importância na problemática do reconhecimento do falante através de um sistema automático. Ao telefone, devido à função de filtragem, nem todas as frequências que constituem o sinal de voz são preservadas na banda passante do canal telefónico, tipicamente entre 300 Hz e 3400 Hz, o que faz com que se tenha menos características para se proceder à comparação de vozes. Neste contexto também, o tempo é um fator crucial, na medida em que extrair particularidades e comparar estatisticamente amostras do discurso é um processo moroso. Em vários países, como na Alemanha e na Suécia, o foneticista forense é-lo a tempo inteiro, ao contrário do que acontece na Austrália ou no Reino Unido, em que a análise é feita, normalmente, por académicos foneticistas que lecionam, fazem investigação e orientam projetos [11].

Seguindo as conclusões de Rose [11], a distorção, frequente nas comunicações efetuadas através de canal telefónico, é outro fator que faz parte das condições reais com que o perito tem de lidar no reconhecimento automático de falante. Pode, do mesmo modo, resultar de pobres condições de gravação, má qualidade do microfone ou de salas com efeito substancial de reverberação.

Outra questão relacionada com as condições ideais do reconhecimento automático de orador é a representatividade, designadamente, o quão representativos são os dados disponíveis sobre determinada voz. Com efeito, a robustez da prova de voz, em casos forenses, depende da maneira em como a amostra A, a do suspeito, e a amostra B, a amostra de comparação, refletem a produção articulatória padrão de cada um deles. Esta questão está dependente dos dados que estão disponíveis: número de amostras e de parâmetros acústicos envolvidos e a sua dimensão [11].

Todos estes fatores reais podem afetar as condições que seriam desejáveis, mas que não acontecem, uma vez que a voz humana é caracterizada por variáveis que diferem de indivíduo para indivíduo e entre vários oradores, como verificámos nas secções anteriores, além de poderem ser contaminadas por influências do canal de comunicação e das condições acústicas ambientais.

2.4 Fonética forense

A fonética forense faz parte da ciência forense e é o principal método pelo qual se identificam suspeitos de um delito com a ajuda de um perito ou de um foneticista forense. Assim, a finalidade da fonética forense é a de fornecer teorias e métodos altamente especializados com aplicações judiciais. De facto, nos últimos anos, a análise e comparação de vozes tem aumentado no âmbito das investigações criminais [67] e com o desenvolvimento das tecnologias de voz.

O principal objetivo desta ciência, considerada um ramo da linguística forense, é o reconhecimento de falantes; contudo, também tem como papel autenticar gravações, reconhecendo que a amostra não foi alterada de alguma maneira, ou identificar conteúdos de uma amostra [57].

Além do reconhecimento e da identificação de orador, a fonética forense também é transversal a áreas como i) a caracterização socioeconómica, por exemplo, do falante, com recurso ao estudo do dialeto, ii) a construção de alinhamentos de voz (*voice line-ups*) e iii) a identificação do conteúdo, quando a gravação é de má qualidade e é necessário distinguir o que foi dito ou a voz é patológica, está disfarçada ou tem marcas de um sotaque estrangeiro [11]. Na prática, de acordo com Nolan [68], os casos típicos incluem chamadas falsas para os serviços de emergência, chamadas telefónicas com conteúdo obsceno e de assédio, negócios fraudulentos ou pedidos de resgate.

2.4.1 Âmbito e expressão em Portugal

De acordo com Martins *et al.* [69], o órgão que atribui a importância efetiva da fonética forense na produção de prova, em Portugal, é o tribunal. O papel do perito forense, hoje em dia, é resultado do crescimento tecnológico e da sociedade moderna em que imperam as gravações de voz e de imagem com recurso a telemóveis, principalmente. Para responder a este progresso repentino das tecnologias, é necessário desenvolver metodologias cada vez mais fiáveis e, para isso, é imprescindível apostar na formação académica, nesta área.

No nosso país, o papel do perito de fonética forense é o de auxiliar o juiz a tomar uma deliberação a partir das provas apresentadas e da apreciação do profissional, de acordo com as técnicas forenses referidas nas secções anteriores. De facto, o ordenamento do processo penal português beneficia da prova material pericial e da prova material documental; o foneticista forense produz a pericial.

As escutas ou interseções telefónicas, previstas nos artigos 187 a 190 do Código de Processo Penal, edificam o meio legal de obtenção de prova, em Portugal, sendo facilmente alcançadas através dos meios de comunicação telefónicos quando existe um despacho do juiz, que zela sempre pela reserva à vida privada dos cidadãos. Braz [70] considera que este meio de obtenção de prova é um dos mais modernos previstos na lei e um dos que oferece maior robustez contra a criminalidade.

O tribunal português, depois de o relatório pericial ser apresentado, irá apreciá-lo tendo em conta o seu valor técnico-científico. Assim, a sua apreciação está condicionada ao critério legal da livre apreciação da prova (artigo 163, nº1 do Código de Processo Penal). O juiz pode, portanto, discordar do relatório pericial, apesar de o tribunal não possuir, *à priori*, conhecimentos técnicos para formular um juízo em contrário. O perito pode, similarmente, ser solicitado a coadjuvar o tribunal na avaliação de uma questão probatória.

A intervenção de foneticistas forenses ou de peritos nesta área é cada vez mais necessária e exigente, devido à difusão e sofisticação dos meios de comunicação. De salientar que, em abril de 2017, fomos solicitados para participar numa perícia que será objeto de discussão no Capítulo 7. Com efeito, apenas com investigação e validação aplicada é exequível criar parâmetros mais sólidos que permitam identificar um suspeito e, em última análise, condená-lo, em tribunal. Cada vez mais os peritos forenses são requeridos, em processos judiciais, devido à legalização das interseções telefónicas como forma legal de obtenção de prova e de produção de prova em juízo. Assim, o papel da fonética forense é, tendencialmente, cada vez mais necessário e visível, especialmente se fundamentado em práticas atualizadas e credíveis. Seguramente que as perícias serão mais robustas se os conhecimentos sobre os parâmetros da voz, das

características da linguagem humana e de reconhecimento automático de orador progredirem e prosperarem [69].

2.4.2 A pertinência do estudo de vozes de gémeos

Quando nos referimos a gémeos, aceitamos, empiricamente, que as vozes são semelhantes; este aspeto percetivo conduz à admissão de que, percetivamente, não se consiga distinguir este género de irmãos, especialmente se não houver familiaridade com eles.

O estudo de vozes de gémeos é pertinente, na medida em que explora ao máximo limites subtis na variação entre os falantes; se as vozes de gémeos conseguem ser discriminadas forensicamente, de forma robusta, recorrendo a parâmetros acústicos apropriados, então isto pode constituir um ponto seguro de partida para o estudo de vozes de outros falantes que não possuem laços familiares [2].

Ainda neste contexto, é importante recordar que é necessário verificar, aquando da comparação de vozes, se os gémeos partilharam do mesmo meio de desenvolvimento e se viveram juntos. Este género de diferenças permite tirar conclusões bastante pertinentes. De acordo com Pinto [52], várias investigações foram levadas a cabo, envolvendo gémeos dizigóticos, monozigóticos e indivíduos sem parentesco. Conclui-se que é mais fácil distinguir oradores sem relação do que os gémeos e, dentro deste género de irmãos, os monozigóticos são os que tornam mais árdua a tarefa de diferenciação. Em particular, a média da frequência fundamental é mais semelhante nos monozigóticos do que nos dizigóticos, aspeto explicado pela semelhança anatómica do aparelho fonador, explorada na Secção 2.2.3. Ainda de acordo com Pinto [52], a nível de comportamento e de hábitos de fala, a média da frequência fundamental oscila tanto nos gémeos dizigóticos como nos monozigóticos.

Weinrich e Lancia [50] confirmam as conclusões anteriores, uma vez que concluem ser mais complicado distinguir gémeos entre si do que entre oradores sem parentesco; os resultados obtidos entre gémeos monozigóticos e dizigóticos não eram completamente díspares. Asseveraram, também, que a média da frequência fundamental modifica para ambos os tipos de gémeos.

Um caso que evidencia estes aspetos aconteceu na Austrália, encontra-se descrito em [11] e consiste num caso de irmãos em que um deles foi acusado de tráfico de droga. O advogado alegou que as gravações não podiam ser atribuídas ao suspeito por as vozes serem semelhantes. Contudo, a fonética forense conseguiu provar que havia hábitos diferentes de articulação ao nível de uma das vibrantes múltiplas, na medida em que foi altamente discriminativo para a comparação entre amostras conhecidas e desconhecidas. Assim, de acordo com as perspetivas de San Segundo e Gómez-Vilda [71], quanto mais influenciado geneticamente um parâmetro é, mais robusto vai ser para o papel da comparação entre os falantes.

Existem estudos que se focaram em alguns elementos anatómicos dos gémeos, como, por exemplo, os dentes, o maxilar e a língua; contudo, não se centraram na aplicação dos resultados à área da fonética. De facto, os gémeos monozigóticos adotam, segundo esses estudos, hábitos articulatórios mais semelhantes entre si do que os dizigóticos. Existe falta de coordenação entre os estudos fonéticos e articulatórios de gémeos, até porque, na vida real, tipicamente não se utiliza o método invasivo de introduzir um dispositivo na cavidade oral do suspeito para a medir ou verificar os hábitos articulatórios, por exemplo [36].

Por todas estas razões, podemos afirmar que o estudo da voz de gémeos é um desafio para a fonética forense e, mais especificamente, para o reconhecimento automático de voz, na medida em que são genética e anatomicamente idênticos, além de partilharem, em condições ideais e frequentes, o mesmo ambiente de desenvolvimento [71] [72].

2.5 Conclusão

Este capítulo focou-se nos conteúdos fulcrais para a investigação sobre o sistema humano de produção de fala, ou seja, o aparelho fonador, passando pelas características físicas e percetivas da voz e remetendo para o aparelho fonador de oradores gémeos, tendo em conta a perspetiva anatómica e uma breve explicação sobre a formação embrionária dos gémeos monozigóticos e dizigóticos. Este foi o ponto de partida para

uma explicação sobre as características de vozes de indivíduos gémeos. De seguida, refletiu-se sobre a identificação de orador, aludindo aos parâmetros acústicos e percetivos que consideramos pertinentes para os resultados da dissertação. Na Secção seguinte, destacou-se a importância e o desafio que é o reconhecimento automático de falante, considerando as condições reais e as condições ideais. Para concluir o capítulo, é referida a fonética forense e o seu âmbito e expressão, em Portugal, assim como a relevância do estudo das vozes de gémeos.

Verifica-se que o presente capítulo se apresenta como o pilar para o desenvolvimento da dissertação e mostra-se crucial para a investigação e a aprendizagem. Dado que o trabalho, nesta dissertação, aborda temáticas diferentes, como o estudo de parâmetros da voz relacionados com perturbação e qualidade, com ressonâncias do trato oral (formantes) e com abordagens de reconhecimento automático do falante, considerações mais específicas sobre o estado-da-arte em cada uma das temáticas serão apresentadas nos capítulos em que estas são tratadas.

Capítulo 3 – Planeamento e estruturação da base de dados

3.1 Introdução

Este capítulo introduz pontos fulcrais na nossa investigação, como os procedimentos adotados no recrutamento de voluntários gémeos e familiares, através de envio de *e-mail* institucional da Universidade do Porto, e justifica a utilização de um diálogo orientado e conduzido no momento de recolha das amostras de fala, assim como a gravação de vogais sustentadas e da leitura em voz alta do texto foneticamente equilibrado “O vento norte e o sol” [80]. Explica-se, resumidamente, a escolha das modalidades de comunicação GSM e VoIP com o intuito de refletir sobre o seu impacto na degradação das vozes, aferido através da comparação de parâmetros de voz ou de comparação de vozes. Descreve-se, ainda, a construção das diferentes bases de dados, assim como se justifica a tarefa de isolamento das vogais sustentadas.

O planeamento e a estruturação da base de dados refletem todo o processo de análise estatística que será debatido no Capítulo 4. São as bases de dados que permitirão o cálculo dos rácios de verosimilhança¹² (*Likelihood Ratios* – LR) para cada membro de pares de gémeos, sendo usados todos os outros membros dos pares de gémeos para formar as curvas H_0 e H_1 . Além disso, recorrendo às outras bases de dados que construímos e que são alvo de caracterização neste capítulo, iremos, posteriormente, nos Capítulos 5 e 6, refletir sobre um dos principais desafios na tarefa de identificação de orador.

3.2 Cenário para recolha de dados

¹² **Rácios de verosimilhança** – Traduzem a robustez dos resultados da tarefa de reconhecimento de voz de orador, através de cálculos determinados pela confiança probabilística.

Em janeiro de 2014, no seguimento de um instrumento de avaliação da unidade curricular de Processamento de Sinais Fisiológicos, frequentada no ano letivo de 2011/2012 na Faculdade de Engenharia da Universidade do Porto, foi lançado um apelo via *e-mail* institucional (ver Anexo 1) com o objetivo de reunir voluntários para um estudo de voz telefónica e que reunissem certos requisitos, designadamente, serem gémeos do mesmo género, e serem acompanhados por um familiar próximo, como pai, mãe, irmão, irmã, primo ou prima, com disponibilidade para fazer parte do estudo. Considerando as escassas respostas que se obteve, na altura, lançou-se novo apelo, em junho de 2015, através do mesmo meio de comunicação (ver Anexo 2). Assim, depois de reformulações de agendamento por alguns dos contactos, foi possível reunir 37 indivíduos, concretamente, 20 do género feminino e 17 do sexo masculino. Por razões anatómicas do trato vocal, na Secção 3.5, iremos caracterizar a amostra reunida tendo em conta esta demarcação de géneros. A recolha das amostras de fala foi realizada no edifício de Engenharia Eletrotécnica da Faculdade de Engenharia da Universidade do Porto, entre novembro de 2015 e março de 2016. Recorremos a um diálogo orientado, construído por nós, e que incluía também, em particular, exercícios de vogais sustentadas assim como a leitura do texto foneticamente equilibrado “O vento Norte e o Sol” [80] (ver Anexo 3).

O objetivo da elaboração de um diálogo orientado é também o de, através de um *corpus* previamente definido, identificar tipos de palavras que expõem a identidade sonora do orador, estando, por isso, relacionadas com a complexidade articulatória. Os temas neutros, como o clima ou a ocupação profissional, o percurso académico e o uso de respostas “sim” e “não”, assim como o abrangimento de vogais sustentadas – facilitando o uso de medidas objetivas [77] (segundo análise acústica convencional) – foram o foco da interação entre o entrevistador e o indivíduo. O interlocutor necessitou de ter aptidão para levar o orador a produzir, de modo natural, os vocábulos que foram considerados essenciais, listados, abaixo. O objetivo era que fossem articulados por todos os falantes do estudo:

- Namorado(a);

- Horas;
- Telefone;
- Pai / Mãe;
- Escola + Estudar;
- Profissional;
- Vizinho(a);
- Sim / Não.

Além disso, os oradores gémeos verbalizaram o nome completo, idade, local de nascimento e profissão. Quanto aos familiares próximos, foi imprescindível referirem os nomes dos parentes (gémeos).

O papel de interlocutor foi desempenhado por nós, uma vez que estamos familiarizados com discursos orientados, ou seja, discursos que conduzam o orador para proferir a palavra que se deseja. A título de exemplo, à questão “Que horas são?”, o indivíduo pode responder, imediatamente, “São quatro”, omitindo a palavra “horas”. A função do interlocutor será a de conseguir com que o orador profira, expressamente, “horas”.

Na revisão de bibliografia relacionada com análise de vozes de gémeos, verificámos algumas metodologias no que diz respeito à obtenção de amostras para análise. Fernández [81] recolheu duas gravações, realizadas com duas ou quatro semanas de intervalo, conversas espontâneas semiestruturadas entre os pares de gémeos. Debruyne *et al.* [45] solicitaram aos oradores que lessem um texto previamente selecionado (“Papa en Marloes”). Lierde *et al.* [40] estudaram a vogal sustentada /a/, depois de aquecimento vocal. De seguida, analisaram a mesma vogal em “Ik zie een grote <a>” (*Eu vejo um grande <a>*). De forma similar, Weinrich e Lancia [50] utilizaram a frase “Ich wasche Haku im Garten” (*Eu lavo o Haku no jardim*) para analisarem /ˈvaʃə/. Loakes [82] serviu-se de conversas telefónicas para investigar os encontros consonânticos /str/, /tr/ e /tʃ/, nos dialetos do inglês australiano. Por seu turno, Whiteside e Rixon [9] optaram pela leitura de uma lista de palavras monossilábicas previamente concebida, sendo que os investigadores estariam focados em vocábulos

cuja consoante tivesse pontos de articulação bilabiais, alveolares, velares e glotais (/b d g h/). Loakes [2] optou por conduzir uma entrevista em que o orador argumentava sobre aspetos da vida, como o percurso académico ou a ocupação profissional. De seguida, o voluntário leu uma lista de palavras. Verificou-se, também, um conjunto de tópicos em que os indivíduos escolhiam um sobre o qual deveriam produzir discurso, fundamentando as suas opiniões. Numa segunda sessão, passados meses, tiveram direito a escolher, novamente, um tema. O objetivo era analisar os onze monotongos do inglês australiano, recorrendo às frequências de formante F1, F2, F3 e F4. Os oradores do estudo de Weinrich [43] repetiram as vogais /a i u/ quarenta vezes, em frases diferentes. Weinrich e Fuchs [83] recorreram à articulografia eletromagnética e à eletropalatografia com o objetivo de analisarem o palato dos gémeos na produção de /sʃ/. Künzel [8] optou por pedir aos participantes que lessem a versão em alemão de “O Vento do norte e o vento do sul” e, posteriormente, que se referissem a aspetos da vida, como a ocupação profissional, os *hobbies* ou o facto de terem um irmão gémeo. Johnson e Azara [84] criaram, da mesma forma, uma lista de palavras. Por seu turno, a leitura de um poema (“I wandered lonely as a cloud”, de William Wordsworth) foi ferramenta de Ariyaeeni [85], assim como a verbalização da data de nascimento de cada indivíduo.

O diálogo orientado conduzido por nós iniciou-se com um cumprimento entre os interlocutores. De seguida, solicitou-se aos oradores que verbalizassem o nome completo, o local e a data de nascimento, assim como a idade, o nome dos pais, do irmão gémeo e do familiar, caso fizesse parte do estudo. Depois, o orador dizia que dia e horas eram, o clima favorito e o estado civil. Com o objetivo de se obterem respostas mais extensas, foram pedidas as definições, para os oradores, do que significam as palavras “namorado” e “vizinho”. Seguidamente, perguntou-se qual a última escola que tinham frequentado, se tinham gostado de estudar, ou não, e pedimos que justificassem. Questionou-se qual a profissão do orador e do irmão gémeo e, em seguimento, o que definia um bom e um mau profissional. Pretendeu-se obter as palavras “não” e “sim”; por conseguinte, questionámos se gostavam de chocolate e de beterraba. Caso o orador respondesse só que “sim” ou que “não” a ambas as perguntas, improvisávamos, utilizando outro alimento. A penúltima parte do diálogo conduzido consistiu em

solicitar que o orador produzisse, sustentadamente, as cinco vogais tónicas¹³. Para terminar, era lido o texto “O vento norte e o sol”.

Com os oradores que eram menores, foi utilizado o mesmo protocolo. Mudaram-se, apenas, as questões relacionadas com namorados e vizinhos, sendo substituídas por questões sobre os passatempos favoritos e os locais para os quais gostavam de viajar. Do mesmo modo, questionou-se qual era o cargo profissional dos pais e que profissão gostariam de exercer, no futuro (ver Anexo 4).

3.3 Tipos de canal e modalidade de gravação

No nosso estudo, foram utilizados dois tipos de canais para gravação simultânea do mesmo diálogo, designadamente, o canal de alta qualidade e o de qualidade telefónica. Foram também usadas modalidades distintas de conversação telefónica, em dois diálogos diferentes, mas comparáveis do ponto de vista do conteúdo, um consistindo na modalidade de comunicação GSM (*Global System for mobile Communications*) – i.e. através de telefone de rede móvel – e um outro consistindo na modalidade de comunicação VoIP (*Voice over Internet Protocol*) – i.e. através de uma chamada *Skype* – com base num terminal físico de um lado (i.e. em *hardware*) e com base num terminal em *software* do outro lado. Estas duas modalidades, por serem relativamente comuns na atualidade, foram escolhidas para estudar e compreender o impacto da degradação das vozes em tarefas de comparação de vozes e identificação de orador.

A comunicação GSM tem a função de conduzir serviços móveis de dados e de voz e é uma tecnologia digital aberta. A VoIP, por sua vez, possibilita as chamadas telefónicas através de uma rede de dados, na medida em que os sinais digitalizados da fala são expedidos através de uma conexão à internet por meio de pacotes com endereço IP [52]. Com o objetivo de simularmos interações telefónicas e de captarmos, também, para referência, as versões de alta qualidade de uma chamada, foram produzidos dois tipos de registos concomitantes para o mesmo diálogo telefónico, nomeadamente, como

¹³ Vogais tónicas – /a/, /e/, /i/, /o/, /u/.

acima referido, o registo de voz de alta qualidade – a que associaremos a sigla HQ – e registo de voz de qualidade telefónica – a que associaremos a sigla TL.

Nas Figuras 11 e 12, podemos verificar uma representação dos cenários utilizados durante a recolha das amostras de fala.

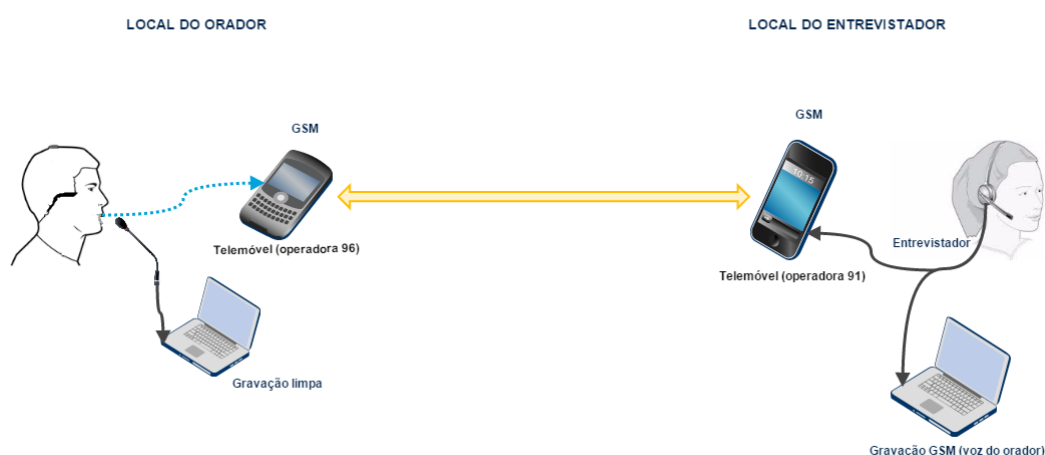


Figura 11 - Representação da ligação na recolha de amostras de fala com modalidade de comunicação GSM, em ambos os tipos de gravação: HQ e TL.

Todas as conversações telefónicas foram gravadas em tempo real. Deve realçar-se que os suportes na emissão e receção eram distintos, na medida em que, embora se encontrassem no mesmo edifício do Departamento de Engenharia Eletrotécnica e Computadores da Faculdade de Engenharia da Universidade do Porto, o entrevistador se situava num laboratório de telecomunicações e o orador se encontrava num gabinete, física e acusticamente distante do primeiro.

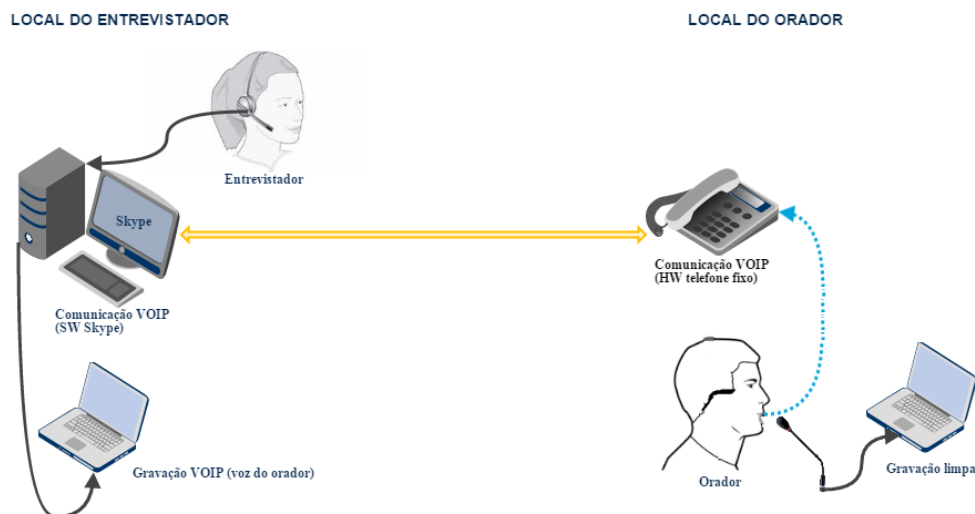


Figura 12 - Representação da ligação na recolha de amostras de fala com modalidade de comunicação VoIP, em ambos os tipos de gravação: HQ e TL.

Para o registo feito no local do orador, usou-se o seguinte equipamento (interface de aquisição de sinal de voz em alta-qualidade):

- Microfone de cabeça Sennheiser, de alta qualidade (Ear Set1-ew3);
- Adaptador Sennheiser (Modelo MZA 900 com *phantom power*);
- Interface áudio de digitalização A/D USB (Cakewalk UA-25EX by Roland);
- Computador pessoal portátil para gravação direta em alta qualidade (HP Compaq 6910p com sistema operativo Windows XP Professional), utilizando o editor de áudio *Adobe Audition* versão 1.5. A qualidade deste registo de áudio é definida por uma frequência de amostragem de 48 kHz e 32 bits de resolução por amostra.
- Para a comunicação celular GSM, o orador usou um equipamento HTC Magic e o seu respetivo auricular. A MEO foi a operadora de telecomunicações (indicativo 96);
- Para a modalidade de comunicação VoIP, foi usado um telefone VoIP físico (Polycom, modelo SoundPoint IP331).

Por sua vez, para o registo no local do entrevistador, usaram-se os seguintes equipamentos:

- Telefone Nokia 108 e o respetivo conjunto original de headphone e de microfone, sendo que a Vodafone (91) foi a operadora de comunicações;
- Computador Sony Vaio (modelo PCG-31311M) com sistema operativo Windows 7 para gravação do sinal de voz GSM ou VoIP, através do *software Audacity* (versão 1.2.6). A qualidade deste registo de áudio é definida por uma frequência de amostragem de 22.05 kHz e 16 bits de resolução por amostra;
- Computador de mesa para executar a aplicação Skype para comunicação VoIP (Intel Core i5 2500K @ 3.3 GHz, 4GB RAM e equipado com interface de áudio Realtek High Definition interface).

3.4 Questionários

Antes de procedermos à recolha das amostras, e depois de mantermos uma conversa de registo informal e descontraído com os participantes, realizámos um pequeno questionário, concebido por nós, com o objetivo de caracterizar os oradores participantes (ver resultados preliminares na Secção 3.5). Assim, as questões de partida estavam relacionadas com dados como i) a data e ii) o local de nascimento, iii) o local de residência (freguesia, concelho e distrito), iv) o nível de habilitações concluído, v) a língua materna dos oradores e dos pais, vi) o grau de parentesco que o familiar próximo partilhava com os gémeos e vii) possíveis patologias que pudessem interferir com a voz. Figurava no questionário, também, um campo em que eram registadas observações que eram tidas como pertinentes aquando da recolha de voz ou da conversa coloquial com os oradores (ver Anexo 5).

Anteriormente à recolha das amostras de fala, foram distribuídos a todos os participantes declarações de consentimento informado (ver Anexo 6) para que tivessem conhecimento, por escrito, do objetivo de estudo, e para que confirmassem que concordavam em participar na investigação, sendo assegurado o anonimato de cada

orador participante. Uma vez que existiram voluntários menores de idade, foram facultados aos encarregados de educação documentos de consentimento semelhantes, mas adaptadas para o efeito (ver Anexo 7).

3.5 Participantes

Dentro do género feminino, existem sete pares de gémeas. De todos, apenas um par não teve presente uma familiar próxima, o que contabiliza 20 oradoras. O gráfico da Figura 13 apresenta a distribuição dos oradores pelos anos de nascimento, sendo que os agregámos por décadas, de 1931 até 2010 (o mesmo procedimento foi adotado nos indivíduos do género masculino).

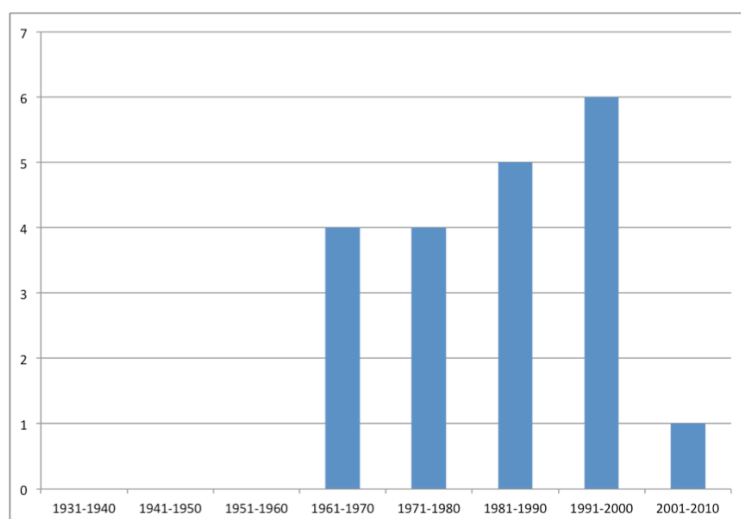


Figura 13 - *Datas de nascimento dos oradores do género feminino, compreendidas entre décadas.*

Verifica-se que a maioria das oradoras nasceu entre 1991 e 2000, seguido da década de 1981-1990. Por questões de maturação do aparelho fonador, consideramos relevante destacar que a única falante da década de 2001-2010 nasceu em 2001.

Os locais de nascimento das oradoras pertencem, na sua grande maioria, à zona Norte do Portugal, estando distribuídos por Aveiro, Porto, Paredes e Vila Nova de Gaia. Duas intervenientes nasceram em Viseu e outras duas em Angola e na Alemanha. Por seu turno, os locais de residência de todas as oradoras fixam-se no grande Porto.

A língua materna (L1) das oradoras e dos pais é o Português e as falantes que nasceram na Alemanha assumiram o Alemão, também, como L1.

Onze oradoras possuem grau académico, entre a licenciatura, o mestrado e o doutoramento. Duas concluíram o ensino básico e sete terminaram o ensino secundário.

No que diz respeito aos graus de parentesco, entrevistaram três mães, uma prima, uma irmã e uma filha de uma das gémeas.

De destacar as observações assinaladas pelas intervenientes no que diz respeito a possíveis doenças crónicas ou condições a apontar no momento da recolha, como tosse, congestionamento nasal, bronquite ou asma. É de realçar, também, que a voluntária número 19 (FEM19) nos comunicou que tinha a perceção de que a sua voz tinha um registo mais agudo do que o da irmã gémea.

Procedeu-se à recolha de sete pares de gémeos do género masculino. Três deles não tiveram a oportunidade de se fazerem acompanhar de um familiar próximo, o que perfaz dezassete oradores. De acordo com o gráfico seguinte, os anos de nascimento dos oradores distribuem-se, apenas, por quatro décadas, o que significa que o grupo é mais homogéneo a nível de idades do que o dos oradores femininos. Assim, verifica-se que a maioria dos oradores masculinos nasceu entre 1991 e 2000.

No que diz respeito ao local de nascimento, a maior parte dos intervenientes provém do grande Porto (Porto, Vila Nova de Gaia, Vila Nova de Famalicão e Guimarães). Existem cinco indivíduos que nasceram em Coimbra, Viana do Castelo e Chaves. Todos moram no grande Porto.

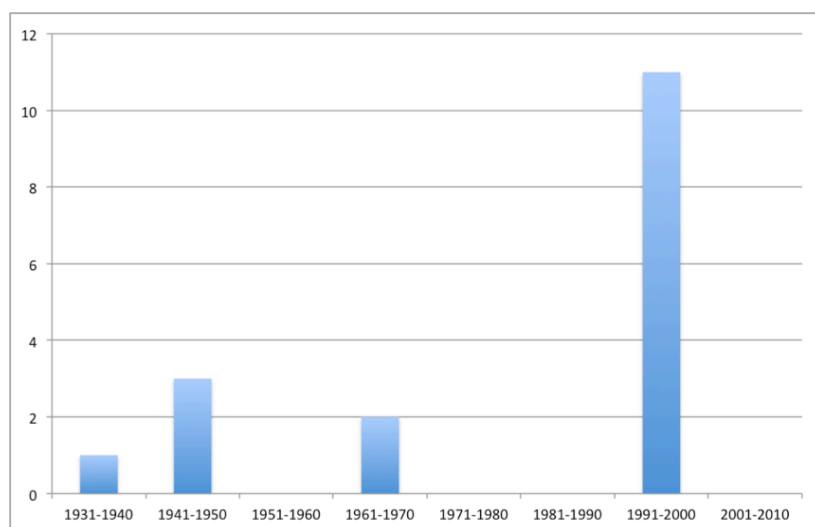


Figura 14 - *Datas de nascimento dos oradores do género masculino, compreendidas entre décadas.*

Os oradores participantes afirmaram que a L1 deles e dos pais é a mesma, especificamente, o português. Dos dezassete participantes, nove são licenciados ou mestres e os restantes concluíram o 12º ano ou o ensino básico.

As relações de parentesco dos familiares próximos ficam-se por dois primos e um pai.

Apenas o orador masculino número 05 (MAL 05) referiu que era fumador há três anos. Foi observado, no momento da condução dos diálogos e em conversa informal nos momentos prévios à gravação que o orador 12 (MAL 12) se apresentava muito comunicativo e fornecia respostas mais longas e detalhadas, ao contrário do irmão, o falante 13 (MAL13), que era introvertido e menos conversador, deixando o irmão responder às questões, sem intervir de alguma forma. Quando o fazia, era através de respostas curtas. Também foi notória a prosódia de ambos, na medida em que diferia perceptivamente.

3.6 Base de dados das gravações

Nesta secção, enunciar-se-ão os processos pelos quais passaram as amostras para se adequarem aos diferentes estudos que se pretendia realizar usando os *software* que estarão em discussão no Capítulo 4. Como referido, todas as recolhas de voz foram feitas recorrendo a uma versão de alta qualidade (HQ) e a uma versão de qualidade telefónica (TL), em ambas as modalidades de comunicação GSM e VoIP, com frequência de amostragem de 22.05 kHz e 16 bits de resolução por amostra (versão telefónica) e de 48 kHz e 32 bits de resolução por amostra (versão de alta qualidade).

3.6.1 Base de dados inicial

Cada uma das gravações dos 37 oradores permitiu que se produzissem quatro gravações de áudio, nomeadamente, versões concomitantes de qualidade telefónica e de alta qualidade, em GSM e VoIP, o que perfaz 148 ficheiros (GSM_HQ – modalidade GSM e canal de alta qualidade; GSM_TL – modalidade GSM e canal de qualidade telefónica; VOI_HQ – modalidade VoIP e canal de alta qualidade; VOI_TL modalidade VoIP e canal de qualidade telefónica). Para ser expedita a caracterização da amostra e identificação de orador, organizou-se uma tabela com o número atribuído a cada orador, a idade no momento da recolha e os graus de parentesco. “FEM” corresponde ao género feminino e “MAL” diz respeito ao masculino. A partir dessa tabela, criou-se um ficheiro com os nomes, para o falante ser de fácil identificação, mas, por questões de anonimato, não fará parte da dissertação. Na Tabela 1, detalham-se a designação de cada orador, a idade e a relação de parentesco.

Uma vez que partilhamos a nossa base de dados com uma estudante de mestrado da Faculdade de Engenharia da Universidade do Porto [52], retirámos do seu estudo o quadro relativo à duração, em minutos e segundos, dos quatro registos de cada orador. Esta informação encontra-se detalhada na Tabela 2.

Tabela 1 - Dados relativos à idade e ao grau de parentesco dos oradores.

Orador	Idade	Grau de parentesco
FEM01/FEM02/FEM03	22/22/45	Gémea-gémea-mãe
FEM04/FEM05/FEM06	27/27/32	Gémea-gémea-prima
FEM07/FEM08/FEM09	38/34/34	Irmã-gémea-gémea
FEM10/FEM11/FEM12	21/55/21	Gémea-mãe-gémea
FEM13/FEM14/FEM15	22/49/22	Gémea-mãe-gémea
FEM16/FEM17/FEM18	49/15/49	Gémea-filha-gémea
FEM19/FEM20	36/36	Gémea-gémea
MAL01/MAL02/MAL03	71/71/71	Gémeo-primo-gémeo
MAL04/MAL05	21/21	Gémeo-gémeo
MAL06/MAL07/MAL08	77/52/52	Pai-gémeo-gémeo
MAL09/MAL10/MAL11	24/24/24	Gémeo-gémeo-primo
MAL12/MAL13	16/16	Gémeo-gémeo
MAL14/MAL15	16/16	Gémeo-gémeo
MAL16/MAL17	19/19	Gémeo-gémeo

Através do editor de áudio *Adobe Audition*, a versão de alta qualidade foi modificada para a frequência de amostragem de 22.05 kHz e 16 bits de resolução por amostra, de modo a se obter o mesmo formato de representação para ambas as versões (alta qualidade e qualidade telefónica).

Tabela 2 - Duração, em minutos e segundos, dos quatro registos de cada orador [52].

Oradores	Duração dos registos			
	GSM_HQ	GSM_TL	VOI_HQ	VOI_TL
FEM01/FEM02/FEM03	03:07/03:01/03:02	03:07/03:01/03:02	03:10/02:37/04:06	03:10/02:37/04:07
FEM04/FEM05/FEM06	03:33/03:14/02:36	03:32/03:12/02:30	04:16/04:28/03:08	04:16/04:28/03:06
FEM07/FEM08/FEM09	02:43/04:05/03:33	03:58/04:04/03:35	03:07/04:52/03:07	03:07/04:50/03:06
FEM10/FEM11/FEM12	02:38/02:30/02:56	02:38/02:32/02:56	02:36/02:50/03:34	02:37/02:53/03:33
FEM13/FEM14/FEM15	02:30/02:43/02:38	02:30/02:45/02:40	02:32/02:58/02:36	02:32/03:00/02:34
FEM16/FEM17/FEM18	03:36/02:50/03:46	03:39/02:38/03:44	03:13/02:17/03:22	03:12/02:18/03:23

FEM19/FEM20	05:25/04:23	05:28/04:23	04:06/03:12	04:05/03:11
MAL01/MAL02/MAL03	04:07/02:59/03:35	04:07/03:03/03:38	07:03/04:10/09:31	07:05/04:10/09:31
MAL04/MAL05	04:25/03:53	04:25/03:51	03:18/03:41	03:17/03:58
MAL06/MAL07/MAL08	03:57/03:48/03:02	03:57/03:48/03:01	03:19/03:14/02:44	03:18/03:16/02:42
MAL09/MAL10/MAL11	03:10/02:51/02:35	03:10/02:51/02:34	03:36/03:51/02:49	03:37/03:50/02:49
MAL12/MAL13	03:15/03:15	03:15/03:10	02:30/02:40	02:30/02:39
MAL14/MAL15	02:16/02:15	02:16/02:16	02:11/02:08	02:11/02:08
MAL16/MAL17	03:03/02:56	03:02/02:54	03:25/03:18	03:23/03:17

3.6.2 Isolamento de regiões de vogais sustentadas

Todas as regiões das várias gravações contendo vogais sustentadas foram segmentadas manual e individualmente, recorrendo ao editor de áudio *Adobe Audition*. Posteriormente, identificaram-se as que teriam qualidade para se extraírem parâmetros que pudessem ser analisados estatisticamente. Assim, verificou-se que algumas vogais apresentavam ruído excessivo ou perturbações do sinal que podiam conduzir a resultados errados ou valores estatísticos descabidos, mormente em consequência de anomalia no funcionamento dos algoritmos de processamento de sinal.

Seguidamente, definimos que a duração mínima de cada vogal sustentada seria de um segundo para ser exequível extraírem-se os parâmetros na região mais estável do respetivo espectrograma. Para isso, e durante a segmentação, eliminaram-se as zonas mais instáveis, especificamente, o início e o fim da vogal. Identificaram-se, portanto, os segmentos de vogais considerados problemáticos por não cumprirem os requisitos da duração mínima de um segundo – eram mais curtos – ou cuja frequência fundamental não era estável, subindo ou descendo abruptamente. Por conseguinte, de modo a dispor-se de uma base de dados equilibrada de vogais sustentadas para os quatro tipos de gravações e para todas as cinco vogais, aquela ficou circunscrita a 19 oradores, concretamente, 10 do género feminino e 9 do género masculino (ver na Tabela 1 os oradores destacados a negrito).

3.6.3 Remoção do segundo orador e conversão da frequência de amostragem

Em resultado de exigências decorrentes dos estudos discutidos na Secção 6.3, procedeu-se a uma operação de pré-processamento manual, através do editor áudio *Adobe Audition*, das gravações de voz de cada orador. O propósito era o de remover a presença residual da voz do interlocutor – isto é, do entrevistador, sem afetar o registo de voz do orador – de modo a se conseguir uma modelização de orador o mais fidedigna possível. É essencial lembrar que as amostras de fala a nível forense possuem, por vezes, não só os registos da voz do orador que se quer examinar, mas também as de um segundo interlocutor, o que perturba os objetivos de análise e modelização.

As ondas osciloscópicas (Figura 15) e os respetivos espectrogramas (Figura 16) ilustram as diferenças acústicas entre a presença do interlocutor e a sua remoção.

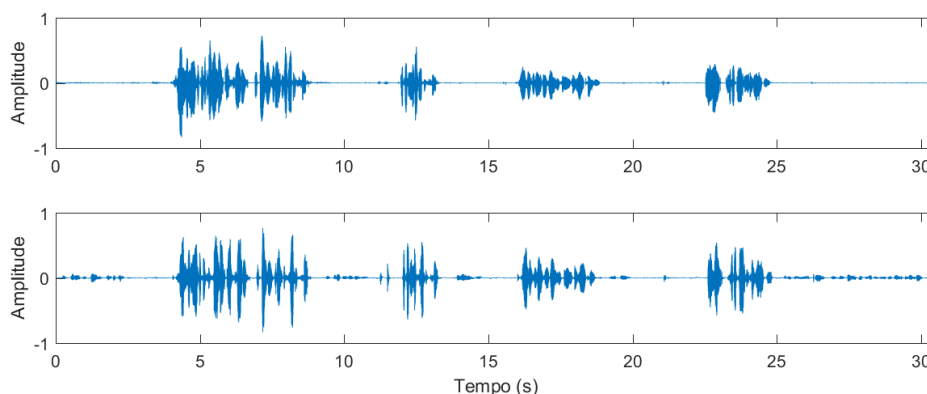


Figura 15 - O painel superior, nesta figura, representa a onda osciloscópica original, na emissão, e o painel inferior representa a onda osciloscópica na receção e denotando a presença da voz do interlocutor.

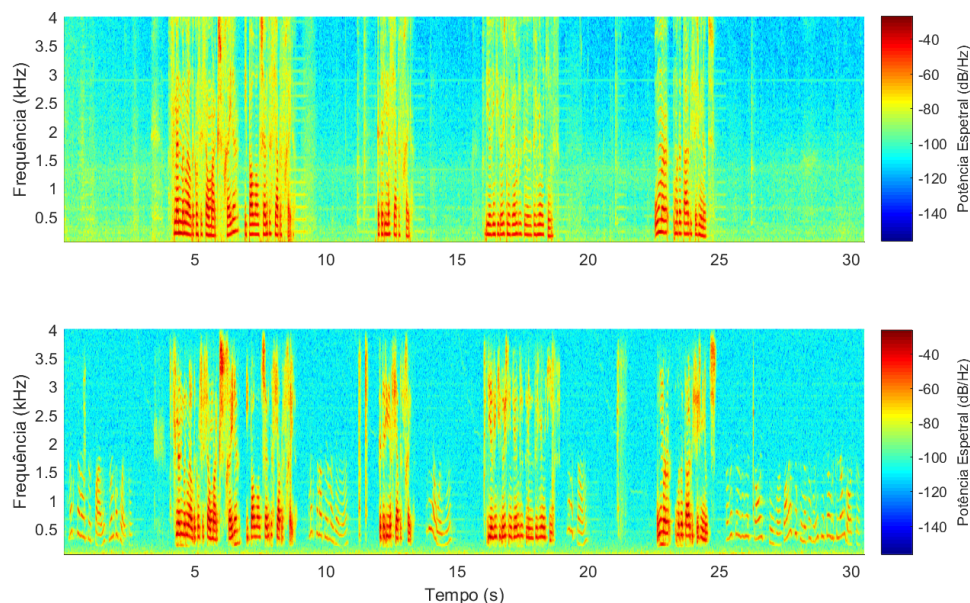


Figura 16 - O painel superior, nesta figura, representa registo espectrográfico original, na emissão, e o painel inferior representa o registo espectrográfico na receção e denotando a presença da voz do interlocutor.

Como será detalhado no Capítulo 6, a frequência de amostragem mais adequada dos registos de voz para o propósito de reconhecimento automático de orador através do *software* utilizado, o *iVocalise*, é 8 kHz. Neste sentido, foi necessário proceder-se à conversão da frequência de amostragem de todos os ficheiros das diversas gravações, de 22.05 kHz para 8 kHz. Esta conversão foi executada recorrendo, também, ao *Adobe Audition*. Garante-se, assim, que se torna possível beneficiar do mesmo formato de representação para as duas versões das gravações contemporâneas, designadamente, a de alta qualidade e a de qualidade telefónica e que, em consequência, os correspondentes modelos de orador são construídos com base nos mesmos pressupostos.

3.6.4 Subdivisão das gravações na base de dados

As condições de gravação para cada orador, contemplando dois tipos de canal (HQ e TL) e duas modalidades de comunicação (GSM e VoIP), não permitem comparações de modelos de orador construídos com base em registos não concomitantes, tal como se ilustra na Figura 17.

	GSM_HQ	GSM_TL	VOI_HQ	VOI_TL
GSM_HQ				
GSM_TL				
VOI_HQ				
VOI_TL				

Figura 17 - Condições de gravação para cada orador, contemplando os dois tipos de canal (HQ e TL) e as duas modalidades de comunicação (GSM e VoIP).

De facto, tal como se ilustra, nesta figura, através de células mais fortemente sombreadas, as condições de comparação segundo a diagonal principal, implicam registos de fala que são exatamente os mesmos, o que retira pertinência à comparação. Por outro lado, as células mais levemente sombreadas na Figura 17, implicam diferentes versões (HQ e TL) da mesma conversação, o que essencialmente significa que os correspondentes modelos de orador só diferem devido à modificação do sinal de fala imposta pelo canal de comunicação. Idealmente, desejar-se-ia que o modelo de orador para cada uma das condições disposta quer na vertical, quer na horizontal da Figura 17, se baseasse num único registo de conversação telefónica, não repetido em qualquer outra condição.

Por outro lado ainda, cada modelo de orador obtido como se sugere na Figura 17, apenas pressupõe uma só condição de canal (GSM ou VoIP). Tal como se discutirá no Capítulo 6, há pertinência em construir modelos de orador que combinem, para o mesmo orador, diferentes condições de canal, assegurando-se mútua exclusividade, não só para os registos de fala de várias – pelo menos duas – condições de canal que são usados para construir um só modelo de orador, quer para todos os que intervêm na comparação entre dois modelos de orador.

De modo a dar resposta às duas preocupações acima apresentadas, foi necessário realizar a divisão dos registos existentes, em duas metades, de modo a se preservar, em cada uma, a mesma estrutura de conversação, mais concretamente, uma parte de diálogo e uma parte de leitura. Esta operação de divisão encontra-se ilustrada na Figura 18 em que se exemplifica a separação de um ficheiro de fala GSM_HQ em duas metades, GSM_HQ_1h e GSM_HQ_2h, com durações temporais comparáveis e com estrutura de conteúdo – entre diálogo e leitura – idêntica à do ficheiro inicial.

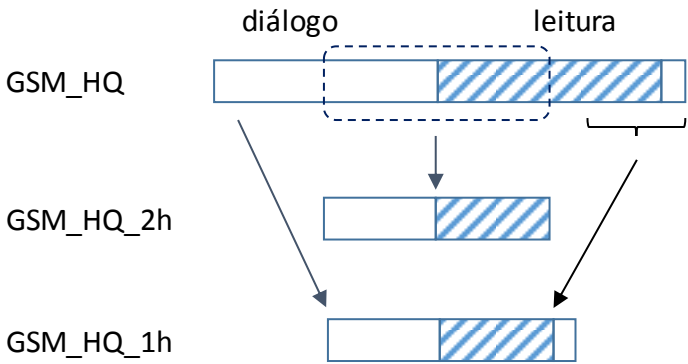


Figura 18 - Operação de divisão dos registos existentes na base de dados inicial.

A divisão foi realizada manualmente através do editor *Adobe Audition*, para todos os registos de cada orador (1ª metade e 2ª metade de GSM_HQ, GSM_TL, VoIP_HQ e VoIP_TL). Tal como se ilustra na Figura 19, tirando partido da referida divisão, mesmo as comparações para o mesmo orador e condição de canal, fundamentam-se em modelos de orador construídos com base em registos de fala mutuamente exclusivos.

	GSM_HQ_2h	GSM_TL_2h	VOI_HQ_2h	VOI_TL_2h
GSM_HQ_1h				
GSM_TL_1h				
VOI_HQ_1h				
VOI_TL_1h				

Figura 19 - Esquema dos modelos de orador construídos com base em registos de fala não concomitantes.

Os vários estudos de comparação realizados encontram-se descritos no Capítulo 6 e decorrem de resultados obtidos com o *software iVocalise* (ver Secção 4.4). Este *software* exige que a duração mínima de voz ativa se localize entre 15 a 30 segundos para ser exequível construir um modelo de orador e proceder à comparação de várias gravações [87]. Para se obter a voz ativa, eliminam-se as regiões de silêncio das gravações, o que corresponde a uma função automática disponível no *software iVocalise*. Mesmo nestas condições, a duração dos ficheiros divididos cumpre o requisito de duração mínima.

Através da combinação de ficheiros relativos ao mesmo orador mas refletindo diferentes condições de canal, pretende-se, como expresso acima, aferir qual a sensibilidade do *software iVocalise* à diversidade de canal, na modelização de dados e orador.

Por conseguinte, foram criadas seis novas bases de dados, combinando, cada uma, os registos de voz de meia duração, de forma diferente, de modo a criar condições para estudos de comparação em que os ficheiros implicados sejam mutuamente exclusivos. A extensão de cada ficheiro subdividido é “1h” ou “2h”, em que “1h” significa *1st half* e “2h” designa *2nd half*. A Figura 20 esquematiza o modelo de construção das seis bases de dados.

		Designação da nova base de dados
GSM_HQ_1h	VOI_TL_2h	1 – GSM_HQ_1h_VOI_TL_2h
GSM_HQ_2h	VOI_TL_1h	2 – GSM_HQ_2h_VOI_TL_1h
VOI_HQ_1h	GSM_TL_2h	3 – VOI_HQ_1h_GSM_TL_2h
VOI_HQ_2h	GSM_TL_1h	4 – VOI_HQ_2h_GSM_TL_1h
GSM_TL_1h	VOI_TL_2h	5 – GSM_TL_1h_VOI_TL_2h
GSM_TL_2h	VOI_TL_1h	6 – GSM_TL_2h_VOI_TL_1h

Figura 20 - Método de construção de seis novas bases de dados.

Como se discutirá no Capítulo 6, o *software iVocalise* foi usado para comparar i) a base de dados 1 *versus* a base 2, ii) a base 3 *versus* a base 4 e iii) a base 5 *versus* a base 6. Destas bases de dados, obter-se-ão três conjuntos de testes, na medida em que cada conjunto produz resultados para oradores do tipo FEM, outros para oradores do tipo MAL e outros ainda para a combinação MAL+FEM. Posteriormente, conduziram-se novos testes que serão detalhados no Capítulo 6. O objetivo era o de averiguar se o modelo de falante obtido com várias metades de diferentes gravações de fala, conduzia, ou não, a resultados mais robustos de identificação de orador.

3.7 Conclusão

Este capítulo debruçou-se sobre a base de dados original, que foi construída através da recolha de vozes contemplando diversidade nas condições de tipos de canal gravado (alta qualidade *versus* telefónico) e modalidades de comunicação telefónica (GSM *versus* VoIP). Assim, discutiram-se, brevemente, aspetos implicados no planeamento e recolha de dados, como a angariação de gémeos e familiares voluntários, a fundamentação do uso dum diálogo orientado, da gravação das vogais sustentadas e da leitura de um texto foneticamente equilibrado. Clarificaram-se quais os tipos de canal utilizados e as modalidades de gravação que se considerou que replicavam as condições típicas das interseções telefónicas. Fez-se menção aos equipamentos utilizados, aos questionários aos participantes, ao isolamento das vogais sustentadas, à remoção do segundo orador e à necessidade de conversão da frequência de amostragem dos ficheiros de voz. Caracterizaram-se, também, sucintamente, os oradores, sem prejuízo para o anonimato dos participantes implicados. Por último, justificou-se a necessidade de subdivisão dos ficheiros de voz gravada, e referiu-se a estrutura das novas bases de dados assim resultantes. Este tema será retomado e detalhado no Capítulo 6, com destaque para a Secção 6.4.

Capítulo 4 – Ambientes e ferramentas de processamento

4.1 Introdução

Nas tarefas de reconhecimento automático de orador, o recurso a *software* especializado é imprescindível para se obterem resultados de análise e comparação úteis, objetivos e reportáveis. Assim, este capítulo introduz conceitos chave sobre algumas aplicações de *software* que se revestiram de especial importância para a nossa recolha, edição e análise das amostras de voz, incluindo a análise estatística.

4.2 *Software Audacity e Adobe Audition*

O *Audacity* é um *software* livre para edição digital de áudio, tendo várias funções, como a gravação e reprodução de sons, a importação e exportação em formato não-comprimido *wav* e formato comprimido *mp3*. Em concreto, é exequível recorrer à edição simplificada como cortar, colar, copiar ou apagar, remover ruídos ou utilizar a função de igualização. Na nossa investigação, utilizou-se a versão 1.2.6 para se gravarem as amostras de voz telefónica, no computador, como se ilustrou nas Figuras 11 e 12. Na Figura 21, é ilustrada a interface deste *software*.

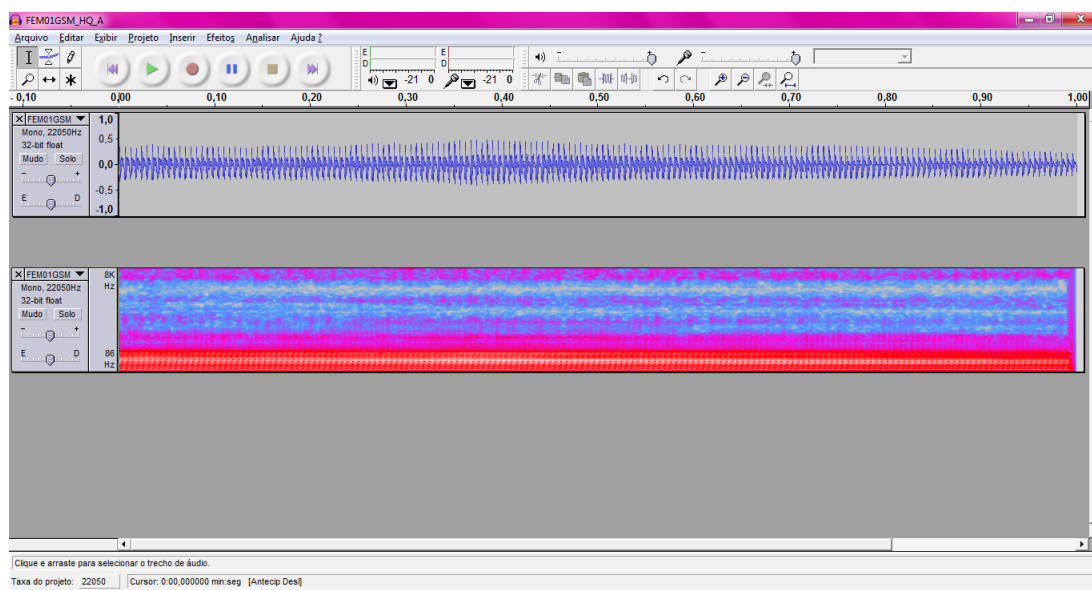


Figura 21 - Interface do Audacity (análise acústica da vogal sustentada /a/ de FEM01, em GSM_HQ).

O *Adobe Audition* é também um editor de ficheiros de áudio – a nível profissional – e, para além da análise de certas características de gravações de voz, possui muitas funções especializadas de pós-processamento como filtragem e síntese de sons, com uma clara vocação para produção musical. Permite, ainda, detetar a presença de ruídos e de vozes que não sejam as do falante.

Durante a investigação, considerou-se que a utilização deste editor era intuitiva e agradável. De facto, é possível personalizar o ambiente de trabalho, podendo, a título de exemplo, agruparem-se painéis com as funções de processamento que são essenciais, trabalhar-se com múltiplos monitores ou usar os efeitos de reverberação para simular vários ambientes em espaços fechados. Contudo, o *Adobe Audition* não disponibiliza ferramentas com funcionalidade de extração de parâmetros específicos para voz, como as frequências formantes.

A Figura 22 representa a interface do *Adobe Audition*, na análise acústica de uma vogal sustentada.

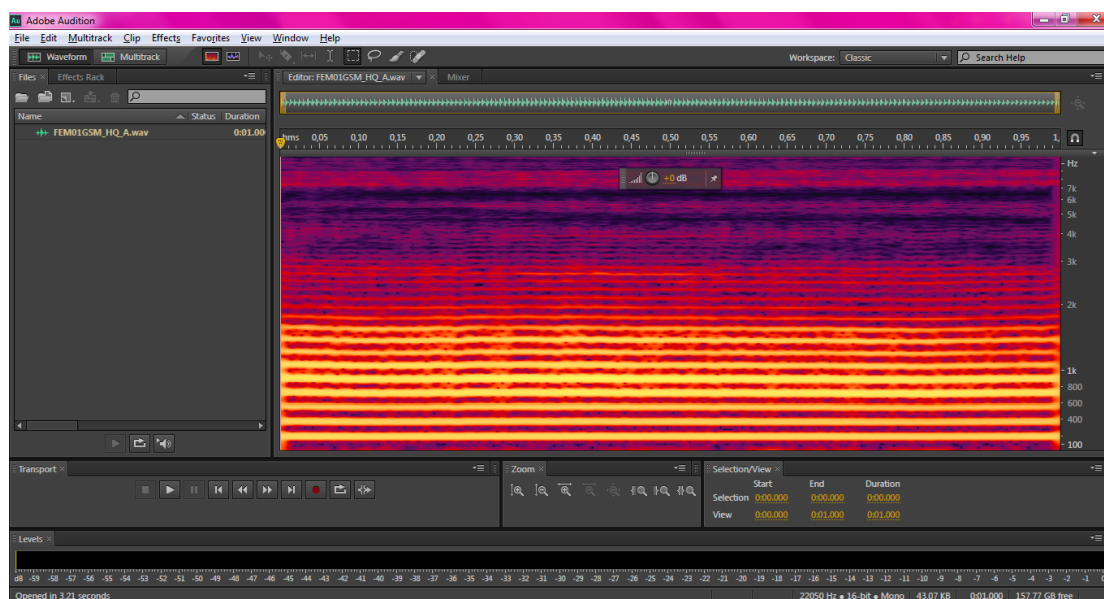


Figura 22 - Interface do Adobe Audition (análise acústica da vogal sustentada /a/ de FEM01, em GSM_HQ).

No âmbito da dissertação, foi utilizada a versão 1.5 do editor *Adobe Audition*.

4.3 Software Praat

O *Praat* é um *software* de ambiente académico para análise acústica de voz e foi criado por Paul Boersma e David Weenink, do *Institute of Phonetic Sciences*, da Universidade de Amsterdão, nos Países Baixos [86]. Abrange várias funcionalidades de medição de parâmetros de voz, como os de qualidade, por exemplo, através do menu “Pulses”, em que se extraem medidas que calculam as irregularidades na frequência – *jitter* – e na amplitude – *shimmer*. Clicando no separador “Voice Report”, obtêm-se os valores para os vários tipos de *jitter* – incluindo o referido nesta dissertação (PPQ5) – de *shimmer*, como o APQ5, a média e o desvio da frequência fundamental e o parâmetro HNR. Através do *Praat*, é igualmente possível medir a amplitude e os valores das diferentes frequências formantes. Outras funcionalidades incluem segmentar e anotar um sinal e, de seguida, guardar cada um dos segmentos em ficheiros individuais. Permite, igualmente, sintetizar voz. É também exequível programar funções de

processamento pelo utilizador, como revelar, apenas, a frequência fundamental e/ou as formantes ou escolher o método de análise da frequência fundamental (“autocorrelation” ou “crosscorrelation”).

De especial importância para a nossa investigação foi a funcionalidade que permitiu extrair as frequências formantes F1, F2, F3 e F4, através do algoritmo *Burg*. Depois de seleccionar o sinal acústico relevante, acede-se a “Formant” e, seguidamente, clica-se em “Formant listing”. A nível de análise espectrográfica, é possível analisar os ficheiros acusticamente através de espectrogramas de banda larga e de banda estreita.

A Figura 23 representa a interface do *Praat* na análise da vogal sustentada /a/, de uma oradora, em registo GSM_HQ. A utilização do *software* revelou-se amigável e intuitiva.

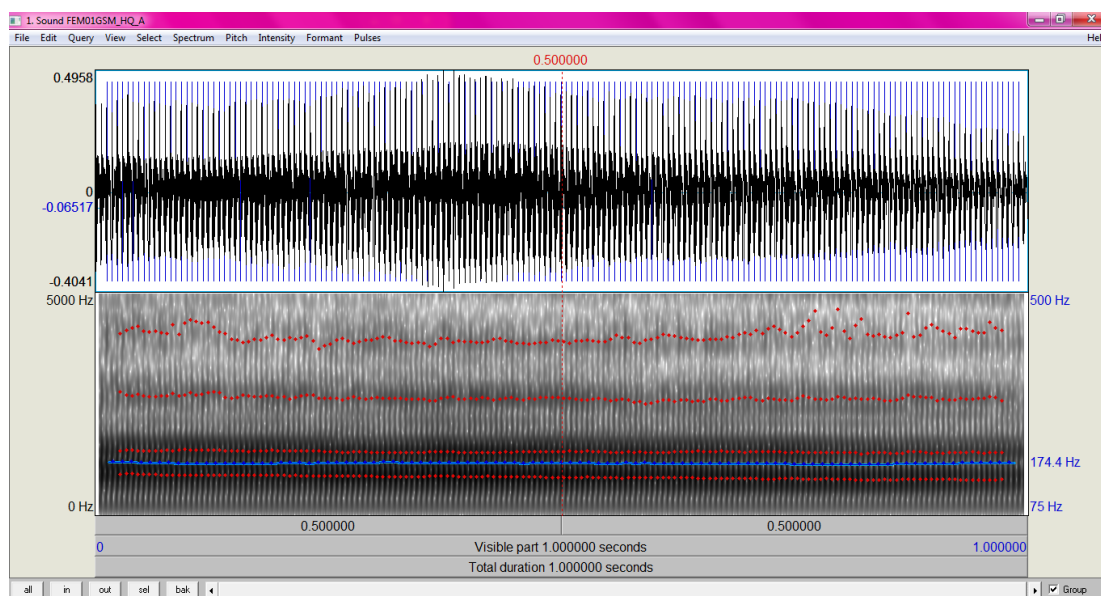


Figura 23 - Interface do *Praat* (análise acústica da vogal sustentada /a/ de FEM01, em registo GSM_HQ).

Em termos práticos, consiste num ambiente de trabalho útil para áreas clínicas como a Terapia da Fala, em que é utilizado para analisar perturbações da fonação, como a disfonia, por exemplo [29].

A versão do *software Praat* utilizada na nossa investigação foi a 6-0-23.

4.4 Software *iVocalise*

Este *software* foi desenvolvido pela empresa *Oxford Wave Research* e tem como propósito ser um sistema de reconhecimento automático de orador que oferece ferramentas de análise de sinal de fala, extração de características fonéticas (*e.g.* frequências formantes) e espectrais (*e.g.* coeficientes MFCC), modelização estatística, criação de modelos de orador e cálculo de distâncias entre modelos de orador. Permite, assim, ao perito, emitir um parecer fundamentado, objetiva e quantitativamente, sobre a plausibilidade da voz de determinado falante – um suspeito – corresponder, ou não, à voz gravada de um orador desconhecido e implicado, por exemplo, numa situação de crime.

O *software* foi desenvolvido de acordo com um sistema *openbox*, ou seja, o foneticista forense tem a possibilidade de escolher os parâmetros sobre os quais quer basear a análise e modelo de orador. Ilustra-se, na Figura 24, a principal interface do *iVocalise*.

Para além de permitir avaliar a distância entre modelos de dois oradores distintos, isto é, efetuar a comparação de um-para-um, permite também efetuar a comparação de um-para-muitos, que é a circunstância ilustrada na Figura 24 em que as várias distâncias são apresentadas numa lista e têm o valor numérico assinalado na coluna de *scores*. No caso mais ambicioso, em termos práticos, permite também a comparação de muitos-para-muitos em que os vários *scores*, ou pontuações, são apresentados numa matriz que é, posteriormente, fornecida a um outro *software*, o *Biometrics* (ver Secção 4.5), para análise de resultados e representação gráfica. Esta modalidade de comparação de distâncias entre gravações de um grupo de oradores, e um outro grupo de oradores, que pode ter um número diferente do primeiro¹⁴, é especialmente importante para definir estatisticamente o perfil de pontuações correspondentes a comparações entre diferentes registos de fala mas pertencentes ao mesmo orador, e o perfil de pontuações correspondentes a comparações entre registos

¹⁴ Se o número de oradores do primeiro grupo (grupo de análise) e do segundo (grupo de comparação) for o mesmo, a matriz de resultados é quadrada, se for diferente, a matriz de resultados é retangular.

de fala de oradores diferentes (ver Secção 4.5) [65] [87]. É esta caracterização estatística que permite estabelecer a gama dinâmica das pontuações (entre a situação de mesmo orador e a situação de orador diferente) e que permite também calcular os rácios de verosimilhança (*Likelihood ratios*), como se detalhará na Secção 4.5.

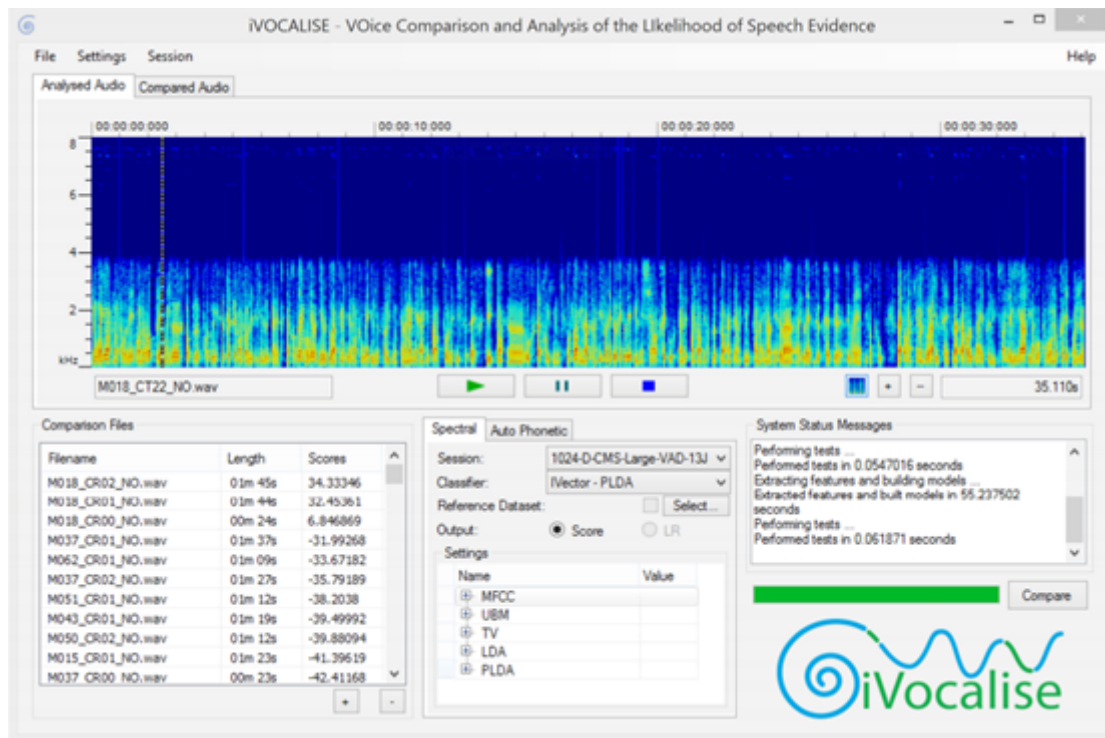


Figura 24 - Interface do iVocalise [91].

Os algoritmos de classificação utilizados pelo *iVocalise* e necessários para o cálculo de distâncias entre modelos de orador são o *Gaussian Mixture Model* (GMM) com ou sem adaptação *Maximum a Posteriori* (MAP), o *i-vector* com *Probabilistic Linear Discriminant Analysis* (PLDA) e o *i-vector* usando a comparação da distância do cosseno, independentemente de existir ou não o método *Linear Discriminant Analysis* (LDA). Na nossa investigação, foi utilizado o *i-vector* PLDA, na medida em que é um método de modelização e classificação mais recente e que produziu melhores resultados em estudos prévios [52].

Para a análise no *iVocalise*, utilizou-se uma das sessões do modo *Spectral* (que a empresa produtora assegura que é a alternativa mais eficaz). Os parâmetros espectrais são extraídos, automaticamente, de uma amostra de fala e estão relacionadas com atributos do sinal no domínio da frequência, num pequeno intervalo de tempo [52]. Esta abordagem de parametrização é bastante popular já que o seu uso é bastante recorrente nas tarefas de reconhecimento de fala e de orador [89].

De acordo com Pinto [52], o *iVocalise* suporta características espectrais flexíveis (MFCCs) já que i) a seleção da banda de frequência é ajustável, ii) os coeficientes de energia, *delta* e *delta-delta* são facultativos e iii) são admitidas técnicas de normalização de características, tais como a *Mean Subtraction* e *Mean Variance Normalization*.

Em particular, o modo *Spectral* suporta três sessões, nomeadamente:

- i) 2016A-Adaptable-1024-D-CMS;
- ii) 2016A-Fixed-1024-D-CMS;
- iii) SREPRISM_27k_2016C_TEL_1024DD-AnilBuild.

Uma vez que os resultados das duas primeiras sessões não diferem muito, optou-se por utilizar a primeira para ser exequível uma comparação de resultados com a terceira sessão SREPRISM_27k_2016C_TEL_1024DD-AnilBuild, em especial, porque possibilita uma melhor adaptação do UBM às amostras da base de dados (ver Secções 2.3 e 2.3.4) [52].

A grande diferença entre a sessão 2016A-Adaptable-1024-D-CMS e a SREPRISM_27k_2016C_TEL_1024DD-AnilBuild está relacionada com o acréscimo do conjunto de dados de treino dos modelos internos de orador, e no uso de coeficientes *delta-delta* por parte da terceira sessão, sendo que a primeira, a 2016A-Adaptable-1024-D-CMS, só utiliza coeficientes *delta* [52].

Outra função importante no *iVocalise* é a *Voice Activity Detection* (VAD), que permite que se eliminem, automaticamente, momentos de silêncio ou de baixa energia no sinal em análise [52] [89]. Este aspeto será retomado na Secção 6.2. De realçar outra funcionalidade, uma ferramenta 3D que possibilita ao perito a observação do

posicionamento relativo dos oradores no espaço *i-vector* reduzido às primeiras três dimensões, o que é útil nos casos em que existem poucos oradores e é necessário verificar a distância entre eles, no espaço [52] [29].

Em suma, o *iVocalise* só avalia, numericamente, a distância entre modelos de orador extraídos de registos de voz, produzindo um resultado que exprime a distância entre o orador A e o orador B. Esta avaliação pode ser executada entre dois oradores, entre um orador e vários oradores e entre vários oradores. Consequentemente, torna-se pertinente estudar as distâncias entre os oradores gémeos, entre eles e o familiar e entre os gémeos e todos os outros oradores voluntários. Este estudo particular foi iniciado no trabalho de dissertação anterior [52] mas não foi aprofundado na presente investigação devido à ampliação dos cenários de teste envolvendo múltiplas condições de canal, tal como se reporta no Capítulo 6.

Na nossa investigação recorreu-se à versão 2.4.0.1532 do *software iVocalise*.

4.5 Software Biometrics

O *Biometrics* é um *software* que processa os resultados dispostos nas matrizes de *scores* (pontuações) ou distâncias, geradas pelo *iVocalise*, de modo a produzir estatísticas e parâmetros globais de qualidade dos dados processados, de distâncias entre oradores, ou de razões de verosimilhança (*Likelihood Ratios – LR*). Para que isto seja possível, é necessário um conjunto de pontuações alcançadas a partir da confrontação de características obtidas de diferentes gravações, mas com origem no mesmo orador, assim como um conjunto de pontuações adquiridas da comparação de características extraídas de gravações com falantes distintos [52]. Segundo Drygailo *et al.* [64], a metodologia utilizada para obter os resultados divide-se em dois momentos. No primeiro, a tarefa consiste em usar os *feature vectors*, os modelos do orador (*i-vector* PLDA, nesta investigação) e calcular os *scores* decorrentes quer das distâncias entre diferentes gravações pertencentes ao mesmo orador, quer das distâncias entre gravações pertencentes a oradores distintos. A segunda parte do processo estrutura os *scores* obtidos em duas distribuições univariadas, a primeira consistindo na função de

densidade de probabilidade das pontuações comparando gravações do mesmo orador, representada por H_0 , e a segunda consistindo na função de densidade de probabilidade das pontuações comparando gravações entre diferentes oradores, representada por H_1 .

O *Likelihood Ratio* é obtido com base na pontuação da evidência, que representamos por E , da comparação entre o modelo de orador resultante de uma gravação de fala de um indivíduo desconhecido (*i.e.* em forte ligação com um crime) e o modelo do orador resultante de uma gravação de fala de um indivíduo suspeito e conhecido. A Figura 25 ilustra o método de obtenção do LR. Neste exemplo, a *evidence score* é $E=58$. Dividindo a probabilidade de E resultar da distribuição H_0 (ver Secção 2.3.4) pela probabilidade de E resultar da distribuição H_1 , perfaz $LR=25$. Assim, verifica-se que é 25 vezes mais provável de observar a *evidence score* (E) dada a H_0 , ou seja, a amostra do orador suspeito e a amostra do orador desconhecido terem a mesma voz, do que observar E dada a H_1 , *i.e.* as amostras conterem registos de voz diferentes, na medida em que o orador não é o mesmo [64].

Genericamente, a análise e visualização de resultados produzidos pelo *software iVocalise*, pode ser conseguida recorrendo aos seis tipos de gráficos que o *Biometrics* utiliza [52]. No âmbito do nosso trabalho, ilustram-se, apenas, os três primeiros que derivaram da obtenção dos resultados especialmente significativos:

- *Likelihood ratio plot (LR plot)* – Figura 26;
- *Equal Error Graph* – Figura 27;
- *Detection Error Trade-off (DET plot)* – Figura 28;
- *Receiver Operating Characteristic plot (ROC plot)*;
- *Zoo plot*;
- *Scatter plot*.

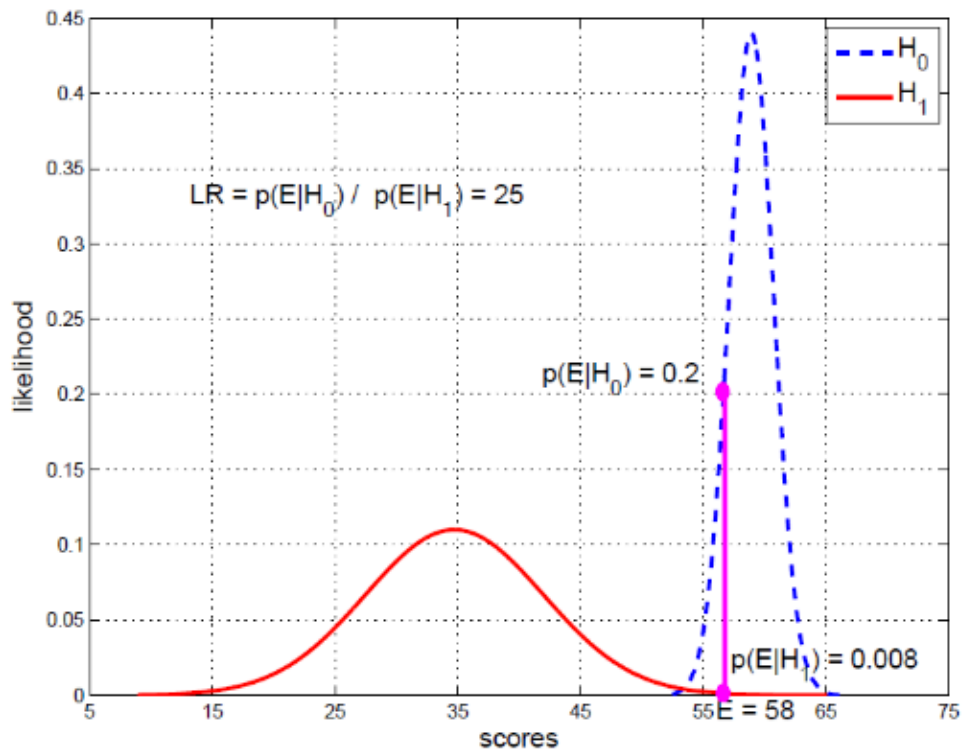


Figura 25 - Representação do método de obtenção do Likelihood Ratio [64].

A Figura 26 ilustra o LR Plot do *software Biometrics* e facilita o cálculo do rácio LR de modo idêntico à situação já ilustrada no gráfico da Figura 25. Estima, portanto, em função da distribuição H_0 , relativa às comparações entre o mesmo orador, e a distribuição H_1 , relativa às comparações entre diferentes oradores, qual o rácio LR para um caso em que a pontuação da evidência é $E=4.44$. No caso ilustrado, esse valor é $LR=55.1$.

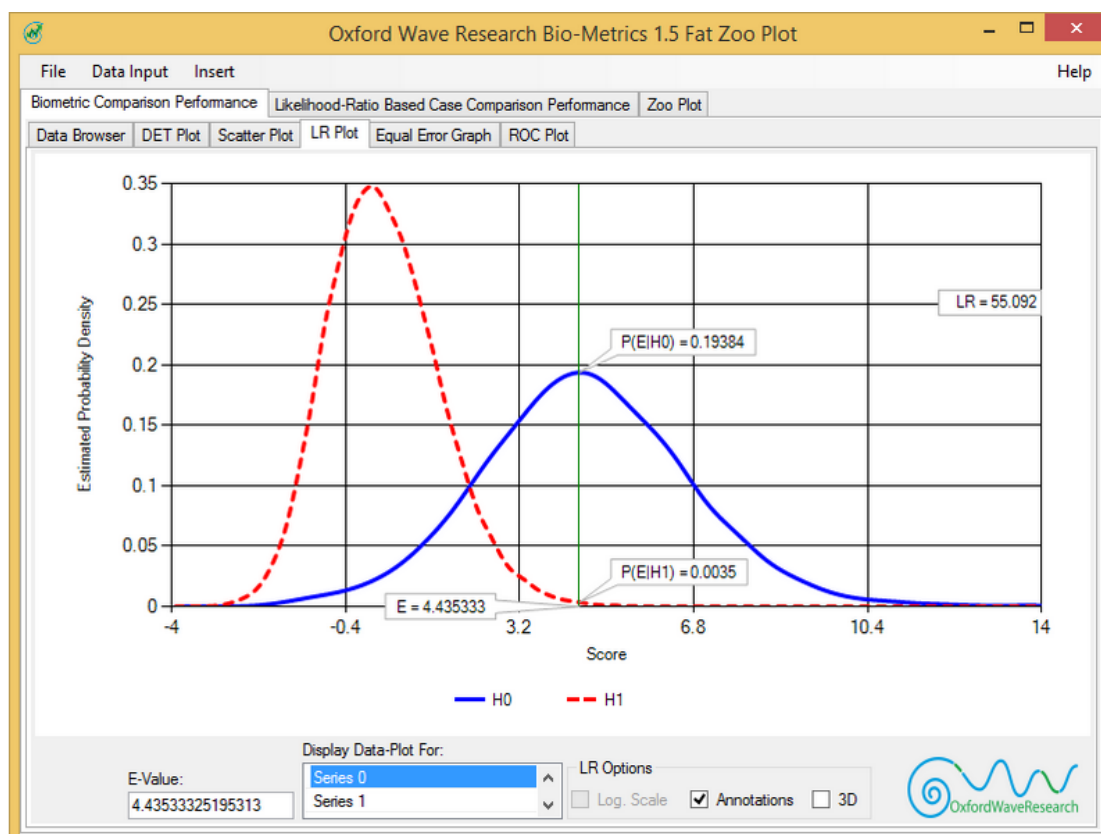


Figura 26 - Resultados do iVocalise no LR Plot [90].

A Figura 27 ilustra o *Equal Error Graph* (EEG) a partir dos conceitos de *False Acceptance Rate* (FAR) e de *False Rejection Rate* (FRR). Estes conceitos decorrem da integração das curvas de distribuição H_0 e H_1 , para um dado limiar que seja definido para a pontuação (*score*) que conduz à afirmação de que um dado valor de evidência E exprime um resultado de mesmo orador – caso E seja superior a esse limiar –, ou que conduz à afirmação de que um valor de E exprime um resultado de orador diferente – caso E seja inferior a esse limiar. O FAR decorre da integração da distribuição H_1 para valores superiores ao limiar e mede a probabilidade acumulada de o sistema errar na conclusão de mesmo-orador. O FRR decorre da integração da distribuição H_0 para valores inferiores ao limiar e mede a probabilidade de o sistema errar na conclusão de orador-diferente. O ponto em que as curvas se cruzam é o EER. Este é um parâmetro global de desempenho ou eficácia de um sistema de reconhecimento de orador. A Figura 27 ilustra as curvas FAR e FRR para três situações de distribuição H_0 e H_1 .

Intuitivamente, é compreensível que quanto mais baixo for o limiar menos exigente o sistema será na correta identificação, pelo que o FAR tenderá a ser alto. Ao invés, quanto mais alto for o limiar mais exigente o sistema será na correta identificação, pelo que o FRR tenderá a ser alto.

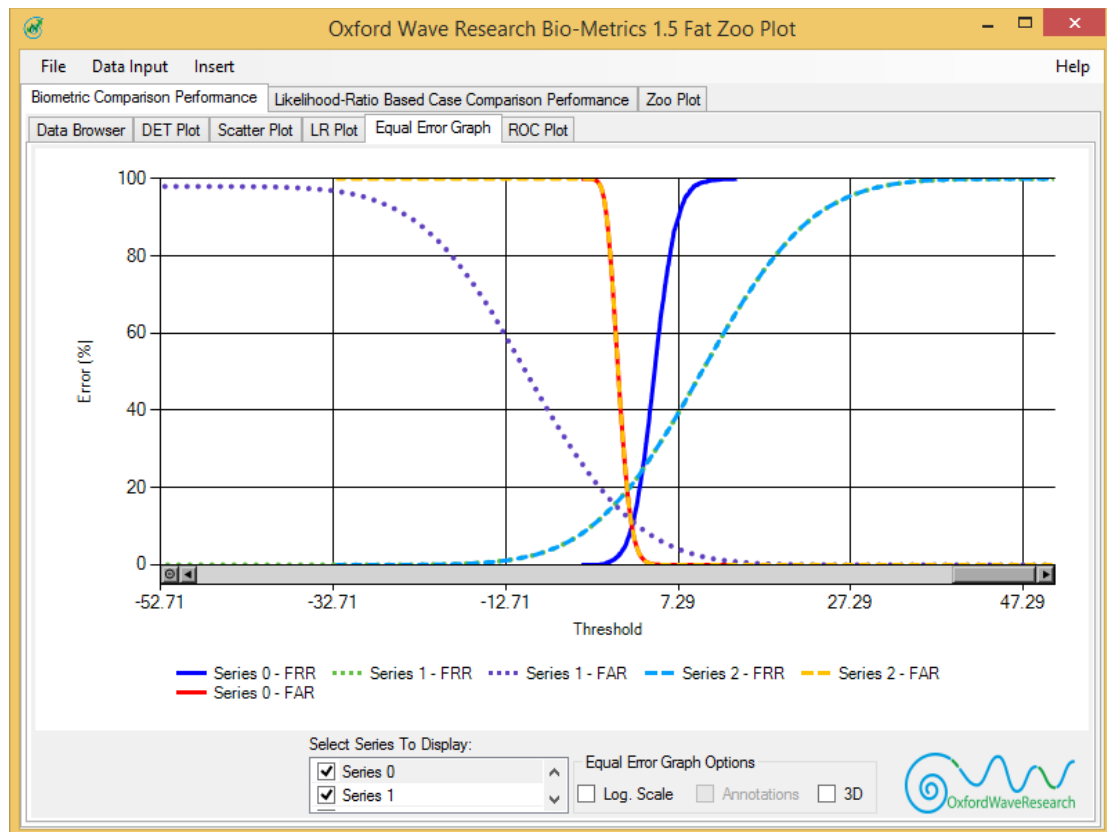


Figura 27 - Resultados do *iVocalise* no EEG [90].

Assim, dado que a um aumento da FAR decorre uma diminuição da FRR, é útil a representação gráfica de uma em função da outra, fazendo abstração do limiar. Esta representação é dada pelo gráfico DET *Plot* como se ilustra na Figura 28.

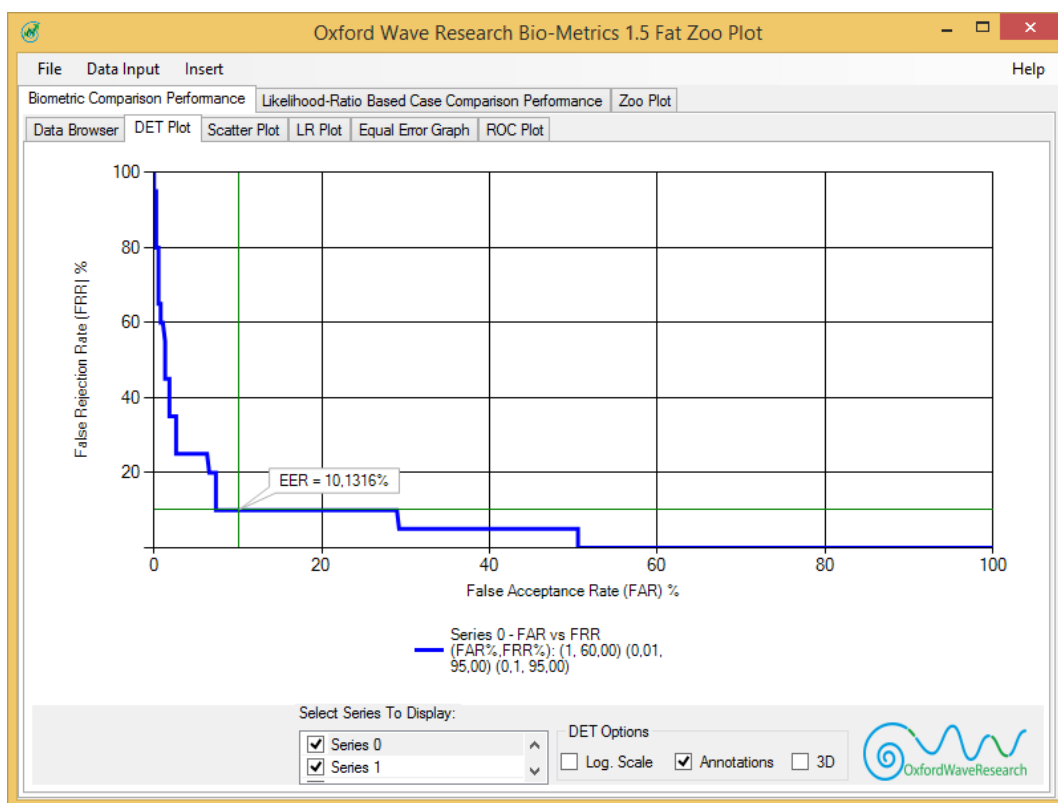


Figura 28 - Resultados do *iVocalise* no DET Plot [90].

O DET *Plot* indica assim qual o desempenho do sistema ao traçar os *false match rate* contra os *false non-match rate* numa gama para os valores do limiar. Tem especial interesse a sinalização neste gráfico do ponto EER, que se verifica quando o *false match rate* (FAR) equivale ao *false non-match rate* (FRR). Normalmente, os eixos do DET *Plot* encontram-se numa escala logarítmica, mas também podem ser encontradas escalas lineares [90]. O sistema de classificação é mais preciso quanto menor for o valor do EER, o que conduz a que os dois tipos de erros enunciados sejam de baixa ocorrência [52] [88]. O EER é assim um parâmetro holístico de desempenho que é muito útil para comparar diferentes sistemas de classificação de orador.

A versão do *software Biometrics* utilizada foi a 1.5.0.637.

4.6 Software *VoiceStudio*

O *VoiceStudio* consiste num programa de análise e processamento de sinais, direcionado para análise e diagnóstico de sinais de voz, seja com finalidades clínicas, forenses ou privadas. Apresenta-se em português e é de acesso comercial. Foi criado pela *spin-off* universitária *SEEGNAL Research*. A Figura 29 ilustra a interface deste *software*.

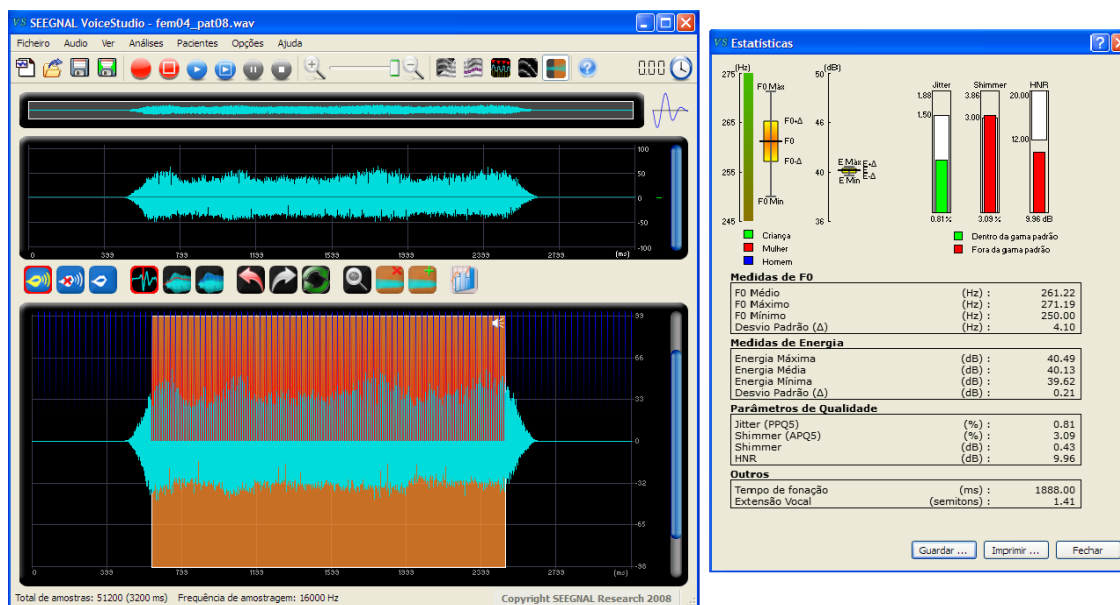


Figura 29 - Interface do VoiceStudio (análise acústica de uma vogal de uma oradora feminina).

Para além das funções habituais de um editor de áudio, incluindo a gravação, arquivo e reprodução, este *software* permite também a análise interativa de espectrograma de banda larga e de banda estreita, a avaliação da frequência fundamental sobreposta à análise do espectrograma, assim como a análise interativa combinada de sinal temporal, de espectro e de espectrograma, de cepstrograma e do vozeamento do sinal. A maior parte das funções de análise é feita em tempo real (*on-the-fly*), durante a reprodução e a gravação.

O *VoiceStudio* permite também a decomposição automática e reprodução seletiva de regiões vozeadas da fala nas suas componentes de harmónicos (ondas sinusoidais) e ruído, assim como a obtenção dos principais parâmetros objetivos de qualidade de voz, como a frequência fundamental, o *jitter*, o *shimmer* e o HNR.

No nosso trabalho, a versão do *VoiceStudio* utilizada foi a VS2009.01.

4.7 Análise estatística

Toda a análise estatística desenvolvida no âmbito do nosso trabalho foi realizada usando funções nativas da plataforma *Microsoft Office Excel* (versão 15.30), um editor de cálculo, assim como funções de extensão disponibilizadas através do pacote livre *Real Statistics*¹⁵. Referimos aqui alguns conceitos base que se prendem com funções estatísticas mais relevantes que foram utilizadas no decurso da nossa investigação, especialmente em relação a resultados que se apresentam nos Capítulos 5 e 6.

Considerando um conjunto de resultados relativamente a uma experiência A (*e.g.* vários valores de HNR decorrentes de uma sequência definida de vogais sustentadas produzidas por um orador, num dado dia) e um outro conjunto de resultados relativamente a uma experiência B (*e.g.* vários valores de HNR decorrentes da mesma sequência de vogais sustentadas produzidas pelo mesmo orador uma semana depois), afirma-se que há significância estatística na comparação de atributos, *i.e.* parâmetros estatísticos, daqueles dois conjuntos de resultados, quando há fundamento para rejeitar a designada hipótese nula. A hipótese nula, habitualmente representada por H_0 (não confundir com H_0 na Secção 4.5 que tem um outro significado) exprime um teste, ou condição, que se pretende avaliar, sendo a sua oposta designada hipótese alternativa, ou H_1 (não confundir com H_1 na Secção 4.5 que tem um outro significado). Diz-se que p designa a probabilidade de a hipótese nula ser verdadeira devido a razões fortuitas ou por mero acaso. Habitualmente, tal como também assumimos nesta dissertação, se $p < 0.05$, existe fundamento estatístico para se rejeitar a hipótese nula e, em consequência, aceitar a hipótese alternativa. Com base neste conceito, têm particular importância na nossa investigação os testes listados a seguir.

- O *t-test* é utilizado para testar a hipótese nula de que as médias de duas populações são iguais, ou seja, que a sua diferença é zero. A hipótese H_1 afirma,

¹⁵ Disponível em <http://www.real-statistics.com>. Último acesso em janeiro de 2018.

assim, que as médias são diferentes. Usando o exemplo acima, aplicando o *t-test* às médias dos resultados das duas experiências, A e B, caso $p < 0.05$, pode afirmar-se que as duas médias diferem de forma estatisticamente significativa, isto é, aceita-se H_1 , a hipótese alternativa.

- O *F-test* é utilizado para testar a hipótese nula de que as variâncias de duas populações são idênticas, isto é, que a sua relação é unitária. A hipótese alternativa afirma que as variâncias não são idênticas. Usando o exemplo acima, aplicando o *F-test* às variâncias dos resultados das duas experiências, A e B, caso $p < 0.05$, pode afirmar-se que as duas variâncias diferem de forma estatisticamente significativa, isto é, aceita-se H_1 , a hipótese alternativa.
- A correlação normalizada, também denominada usualmente por correlação de Pearson, é um parâmetro que exprime a relação linear entre duas variáveis aleatórias, ou entre duas sequências de dados. Varia entre -1, caso que significa que há dependência total na variação oposta – um aumento numa provoca uma correspondente diminuição na outra –, e entre +1, significando dependência total no mesmo sentido – um aumento numa implica um correspondente aumento na outra. A hipótese nula, no cálculo da correlação normalizada, é que o seu valor é zero. Assim, se o valor de p associado ao cálculo da correlação cruzada for $p < 0.05$, isso significa que só nesse caso é que o cálculo é estatisticamente significativo. Usando o exemplo acima, se o resultado da correlação normalizada entre as duas sequências de resultados for, por exemplo, 0.4 e o valor $p = 0.1$, conclui-se que não é possível rejeitar a hipótese nula, pelo que o valor da correlação não é estatisticamente significativo.
- Dada uma sequência de valores, decorrentes da experiência A ou B acima, define-se intervalo de confiança para a média desses valores, como o intervalo entre um valor máximo e um valor mínimo correspondente à hipótese alternativa (H_1), sendo a hipótese nula a média situar-se fora daquele intervalo. Se o limiar de significância for $p = 0.05$, a rejeição da hipótese nula indica que aquele intervalo contém a média com uma confiança de $1 - p = 0.95 = 95\%$.

4.8 Conclusão

Este capítulo focou-se nas aplicações de *software* utilizadas durante a investigação para recolher, editar, criar novas bases de dados e as analisar, estatisticamente. Antecipa-se, portanto, que o conjunto de *software* que se utilizou se revelou de uma ajuda imprescindível para se obterem os resultados apresentados e discutidos nos Capítulos 5 e 6. Afirmar-se, da mesma forma, que os peritos de voz e os foneticistas forenses que tenham como função executar tarefas de reconhecimento de orador têm à sua disposição ferramentas complexas que, para além de apoiarem em estudos de perturbações de voz, produzem também resultados muito úteis numa perspetiva prática e até formal, porque são eminentemente objetivos, documentáveis e comunicáveis.

A maior parte das aplicações de *software* que auxiliaram na nossa investigação é de utilização intuitiva e clara, o que contribuiu para a obtenção dos resultados. A diversidade de funções de medição permitiu que os resultados dos capítulos 5 e 6 beneficiassem de um conjunto rico de características e que abrangessem múltiplos objetivos, como a extração de características fonéticas e espectrais, ou mesmo a análise estatística, depois da tarefa de análise e de processamento de sinais.

Capítulo 5 – Estudos com base em parâmetros acústicos

5.1 Introdução

O presente Capítulo foca-se no estudo das vogais sustentadas e dos parâmetros acústicos seleccionados para a análise de voz, concretamente, a frequência fundamental, o *jitter*, o *shimmer*, o parâmetro HNR e valores das frequências formantes. O estudo apoia-se sobretudo em estimativas, para aqueles parâmetros, de média e desvio padrão, intervalos de confiança da média e coeficientes de correlação. Algumas destas estimativas caracterizam e comparam resultados obtidos através do *software Praat* e do *software VoiceStudio*.

5.2 Vogais e parâmetros acústicos seleccionados

Pediu-se aos oradores gémeos e aos familiares que, no âmbito dos diálogos telefónicos conduzidos, produzissem as vogais sustentadas tónicas /a/, /e/, /i/, /o/ e /u/ para posterior análise (ver Secções 3.1 e 3.6.2). Após serem isoladas manualmente, através do *software Adobe Audition*, e de serem segmentadas manual e individualmente, triaram-se os oradores que produziram vogais que foram consideradas aptas para o estudo (ver Secção 3.6.2).

Os parâmetros acústicos seleccionados, designadamente, a média e o desvio padrão da frequência fundamental, o *jitter*, o *shimmer*, o HNR, assim como os valores das frequências formantes, são tipicamente utilizados para análise de voz normal e, também, para diagnóstico de voz patológica. Contudo, dado que veiculam também informação sobre características distintivas do orador, encerram um potencial interessante para a identificação de orador, tal como se referiu no Capítulo 2, razão pela qual são referidos e estudados nesta dissertação, como será detalhado nas secções seguintes.

5.2.1 Vogais sustentadas

As vogais sustentadas produzidas com voz saudável são objeto de análise na acústica e comparação de vozes, na medida em que contêm um espectro ruidoso muito reduzido e restrito [18]. Assim, existem vários argumentos que sustentam a conveniência da medição de vogais sustentadas, nomeadamente, o facto de que pouco variam temporalmente e de não terem influências coarticulatórias, ou seja, influências de contexto fonético em que possam estar inseridas. Além disso, não são afetadas pelos elementos de prosódia, como a entoação, a acentuação ou a cadência silábica. São também fáceis de produzir e de controlar, durante a fonação, correspondendo, por isso, a um regime fonatório controlado. Por fim, não contêm, normalmente, marcas dialetais do orador que estão presentes no discurso contínuo. Outra vantagem relacionada com o estudo das vogais sustentadas é o facto de que a ausência relativa de influências prosódicas e articulatórias pode auxiliar na avaliação percetiva, na medida em que o avaliador se pode focar em parâmetros relacionados com a qualidade da fonte de voz [92]. De facto, de acordo com Ferreira [18], os fones produzidos sustentadamente, pelo mesmo orador, implicam, apenas, modificação dos filtros e dos ressoadores do trato vocal, uma vez que a fonte – decorrente sobretudo da ação das pregas vocais – não se altera significativamente. Sousa [29] afirma que, na voz sustentada, é gerado um som complexo periódico que é, caracteristicamente, composto por mais do que uma componente sinusoidal. Existe, também, uma correspondência com o modo vibratório das pregas vocais que, no essencial, se caracteriza por uma grande estabilidade.

As vogais são produzidas com ausência de obstrução no trato vocal, e as vogais sustentadas apresentam, em particular, quasi-periodicidade e são geradas por impulsos sucessivos também quasi-periódicos que resultam das irregularidades intrínsecas à vibração das pregas vocais [29].

Na Figura 30, podemos observar a representação da vogal sustentada /a/ de um indivíduo do género feminino, através de uma onda osciloscópica e de um espectrograma, obtidos no *software Praat*.

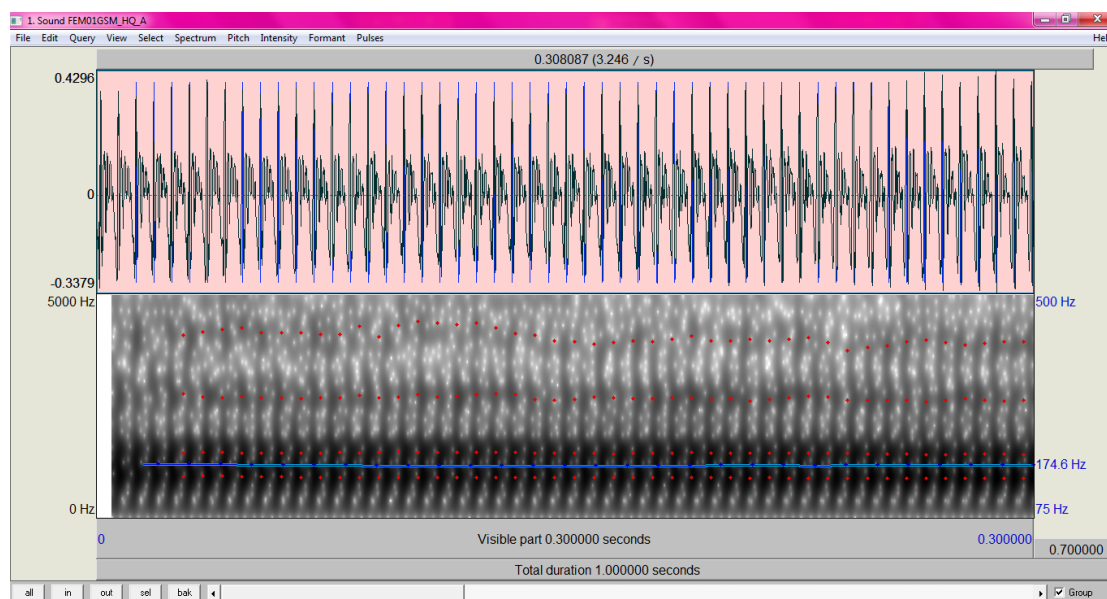


Figura 30 – Representação na onda osciloscópica e no espectrograma da vogal sustentada /a/ de uma oradora, dos 0 aos 0.30 segundos.

Através desta representação acústica, verifica-se que as regiões das frequências formantes são estáveis ao longo do tempo, não oscilando. Este fone é o mais utilizado na avaliação perceptiva [29] e é a vogal que demonstra mais facilmente a presença de disfonia [93].

Por todas estas razões, as vogais sustentadas revelam a robustez da voz e são, inclusive, alvo de análise acústica recorrente em Terapia da Fala e no ensino do canto [29]. O *jitter* e o *shimmer*, parâmetros de perturbação de avaliação acústica que foram introduzidos no Capítulo 2, são medidas utilizadas para se estudar perturbações que afetam as vogais sustentadas. Em especial, permitem avaliar os desvios que possam existir em relação aos valores de uma voz normal quando a amostra de análise corresponde a uma voz patológica.

Para o presente estudo, utilizaram-se as vogais sustentadas dos oradores gémeos e dos familiares que fazem parte da investigação, em versão de alta qualidade e de qualidade telefónica. As amostras foram segmentadas manualmente no *software Adobe Audition* para terem todas um segundo de duração. Apenas foram consideradas as vogais sustentadas que apresentavam regiões de formantes estáveis e claras, sem

quebras de voz ou frequência fundamental que oscilasse; caso contrário, seria provável que as regiões instáveis provocassem mau funcionamento dos algoritmos de estimação e, conseqüentemente, valores de medição errados. Neste sentido, foram excluídas, especificamente, as vogais de sete oradores femininos e de quatro oradores masculinos – a base de dados inicial era de vinte oradores femininos e dezassete oradores masculinos. Concluiu-se, do mesmo modo, que as estimativas de erro da frequência fundamental tinham bastante influência na avaliação de outros parâmetros acústicos. Conseqüentemente, sempre que se verificava esta eventualidade, as correlações normalizadas entre os pares de alta qualidade e os de qualidade telefônica eram severamente afetadas. Assim, averiguaram-se os erros de estimativa da frequência fundamental que estavam relacionados com o *Praat* e o *VoiceStudio* e excluíram-se mais três oradores femininos e quatro oradores masculinos. No total, permaneceram dezanove oradores, concretamente, dez femininos e nove masculinos, como ilustra a Tabela 3.

Tabela 3 - Idade e graus de parentesco dos oradores cujas vogais sustentadas foram analisadas acusticamente.

Orador	Idade	Grau de parentesco
FEM01	22	Gémea
FEM04/FEM05	27/27	Gémea-gémea
FEM10/FEM11/FEM12	21/55/21	Gémea-mãe-gémea
FEM13	22	Gémea
FEM16/FEM17	49/15	Gémea-filha
FEM19	36	Gémea
MAL01/MAL02	71/71	Gémeo-primo
MAL05	21	Gémeo
MAL06/MAL07/MAL08	77/52/52	Pai-gémeo-gémeo
MAL09	24	Gémeo
MAL16/MAL17	19/19	Gémeo-gémeo

Uma vez que para as diferentes modalidades de comunicação (alta-qualidade e telefónica) as vogais foram recolhidas utilizando diferentes equipamentos, a segmentação das amostras com qualidade telefónica e a segmentação das amostras de alta qualidade foram realizadas simultaneamente com o objetivo de facilitar o alinhamento temporal.

5.2.2 Parâmetros acústicos relevantes na comparação de vozes

Como foi expresso ao longo da dissertação, e segundo Fernandes e Ferreira [100], os parâmetros acústicos são utilizados para análise de voz normal e diagnósticos de voz patológica, por especialistas da fala, e para a comparação de amostras de voz. São, também, usados em reabilitação. Acresce que também transportam informação idiossincrática da voz, o que estimula a sua análise em contexto forense.

Tanto quanto é do nosso conhecimento, até ao momento, nenhum estudo se focou no impacto da degradação dos valores dos parâmetros acústicos após as comunicações de canal de GSM e VoIP, em relação aos obtidos nas gravações de voz de alta qualidade [100]. Por estas mesmas razões, considerou-se pertinente a escolha da análise i) da média e do desvio padrão da frequência fundamental, ii) da variação de curto termo do período glotal, através do *jitter*, iii) da variação de curto termo da amplitude glotal através do *shimmer*, e iv) do rácio entre a potência média do conteúdo harmónico de um sinal vozeado e a potência média do seu conteúdo de ruído (*Harmonics-to-Noise Ratio* – HNR). Incluímos, igualmente, a análise acústica dos valores das formantes, o que será tratado na Secção 5.3.6.

As medidas que analisam, acusticamente, o discurso contínuo – ou fala corrida – não são, habitualmente, as medidas de perturbação de qualidade referidas no parágrafo anterior, uma vez que este género de amostra implica aspetos de co-articulação, acentuação, prosódia e gestos articulatórios relacionados com os hábitos articulatórios do indivíduo [92]. Contudo, de acordo com Segundo *et al.* [77], é comum existirem na fala corrida hesitações de discurso como os arrastamentos silábicos e interjeições de ligação que se assemelham à produção de vogais sustentadas. Estas ocorrências são

relativamente comuns nas interceções telefônicas, o que as torna interessantes pois incluem palavras ou monossílabos cujas vogais têm uma duração útil além do normal, o que significa que incluem vários pulsos glotais num regime que se pode considerar quase estacionário, comparável, portanto, com o das vogais sustentadas [100].

Em resultado da influência do trato vocal, as formantes traduzem concentração de energia acústica à volta de estreitas e específicas faixas de frequências da onda sonora produzida e são estimadas através de um procedimento computacional – ou algoritmo. As primeiras formantes, F1 e F2, são determinantes na definição da identidade e qualidade das vogais e estão sobretudo relacionadas com a abertura e o fechamento do maxilar inferior, a anterioridade ou posterioridade da posição da língua em relação ao palato e a configuração labial [97] [98]. Por seu turno, as formantes F3 e F4 estão especialmente ligadas às qualidades espectrais da voz cantada, pois revelam uma estrutura bastante mais vincada e robusta nessa atividade no que na produção de discurso falado [97] [98]. Uma vez que na maior parte das tarefas de comparação de voz apenas a conversação telefônica está disponível, considerou-se desafiante avaliar como é que as frequências formantes são modificadas devido à influência dos canais de comunicação VoIP e GSM, tendo-se por referência as gravações de alta qualidade.

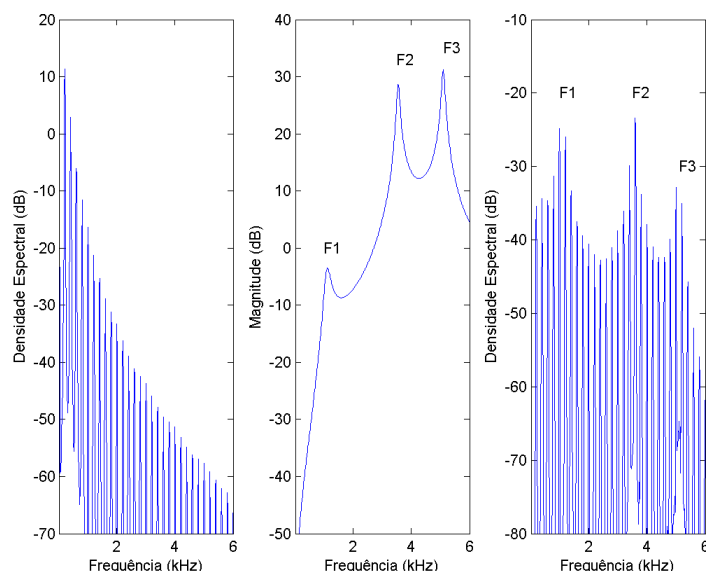


Figura 31 - Ilustração da influência das frequências formantes sobre o espectro do sinal de fonte laríngea [18].

A Figura 31 ilustra a influência das formantes sobre o espectro do sinal de fonte laríngea, cuja densidade espectral decresce à razão de 12 dB por oitava. O espectro à esquerda nesta figura representa o sinal de fonte, gerado pelas pregas vocais. A imagem ao centro representa um hipotético filtro que é aplicado em consequência da forma e do volume dos ressoadores e da configuração dos órgãos articuladores. O espectro à direita evidencia o resultado da influência do filtro sobre o espectro do sinal da fonte laríngea. Destaca-se, em especial, que as frequências formantes que se enquadram dentro dos limites da banda passante do filtro e identificadas na Figura 31 através dos picos F1, F2 e F3, moldam o espectro do sinal de fonte, dando-lhe identidade linguística, mas também uma assinatura idiossincrática.

5.3 Análise através de parâmetros acústicos

Nas Secções subsequentes, apresentam-se os resultados da análise com recurso a parâmetros acústicos como a média e desvio padrão da frequência fundamental, o *jitter*, o *shimmer* e o HNR. Serão abordados os resultados obtidos no ambiente *Praat* e os resultados deste *software* em comparação com os resultados obtidos no ambiente *VoiceStudio*. Optou-se por apresentar, primeiramente, os resultados decorrentes do *Praat* e, de seguida, os resultados comparativos de ambos os *software*. O objetivo é aferir até que ponto os valores podem diferir tendo em consideração as diferentes ferramentas de análise dos parâmetros acústicos.

5.3.1 Análise com base na frequência fundamental

A Figura 32 representa a frequência fundamental média, obtida no ambiente *Praat*, das gravações das cinco vogais sustentadas para cada um dos dezanove indivíduos. Verifica-se que os valores médios da frequência fundamental dos oradores do sexo feminino são uma oitava mais elevados do que os dos masculinos, como expectável. Consequentemente, no que diz respeito à análise estatística, investigar as alterações observadas nas versões telefónicas (TL) de todo o conjunto das vogais,

relativamente às de alta qualidade (HQ) dos mesmos sinais, não pareceu adequado, na medida em que era pretendida a média. Por conseguinte, separaram-se os sujeitos femininos dos masculinos no que concerne a média da frequência fundamental.

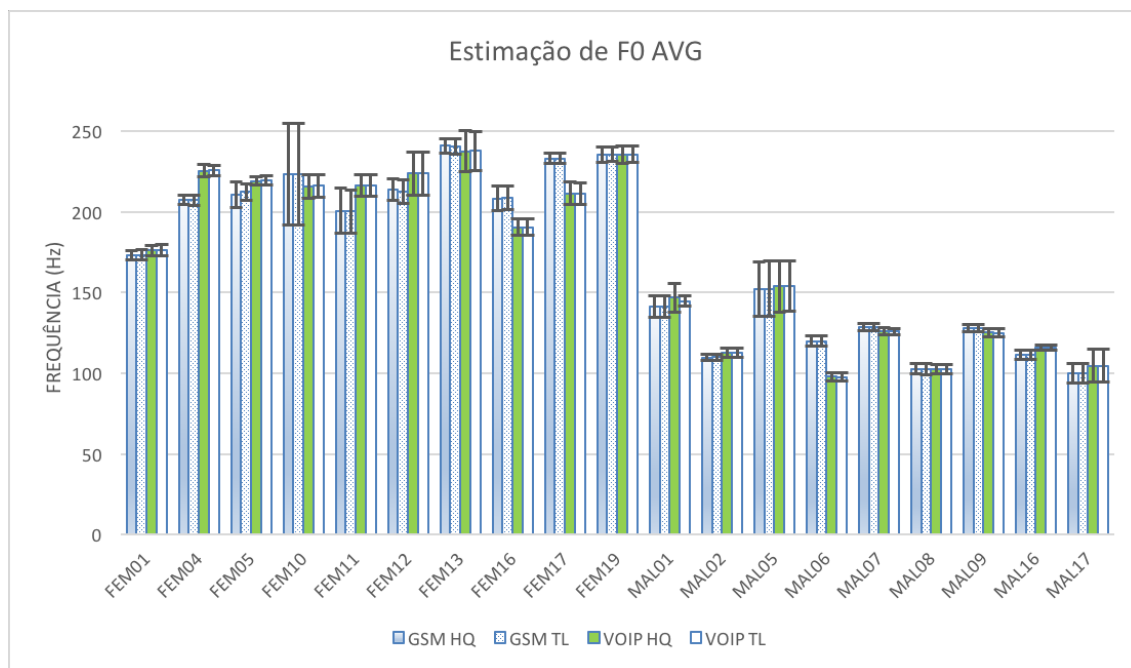


Figura 32 - Resultados da média estimada de F0, incluindo os intervalos de confiança de 95%.

Na Figura 32, observam-se 10 oradores femininos (FEM) e 9 do género masculino (MAL). Constata-se que a média da frequência fundamental das mulheres é mais elevada do que a dos homens, como é definido pela literatura [92], o que justifica, como referido atrás, a opção de se constituírem dois grupos diferentes: homens e mulheres. Além disso, na maioria dos casos, o Intervalo de Confiança, que representa o intervalo estimado em que a média de um parâmetro de uma amostra tem maior probabilidade de ocorrer, é pequeno, o que revela que o parâmetro acústico é bastante estável.

É possível, ainda, apreciar o comportamento da média da frequência fundamental de um só indivíduo em diferentes circunstâncias, contemporâneas e não-

contemporâneas, na medida em que existem dados para os canais GSM (TL e HQ) e VoIP (TL e HQ).

Considerou-se importante verificar se as versões TL das gravações de voz contemplam ou não diferenças estatísticas significativas relativas às versões em HQ no parâmetro acústico que é a média da frequência fundamental. Com esse propósito, aplicou-se o teste *t* de Student (para amostras emparelhadas em duas caudas, *i.e.* bilateral, e nível de significância de 0.05), avaliando a hipótese nula de médias idênticas para as distribuições de cada par HQ-TL em relação ao parâmetro da média da F0 e diferenciando os géneros masculino (9 indivíduos) e feminino (10 indivíduos). Os resultados não exibiram diferenças estatisticamente significativas para todas as possibilidades de modalidades de comunicação ($n=10$, $p>0.18$ para os sujeitos FEM e $n=9$, $p>0.26$ para os MAL).

Da mesma forma, recorreu-se a um teste F com o objetivo de estudar as variâncias das distribuições, sendo que, neste caso, avaliou-se a proporção das variações obtidas em cada par HQ-TL. Tal como no teste *t*, os resultados não revelaram dissemelhanças significativas nos desvios ($n=10$, $p>0.49$ em FEM e $n=9$, $p>0.48$ em MAL).

Tendo em conta o conjunto completo dos 19 oradores, foram calculadas as correlações normalizadas (*r*) para os valores da média da frequência fundamental entre os pares HQ-TL, para cada modalidade de comunicação, sendo que os resultados obtidos no *Praat* são apresentados na Tabela 4.

Tabela 4 - Correlações normalizadas da média da F0 ($n=19$, todos $p<<0.05$).

<i>r</i>	GSM HQ	GSM TL	VOIP HQ	VOIP TL
GSM HQ	1.0			
GSM TL	0.99995	1.0		
VOIP HQ	0.97712	0.97693	1.0	
VOIP TL	0.97730	0.97712	0.99996	1.0

Estes resultados, além de estatisticamente significativos, são também consistentes, na medida em que indicam, como esperado, que a máxima correlação é obtida para as gravações dos pares contemporâneos HQ-TL, uma vez que correspondem a versões diferentes do mesmo sinal de voz. Correlações elevadas também se verificaram nas gravações não contemporâneas, uma vez que traduzem discurso proferido pelo mesmo orador em sinais acústicos distintos.

Partindo destes resultados, conclui-se que a média da frequência fundamental é um parâmetro acústico fulcral e robusto a ser estudado em contexto áudio-forense, na medida em que não é significativamente afetado em condições normais de conversação, mesmo considerando diferentes canais de comunicação, como GSM ou VoIP.

Na Figura 32, os resultados também evidenciam que para certos pares de oradores gémeos, os intervalos de 95% de confiança têm um alinhamento assinalável, como, por exemplo, nos casos FEM04-FEM05 e MAL16-MAL17, o que era expectável. Contudo, para outros casos de oradores gémeos, como por exemplo FEM10-FEM12 e MAL07-MAL08, este franco alinhamento de intervalos não se verifica, o que poderia indicar que estão envolvidos oradores diferentes. Podemos assim concluir que apesar do parâmetro da frequência fundamental ser, potencialmente, mais interessante do ponto de vista idiossincrático, ou seja, relacionado com o comportamento e o temperamento próprios de um indivíduo, o seu potencial na discriminação de vozes depende, em grande medida, da quantidade de sinal envolvido na estimação deste parâmetro. Uma vez que os nossos dados se focam em vogais sustentadas com a duração de cerca de um segundo, admitimos que as médias da frequência fundamental, apesar da sua importância idiossincrática, também refletem condições situacionais e não constituem, por isso e por si só, um fator discriminatório.

Também os resultados obtidos no *software VoiceStudio* revelam que existe uma forte correlação, para o mesmo orador, entre as amostras e os pares GSM HQ – GSM TL, e VoIP HQ – VoIP TL, relativamente à média da frequência fundamental. Nos casos em que as gravações não são contemporâneas, este factor é indicativo de que os sinais envolvidos nas correlações foram produzidos pelo mesmo orador.

Adicionalmente e considerando que amostras há que foram produzidas em conversas diferentes com uma pausa de cerca de vinte minutos entre elas, é de reforçar o facto de que, de um modo geral, a média da frequência fundamental é um parâmetro acústico bastante robusto na área da investigação acústica forense. A título de exemplo, a Figura 33 mostra a média da frequência fundamental para todos os 19 oradores, no que diz respeito à vogal /u/ e diferentes canais de comunicação.

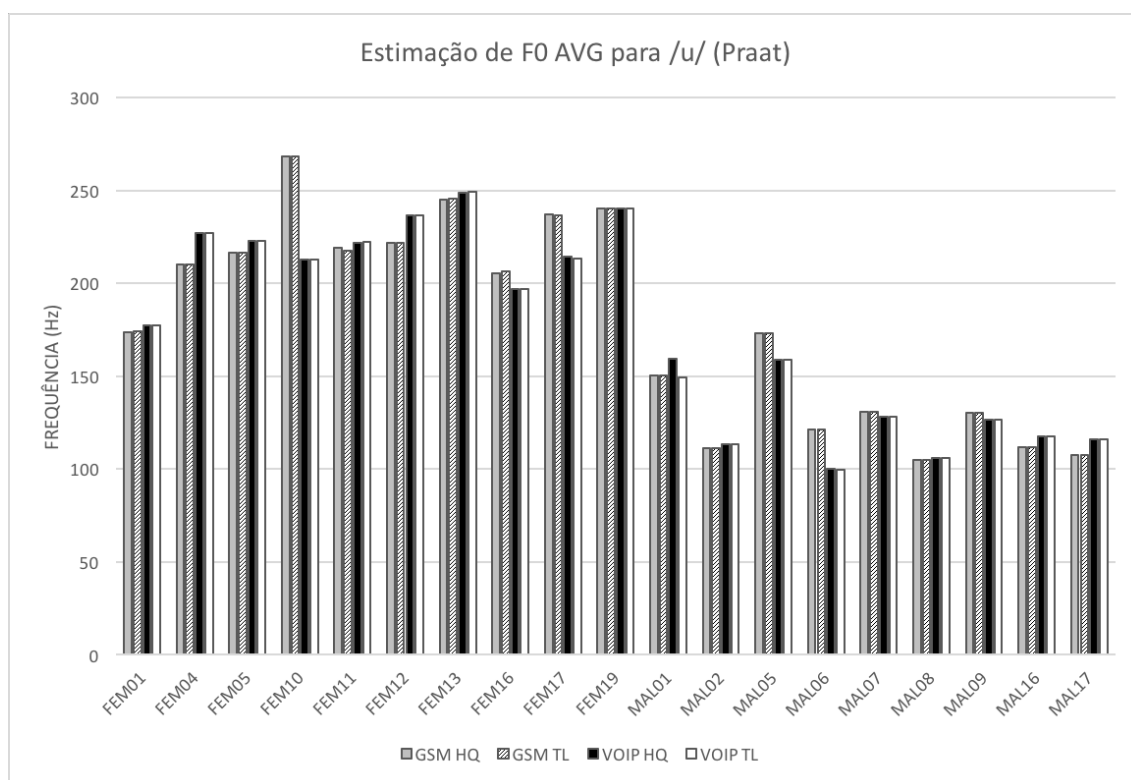


Figura 33 - Média da F0 da vogal /u/ dos 19 oradores, no Praat.

O exemplo FEM10 sugere, contudo, que se deve usar alguma prudência quando se usam registos de muito curta duração, na ordem de poucos segundos. Os resultados obtidos no ambiente *VoiceStudio* são similares.

5.3.2 Análise com base no desvio padrão da frequência fundamental

Ao analisar o desvio padrão da frequência fundamental no *Praat*, procurou-se averiguar se o tipo de canal de comunicação, GSM ou VoIP, influenciaria, sistematicamente, os valores deste parâmetro nas gravações de qualidade telefônica em relação aos das de alta qualidade, além de se investigar a aptidão deste parâmetro para a discriminação de oradores.

A Figura 34 representa o desvio padrão da frequência fundamental da média das cinco vogais de cada sujeito e os intervalos de confiança de 95%. Conclui-se que os valores do desvio padrão da F0, em ligação com os correspondentes intervalos de confiança, revelam que existem poucas oportunidades para a discriminação de orador, na medida em que o alcance médio para a maioria dos oradores é entre 1 Hz e 2 Hz.

Observa-se, também, que, em alguns casos, o intervalo de confiança é comensurável ou até mesmo excede o valor médio do desvio padrão da frequência fundamental, o que sugere que pode existir, aqui, interferência do canal de comunicação ou, inclusivamente, do registo emocional do orador.

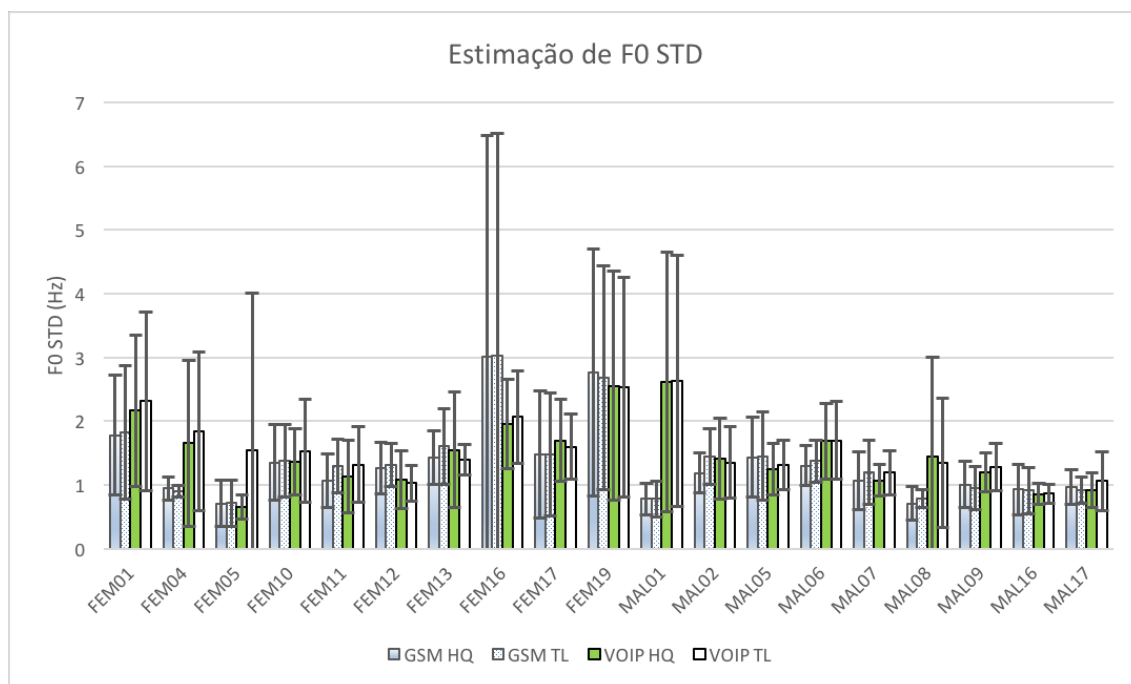


Figura 34 - Resultados estimados do desvio padrão da F0, incluindo os intervalos de confiança de 95%.

Um teste t de Student que avaliou a hipótese nula de desvios padrões idênticos para cada par das distribuições HQ-TL indicou que as diferenças não eram, novamente, estatisticamente significativas em relação a todas as vogais e modos de comunicação ($p>0.08$), com a exceção da vogal /u/ ($p=0.016$).

Nas gravações não contemporâneas, que são especialmente relevantes para o contexto forense, o desvio padrão da frequência fundamental pode variar bastante. Na Tabela 5, encontramos as correlações que podem justificar esta conclusão.

Contrariamente à correlação verificada para o par HQ-TL, a correlação das amostras não contemporâneas é pouco significativa, pois os valores- p são pouco inferiores a 0.05.

Tabela 5 - Correlações normalizadas do desvio Padrão da F0 ($n=19$).

<i>r</i>	GSM HQ	GSM TL	VOIP HQ	VOIP TL
GSM HQ	1.0			
GSM TL	0.98852 ($p<0.05$)	1.0		
VOIP HQ	0.54663 ($p<0.016$)	0.52663 ($p<0.021$)	1.0	
VOIP TL	0.49663 ($p<0.031$)	0.46152 ($p<0.047$)	0.91365 ($p<0.05$)	1.0

Na Figura 34, observam-se discrepâncias nos valores obtidos nas modalidades GSM e VoIP, como, por exemplo, em FEM04, FEM16, MAL01 e MAL08. Apesar de, na maior parte dos casos, os resultados HQ-TL serem consistentes, em regra, o desvio padrão da frequência fundamental não é peculiar, refletindo, deste modo, situações pontuais, como o stress ou a ansiedade do orador, o que conduz ao facto de o seu interesse para a fonética forense ser limitado e não constituir uma característica confiável de determinado orador.

Ao contrário do referido para os resultados do desvio padrão da frequência fundamental obtidos através do *Praat*, em que se verificava uma correlação sustentada nos pares HQ-TL, os valores resultantes da análise do *VoiceStudio* não são estatisticamente significativos. De facto, em alguns casos de qualidade telefónica, o valor estimado foi mais elevado do que 30 ou mesmo 50 Hz, o que pode ser explicado por uma elevada sensibilidade do algoritmo às alterações do sinal devido à influência do canal de comunicação. A título de exemplo, no que diz respeito à vogal /a/, os intervalos de confiança de 95% que caracterizam essas diferenças revelam que, enquanto que para o *Praat*, essas dissemelhanças se aproximavam do zero (média de 0.044Hz para o par GSM TL-HQ e 0.085 Hz para o VoIP TL-HQ), o valor médio e a variância, no *VoiceStudio*, eram bastante díspares (3.037Hz para o par GSM TL-HQ e 2.838HZ para o VoIP TL-HQ).

Estes valores atípicos – *outliers* – parecem depender do orador e da vogal, o que não pode ser encarado como significativo. A explicação mais plausível para o facto de o *Praat* e o *VoiceStudio* não apresentarem resultados aproximados em relação a este parâmetro acústico pode justificar-se pelo facto de o *VoiceStudio* não proporcionar estimativas confiáveis quando estão envolvidas gravações de voz em qualidade telefónica.

5.3.3 Análise com base no parâmetro *jitter*

Pela sua importância idiossincrática e por ser o resultado de uma variação microscópica e involuntária da atividade periódica glotal, consideramos que este parâmetro é relevante para o estudo. Relembramos que o *jitter* PPQ5 de uma voz sem patologias é menos do que 1.5%, tipicamente.

Depois da análise no *Praat*, um teste *t* de Student avaliou a hipótese nula de médias idênticas para cada par de distribuição alta qualidade-qualidade telefónica e demonstrou que, no caso do canal GSM, as diferenças não são estatisticamente significativas em todas as vogais ($p \geq 0.05$), à exceção de /u/, na medida em que $p = 0.003$. No que diz respeito à comunicação através de VoIP, todas as vogais exibem diferenças estatisticamente significativas ($p < 0.004$), excetuando /e/ ($p = 0.064$).

A Figura 35 sintetiza a média do *jitter* PPQ5 de todas as vogais de cada indivíduo, assim como os intervalos de confiança de 95%.

Verifica-se que os resultados relativos ao canal GSM são consistentes, na medida em que é sempre introduzida uma pequena degradação através desta via. No que diz respeito às gravações não contemporâneas, o mesmo é observável, à exceção dos casos FEM19, MAL02 e MAL08.

No que concerne o canal VoIP, o nível absoluto e a variância degradam-se, sendo em certos casos o intervalo de confiança tão extenso que denuncia uma propensão nefasta deste tipo de canal. A Tabela 6, através dos valores de correlação normalizada, reforça estas conclusões. Sugere que os valores de correlação *jitter* PPQ5 para o par GSM HQ-TL não são estatisticamente significativos. Por seu turno, os valores de correlação entre as gravações GSM_HQ e VoIP_HQ são estatisticamente significativos, o que sugere relevância idiossincrática do parâmetro presentemente estudado. Contudo, normalmente as gravações de alta qualidade não estão disponíveis para análise forense. Habitualmente, os peritos dispõem, apenas, de amostras nos formatos VoIP TL ou GSM TL. Nestes casos, não se encontrou correlação relevante.

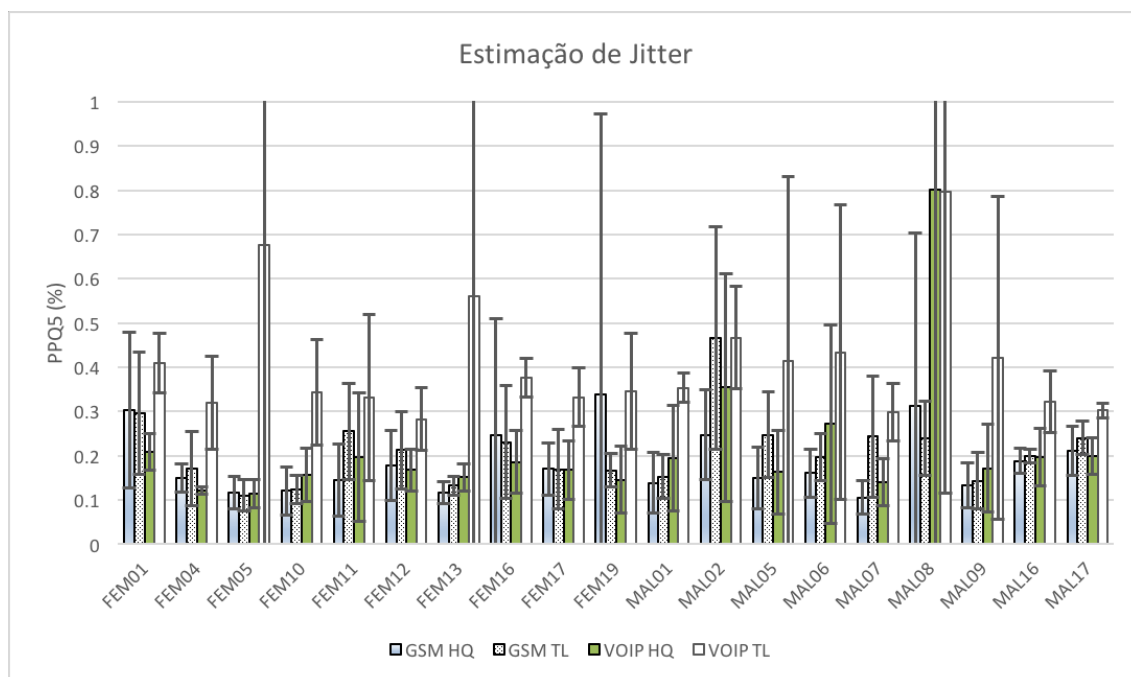


Figura 35 - Resultados estimados do *jitter* PPQ5, incluindo os intervalos de confiança de 95%.

Na Tabela 6, também se verifica que a correlação dos pares VoIP HQ-TL é estatisticamente significativa, o que indicia que, genericamente, a degradação do *jitter* introduzida pelo canal VoIP é sistemática.

Tabela 6 - Correlações normalizadas do jitter PPQ5 (n=19).

<i>r</i>	GSM HQ	GSM TL	VOIP HQ	VOIP TL
GSM HQ	1.0			
GSM TL	0.44249 ($p>0.058$)	1.0		
VOIP HQ	0.50984 ($p<0.027$)	0.35531 ($p>0.134$)	1.0	
VOIP TL	0.18405 ($p>0.450$)	-0.02808 ($p>0.908$)	0.65058 ($p<<0.004$)	1.0

Concluimos, portanto, que o *jitter* não é um parâmetro que usufrua de especial utilidade para a análise forense, excetuando quando o perito tem a possibilidade de analisar a gravação do canal de comunicação GSM.

Examinando os resultados do *jitter* obtidos pelo *Praat* e pelo *VoiceStudio*, podemos concluir que a correlação que existe entre as versões correspondentes de qualidade telefônica e de alta qualidade varia substancialmente tendo em conta a vogal estudada e a modalidade de comunicação. A título de exemplo, no caso do *Praat*, a correlação de *Pearson* varia entre -0.11 e 0.90, com cerca de metade dos valores a não serem estatisticamente significativos; com o *VoiceStudio*, os valores colocam-se entre -0.33 e 0.80 e mais de metade das correlações não é estatisticamente significativa. Além disso, enquanto que, no *Praat*, a degradação mais evidente do *jitter* é detetada no canal VoIP, no *VoiceStudio*, verifica-se essa questão nas comunicações GSM.

Como conclusão, inferimos que os resultados do *jitter* não foram sistemáticos mesmo a nível intra-orador, decorrendo aleatoriamente, o que pode ser causado pela sensibilidade dos algoritmos em cada um dos *software* ou pelas condições de flutuação

do canal. Assim, é observável que as medidas do *jitter* não relevam idiossincrasia quando estão a ser avaliadas amostras de voz recolhidas através dos canais GSM ou VoIP.

5.3.4 Análise com base no parâmetro *shimmer*

As variações involuntárias da amplitude do movimento glotal são traduzidas através do *shimmer*. Uma voz saudável, normalmente, caracteriza-se por um valor inferior a 3% no que diz respeito ao *shimmer* APQ5.

No *Praat*, tal como nos parâmetros anteriores, encontramos diferenças relevantes entre os pares HQ-TL, tanto no canal GSM como no VoIP, em todas as vogais ($p<0.02$). A Figura 36 representa o parâmetro acústico APQ5 para todas as vogais dos oradores intervenientes, com o intervalo de confiança de 95%.

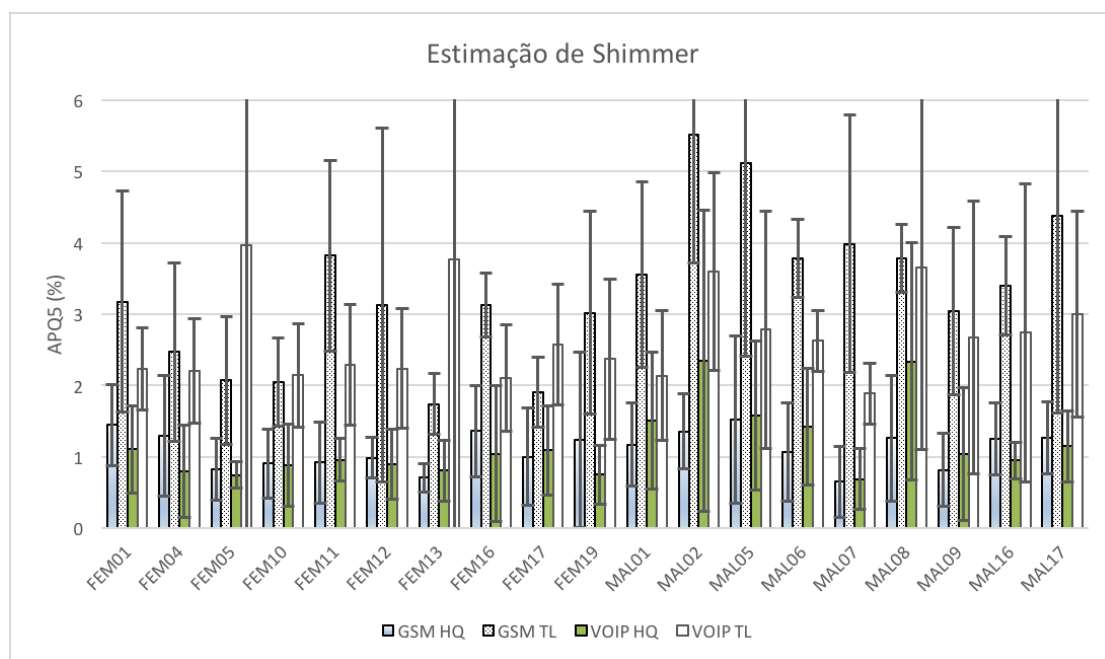


Figura 36 - Resultados estimados do shimmer APQ5, incluindo o intervalo de confiança de 95%.

Através deste gráfico, observa-se que existe concordância entre os valores de GSM HQ e de VoIP HQ, o que, à partida, indicia idiosincrasia. Não obstante, os valores obtidos para as amostras de qualidade telefónica divergem e parecem exacerbados. Para auxiliar na avaliação destes dados, a Tabela 7 apresenta as correlações encontradas para os quatro tipos de gravações.

Tabela 7 - Correlações normalizadas para o shimmer APQ5 ($n=19$).

<i>r</i>	GSM HQ	GSM TL	VOIP HQ	VOIP TL
GSM HQ	1.0			
GSM TL	0.53182 ($p<0.020$)	1.0		
VOIP HQ	0.52809 ($p<0.021$)	0.63616 ($p<0.004$)	1.0	
VOIP TL	-0.00489 ($p>0.983$)	0.02820 ($p>0.908$)	0.41149 ($p>0.079$)	1.0

Assim, no que diz respeito ao *shimmer*, a versão VoIP TL não está profundamente correlacionada com outro tipo de gravação de voz, o que é justificável devido ao facto de o canal VoIP introduzir uma elevada compressão de amplitude não linear. De realçar também que, apesar de algumas das correlações encontradas serem estatisticamente significativas, os intervalos de confiança de 95% das médias sobrepõem-se fortemente para quase todos os sujeitos. Este facto releva que o *shimmer* não é, por si só, um parâmetro de orador discriminatório. Assim, estas diferenças relativas aos diferentes oradores são uma prova de conjunturas de circunstância que se podem alterar de forma imprevisível.

Deste modo, o interesse forense no *shimmer* deve ser questionado, na medida em que a sua capacidade de discriminação é baixa e é plausível que os canais telefónicos introduzam alterações neste parâmetro de maneiras imprevisíveis. Decorre, portanto, que o seu estudo em cenários forenses, com GSM ou VoIP, não é verdadeiramente útil.

Em reforço desta perspectiva, realizámos um estudo das diferenças no parâmetro *shimmer* para os pares TL-HQ das mesmas gravações, quer através do *Praat* quer através do *VoiceStudio*. A Figura 37 representa a média destas diferenças, assim como os intervalos de confiança associados, considerando todos os oradores, para cada vogal e modalidade de comunicação.

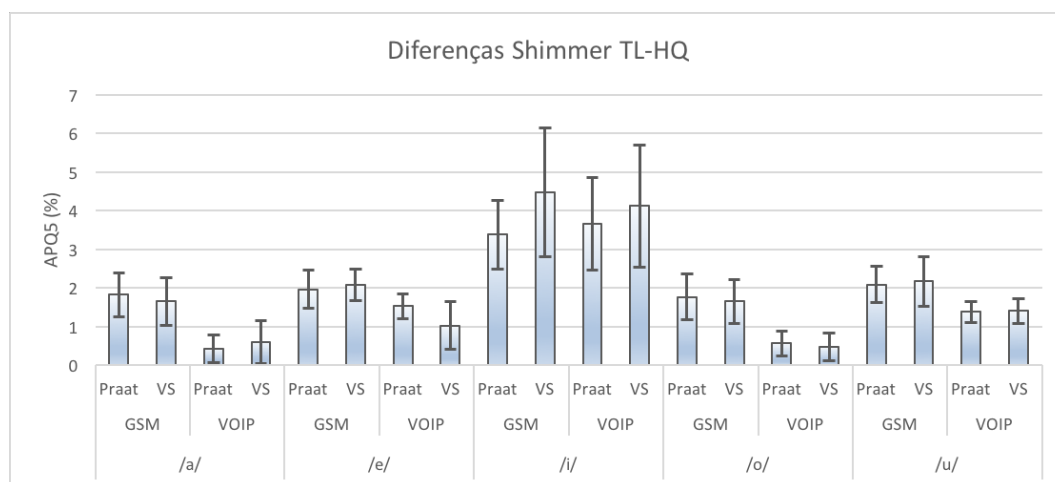


Figura 37 - Médias e intervalos de confiança de 95% do shimmer entre os pares TL e HQ, no Praat e no VoiceStudio.

Observa-se que a consistência entre o *Praat* e o *VoiceStudio* para a média do shimmer é bastante alta. Em média, o canal GSM agrava o shimmer em cerca de 2.3% e o VoIP em cerca de 1.5%. Contudo, estes valores são comensuráveis com os valores típicos de uma voz saudável, além de que este agravamento pode ocorrer aleatoriamente. Assim, reforça-se a conclusão de que o shimmer é bastante vulnerável às influências relacionadas com flutuações de canal, o que confirma o seu reduzido interesse em estudos forenses de comparação de vozes através dos canais GSM e VoIP.

5.3.5 Análise com base no parâmetro HNR

O HNR (*Harmonics-to-Noise Ratio*) é um parâmetro acústico de análise da qualidade de voz que expressa a proporção entre a potência média de sinusoides

harmonicamente relacionadas de uma vogal sustentada e a potência média da componente aperiódica do sinal – ou ruído. Considera-se que o HNR de uma vogal sustentada produzida por um indivíduo com uma voz saudável iguala ou excede os 12 dB.

Foi realizado um teste *t* de Student para avaliar a hipótese nula de médias idênticas no parâmetro HNR entre as versões de alta-qualidade (HQ) e telefônica (TL) dos registos de vogais sustentadas. Este teste revelou que as diferenças entre os valores do parâmetro HNR para as duas versões das vogais de todos os oradores, quer no canal GSM quer no canal VoIP, são estatisticamente significativas ($p < 0.0002$, para dados obtidos no *Praat*).

A Figura 38 representa o parâmetro acústico HNR considerando todas as vogais dos oradores intervenientes, assim como o associado intervalo de confiança de 95%.

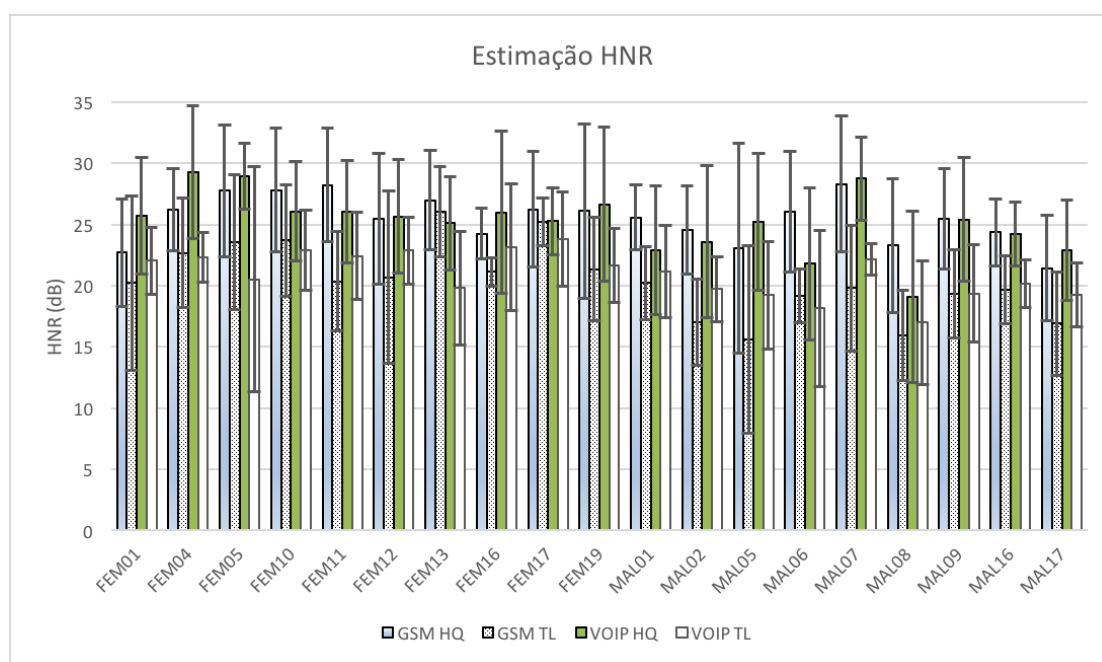


Figura 38 - Resultados estimados do HNR, incluindo os intervalos de confiança de 95%.

Como exemplo concreto, a Figura 39 representa todos os valores de HNR para a vogal /u/ obtidos com o *VoiceStudio*, em todas as modalidades de comunicação.

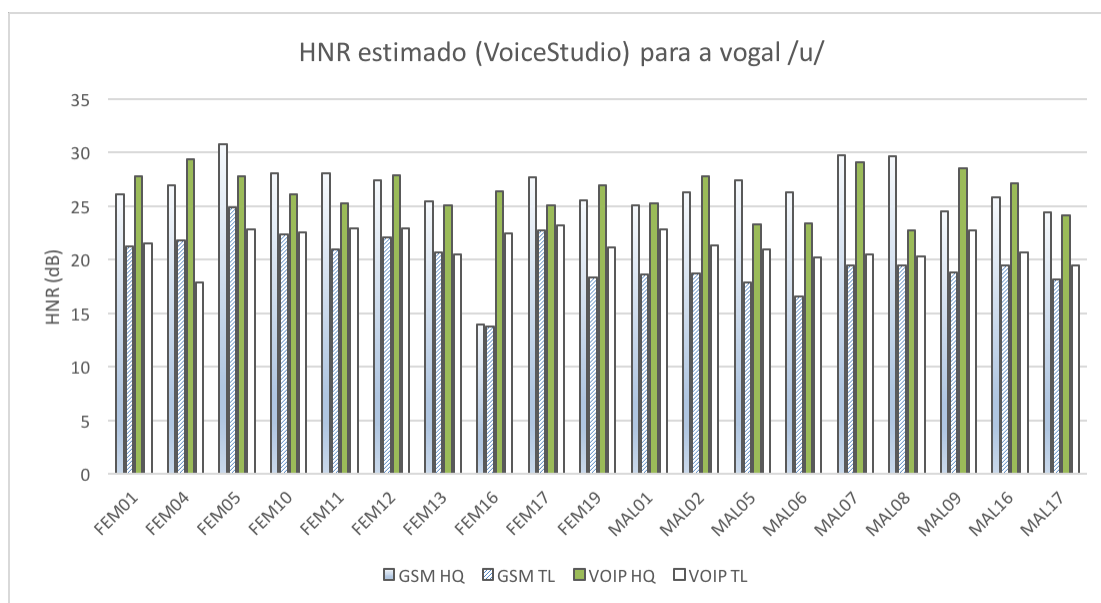


Figura 39 - Resultados do HNR da vogal /u/ no VoiceStudio.

Pode confirmar-se que, para o mesmo orador, os valores obtidos para os registos HQ são condizentes e que, tendencialmente, a redução verificada nos valores de HNR é mais acentuada nas versões TL da modalidade de comunicação GSM do que o é na modalidade VoIP. Este resultado é representativo dos resultados obtidos com o *Praat*, apesar de os valores neste *software* serem, normalmente, ligeiramente mais elevados do que os obtidos no *VoiceStudio*.

Uma forma alternativa e mais útil de caracterizar estes resultados é através da observação das diferenças dos valores de HNR entre as versões HQ e TL de cada vogal. A Figura 40 ilustra a média para essas diferenças e seu intervalo de confiança, entre todos os oradores, para cada *software*, vogal e modalidade de comunicação.

Assim, considerando todas as vogais, pode concluir-se que os resultados no *Praat* e no *VoiceStudio*, representando as diferenças entre os pares HQ-TL e os intervalos de confiança de 95%, são muito semelhantes, sendo, no entanto, que a vogal /i/ é a que representa as maiores dissemelhanças, não sendo, contudo, significativa esta assimetria.

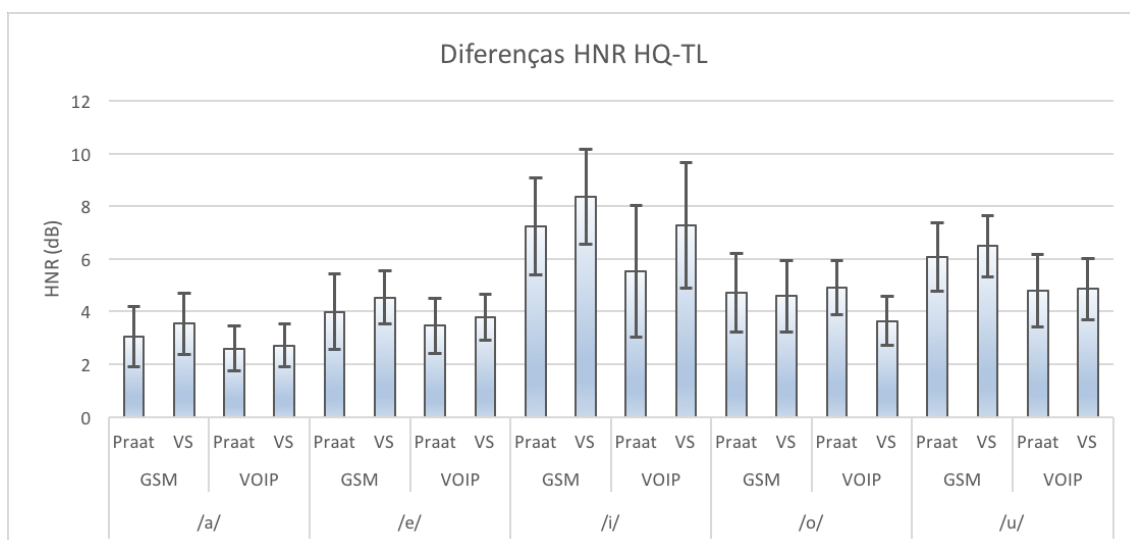


Figura 40 - Média e intervalo de confiança (95%) para a diferença de valores HNR obtidos pelo Praat e VoiceStudio.

A degradação média para o canal GSM é assim de 5.50 dB (*VoiceStudio*) e 5.01dB (*Praat*). Estes valores estão em concordância e sugerem que o canal GSM degrada o HNR em 5.25 dB, em média. Por seu turno, no que diz respeito ao canal VoIP, a média é de 4.45 dB, quando se refere ao *VoiceStudio*, e de 4.26 dB, quando se analisam os valores do *Praat*. Mais uma vez, estes valores estão em conformidade e sugerem que a média de degradação do canal VoIP é de 4.35 dB.

Estes resultados, que estão abaixo do valor mínimo do HNR no que diz respeito a uma voz saudável (12dB), sustentam a conclusão de que este parâmetro acústico pode ser útil em contexto forense de análise de voz.

A Tabela 8 revela que as correlações entre todos os tipos de gravações das amostras são estatisticamente significativas, com a exceção do par GSM HQ-VoIP TL. As correlações mais elevadas são obtidas por ambos os pares HQ-TL, o que confirma que o HNR é um parâmetro acústico informativo. Por outro lado, a correlação média verificada entre os registos VoIP TL e GSM TL, que correspondem a registos plausíveis em situações reais de investigação, é muito expressiva e sugere que o parâmetro HNR pode ser relevante em investigações forenses.

Tabela 8 - Correlações normalizadas de HNR (n=19).

<i>r</i>	GSM HQ	GSM TL	VOIP HQ	VOIP TL
GSM HQ	1.0			
GSM TL	0.64307 (<i>p</i> <0.003)	1.0		
VOIP HQ	0.53973 (<i>p</i> <0.017)	0.52881 (<i>p</i> <0.02)	1.0	
VOIP TL	0.41714 (<i>p</i> >0.075)	0.60049 (<i>p</i> <0.007)	0.65936 (<i>p</i> <0.002)	1.0

Por último, correlacionaram-se, para cada parâmetro acústico, os resultados de estimação obtidos pelos dois *software* e para todas as vogais, modalidades de comunicação (GSM e VoIP) e tipos de gravação (HQ e TL), dos dezanove oradores selecionados. Assim, usaram-se, no total, vinte valores de correlação *Praat-VoiceStudio* para cada parâmetro acústico. Na Figura 41, observam-se os valores médios das correlações e os respectivos intervalos de confiança de 95%.

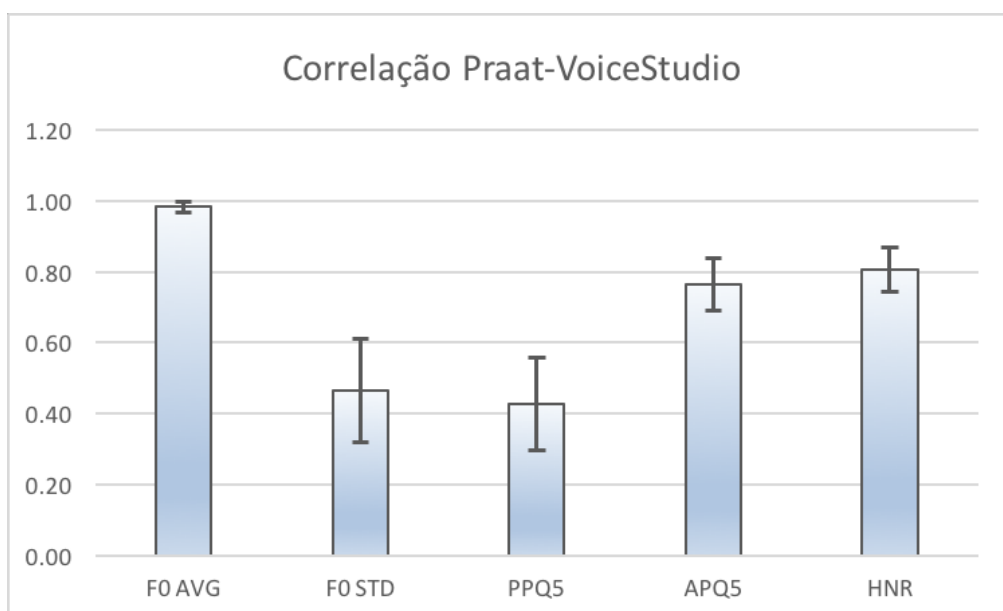


Figura 41 - Correlação entre o Praat e o VoiceStudio para os diferentes parâmetros acústicos (n=20), a média e o intervalo de confiança de 95%.

Conclui-se que as médias das correlações relativas quer ao desvio padrão da frequência fundamental, quer ao *jitter* PPQ5, são muito modestas e que os respetivos intervalos de confiança são os mais dilatados. Realizou-se um teste *t* de Student que confirmou que estes resultados das correlações não são estatisticamente significativos ($p > 0.05$). Ao contrário, no caso da média da frequência fundamental, do *shimmer* APQ5 e do HNR os resultados são mais interessantes e estatisticamente significativos na medida em que para estes casos $p < 0.05$. A Tabela 9 explicita os valores do teste *t*, sendo que os valores-*p* refletem os piores valores obtidos para as vinte correlações possíveis.

Tabela 9 - Correlações Praat-VoiceStudio consideradas para os 19 oradores ($n=20$).

Média F0	Desvio padrão F0	<i>Jitter</i> PPQ5	<i>Shimmer</i> APQ5	HNR
0.99876	0.51123	0.42166	0.78681	0.82367
$p < 0.05$	$p > 0.05$	$p > 0.05$	$p < 0.04$	$p < 0.04$

Estes resultados suscitam várias hipóteses para a fraca consistência que se verifica em alguns casos (desvio padrão da frequência fundamental e *jitter* PPQ5). Em especial, é plausível a hipótese de estarem relacionados com um mau funcionamento dos algoritmos de estimação ou com uma elevada sensibilidade destes às condições adversas do canal de comunicação.

5.3.6 Análise com base nas frequências das formantes

Como já referido na Secção 5.2.2, a análise das frequências formantes é muito importante, dado que veiculam informação linguística e também de orador. No contexto da acústica forense, dado que tipicamente os únicos registos disponíveis são de voz telefónica, importa compreender se, em relação aos registos de alta qualidade, existe alteração relevante das frequências formantes devido à influência dos canais de

comunicação GSM e VoIP. Em concreto, o nosso principal objetivo era o de responder às seguintes questões:

- O grau de modificação é uma função da ordem das formantes?
- A modificação das formantes é mais severa no caso da comunicação GSM ou da comunicação VoIP?
- A modificação é mais vincada nos casos de voz feminina ou de voz masculina?
- Como se compara o triângulo das vogais dos oradores desta investigação com os dos outros estudos?
- Como se compara a modificação das formantes que é detetada na nossa investigação com a reportada por outros estudos?

Para esta tarefa, incluímos os oradores que formam um subconjunto da base de dados, uma vez que esta apenas inclui as amostras dos indivíduos para os quais não existiram erros na deteção das frequências formantes no *software Praat*. Consequentemente, tal como a Tabela 10 explicita, o painel de oradores para este estudo compreende 17 oradores femininos e 15 oradores masculinos, incluindo gémeos. Validaram-se dois diálogos por indivíduo correspondendo às duas modalidades de comunicação – GSM e VoIP – (quatro gravações por orador em HQ e TL) e cada uma das conversas incluiu, igualmente, cinco vogais sustentadas, sendo que foi a partir delas que se extraíram e analisaram as formantes.

A análise das formantes foi efetuada através do algoritmo *Burg* que é executado no *Praat*. As frequências que se pretendiam estudar confinam-se na gama entre 50Hz e 5500Hz. De ressaltar que as definições daquele algoritmo permitem, no máximo, o alcance de cinco formantes.

As Figuras 42 e 43 ilustram, para oradores femininos e masculinos, respetivamente, a média da formante F1 para as várias situações de vogal e canal de comunicação, assim como o respetivo intervalo de confiança de 95%. Para as várias vogais, verifica-se que os intervalos de confiança para as médias da formante F1 nas quatro variantes de comunicação possuem uma assinalável sobreposição. De modo a

facilitar a caracterização dos desvios no valor estimado de F1 entre as versões telefônicas e de alta-qualidade dos registos, calculou-se o valor médio dos desvios relativos tendo por referência os valores dos registos de alta-qualidade. Estes resultados encontram-se representados na Figura 44.

***Tabela 10** - Idade, género e grau de parentesco dos oradores cujas frequências formantes foram extraídas e analisadas a partir das vogais sustentadas.*

Orador	Idade	Género	Gémeo/Parente
FEM 01/02/03	22/22/45	F	Gémeas/Mãe
FEM 05/06	27/32	F	Gémea/Prima
FEM 07	38	F	-
FEM 10/11/12	21/55/21	F	Gémea/Mãe/Gémea
FEM 13/14/15	22/49/22	F	Gémea/Mãe/Gémea
FEM 16/17/18	49/15/49	F	Gémea/Filha/Gémea
FEM 19/20	36/36	F	Gémeas
MAL 01/02/03	71/71/71	M	Gémeo/Primo/Gémeo
MAL 04/05	21/21	M	Gémeos
MAL 06/07/08	77/52/52	M	Pai/Gémeos
MAL 09/10/11	24/24/24	M	Primo/Gémeos
MAL 12	16	M	-
MAL 15	16	M	-
MAL 16/17	19/19	M	Gémeos

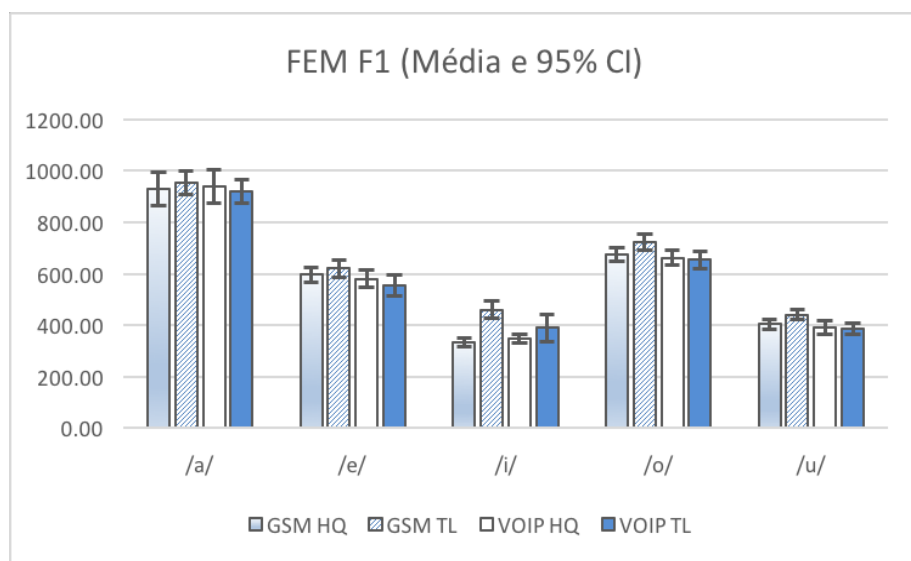


Figura 42 - Média da F1 e respetivo intervalo de confiança de 95% (oradores femininos).

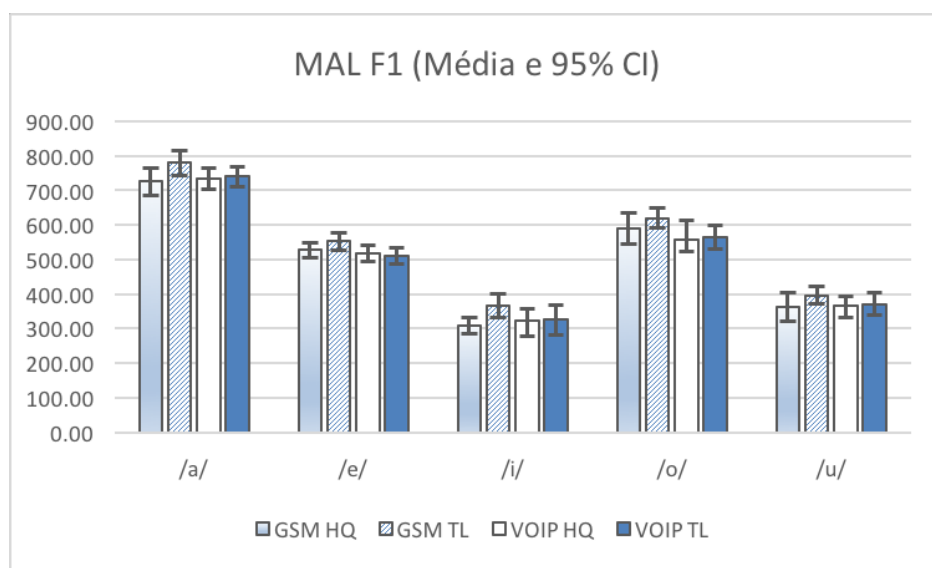


Figura 43 - Média da F1 e respetivo intervalo de confiança de 95% (oradores masculinos).

Contrariamente às restantes figuras, esta figura tem a particularidade de sinalizar os desvios relativos que são estatisticamente significativos, através de cor sólida (azul). Assim, conclui-se desde logo que, no que diz respeito à formante F1 e como é ilustrado

na Figura 44, o desvio relativo mais expressivo e estatisticamente significativo ocorre no caso das comunicações GSM. Além disso, é possível observar que o desvio relativo mais acentuado acontece nas vogais cuja média de F1 é mais baixa, como no caso de /i/ e /u/.

Estas constatações são facilmente compreendidas considerando que o desvio é consequência da limitação de banda imposta pelo canal de comunicação telefónica que rejeita as frequências de sinal abaixo de 300 Hz. Por consequência, o algoritmo que calcula o valor das frequências formantes tem tendência para as sobrestimar no caso dos registos telefónicos. Como última nota, na Figura 44, observa-se ainda que, no pior caso, o desvio relativo da frequência F1 se aproxima dos 40%.

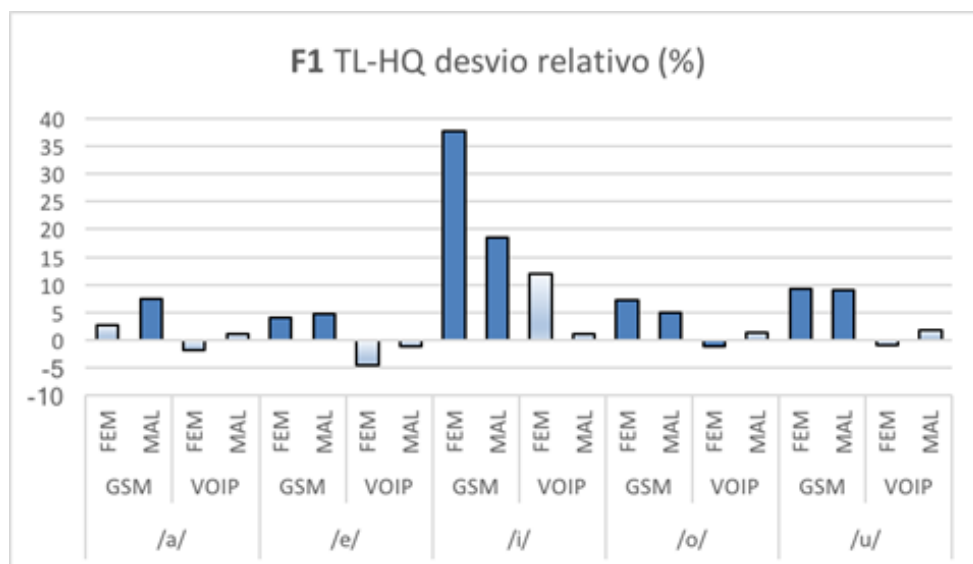


Figura 44 - Médias dos desvios relativos de F1 entre as versões TL e HQ. Os resultados com significância estatística assinalam-se a cor sólida (azul).

De modo análogo ao caso anterior, as Figuras 45 e 46 revelam a média da formante F2 das amostras produzidas pelos oradores masculinos e femininos, respetivamente, assim como o respetivo intervalo de confiança de 95%. Verifica-se, neste caso, que os valores obtidos para as versões de alta-qualidade e qualidade telefónica do mesmo sinal de vogal são muito condizentes, pelo que uma representação

dos desvios relativos pode ajudar a um melhor esclarecimento dos resultados. Esta representação encontra-se na Figura 47.

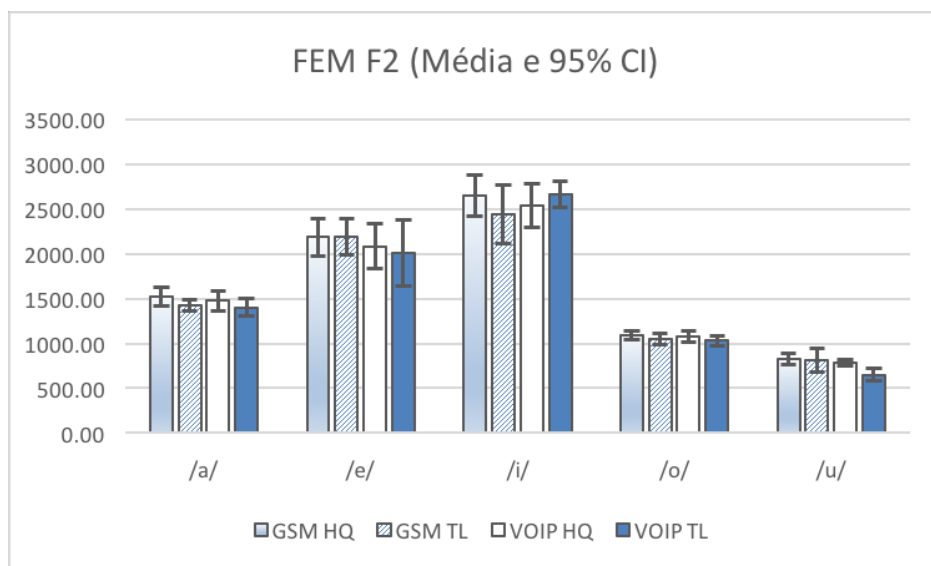


Figura 45 - Média da F2 e respetivo intervalo de confiança de 95% (oradores femininos).

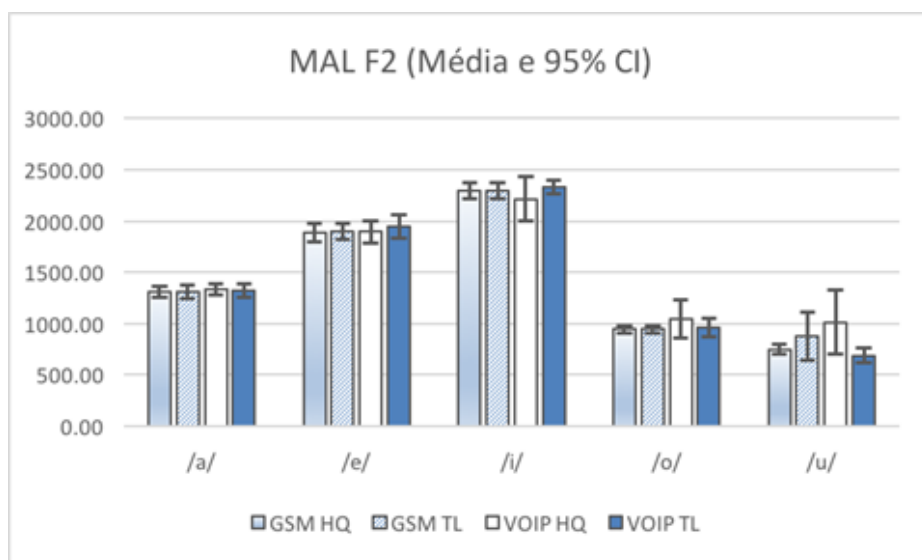


Figura 46 - Média da F2 e respetivo intervalo de confiança de 95% (oradores masculinos).

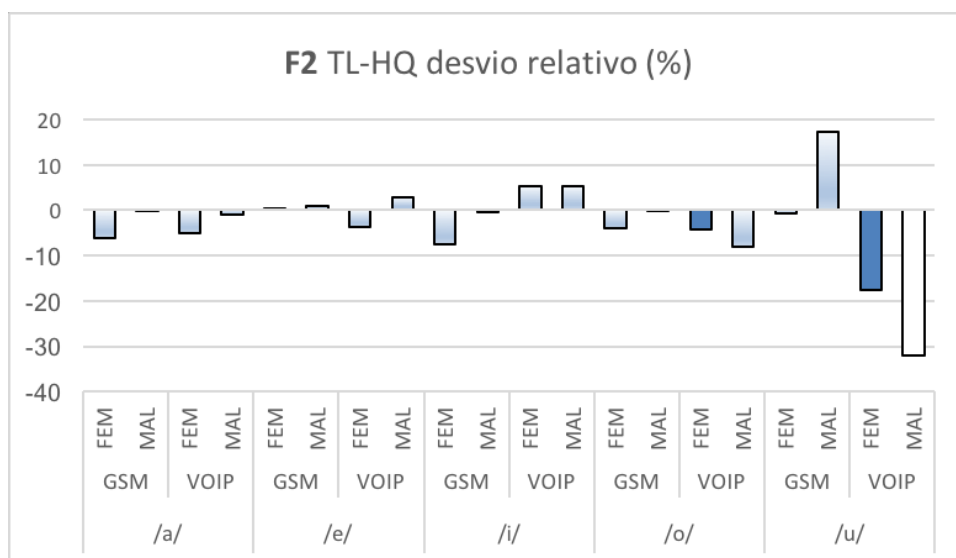


Figura 47 - Médias dos desvios relativo de F2 entre as versões TL e HQ. Os resultados com significância estatística assinalam-se a cor sólida (azul).

Esta figura revela que foram encontrados apenas dois desvios estatisticamente significativos da formante F2, nos oradores femininos. Verifica-se, igualmente, um resultado marginalmente não estatisticamente significativo assinalado pela coluna branca, em que $p=0.051$. O valor máximo de desvio da F2 aproxima-se dos 20%. É de realçar que estas conclusões aplicam-se às vogais /o/ e /u/ que, entre todas as consideradas, possuem o valor mais baixo para a média da frequência F2, para ambos os géneros de oradores. Deste modo, pode associar-se como causa provável o efeito de limitação de banda do canal telefónico.

De uma maneira geral, pode concluir-se que no caso da frequência formante F2, a maior parte das diferenças não são estatisticamente significativas, o que indica que a análise da formante F2 é relativamente robusta e segura, na maior parte dos casos.

As Figuras 48 e 49 apresentam os resultados obtidos com a análise da formante F3, incluindo o intervalo de confiança de 95%. Contrariamente ao observado para as formantes F1 e F2, verifica-se que a dispersão da formante F3 entre as várias vogais, quer para oradores femininos quer para masculinos, é reduzida, o que significa que a formante F3 não oferece potencial discriminatório. Por outro lado, considerando as várias vogais, as médias da formante F3 para as versões telefónicas diferem mais das

versões de alta qualidade, no caso dos oradores femininos, especialmente para as vogais /a/ e /o/. Aqui também uma análise dos desvios relativos ajuda a caracterizar esta constatação.

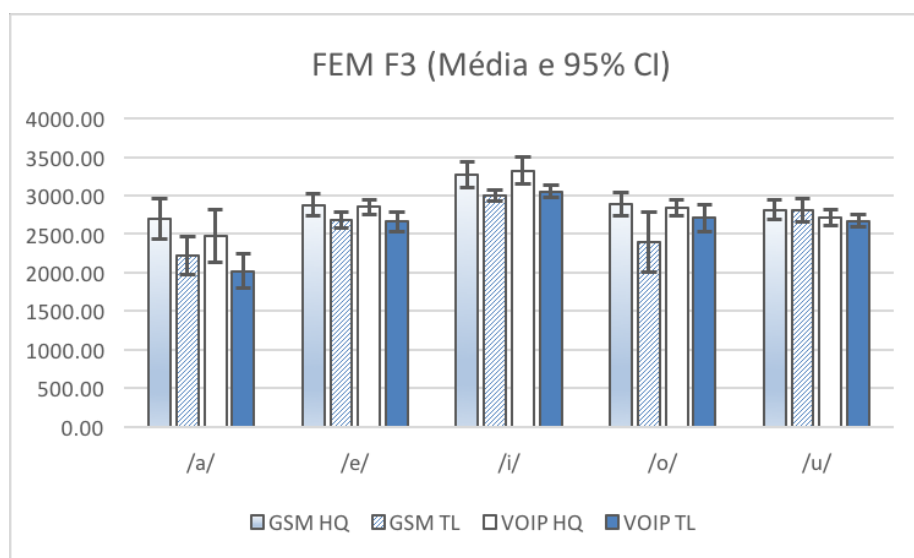


Figura 48 - Média da F3 e respetivo intervalo de confiança de 95% (oradores femininos).

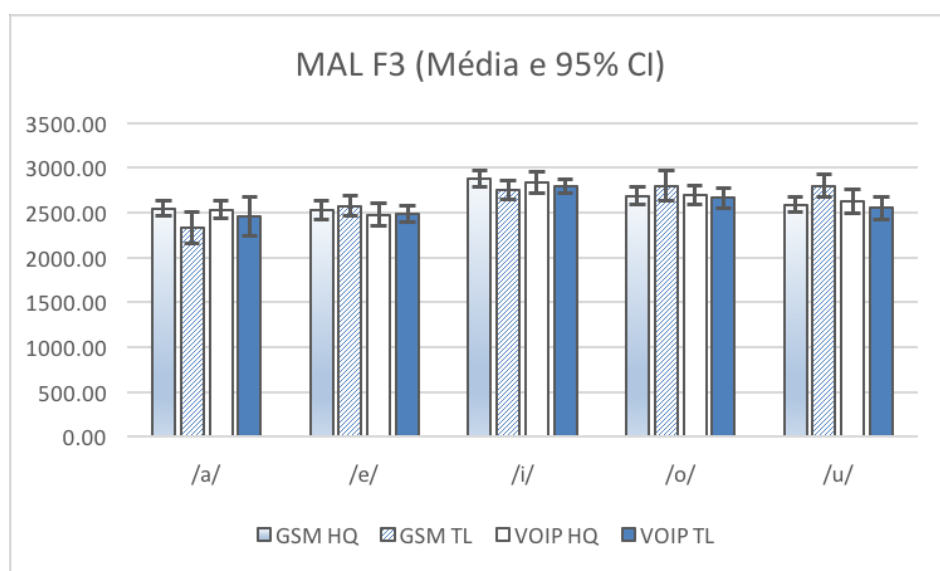


Figura 49 - Média da F3 e respetivo intervalo de confiança de 95% (oradores masculinos).

A Figura 50 ilustra a média do desvio relativo da frequência formante F3 entre as versões TL e HQ dos registos de vogais sustentadas, em função da vogal, do género do orador (FEM e MAL), e da modalidade de comunicação (GSM e VoIP).

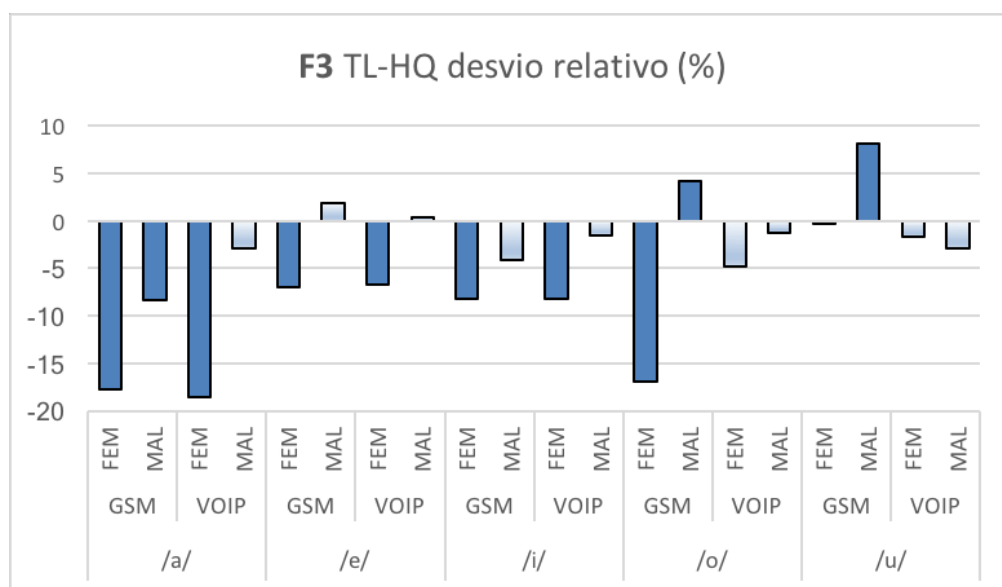


Figura 50 - Médias dos desvios relativos de F3 entre as versões TL e HQ. Os resultados com significância estatística assinalam-se a cor sólida (azul).

As principais conclusões extraídas com os resultados de F3 prendem-se com o facto de que o desvio mais destacado e estatisticamente significativo ocorre no caso das amostras de vogais dos oradores FEM, uma vez que os valores que são apresentados são superiores em relação aos valores extraídos das gravações dos oradores MAL. Este desvio advém do limite de banda do canal de comunicação que rejeita as frequências do sinal na proximidade de 3400 Hz e acima desta frequência. Deste modo, o algoritmo de estimação subestima as frequências das formantes, com coerência com o facto do desvio ser tendencialmente negativo. Esta tendência é particularmente visível nas vozes femininas porque a frequência fundamental é cerca de uma oitava superior à das vozes masculinas, o que significa que o enfraquecimento dos harmónicos é maior, o que tem grande impacto nos resultados de estimação uma vez que dependem da energia dos harmónicos.

Observa-se, do mesmo modo, que o desvio máximo mais significativo da F3 é de 20%. Esta formante não varia muito entre as vogais, o que manifesta que, como referido acima, não contribui para as marcas espectrográficas relevantes das vogais, ou seja, a sua identidade, ao contrário de F1 e de F2.

Por último, nas Figuras 51 e 52, ilustram-se os resultados relativos à frequência formante F4 nos oradores femininos e masculinos, incluindo os intervalos de confiança de 95%. Como seria fácil de antecipar, os resultados da média estimada para F4 exprimem uma tendência mais pronunciada da já verificada para os resultados da formante F3. Com efeito, dado que por definição a formante F4 tem frequência superior a F3 e dado que a limitação de banda do sinal de voz imposta pelo canal telefónico é a mesma, a formante F4 é mais fortemente subestimada nas versões TL dos registos em relação às versões HQ. Esta conclusão é especialmente notória para as vogais /a/, /e/ e /i/, quer para oradores FEM quer para MAL, na medida em os valores médios estimados para F4 são sistematicamente inferiores nas versões TL em relação às versões HQ. Estas alterações são mais facilmente caracterizadas através da representação dos desvios relativos.

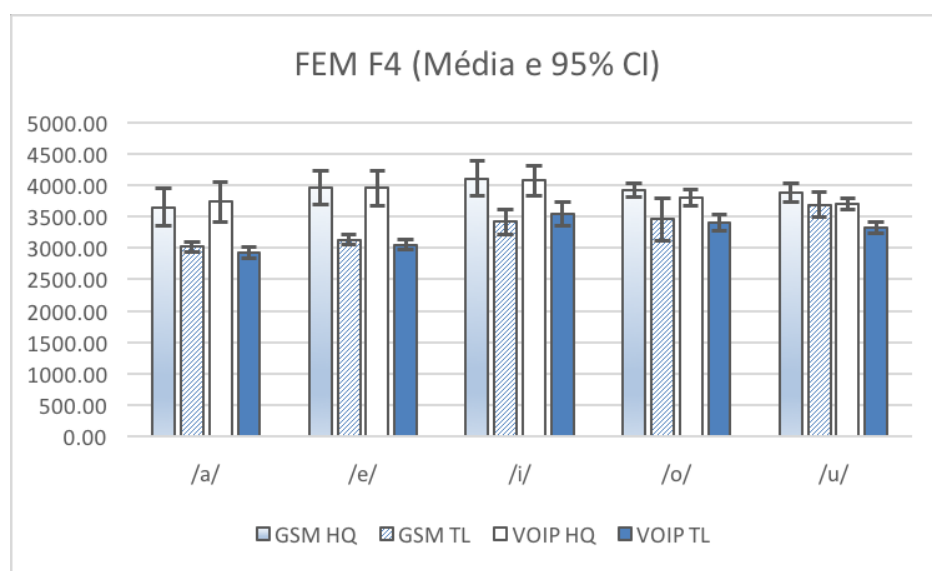


Figura 51 - Média da F4 e respetivo intervalo de confiança de 95% (oradores femininos).

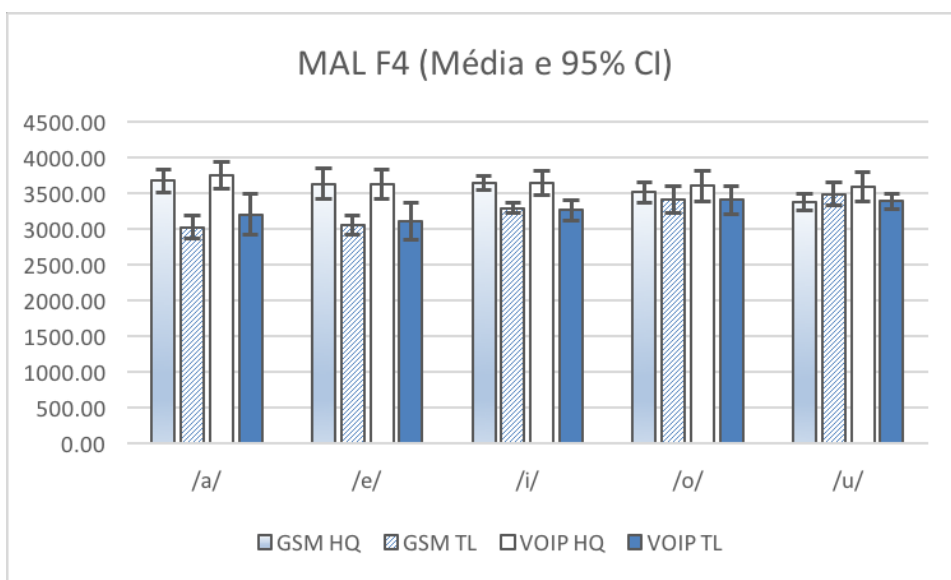


Figura 52 - Média da F4 e respetivo intervalo de confiança de 95% (oradores masculinos).

A Figura 53 ilustra o desvio relativo da formante F4 da versão TL em relação à versão HQ, para todos os oradores, FEM e MAL, quer em GSM quer em VoIP.

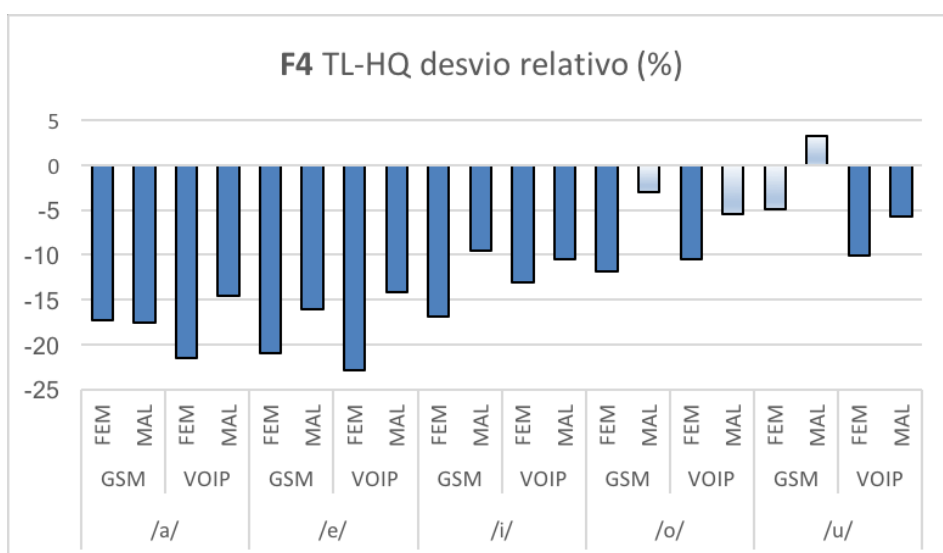


Figura 53 - Médias dos desvios relativos de F4 entre as versões TL e HQ. Os resultados com significância estatística assinalam-se a cor sólida (azul).

Esta figura revela que quase todos os desvios de F4 são substanciais e estatisticamente significativos. Tal como em F3, o desvio em F4 é consequência do limite de banda do canal de comunicação que rejeita as frequências do sinal na proximidade de 3400 Hz e acima deste valor. Assim, F4, cuja média se situa na ordem de 3500 Hz, é especialmente afetada, o que induz o algoritmo a subestimar esta formante, como referido acima. Por seu turno, o desvio máximo de F4 aproxima-se de 25%. Tal como em F3, não existe para F4 uma significativa dispersão de resultados entre as vogais, o que significa que F4 não contribui, de modo relevante, para a identidade das vogais.

Os resultados aqui apresentados sugerem também que as estimativas das formantes são particularmente consistentes entre as gravações de voz de alta qualidade, *i.e.* GSM-HQ e VoIP-HQ. Assim, tornou-se pertinente comparar as médias de F1 e de F2, que se extraíram dos registos HQ, com os resultados de outros autores como Escudero *et al.* [94]. As Figuras 54 e 55 ilustram essa comparação (oradores femininos e oradores masculinos, respetivamente). A linha interrompida corresponde aos resultados de [94] e a contínua, a cor azul, ilustra os obtidos na nossa investigação.

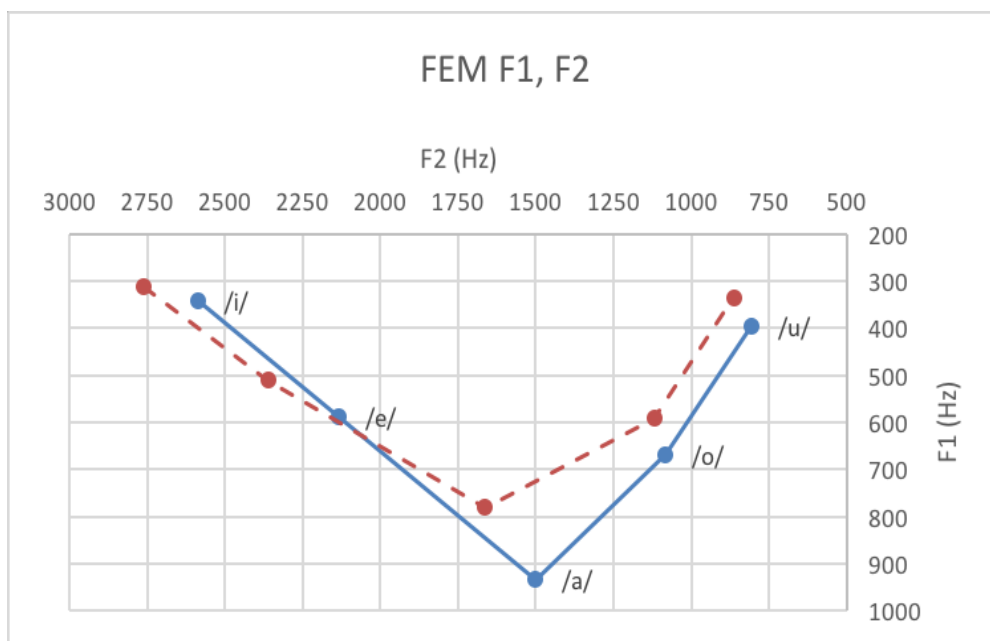


Figura 54 - Comparação entre as médias de F1 e de F2 (oradores femininos). A linha contínua (azul) exprime os resultados do nosso estudo e a interrompida (castanho) representa resultados de [94].

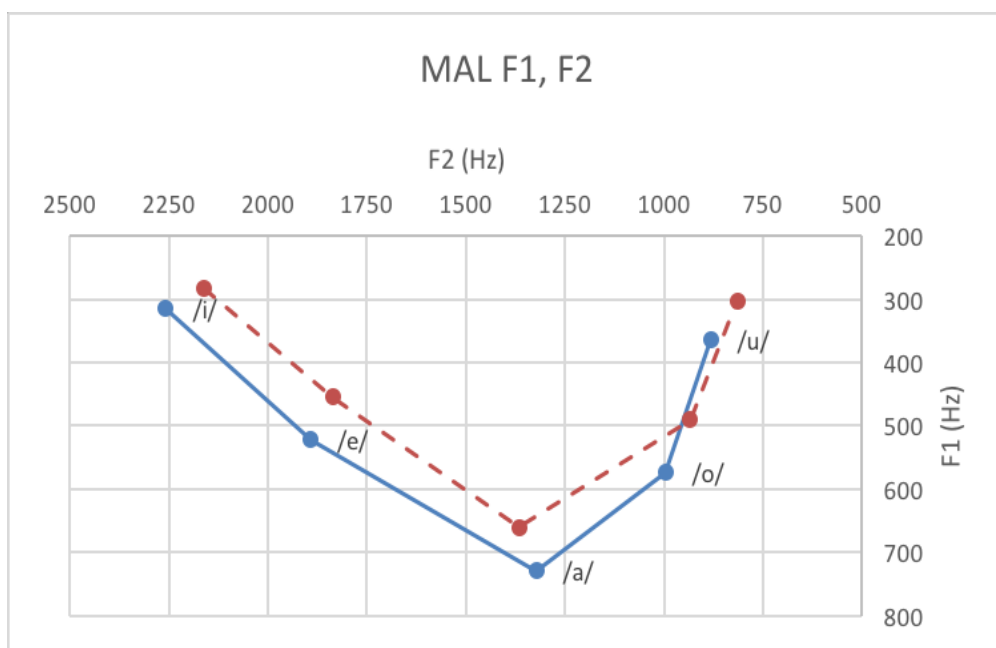


Figura 55 - Comparação entre as médias de F1 e de F2 (oradores masculinos). A linha contínua (azul) exprime os resultados do nosso estudo e a interrompida (castanho) representa resultados de [94].

Através destes gráficos observa-se que, apesar de não exatamente coincidentes, os resultados são consistentes, desde logo porque o triângulo das vogais é mais expandido no caso dos oradores femininos do que nos masculinos. Por outro lado, os nossos dados também corroboram a tese de Escudero *et al.* [94], segundo a qual cada vogal recuada (/u/ e /o/) tem uma média de F1 mais elevada do que as anteriores correspondentes (/i/ e /e/).

É oportuno ter em consideração que as diferenças entre a análise de Escudero *et al.* [94] e as alcançadas nesta dissertação podem prender-se com as diferentes bases de dados. Com efeito, em [94], todos os oradores têm menos de trinta anos de idade e a base de dados é constituída por 10 oradores FEM e 10 MAL. Por sua vez, a base de dados da nossa investigação inclui 17 oradores FEM e 15 MAL numa faixa etária que compreende idades entre os dezasseis e os setenta e um anos. Adicionalmente, nas amostras de [94], as vogais são produzidas recorrendo à estrutura dissilábica CVCV

(consoante-vogal-consoante-vogal); as vogais do estudo desta dissertação, como referimos, são sustentadas e têm cerca de um segundo de duração.

Por outro lado ainda, as diferenças entre os resultados de Escudero *et al.* [94] e os resultados desta dissertação são consistentes com a indicação de que o diagrama de F1 e F2 pode desviar cerca de 20% da média, especialmente no que concerne F2.

Por último, recorreu-se às vogais sustentadas do nosso estudo para avaliar, também, o resultado dos desvios das formantes F1 e F2 no triângulo acústico tradicional, ilustrado na Figura 56 para o caso dos oradores do género feminino. Trata-se de uma alternativa de apresentação dos resultados já discutidos anteriormente mas permitindo uma leitura abrangente.

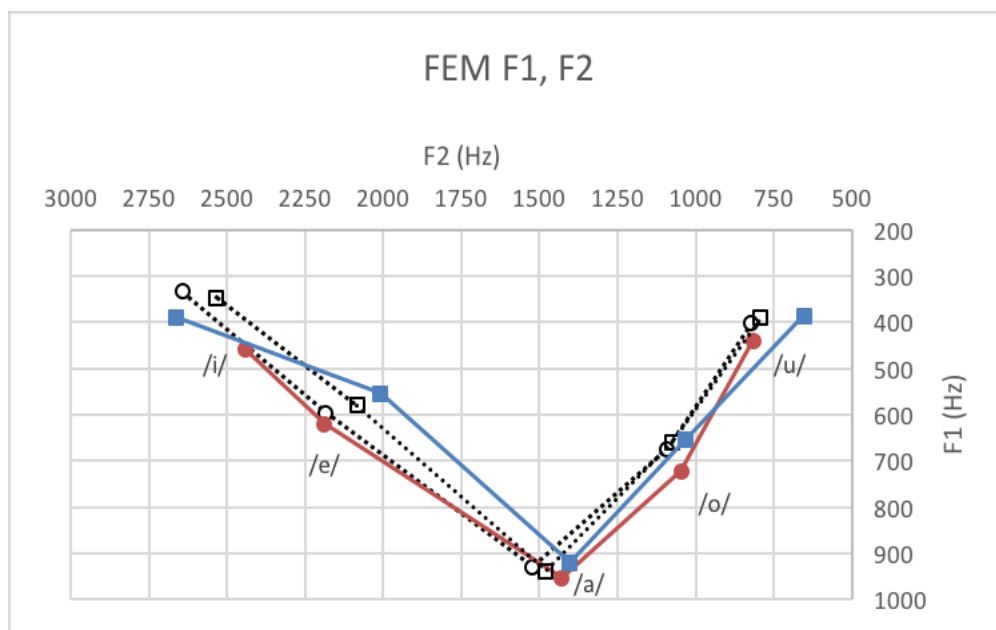


Figura 56 - Resultados dos desvios de F1 e de F2 no mapa dos triângulos acústico das vogais (oradores femininos). (GSM HQ – linha pontilhada redonda; GSM TL – linha sólida redonda; VoIP HQ – linha pontilhada quadrada; VoIP TL – linha sólida quadrada).

Com efeito, é notório na Figura 56 que a separação entre as linhas pontilhadas é menor do a separação entre as linhas com símbolos redondos, ou entre as com símbolos quadrados. Decorre assim que os canais de comunicação GSM e VoIP podem dar aso a

diferenças na estimação do valor das formantes que excede a variação natural intrafalante, o que sugere que usar as frequências formantes em tarefas de comparação de orador deve ser acompanhado de precaução, porque pode conduzir a conclusões incorretas.

A Figura 57 ilustra os casos das amostras de voz masculina.

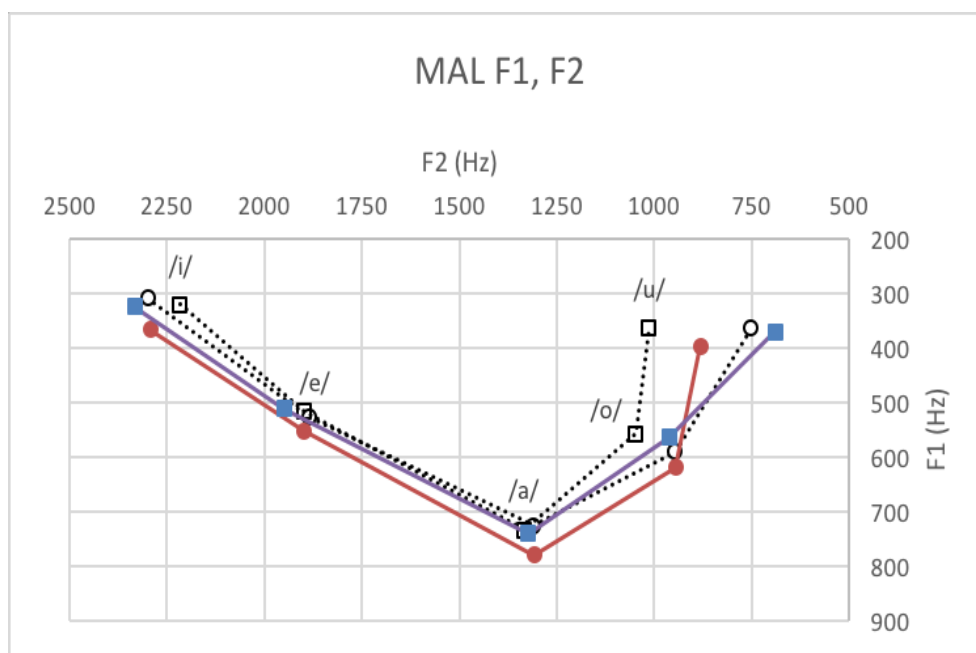


Figura 57 - Resultados dos desvios de F1 e de F2 no mapa dos triângulos acústico das vogais (oradores masculinos). (GSM HQ – linha pontilhada redonda; GSM TL – linha sólida redonda; VoIP HQ – linha pontilhada quadrada; VoIP TL – linha sólida quadrada).

No caso das vogais dos oradores masculinos, as vogais recuadas /u/ e /o/ aparentam ser especialmente afetadas pela diferença de estimativa de F2 entre a versão de alta qualidade e a versão de qualidade telefônica. Neste diagrama, é possível, então, confirmar que os canais de comunicação GSM e VoIP têm uma propensão para alterar os valores das formantes, que podem exceder a variação natural intrafalante.

O último teste executado com as vogais sustentadas relaciona-se com os resultados agregados em função de canal GSM e VoIP, por um lado e, por outro lado, de género (FEM e MAL). A Tabela 11 expõe os valores médios, em percentagem, dos

resultados agregados entre oradores e vogais, para os desvios relativos das frequências formantes, em função do tipo de canal.

Tabela 11 - Valores médios dos resultados agregados, em percentagem, para os desvios relativos das frequências formantes, em função do tipo de canal.

	GSM	VOIP	(%)
F1 AVGdev	10,56	0,76	
F2 AVGdev	-0,09	-5,82	
F3 AVGdev	-4,86	-4,83	
F4 AVGdev	-11,49	-12,86	

Verifica-se, na Tabela 11, que as formantes F1 e F4 estão mais expostas aos efeitos da limitação de banda dos canais de comunicação e, por isso, mais suscetíveis a erros de estimação, por oposição a F2 e a F3, uma vez que, por estarem dentro dos limites da banda, são apenas ligeiramente afetados.

Além disso, através da Tabela 11, verifica-se que enquanto em F1 é tipicamente sobrestimada, as formantes F3 e a F4 são subestimadas devido aos limites de banda do canal. O erro relativo de F4 é mais severo, em termos absolutos, do que o erro relativo que diz respeito à F1. Globalmente, o canal de comunicação GSM produz mais erro na estimação das formantes do que o canal VoIP.

A Tabela 12 apresenta os valores médios, em percentagem, dos resultados agregados entre canais de comunicação e vogais, para os desvios relativos das frequências formantes, em função do género do orador.

Tabela 12 - Valores médios dos resultados agregados, em percentagem, para os desvios relativos das frequências formantes, em função do género do orador.

	FEM	MAL	(%)
F1 AVGdev	6,45	4,86	
F2 AVGdev	-4,38	-1,54	
F3 AVGdev	-9,02	-0,67	
F4 AVGdev	-14,98	-9,38	

Conclui-se que, também devido à limitação de banda do canal, a F1 é tipicamente sobrestimada, ao contrário do que acontece com F2, F3 e F4. Além disso, os erros de estimação são mais severos nos casos das vozes femininas do que nas amostras de voz masculinas, o que pode ser explicado pelo facto de as vozes das mulheres possuírem um espectro de magnitudes mais esparso do que os dos homens. A principal razão decorre da frequência fundamental ser, tipicamente, uma oitava mais elevada do que a F0 dos homens e, por isso, os harmónicos são mais fortemente afetados pela limitação de banda o que penaliza a estimação de formantes que se situam precisamente nos extremos da banda.

Com base noutros estudos, por exemplo [95], concluímos que, em coerência com os nossos resultados, a formante F1 é classicamente sobrestimada. Os autores afirmam que, em média, esta formante é sobrestimada em 29% e que a F3 é subestimada em 3,4% para as amostras de vozes masculinas e 1,9% para as gravações de voz femininas. Joerg Bitzer *et al.* [96] afirmam que a média estimada do desvio de frequência é de 1%, sensivelmente. Apesar dos resultados numéricos serem naturalmente diferentes dos nossos, atendendo sobretudo às diferentes bases de dados, há uma clara concordância na conclusão global de que o canal de comunicação provoca alterações estatisticamente significativas na estimação das formantes.

5.4 Conclusão

Neste capítulo, apresentaram-se os estudos com base em parâmetros acústicos seleccionados, como a frequência fundamental, o desvio padrão da frequência fundamental, o *jitter*, o *shimmer*, o HNR e as frequências formantes, através da análise das vogais sustentadas. Por conseguinte, este capítulo traduz resultados de investigação que ocuparam grande parte da duração da investigação. Seguidamente, sintetizam-se algumas das conclusões principais.

Através dos resultados do *software* utilizado de avaliação acústica, os parâmetros que se revelaram mais interessantes do ponto de vista forense foram a média

da frequência fundamental e o HNR (*Harmonics-to-Noise Ratio*), na medida em que revelaram uma correlação mais elevada e estatisticamente significativa entre os estudos de vozes de cada orador com amostras obtidas através dos canais GSM e VoIP.

O desvio padrão da frequência fundamental e o *jitter* PPQ5 foram considerados o segundo conjunto mais pertinente na análise forense de amostras de voz. De facto, ambos os parâmetros revelaram baixa capacidade de discriminação; podem ser afetados por fatores situacionais – não estruturais – como a ansiedade ou o stress; além disso, são vulneráveis aos efeitos dos canais de comunicação. Para serem úteis em contextos forenses, deve existir diversidade de amostras de voz que mitiguem perturbações de sinal devidas ao canal ou ao orador.

Finalmente, concluiu-se que o *shimmer* APQ5 se revelou irrelevante no que diz respeito à análise de voz em contexto forense, uma vez que os resultados, em termos de comparação de vozes que são afetadas pelos canais GSM e VoIP, foram aleatórios. Pensa-se que estes resultados refletem as condições flutuantes das condições dos canais, além de que o nível desses efeitos claramente excede o alcance do *shimmer* que é considerado normal para vozes humanas saudáveis.

Depois de se proceder a um estudo comparativo entre os resultados de ambos os *software* utilizados, *Praat* e *VoiceStudio*, pudemos concluir que, tal como no estudo anterior, a média da frequência fundamental continua a ser o parâmetro acústico mais relevante e fidedigno quando nos referimos a análise de voz em contexto forense, na medida em que é, essencialmente, imune às degradações introduzidas pelo canal telefónico. Assim, concluímos que revela concordância entre os pares GSM e VoIP para o mesmo orador e revela consistência entre os valores quando se recorre a *software* diferentes. Portanto, infere-se que se não houver grande diversidade de gravações de voz, este parâmetro é muito útil.

No que diz respeito ao desvio padrão da frequência fundamental, os resultados obtidos com o *VoiceStudio* exibiram variações que não se comprovaram ser consistentes com os resultados do *Praat*. Este último revelou boa consistência entre as gravações HQ-TL, mas as que são não contemporâneas não confirmam essa propriedade a nível

estatístico. Isto indica que o desvio padrão da F0 não é uma característica com idiossincrasia permanente e pode refletir situações pontuais.

O *Praat* revelou também melhores resultados do que o *VoiceStudio* no que diz respeito ao *jitter*. Não obstante, as perturbações do *jitter* não se afiguraram ser sistemáticas, o que pode indicar suscetibilidade a oscilações do canal e a fatores emocionais. Este parâmetro, tal como o desvio padrão da frequência fundamental, pode ajudar a reforçar as conclusões obtidas através da média da frequência fundamental, quando há diversidade suficiente de amostras de voz.

Foi observada uma boa consistência *Praat-VoiceStudio* no que diz respeito ao *shimmer*, mas o elevado nível de perturbação introduzida pelos canais, assim como o seu perfil de variação ao longo do tempo, reduzem o interesse deste parâmetro no âmbito forense de estudo de voz.

Os resultados com base no parâmetro HNR estimado em ambos os *software* foram muito consistentes e denotaram uma influência sistemática do canal que pode ser característica do GSM e do VoIP, o que pode ser vantajoso em estudos forenses. Tal como a média da frequência fundamental, torna-se útil quando não existe grande diversidade de gravações.

O maior desafio, na análise das frequências formantes, foi clarificar as questões centrais relacionadas com a influência do canal de comunicação na estimação das formantes. Foram apresentados resultados quantitativos detalhados que suportam conclusões pertinentes e implicações para os desafios de comparação de voz, numa perspetiva forense, usando discurso telefónico. Em concreto, concluiu-se que devido à limitação de banda imposta pelo canal GSM e VoIP, a formante F1, situada no limite inferior da banda de voz, é sobrestimada em cerca de 11% (com as gravações GSM). A formante F3 e, particularmente, F4, situadas no limite superior da banda de voz imposta por ambos os canais de comunicação, são naturalmente subestimadas em cerca de 12%.

Concluiu-se também que as vozes femininas são mais afetadas (cerca de 15%) do que as vozes masculinas (cerca de 9%) pelos erros de estimação da formante F4. Adicionalmente, concluiu-se também que o desvio das formantes devido aos canais de comunicação pode exceder a normal variação intrafalante, o que sugere que o uso da

estimativa das frequências formantes nas tarefas de comparação de voz, num cenário forense, devem ser executadas com precaução e a sua análise deve ser combinada, de preferência, com parâmetros de voz alternativos.

Estas conclusões vão ao encontro dos resultados de outros estudos, designadamente o de Lal *et al.* [99], que demonstram que os telemóveis interferem consideravelmente com as propriedades acústicas dos sinais de fala, incluindo nas medidas das frequências formantes.

Capítulo 6 – Estudos com base no reconhecimento automático de orador

6.1 Introdução

Neste capítulo descrevemos estudos de reconhecimento de orador que realizámos através do *software* completamente automático de análise e comparação de oradores que já foi sucintamente descrito na Secção 4.4, o *iVocalise*. Este *software* possui uma arquitetura que será apresentada mais adiante nesta secção. Antes, porém, introduzimos aqui algumas noções enquadradoras que tentam descrever a evolução histórica que as abordagens totalmente automáticas de reconhecimento de orador revelaram ao longo do tempo.

A voz consiste num sinal biométrico que é útil para distinguir um orador do outro, pois reflete propriedades fisiológicas e comportamentais que representam a assinatura acústica, ou identidade de voz, de um indivíduo [104]. Muito especificamente, a assinatura acústica (ou sonora) de um orador resulta quer da sua idiossincrasia anatómica – tecidos das pregas vocais, volume da faringe e das cavidades oral e nasal, densidade dos tecidos do trato vocal – quer do gesto articulatório do orador – grau de nasalidade, forma da língua, dentes e lábios usados nos movimentos articulatórios, dicção –, que introduzem marcas temporais e espectrais no sinal de voz que são pessoais e, portanto, potencialmente identificadoras da identidade do falante [104]. A título de exemplo, as trajetórias das frequências de ressonância devidas ao trato vocal e ao trato nasal podem ser facilmente identificadas, por exemplo, em espectrogramas de ditongos. Os gestos coarticulatórios, com destaque para os produzidos em contexto CVC ou VCV, são também fonte de informação relevante para a identificação do orador [105]. Estas marcas espectrais constituem pistas muito informativas o que evidencia a relevância da informação no domínio das frequências.

É, portanto, sem surpresa que as marcas espectrais e, em menor grau, as marcas temporais, têm revelado um grande potencial na representação destes aspetos

idiossincráticos na produção de fala por um dado orador. Assim, é comum na maioria dos sistemas de reconhecimento/identificação de orador proceder-se à extração de segmentos do sinal de voz, com a duração entre cerca de 10 ms e 30 ms, o que corresponde à duração média de um fonema em conversação normal (sobretudo fonemas vozeados e alguns desvozeados porque as oclusivas, por exemplo, podem ter uma duração menor), com o objetivo de extrair características acústicas do sinal de voz que reflitam aquelas marcas temporais e espectrais idiossincráticas. Na maior parte dos sistemas, estas características são baseadas em vetores de coeficientes MFCC (*Mel-Frequency Cepstral Coefficients*) que descrevem a estrutura da magnitude espectral de cada segmento de voz, utilizando uma escala não-linear em frequência (a escala Mel) e que está adaptada à resolução espectral natural do sistema auditivo humano [111].

Devemos aqui clarificar que as características acústicas que investigamos no Capítulo 4, são sobretudo vocacionadas para estudos com base em procedimentos de análise não-automáticos porquanto exigem intervenção manual, mormente na segmentação supervisionada do sinal de voz em que aquelas características podem ser extraídas. No contexto deste capítulo, são relevantes as características que podem ser extraídas do sinal de voz de forma não supervisionada e, portanto, automática. As únicas cautelas a ter envolvem a remoção das regiões de silêncio e de ocorrência das vozes de segundos ou múltiplos oradores. Estes aspetos serão referidos na Secção 6.2.

Um vetor de coeficientes MFCC permite uma caracterização espectral do sinal de fala numa perspetiva estática, restrita ao segmento do sinal de voz em que esse vetor é extraído. No entanto, há informação muito útil de orador que decorre da dinâmica temporal das caracterizações espectrais, como referimos acima, concretamente em relação às trajetórias das frequências de ressonância. Para isso, os vetores de coeficientes MFCC são tipicamente aumentados com as diferenças entre vetores MFCC adjacentes, o que se usa para medir ‘velocidade’ na dinâmica temporal, e também com as diferenças entre vetores adjacentes de diferença MFCC (ou seja, são as segundas-diferenças), o que se usa para medir ‘aceleração’ na dinâmica temporal. As primeiras diferenças são tipicamente conhecidas por Delta-MFCC (ou Δ -MFCC) e as segundas diferenças são usualmente conhecidas por Delta-Delta-MFCC (ou $\Delta\Delta$ -MFCC).

Portanto, em praticamente todos os sistemas de reconhecimento/identificação automático/a de orador, existe uma primeira fase de análise de sinal, sucintamente referida de “extração de características” que envolve, além da segmentação, o cálculo os coeficientes MFCC, mas também, quando aplicável, dos coeficientes Δ -MFCC e $\Delta\Delta$ -MFCC, o que presumiremos neste capítulo. Se a dimensão do vector MFCC base for 13, como é típico, o vector completo de características MFCC tem dimensão 39.

No pressuposto de identificação de orador independente do texto (*i.e.* independente da fala), os vetores de características MFCC são usados para construir modelos de orador que consistem em modelos estatísticos de funções densidade de probabilidade. Um tipo de modelo que se tem revelado muito eficaz e popular é o modelo de misturas Gaussianas GMM (do inglês *Gaussian Mixture Model*). Os vetores MFCC extraídos do sinal de fala de um orador, são combinados para estimar a função densidade de probabilidade conjunta que os explica e que tipicamente é multimodal (*i.e.* exibe vários picos locais). Esta função é convertida numa representação paramétrica de várias (*i.e.* de uma mistura de) funções Gaussianas unimodais (posto que cada uma delas tem uma só forma de ‘sino’) e multidimensionais (posto que os vetores MFCC têm vários coeficientes). O algoritmo iterativo que permite obter esta representação paramétrica designa-se *expectation-maximization* (EM) e fornece, para cada orador, um modelo que especifica quer o número de funções Gaussianas unimodais quer, para cada uma destas (sendo o seu índice k), um fator que a pondera (w_k) na mistura, assim como o vetor de média dos coeficientes (μ_k), além de uma matriz de covariância dos coeficientes (Σ_k) que, por razões práticas é, na maior parte das vezes, tomada como uma matriz diagonal. O conjunto destes parâmetros consiste no modelo de orador.

As primeiras formas mais populares de modelização estatística de orador, de forma automática, basearam-se em modelização GMM de vetores de características MFCC. Neste pressuposto, a comparação entre as vozes de dois oradores era realizada de acordo com duas possibilidades. A primeira esquematiza-se na Figura 58 e envolve a criação de um modelo GMM de um orador conhecido.

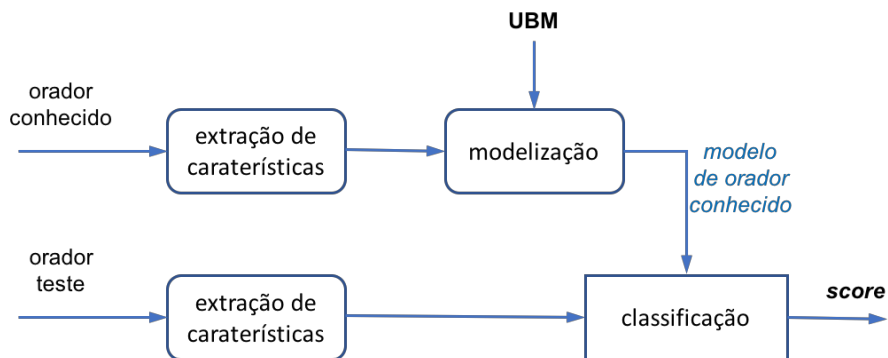


Figura 58 - Ilustração da comparação entre dois oradores em que o sinal de voz do orador de teste é classificado de acordo com um modelo estatístico de um orador conhecido.

Usando este modelo, os vetores MFCC do orador de teste são usados para se calcular uma medida (tipicamente logarítmica) da verosimilhança média (em inglês: *log-likelihood*) dos vetores de teste serem gerados pelo modelo estatístico correspondente ao modelo GMM do orador conhecido [109]. Esta operação é comumente conotada como uma operação de classificação já que o sinal de fala de um orador de teste é classificado de acordo com o modelo estatístico de um orador conhecido. Quanto maior for valor de *log-likelihood*, que aqui designamos simplesmente por *score*, maior é a confiança dos dois oradores serem o mesmo.

Esta primeira possibilidade de comparação de vozes é adequada a requisitos de funcionamento de tempo-real já que pode utilizar um modelo estatístico de orador pré-calculado. Uma outra possibilidade de comparação é computacionalmente mais exigente e ilustra-se na Figura 59.

Esta outra possibilidade de comparação é mais vocacionada para funcionamento *off-line* e envolve a obtenção, em primeiro lugar, dos modelos GMM dos dois oradores e a sua posterior comparação de que resulta um *score* de similaridade [109]. Também aqui, quanto maior for este, maior é a confiança dos dois oradores serem o mesmo.

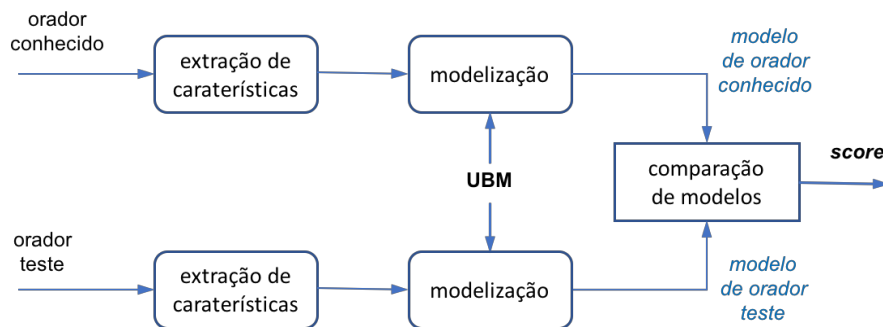


Figura 59 - Ilustração da comparação entre dois oradores em que são primeiro extraídos os seus modelos estatísticos sendo posteriormente comparados para se obter um score de similaridade.

As duas possibilidades aqui referidas são atreitas a interferências várias e que comprometem a sua correta operação e desempenho, designadamente: os efeitos do canal de comunicação, das condições acústicas de captação do sinal de voz ou, ainda, dos efeitos de variabilidade intra-orador. Estes efeitos podem ser de alguma forma acautelados e obviados, quer na fase de modelização, quer na de classificação, inserindo um modelo de referência conhecido por UBM (do inglês *Universal Background Model*) e que consiste no modelo médio de um orador indiferenciado. Este modelo médio é construído com base numa base de dados envolvendo gravações de fala de um elevado número de oradores. As operações matemáticas são assim realizadas numa perspetiva relativa em relação a este modelo médio de referência.

Novas abordagens no reconhecimento de orador emergiram por volta do ano 2000 e baseiam-se no conceito de super-vetores [54] [108]. Este conceito corporiza uma ideia e técnica conhecida em vários problemas de matemática e otimização e que se traduz por dois passos principais. Num primeiro passo, o espaço original de representação de um sinal ou modelo (no nosso caso, os parâmetros GMM do modelo de orador), é expandido para um outro espaço muito dilatado. No nosso caso, consiste em simplesmente concatenar para um único vetor todos os parâmetros de todas as Gaussianas extraídas de um orador. Assim, por exemplo, se a dimensão dos vetores MFCC for 39 e se forem usadas 256 Gaussianas, então a dimensão do super-vetor será 9984. Seguidamente, num segundo passo, este espaço dilatado de representação é

contraído para um outro espaço com uma dimensão entre a do espaço original e a do espaço dilatado, de modo que esta nova representação permita separar mais eficientemente os diferentes oradores (e, portanto, permita uma melhor discriminação) enquanto atenua ou mitiga as influências indesejáveis decorrentes do canal de comunicação, das condições acústicas de captação de voz e da variabilidade natural intra-orador.

As técnicas específicas usadas no âmbito do segundo passo atrás referido evoluíram durante a última década como se descreve a seguir. Um primeiro conjunto de técnicas implementa o princípio de *Factor Analysis* (FA) em que se assume que o espaço total dos super-vetores é criado como uma combinação envolvendo o super-vetor derivado do UBM e uma projeção a partir de dois sub-espacos de dimensão inferior à dos super-vetores, um exprimindo as variações devidas à sessão (*i.e.* aos efeitos do canal telefónico e da variação intra-orador) e outro exprimindo a variação devida aos diversos oradores da base de dados [112]. O primeiro sub-espaço permite identificar e subtrair as influências devidas ao canal telefónico e da variação intra-orador. O segundo sub-espaço é o que permite discriminar os oradores. A identificação dos dois subespaços consiste em identificar duas matrizes de ordem reduzida (em relação à dos super-vetores), e os vetores nestes espaços que explicam o sinal de fala de um orador, permitindo também identificar este [54]. Em particular, o cálculo das matrizes é realizado através de técnicas EM usando dados etiquetados de orador e equivale ao uso de técnicas de Análise em Componentes Principais através da escolha dos vetores próprios que estão associados a valores próprios mais elevados.

Foi posteriormente reconhecido que esta abordagem, apesar de ser computacionalmente muito exigente, não assegura realmente a desejada independência entre os sub-espacos de sessão e de orador [113]. Passou então a ser adotada uma nova abordagem na comunidade científica, desde cerca de 2010, de redução do espaço dos super-vetores, para um único espaço de ordem inferior, mas representando ambos o fatores de variabilidade: devido à sessão e devido ao orador. Este sub-espaço é caracterizado por uma única matriz que exprime esta variabilidade total (TV) e que tem a vantagem de poder ser calculada sem necessidade de dados etiquetados, isto é, pode

ser calculada *off-line* e disponibilizada na forma pré-calculada. O cálculo da matriz TV pode também ser considerado equivalente ao uso de técnicas de Análise em Componentes Principais [54].

Deste modo, um super-vetor é encarado como resultando de uma combinação envolvendo o super-vetor derivado do UBM e a projeção, a partir do sub-espço projetado através da matriz TV, de um vetor intermédio, ou de identidade, abreviadamente representado de *i-vector*, e que identifica o orador (apesar de refletir também influências do canal de comunicação e da variabilidade intra-orador). Tipicamente, um *i-vector* tem uma dimensão entre 400 e 800 [54].

Na literatura, considera-se que o *i-vector* é uma nova forma de característica do orador, apesar do seu significado ser abstrato (*i.e.* não tem por exemplo a interpretação espectral dos coeficientes MFCC), é passível de compensação para atenuar as influências de canal, e pode ser usado para calcular projeções ou distâncias entre oradores [54] [108].

Uma possibilidade para comparar dois oradores através dos seus *i-vectors* é usar uma técnica conhecida por SVM (do inglês *Support Vector Machine*) que permite classificação binária através de vetores de suporte que ajudam a demarcar a separação entre os oradores contando também com a contribuição devido ao super-vetor representativo do UBM [54].

Outra possibilidade, que é usada no ambiente *iVocalise*, consiste em realizar uma nova projeção do espaço agora representado pelos *i-vectors*, para um outro, de dimensão ainda inferior, procurando-se neste processo maximizar a variabilidade inter-oradores (*i.e.* maximizar a capacidade de discriminação) e minimizar a influência da variabilidade intra-orador e dos efeitos do canal de comunicação. Esta técnica é conhecida por LDA (do inglês *Linear Discriminant Analysis*) e consiste numa técnica de redução de dimensionalidade e com semelhanças também com a conhecida técnica PCA atrás referida.

Após a aplicação da técnica LDA aos *i-vectors*, resulta um identificador que é específico para cada orador. A comparação entre oradores pode agora ser realizada ou calculando a distância entre estes identificadores, ou aplicando uma técnica conhecida

por *Probabilistic Linear Discriminant Analysis* e que não envolve redução na dimensão dos vetores. Ambas as possibilidades estão previstas no *iVocalise* como se ilustra através do seu esquema funcional ilustrado na Figura 60.

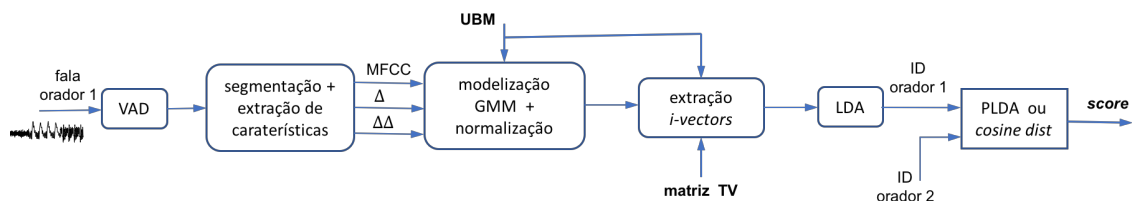


Figura 60 - Diagrama de blocos da estrutura e funcionamento do software *iVocalise*.

A primeira alternativa (distância de cosseno) consiste numa projeção entre os vetores que identificam os dois oradores em comparação. A segunda alternativa segue uma abordagem probabilística para avaliação da verosimilhança dos dois vetores de identidade terem sido gerados pelo mesmo orador [114]. Como se referirá na Secção 6.3, comprovou-se que esta última produz melhor desempenho pelo que é a que se presume na continuação deste capítulo.

6.2 Pré-condicionamento dos dados

A qualidade, fiabilidade e robustez de um modelo de orador depende da qualidade dos dados usados na sua construção. Em concreto, isto significa que os dados, isto é, os registos de voz, devem conter informação só relativa ao orador pretendido. Deste modo, as regiões de sinal contendo silêncio, ruído ambiente e vozes de outros oradores ou outros sons, devem ser removidas. No caso da nossa base de dados, são críticas só as regiões de silêncio e do interlocutor, isto é, do segundo orador.

Como descrito e ilustrado na Secção 3.6.3, a remoção das regiões de sinal contendo a voz do segundo orador, isto é, da entrevistadora, foi realizado em todos os registos de voz telefónica em que tal se revelou necessário. Este trabalho, totalmente manual e muito moroso, foi muito útil, em particular, para os resultados apresentados

em [52] e que constituem a base de partida para a investigação relatada na presente dissertação, tal como se refere na Secção 6.3.

No caso das regiões de silêncio, a sua remoção é automática e realizada pelo próprio ambiente *iVocalise* recorrendo a um módulo de pré-processamento de sinal designado VAD (do inglês *Voice Activity Detector*) que efetua a deteção das regiões de voz ativa, tal como se sugere na Figura 61.

A Figura 61 ilustra um segmento de voz gravada correspondendo à última região de sinal ilustrada na Figura 15. Usando um critério de deteção de voz ativa baseado na energia do sinal, é possível demarcar sub-regiões que correspondem ao sinal de interesse para análise (como se assinala a cor magenta no painel superior da Figura 61) e descartar as restantes correspondendo essencialmente a silêncio ou ruído de baixa intensidade. Após descartar estas sub-regiões, o sinal remanescente é o de voz ativa sendo a sua duração uma fração da do sinal original, tal como se ilustra no painel inferior da Figura 61. Tal como se refere em [109], a utilização do VAD permite descartar cerca de 25% do sinal gravado de voz telefónica. Como se comprova em [52], a utilização do pré-processamento VAD origina uma melhoria significativa do desempenho em tarefas de comparação de vozes, pelo que a sua utilização se torna não só recomendável, como obrigatória.

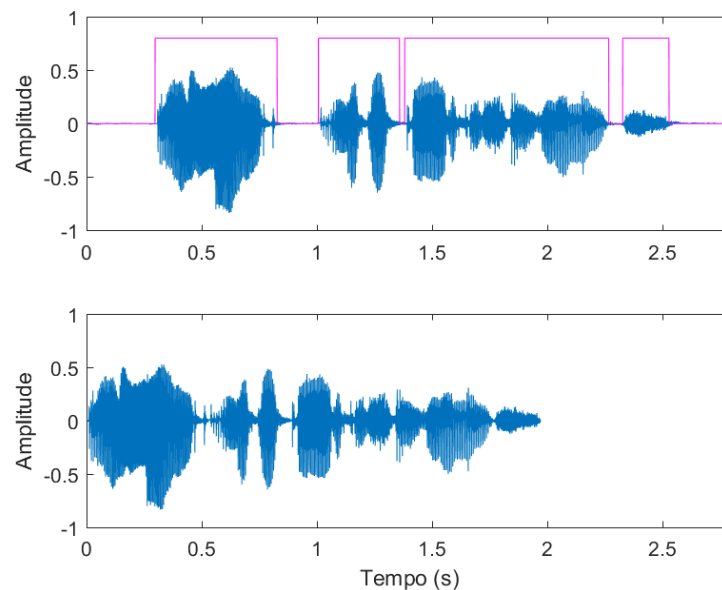


Figura 61 - Ilustração da operação do pré-processamento VAD. O painel superior representa um sinal de voz em que as regiões de voz ativa são assinaladas pelo detetor VAD, neste caso através de linhas de delimitação de cor magenta. Por último, o processador VAD remove as regiões de silêncio e ruído de fundo de baixa intensidade, preservando somente as de voz ativa, tal como se ilustra no painel inferior.

Convém, contudo, assinalar que, de um modo geral, a utilização do pré-processamento VAD consiste numa segmentação agnóstica pois não considera a natureza fonética de cada sub-região do sinal de voz ativa. Trabalho anterior [109] [115] revela, contudo, que se a segmentação do sinal de voz for realizada de forma orientada a algumas classes fonéticas relevantes, consegue-se uma modelização e classificação mais eficiente de sinal de voz.

6.3 Resultados previamente consolidados

O trabalho de investigação realizado no âmbito desta dissertação e com base no *software iVocalise* beneficiou de uma colaboração com um trabalho de dissertação de mestrado integrado em Engenharia Eletrotécnica e Computadores e desenvolvido

durante o segundo semestre do ano letivo 2016-2017 [52]. Este trabalho faz uso da base de dados constituída no âmbito da presente dissertação, como se descreve no Capítulo 3. A colaboração consistiu numa familiarização conjunta com o *software iVocalise* e consistiu na partilha do trabalho manual de remoção do segundo orador em todas as gravações de voz telefónica, como reportado na Secção 3.6.3.

Na presente Secção, descrevemos os principais resultados de investigação alcançados por aquela dissertação de mestrado e que tomamos como ponto de partida para os novos resultados que apresentaremos nas secções seguintes.

Um primeiro conjunto de resultados muito relevante refere-se ao impacto no desempenho automático de orador em consequência da frequência de amostragem do registo de voz, em resultado da remoção, ou não, do segundo orador, e em resultado da utilização, ou não, do pré-condicionamento VAD a que aludimos na secção anterior.

Por um lado, estudou-se o impacto de reduzir a frequência de amostragem original de todas as gravações na nossa base de dados, de 22050 Hz, para a frequência de amostragem de voz telefónica, de 8000 Hz, que é a que é recomendada na utilização do *software iVocalise*. Verificou-se que para as várias combinações das outras condições (remoção ou não do segundo orador e inclusão ou não de VAD), o impacto era negligenciável. Não sendo um resultado esperado, já que a expectativa era de obter melhor desempenho quando a representação de sinal acomoda uma maior largura de banda (o que é a circunstância da frequência de amostragem de 22050 Hz), concluiu-se posteriormente que o *software iVocalise* realiza internamente uma conversão da frequência de amostragem se a fornecida não for 8000 Hz. Compreende-se, deste modo, que o desempenho resulte essencialmente inalterado.

Por outro lado, considerando todos os registos de todos os 37 oradores que compreendem a nossa base de dados, estudou-se o desempenho na classificação de orador, através do parâmetro EER discutido na Secção 4.5, quando se consideram, ou não, os outros dois fatores: pré-processamento VAD e remoção do segundo orador. Os resultados encontram-se ilustrados na Figura 62.

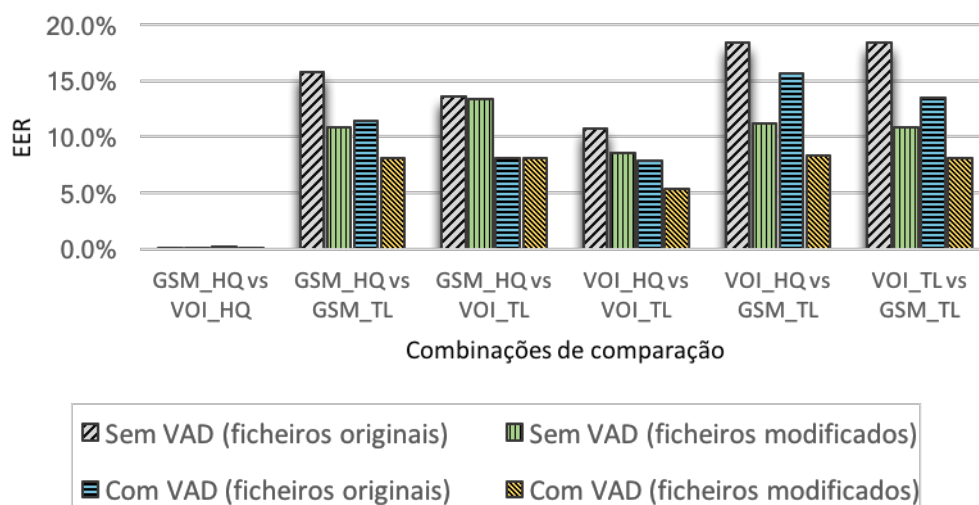


Figura 62 - Resultados de desempenho na identificação de orador, para cada uma das seis situações de teste, quando se inclui, ou não, pré-processamento VAD, e quando se considera, ou não, a remoção do segundo orador (situação em que os ficheiros são modificados).

Esta figura representa resultados individualizados para cada uma das seis combinações de comparação entre tipos de registos: GSM_HQ *versus* VOI_HQ, GSM_HQ *versus* GSM_TL, GSM_HQ *versus* VOI_TL, VOI_HQ *versus* VOI_TL, VOI_HQ *versus* GSM_TL e, finalmente, VOI_TL *versus* GSM_TL. Estas alternativas correspondem às seis células que figuram ou acima da diagonal principal da matriz representada na Figura 17, ou, equivalentemente, abaixo dessa diagonal.

Excetuando a combinação GSM_HQ *versus* VOI_HQ em que os desempenhos são próximo do ideal para qualquer situação, as restantes combinações permitem extrair as seguintes conclusões genéricas:

- a remoção do segundo orador, o que se assinala na Figura 62 através de ‘ficheiros modificados’, é sempre benéfica, podendo a melhoria no desempenho EER, em termos absolutos, ultrapassar mesmo os 5%,
- a inclusão do pré-processamento VAD é sempre benéfica podendo a melhoria no desempenho EER, em termos absolutos, atingir até cerca de 5%.

Esta tendência verificou-se também quando nas comparações se consideram só os registos de oradores femininos, ou quando se considera toda a base de dados, com ambos os géneros.

Assim, apresentam-se na Tabela 13 os melhores resultados alcançados quando a frequência de amostragem é 8000 Hz, quando se usa pré-processamento VAD e quando o segundo orador é removido de todas as gravações. Para além de se indicarem os resultados para cada uma das seis combinações de comparação acima referidas, discriminam-se também os resultados quando se consideram só oradores masculinos, quando se consideram oradores femininos, ou quando se consideram ambos.

Tabela 13 - Melhores resultados de desempenho EER alcançados em [52] para cada uma das combinações de teste e para cada cenário de género de oradores: FEM, MAL e FEM & MAL.

COMBINAÇÕES	FEM & MAL	FEM	MAL
GSM_HQ vs VOI_HQ	0,11%	0,00%	0,00%
GSM_HQ vs GSM_TL	8,11%	10,13%	5,88%
GSM_HQ vs VOI_TL	8,07%	1,18%	0,74%
VOI_HQ vs VOI_TL	5,37%	0,92%	1,10%
VOI_HQ vs GSM_TL	8,33%	15,00%	7,35%
VOI_TL vs GSM_TL	8,15%	11,05%	11,76%

Nesta tabela, os resultados indicados para FEM&MAL são os mesmo assinalados na Figura 62 com as barras com valores de EER mais baixos (e representadas a amarelo).

Deste modo, a Tabela 13 representa os resultados de referência para os novos resultados de investigação a apresentar nas secções seguintes. Deve-se assinalar que por indicação da empresa que produziu o *software iVocalise*, valores de EER até cerca de 6% são considerados aceitáveis e normais para dados reais de algumas dezenas de

oradores diferentes, enquanto que valores de ERR acima deste limiar são indicativos de que pode haver margem, por exemplo ao nível de condicionamento adicional dos registos de voz, para se melhorar o desempenho.

É oportuno retirarem-se algumas conclusões da Tabela 13 que são relevantes para os novos testes a apresentar nas seções seguintes. Em primeiro lugar, é notório que os resultados são, em média, melhores no caso dos testes realizados com os oradores MAL do que os realizados com os oradores FEM. Isto poderá ser uma consequência da maior dificuldade de modelização dos oradores FEM atendendo, nomeadamente, a que a sua frequência fundamental é tipicamente uma oitava mais elevada do que a dos oradores MAL, o que provoca maior indefinição na captação das frequências formantes.

Por outro lado, há dois conjuntos de resultados que causam alguma perplexidade. Um conjunto envolve as combinações de teste GSM_HQ *versus* VOI_TL e VOI_HQ *versus* VOI_TL. De facto, nestes casos e para o cenário de oradores FEM & MAL, o desempenho EER é significativamente pior do que o desempenho separadamente obtido para FEM ou MAL. O facto destas combinações implicarem registos VOI_TL poderá sugerir que há uma dificuldade particular na modelização deste tipo de registos, o que merece esclarecimento adicional. Uma perplexidade em sentido oposto decorre da combinação de teste VOI_TL *versus* GSM_TL. De facto, neste caso, o desempenho EER é significativamente melhor do que desempenho separadamente obtido para FEM ou MAL, o que é intrigante porque corresponde à circunstância de teste mais exigente porquanto estão implicados registos telefónicos em que a degradação de sinal (por exemplo, devidos a distorções e interferência várias) é maior. É algo que se procurará também esclarecer.

6.4 Segmentação e recomposição da base de dados

A segmentação de todas as gravações constantes da nossa base de dados foi executada como descrevemos na Secção 3.6.4. Como aí se refere, a sua necessidade decorre de a organização inicial da base de dados não permitir combinações de registos de voz mutuamente exclusivos (*i.e.* que não contenham o mesmo sinal de fala), o que é

especialmente verdade para as comparações segundo a diagonal principal da Figura 17, em que os registos em comparação e para o mesmo orador são exatamente os mesmos, e o que é também verdade para as combinações assinaladas na Figura 17 a sombreado ligeiro, em que os registos em comparação e para o mesmo orador provêm exatamente da mesma conversação, sendo um deles simplesmente a versão modificada pelo canal telefónico.

Como se descreve na Secção 3.6.4, a segmentação consistiu em subdividir cada gravação inicial em duas metades, mantendo-se em cada uma delas a mesma estrutura de exercício de voz, isto é, uma parte de diálogo e uma parte de leitura. A oportunidade para esta subdivisão decorre da possibilidade de cada metade poder ter, ainda assim, uma duração na ordem ou superior a cerca 20 segundos, que é a duração mínima para que seja possível construir um modelo de orador.

Deste modo e tal como se ilustra na Figura 19, organizando por exemplo os registos da primeira metade segundo a vertical e organizando os registos da segunda metade segundo a horizontal, assegura-se sempre mútua exclusividade dos registos de voz em comparação, mesmo quando pertencem ao mesmo orador. Este é o enquadramento dos resultados que se apresentam na Secção 6.6. Na Secção 6.5 avalia-se o impacto, em relação aos resultados de referência da Tabela 13, de realizar as mesmas comparações que esta traduz, mas na situação em que os registos de voz têm metade da duração.

Por último, observando-se ainda o requisito de mútua exclusividade entre os registos de voz em comparação, pretendeu-se estudar o desempenho na identificação de orador quando os seus modelos são construídos recompondo os registos de voz através da junção, para cada orador, de várias metades dos registos originais de forma a refletir diferentes condições de canal. Esta lógica de recomposição dos registos foi já apresentada na Secção 3.6.4, concretamente através da Figura 20. Os resultados destes estudos apresentam-se nas Secções 6.7 e 6.8 e permitem concluir sobre o impacto da diversidade de condições de canal na construção do modelo de orador.

6.5 Testes 1h-1h + 2h-2h

Nesta secção apresentam-se os resultados no desempenho na identificação de orador, avaliado através do parâmetro EER, em resultado da comparação, através do *software iVocalise*, entre os registos que consistem nas primeiras metades após segmentação, isto é, dos registos com etiqueta ‘1h’, e também, separadamente, entre os registos que consistem nas segundas metades após segmentação, isto é, dos registos com etiqueta ‘2h’. Nesta circunstância, os dados não são mutuamente exclusivos e está-se, por isso, a replicar a circunstância que conduziu aos resultados de referência, apresentados na Secção 6.3. Contudo, isto é deliberado para se aferir a degradação que ocorre quando se comparam registos que têm metade da duração dos registos originais. Dado que as comparações ‘1h-1h’ e ‘2h-2h’ produzem resultados semelhantes para as mesmas condições de teste, apresenta-se só a média destes resultados.

A Tabela 14 permite comparar, para os 20 oradores do género feminino, os resultados de referência relativos aos registos com a duração original e já apresentados na Tabela 13, com a média dos resultados obtidos quando se comparam as primeiras metades e, separadamente, quando se comparam as segundas metades.

Tabela 14 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos dos (20) oradores do género FEM. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos com metade da duração.

FEM	GSM_HQ	GSM_TL	VOI_HQ	FEM	GSM_HQ	GSM_TL	VOI_HQ
GSM_TL	10,13			GSM_TL	10,46		
VOI_HQ	0,00	15,00		VOI_HQ	0,13	17,57	
VOI_TL	1,18	11,05	0,92	VOI_TL	4,54	16,91	0,53

Exceto para a combinação de comparação VOI_TL *versus* VOI_HQ, verifica-se que há sempre uma ligeira degradação quando se comparam registos que têm metade da duração. Destacam-se, contudo, as degradações verificadas para a combinação VOI_TL

versus GSM_HQ (em que o EER aumenta de 1.18% para 4.54%), e para a combinação VOI_TL *versus* GSM_TL (em que o EER aumenta de 11.05% para 16.91%). Não é clara a razão para esta tendência, observa-se contudo que os registos VOI_TL estão implicados nesta constatação, tal como já se havia notado na Secção 6.3 em relação a resultados que aí causaram alguma estranheza.

A Tabela 15 apresenta o mesmo tipo de resultados dos apresentados na tabela anterior, desta vez, porém, para os 17 oradores do género masculino.

Tabela 15 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos dos (17) oradores do género MAL. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos com metade da duração.

MAL	GSM_HQ	GSM_TL	VOI_HQ	MAL	GSM_HQ	GSM_TL	VOI_HQ
GSM_TL	5,88			GSM_TL	6,07		
VOI_HQ	0,00	7,35		VOI_HQ	0,10	8,73	
VOI_TL	0,74	11,76	1,10	VOI_TL	5,61	17,28	0,83

Verifica-se, aqui também, o mesmo tipo de conclusões que atrás extraímos para a tabela anterior. Concretamente, para além da exceção associada à combinação VOI_TL *versus* VOI_HQ, em todos os restantes casos verifica-se uma degradação no valor de EER, sendo especialmente expressiva para a combinação VOI_TL *versus* GSM_HQ (em que o EER aumenta de 0.74% para 5.61%), e para a combinação VOI_TL *versus* GSM_TL (em que o EER aumenta de 11.76% para 17.28%). Mais uma vez, parece confirmar-se que os registos VOI_TL são particularmente problemáticos.

A Tabela 16 apresenta o mesmo tipo de resultados dos apresentados nas duas tabelas anteriores, desta vez para o total de 37 oradores abrangendo ambos os géneros.

Tabela 16 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos de todos os (37) oradores dos dois géneros. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos com metade da duração.

FEM & MAL	GSM_HQ	GSM_TL	VOI_HQ	FEM & MAL	GSM_HQ	GSM_TL	VOI_HQ
GSM_TL	8,11			GSM_TL	6,99		
VOI_HQ	0,11	8,33		VOI_HQ	0,38	9,74	
VOI_TL	8,07	8,15	5,37	VOI_TL	8,28	13,51	5,22

Conclui-se que a comparação de registos com uma duração de metade da duração original acarreta alguma degradação no desempenho EER, exceto para os casos das combinações de comparação GSM_TL *versus* GSM_HQ e VOI_TL *versus* VOI_HQ, embora seja pouco expressiva. Os resultados suportam assim conclusão de que, neste caso em que o número de oradores é mais elevado, a penalização devida à redução da duração dos registos é pouco significativa, o que significa que a modelização de orador não é significativamente afetada.

6.6 Testes 1h-2h

Nesta secção apresentam-se resultados das comparações ‘1h-2h’ e ‘2h-1h’ em que os registos comparados, especialmente para o mesmo orador, são sempre mutuamente exclusivos do ponto de vista da fala neles contida. Por outras palavras e como referido na Secção 6.4, as comparações são efetuadas com base numa combinação dos registos de voz tal como se ilustra na Figura 19. As seis células na matriz/tabela desta figura e correspondentes às combinações acima da sua diagonal principal, podem ser vistas como comparações de ficheiros ‘1h-2h’. As seis células correspondentes às combinações abaixo da sua diagonal principal, podem ser vistas como comparações de ficheiros ‘2h-1h’. As combinações homólogas nestes dois casos (uma combinação homóloga é, por exemplo, a comparação GSM_HQ_1h *versus* GSM_TL_2h em relação à comparação GSM_HQ_2h *versus* GSM_TL_1h), apesar de envolverem registos de voz diferentes, numericamente são muito semelhantes (com exceção de alguns casos muito pontuais) pelo que nesta secção apresenta-se só a sua média para facilitar

comparação com os resultados de referência (Secção 6.3) obtidos com os ficheiros de duração original (*i.e.* não segmentados).

Por outro lado, obviamente que por envolver ficheiros mutuamente exclusivos, são também relevantes os resultados decorrentes das comparações expressas sobre a diagonal principal da matriz/tabela da Figura 19. Como, porém, correspondem a comparações que exprimem as mesmas condições de canal de comunicação, os valores de EER são, como seria de esperar, muito baixos e inferiores a 1%. Indicamos, por isso, somente a sua média. Assim, no caso dos testes de comparação envolvendo registos de oradores femininos, a média é EER=0.35%, no caso de oradores masculinos a média é EER=0.14% e, no caso dos testes de comparação envolvendo registos dos 37 oradores de ambos os géneros, a média é EER=0.20%.

A Tabela 17 apresenta, à sua direita, os resultados médios decorrentes das várias combinações homólogas de comparação com os registos de voz dos 20 oradores FEM da nossa base de dados. Repete-se também a apresentação, à sua esquerda e para facilitar comparação, dos resultados de referência obtidos com os registos de duração original e não mutuamente exclusivos (Tabela 13).

Tabela 17 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos dos (20) oradores do género FEM. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos mutuamente exclusivos e com metade da duração original.

FEM	GSM_HQ	GSM_TL	VOI_HQ	FEM	GSM_HQ	GSM_TL	VOI_HQ
GSM_TL	10,13			GSM_TL	20,07		
VOI_HQ	0,00	15,00		VOI_HQ	0,53	17,83	
VOI_TL	1,18	11,05	0,92	VOI_TL	5,97	16,84	7,17

Conclui-se que todas as combinações de comparação sofrem uma degradação no desempenho avaliado pelo parâmetro EER, muito especialmente no caso da combinação GSM_HQ *versus* GSM_TL em que o EER é elevado e cerca de duas vezes pior do que

que no caso das comparações com ficheiros não mutuamente exclusivos e de duração original.

Quando os resultados da Tabela 17 são comparados com os da Tabela 14, conclui-se que a degradação é especialmente notória nas combinações de comparação GSM_HQ *versus* GSM_TL e VOI_HQ *versus* VOI_TL, o que é perfeitamente compreensível porque é nestas combinações que o impacto da mútua exclusividade dos registos se faz notar (isto é, para estas combinações os registos implicados nos resultados da Tabela 14 não são mutuamente exclusivos enquanto que os registos implicados nos resultados da Tabela 17 são mutuamente exclusivos). Em todo o caso, os piores resultados não descolam dramaticamente dos obtidos nas comparações envolvendo ficheiros não mutuamente exclusivos e com a duração original.

Tal como no caso anterior, a Tabela 18 apresenta, à direita, os resultados médios decorrentes das várias combinações homólogas de comparação com registos dos 17 oradores MAL da nossa base de dados. Também se repete a apresentação, à esquerda e para facilitar comparação, dos resultados de referência obtidos com os ficheiros de duração original e não mutuamente exclusivos (Tabela 13).

Tabela 18 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos dos (17) oradores do género MAL. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos de registos mutuamente exclusivos e com metade da duração original.

MAL				MAL			
	GSM_HQ	GSM_TL	VOI_HQ		GSM_HQ	GSM_TL	VOI_HQ
GSM_TL	5,88			GSM_TL	9,11		
VOI_HQ	0,00	7,35		VOI_HQ	0,47	11,76	
VOI_TL	0,74	11,76	1,10	VOI_TL	3,43	17,56	3,40

Também se conclui neste caso que, como seria de esperar, existe um agravamento nos valores de EER em relação aos resultados das comparações efetuadas com registos não mutuamente exclusivos e de duração original, porém, o agravamento não é tão acentuado quanto o que foi observado nos testes com oradores FEM. Por outro

lado, comparando os resultados da Tabela 15 com os da Tabela 18, constata-se que existe um maior agravamento para as combinações de comparação alinhadas segundo a diagonal principal da tabela, o que se deve, como se viu antes, à circunstância de mútua exclusividade dos ficheiros comparados, porém há um caso que suscita curiosidade. Trata-se da combinação de comparação GSM_HQ *versus* VOI_TL em que, aparentemente, há uma melhoria no desempenho EER. Esta circunstância é algo estranha porque, quer no resultado da Tabela 15, quer no da Tabela 18, os ficheiros comparados são sempre mutuamente exclusivos dado que decorrem de conversações diferentes. Uma possível explicação pode consistir no facto do valor $EER=3.43\%$ decorrer da média de valores díspares: $EER=0.98\%$ e $EER=5.88\%$ sendo que um destes casos envolve um resultado favorável que poderá ficar a dever-se a uma especial, mas casual, consistência entre os registos envolvidos.

Finalmente, a Tabela 19 apresenta os resultados para o conjunto total de 37 oradores de ambos os géneros. Tal como anteriormente, repetem-se à esquerda e para facilitar comparação, os resultados de referência obtidos com os ficheiros de duração original e não mutuamente exclusivos (Tabela 13) e, à direita, apresentam-se os resultados médios decorrentes das várias combinações homólogas de comparação dos registos de voz.

Em relação aos resultados não segmentados e não mutuamente exclusivos (tabela à esquerda na Figura 19), confirma-se que o desempenho EER, quando são comparados registos de voz mutuamente exclusivos e com metade da duração, vem agravado, apesar da diferença não ser muito expressiva. Em relação aos resultados da Figura 16, em que os registos de voz comparados não são todos mutuamente exclusivos, os resultados da Figura 19, em que os registos de voz comparados são todos mutuamente exclusivos, revelam um agravamento do desempenho na identificação de orador (*i.e.* os valores de EER são mais elevados), o que não constitui surpresa. No entanto, o agravamento é relativamente modesto, o que parece confirmar, tal como se havia já concluído em relação aos resultados da Figura 16, que um maior e mais diversificado número de oradores favorece uma mais estável capacidade de discriminação por parte do *software iVocalise*.

Tabela 19 - Resultados de desempenho EER (em %) obtidos para cada uma das combinações de comparação de registos de todos os (37) oradores dos dois géneros. A tabela à esquerda representa os resultados de referência já apresentados na Tabela 13 para os registos com a duração original. A tabela à direita representa a média dos resultados obtidos para os dois conjuntos homólogos de registos de voz mutuamente exclusivos e com metade da duração original.

FEM & MAL	GSM_HQ	GSM_TL	VOI_HQ	FEM & MAL	GSM_HQ	GSM_TL	VOI_HQ
GSM_TL	8,11			GSM_TL	9,98		
VOI_HQ	0,11	8,33		VOI_HQ	1,17	10,94	
VOI_TL	8,07	8,15	5,37	VOI_TL	8,56	12,01	8,11

Note-se, por último, que os resultados sugerem também que flutuações na ordem de 1% ou 2% no EER são normais para o mesmo cenário de comparação de oradores, tudo dependendo dos dados concretos, *i.e.* dos registos de voz concretos, que são utilizados.

6.7 Testes usando diversidade de canal

Os testes descritos nas Secções 6.5 e 6.6 permitem esclarecer, em especial, sobre o impacto que dois fatores importantes exercem no desempenho da identificação de orador através do *software iVocalise*: a duração dos registos de voz e a condição de mútua exclusividade entre quaisquer dois registos de voz que sejam comparados. Permitem também caracterizar o impacto que o tipo de canal de comunicação exerce naquele desempenho. Porém, todos os cenários de comparação até agora considerados presumiram que o modelo de orador que é construído com base em cada registo de voz decorre de uma única condição de canal de comunicação. Isto é, cada registo exprime a voz de um dado orador que é captada numa única condição de canal: ou alta qualidade (HQ) ou telefónica de canal GSM (GSM_TL) ou telefónica de canal VoIP (VOI_TL).

Nesta secção, pretende-se estudar a consequência de haver, num dos registos de voz que são comparados, ou em ambos, o sinal de um dado orador mas captado em duas condições de canal. A segmentação de sinal que realizámos em todos os ficheiros originais, tal como descrevemos na Secção 3.6.4 e na Secção 6.4, permite um tal estudo observando também o requisito de mútua exclusividade entre quaisquer dois registos comparados.

Neste sentido, criámos uma nova versão da base de dados em que, para cada orador, concatenamos duas metades dos registos segmentados, mas exprimindo duas condições de canal diferentes. As várias possibilidades que respeitam este requisito, além do objetivo de mútua exclusividade dos registos (quanto ao conteúdo da fala) encontram-se assinaladas na Figura 63. Em concreto, são excluídas combinações que associam dois registos ‘GSM’ ou dois registos ‘VOI’. São também excluídas combinações que associam dois registos ‘HQ’. Estas exclusões conduzem às seis possibilidades que se assinalam na Figura 63, através do símbolo ‘X’, e correspondem às seis novas bases de dados que foram já identificadas na Secção 3.6.4, na Figura 20.

	GSM_HQ_2h	GSM_TL_2h	VOI_HQ_2h	VOI_TL_2h
GSM_HQ_1h				X
GSM_TL_1h			X	X
VOI_HQ_1h		X		
VOI_TL_1h	X	X		

Figura 63 - Alternativas de combinação de duas metades dos registos de voz que representam tipos de canal diferentes e respeitam o objetivo de mútua exclusividade dos registos voz nos estudos de comparação. Quaisquer duas combinações assinaladas com ‘X’ são viáveis, exceto as seis associações assinaladas através de grafismo oval.

O número total de comparações suscitadas para estas seis possibilidades de combinação é $\binom{6}{2} = \frac{6!}{2!4!} = 15$. Porém, quaisquer comparações de duas combinações adjacentes na vertical, ou na horizontal, ou na diagonal, violariam o requisito de mútua

exclusividade dos registos comparados pelo que são também excluídas. Estas comparações excluídas, no total de 6, identificam-se na Figura G através de grafismo oval. Restam, assim, 9 possibilidades de comparação. Dado que pretendemos efetuar estudos separadamente para oradores FEM, MAL e de ambos os géneros, isto implicaria a realização de 27 conjuntos testes, o que é um número excessivo.

Para tornar o nosso trabalho de investigação mais gerível, optámos por seleccionar cenários de comparação que representam simetria nas combinações que são comparadas através do *software iVocalise*. Estes cenários, no total de 3, correspondem às combinações de 'X' que se situam em posições simétricas em relação à diagonal principal da matriz ilustrada na Figura 63. Usando a numeração das seis bases de dados identificadas na Secção 3.6.4, na Figura 20, isto significa comparar os registos da base de dados 1 com os da base de dados 2, os da base de dados 3 com os da base de dados 4, e os da base de dados 5 com os da base de dados 6. De seguida, apresentamos esses resultados.

A Tabela 20 apresenta os resultados do desempenho na identificação de orador quando são confrontadas as bases de dados 1 e 2. Consideram-se os três casos de género de orador (FEM, MAL e FEM & MAL). Em cada caso e para facilitar a discussão, é ainda apresentada uma matriz, com os resultados anteriormente obtidos com as bases de dados com uma só condição de canal. Deve salientar-se que a base de dados 1 (BD 1) concatena os registos indicados na vertical (*i.e.* segundo as linhas) da matriz de resultados, enquanto que a BD 2 concatena os registos indicados na horizontal (*i.e.* segundo as colunas) da matriz de resultados.

Estes resultados constituíram uma verdadeira surpresa no sentido em que, para os três casos, o resultado é o melhor possível: EER=0%. Por outras palavras, para estas condições de teste, não há erros na identificação de orador. Há, pelo menos, duas hipóteses plausíveis que podem ajudar a explicar os valores nulos de EER obtidos. Por um lado, como todos os registos incluem uma parte em que o canal é HQ, então a modelização de orador é mais eficaz e alavanca os melhores resultados parciais em que os registos comparados têm este mesmo tipo de canal, concretamente, EER=0.63%, EER=0.00% e EER=0.19%. Uma outra hipótese é que como ambas as bases de dados 1

e 2 contêm duas condições de canal para cada orador, então o processo de modelização de orador é mais informado e consegue efetuar uma discriminação mais eficiente entre oradores porque a eliminação das influências de canal é mais eficaz. Por outras palavras, a diversidade de condições de canal poderá ser uma condição necessária para uma correta modelização e eficiente identificação de orador.

Tabela 20 - Resultados de desempenho EER (em %) obtidos com base na comparação das bases de dados 1 e 2 em que os registos de voz incluem duas condições de canal e são mutuamente exclusivos quanto ao conteúdo da fala. Apresentam-se os resultados obtidos para os oradores do género MAL, FEM e junção dos dois. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita os valores previamente obtidos na comparação das bases de dados com uma só condição de canal.

FEM				
	BD 2		GSM_HQ_2h	VOI_TL_1h
BD 1	0,00	GSM_HQ_1h	0,63	4,34
		VOI_TL_2h	4,74	0,39

MAL				
	BD 2		GSM_HQ_2h	VOI_TL_1h
BD 1	0,00	GSM_HQ_1h	0,00	6,25
		VOI_TL_2h	4,96	0,55

FEM & MAL				
	BD 2		GSM_HQ_2h	VOI_TL_1h
BD 1	0,00	GSM_HQ_1h	0,19	8,91
		VOI_TL_2h	7,66	0,49

A Tabela 21 repete as mesmas circunstâncias de teste anteriormente consideradas mas, desta vez, para as bases de dados 3 e 4.

Também neste caso se conclui que ou a diversidade de canal ajudou a uma mais eficaz remoção da influência do canal na modelização de orador e, por conseguinte, a uma mais eficiente identificação de orador ou, então, a circunstância de haver registos

com condição de canal HQ em cada um dos novos registos das duas bases de dados, permite alavancar os melhores resultados obtidos nas comparações de registos com uma só condição de canal.

Tabela 21 - Resultados de desempenho EER (em %) obtidos com base na comparação das bases de dados 3 e 4 em que os registos de voz incluem duas condições de canal e são mutuamente exclusivos quanto ao conteúdo da fala. Apresentam-se os resultados obtidos para os oradores do género MAL, FEM e junção dos dois. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita os valores previamente obtidos na comparação das bases de dados com uma só condição de canal.

FEM				
	BD 4		VOI_HQ_2h	GSM_TL_1h
BD 3	0,13	VOI_HQ_1h	0,26	15,00
		GSM_TL_2h	20,13	0,13

MAL				
	BD 4		VOI_HQ_2h	GSM_TL_1h
BD 3	0,00	VOI_HQ_1h	0,00	5,88
		GSM_TL_2h	11,58	0,00

FEM & MAL				
	BD 4		VOI_HQ_2h	GSM_TL_1h
BD 3	0,08	VOI_HQ_1h	0,08	8,07
		GSM_TL_2h	11,41	0,04

A Tabela 22 repete igualmente as mesmas circunstâncias de teste anteriormente consideradas mas, desta vez, para as bases de dados 5 e 6. Deve realçar-se que, neste caso, estão só envolvidos registos que são influenciados por canais do tipo TL, isto é, não há registos limpos do tipo HQ.

Os resultados que constam desta última tabela são ainda mais surpreendentes do que os resultados apresentados anteriormente. Com efeito, em ambas as bases de dados (5 e 6), estão somente envolvidos registos que são perturbados por canais telefónicos, não havendo qualquer componente de sinal gravado de alta qualidade (HQ). Ainda

assim, constata-se, uma vez mais, que os resultados obtidos para qualquer género de orador (FEM, MAL e FEM & MAL) são idênticos ou melhores do que os resultados obtidos nas comparações entre registos com uma só condição de canal.

Tabela 22 - Resultados de desempenho EER (em %) obtidos com base na comparação das bases de dados 5 e 6 em que os registos de voz incluem duas condições de canal telefónico e são mutuamente exclusivos quanto ao conteúdo da fala. Apresentam-se os resultados obtidos para os oradores do género MAL, FEM e junção dos dois. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita os valores previamente obtidos na comparação das bases de dados com uma só condição de canal telefónico.

FEM				
	BD 6		GSM_TL_2h	VOI_TL_1h
BD 5	0,13	GSM_TL_1h	0,13	13,82
		VOI_TL_2h	20,00	0,39

MAL				
	BD 6		GSM_TL_2h	VOI_TL_1h
BD 5	0,00	GSM_TL_1h	0,00	16,91
		VOI_TL_2h	17,65	0,55

FEM & MAL				
	BD 6		GSM_TL_2h	VOI_TL_1h
BD 5	0,04	GSM_TL_1h	0,04	10,81
		VOI_TL_2h	16,22	0,49

Estes resultados contrariam, assim, a hipótese anterior de que os registos de alta qualidade poderiam estar a condicionar a modelização de orador, dessensibilizando-a em relação aos registos contaminados pela influência dos canais de tipo GSM_TL ou VOI_TL. Sobra assim a hipótese de que será a diversidade das condições de canal nos registos telefónicos que alimentam a modelização de orador, que realmente contribui para uma rejeição eficaz das influências de canal telefónico e, em consequência, para

uma identificação eficiente de orador. Acreditamos que esta é uma conclusão inovadora no contexto do reconhecimento automático de orador.

6.8 Testes com assimetria na influência de canal

Na secção anterior, estudámos o desempenho na identificação de orador quando os registos de voz são enriquecidos com duas condições de canal. Os resultados, quando confrontados com os anteriormente obtidos para registos de voz com uma só condição de canal, foram surpreendentes e muito positivos. Suscitaram também mais um conjunto de testes, não inicialmente previsto, para responder à seguinte questão: se a modelização de orador parece resultar mais informada e eficaz quando os registos de voz contêm mais do que uma condição de canal, será que essa eficácia se manifesta também quando, em vez dos dois, só um dos dois registos que intervêm numa comparação for composto por duas condições de canal? Por outras palavras, os mesmos bons resultados na identificação de orador mantêm-se quando as condições de canal (e duração) nos dois registos de voz que intervêm numa comparação forem assimétricas?

Para responder a esta pergunta, utilizámos três das novas bases de dados com duas condições de canal (BD 1, BD 3 e BD 5) e estudamos, para cada caso de género de orador (FEM, MAL e FEM & MAL) o desempenho na identificação de orador quando são comparadas com bases de dados de registos *telefónicos* de uma só condição de canal e mutuamente exclusivos, isto é, em que não haja qualquer parte comum da fala gravada.

Em primeiro lugar, compararam-se os registos da BD 1 (GSM_HQ_1h_VOI_TL_2h) com os registos de meia duração e correspondentes à base de dados GSM_TL_2h. Repetiu-se o teste em relação à base de dados VOI_TL_1h. Os correspondentes resultados de desempenho na identificação de orador (aferido pelo parâmetro EER), para os três casos de género de orador, apresentam-se na Tabela 23. Esta tabela apresenta também, para referência, resultados anteriormente obtidos na comparação entre bases de dados com registos de voz afetados por uma só condição de canal. A relevância destes resultados de referência advém da circunstância de

exprimirem as mesmas comparações realizadas com uma parte só dos registos contidos nas duas condições de canal. Deste modo, é possível apreciar se a informação adicional proporcionada pelas duas condições de canal é benéfica, ou não, do ponto de vista do desempenho na identificação de orador.

Tabela 23 - Resultados de desempenho EER (em %) obtidos com base na comparação da base de dados 1 (GSM_HQ_1h_VOI_TL_2h), em que os registos de voz incluem duas condições de canal, com a base de dados GSM_TL_2h e também VOI_TL_1h, cujo conteúdo de fala é mutuamente exclusivo. Apresentam-se os resultados obtidos para os oradores do género FEM, MAL e FEM & MAL. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita valores relevantes previamente obtidos na comparação de bases de dados com uma só condição de canal.

FEM					
	GSM_TL_2h	VOI_TL_1h		GSM_TL_2h	VOI_TL_1h
BD 1	15,00	0,26	GSM_HQ_1h	20,13	
			VOI_TL_2h		0,39
MAL					
	GSM_TL_2h	VOI_TL_1h		GSM_TL_2h	VOI_TL_1h
BD 1	11,58	5,15	GSM_HQ_1h	6,45	
			VOI_TL_2h		0,55
FEM & MAL					
	GSM_TL_2h	VOI_TL_1h		GSM_TL_2h	VOI_TL_1h
BD 1	10,85	5,41	GSM_HQ_1h	10,46	
			VOI_TL_2h		0,49

Os resultados patentes na Tabela 23 permitem concluir que para os oradores FEM há melhoria no desempenho. Por outro lado, para os oradores MAL e FEM&MAL, há um agravamento do desempenho, muito especialmente na comparação com os registos VOI_TL_1h. De facto, é particularmente significativo que o desempenho nas comparações entre VOI_TL_2h e VOI_TL_1h, sendo de EER=0.55% para oradores do tipo MAL e EER=0.49% para oradores do tipo FEM&MAL, venha

agravado para EER=5.15% e EER=5.41%, respetivamente, quando a BD 1, que contém a condição de canal VOI_TL_2h, é comparada com as aquelas bases de dados (VOI_TL_2h e VOI_TL_1h). Isto denota que o facto da BD1 ser mais informada porque contém duas condições de canal para cada orador (HQ e VOI_TL), não constitui vantagem nesta tarefa de identificação de orador usando o *iVocalise*. Em consequência, esta constatação aponta para a conclusão de que a assimetria que existe na construção do modelo de orador, nos dois casos, tem um impacto negativo na discriminação.

A Tabela 24 apresenta resultados do mesmo tipo dos da tabela anterior mas em que a base de base contendo registos com duas condições de canal é BD 3 (VOI_HQ_1h_GSM_TL_2h) e as bases de dados com as quais esta é comparada e que contém registos telefónicos com uma só condição de canal, são as VOI_TL_2h e também GSM_TL_1h. Como anteriormente, apresentam-se também, para referência, resultados da comparação entre bases de dados com registos de voz afetados por uma só condição de canal.

Os resultados desta tabela denotam que, para qualquer caso de género do orador, o desempenho na identificação de orador quando a comparação é feita com a base de dados VOI_TL_2h não é significativamente alterado, quer se use a base de dados VOI_HQ_1h, ou quer se use a base de dados BD 3. Porém, quando a comparação é feita com a base de dados GSM_TL_1h, há um agravamento muito significativo no desempenho quando se usa a BD 3 que, supostamente, é mais informada porque reflete duas condições de canal.

A tendência que já havia sido identificada através dos resultados da Tabela 23, também aqui, para os resultados da Tabela 24, parece confirmar-se: a assimetria nas condições de canal que afetam os registos que são usados na comparação parece representar o maior desafio para o *iVocalise* realizar uma modelização de orador eficaz e uma discriminação eficiente. A hipótese da assimetria na duração dos registos comparados poder também afetar negativamente o desempenho não é aqui confirmada porque, como já assinalámos, os testes com a base de dados VOI_TL_2h não denotam diferenças significativas. Assim, decorre como plausível a conclusão de que os bons resultados alcançados na comparação da base de dados GSM_TL_2h com a

GSM_TL_1h se compreende porque estas bases de dados refletem exatamente as mesmas condições de canal de comunicação.

Tabela 24 - Resultados de desempenho EER (em %) obtidos com base na comparação da base de dados 3 (VOI_HQ_1h_GSM_TL_2h), em que os registos de voz incluem duas condições de canal, com a base de dados VOI_TL_2h e também GSM_TL_1h, cujo conteúdo de fala é mutuamente exclusivo. Apresentam-se os resultados obtidos para os oradores do género FEM, MAL e FEM & MAL. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita valores relevantes previamente obtidos na comparação de bases de dados com uma só condição de canal.

FEM					
	VOI_TL_2h	GSM_TL_1h		VOI_TL_2h	GSM_TL_1h
BD 3	9,47	10,00	VOI_HQ_1h	9,34	
			GSM_TL_2h		0,13

MAL					
	VOI_TL_2h	GSM_TL_1h		VOI_TL_2h	GSM_TL_1h
BD 3	6,07	5,15	VOI_HQ_1h	5,33	
			GSM_TL_2h		0,00

FEM & MAL					
	VOI_TL_2h	GSM_TL_1h		VOI_TL_2h	GSM_TL_1h
BD 3	8,41	9,61	VOI_HQ_1h	8,07	
			GSM_TL_2h		0,04

Por último, a Tabela 25 apresenta resultados do mesmo tipo das duas tabelas anteriores mas em que a base de base contendo registos com duas condições de canal é BD 5 (GSM_TL_1h_VOI_TL_2h) e as bases de dados com as quais esta é comparada e que contêm registos telefónicos com uma só condição de canal, são as GSM_TL_2h e também GSM_TL_1h. Este cenário, contrariamente aos das duas tabelas anteriores, é de interesse acrescido porquanto todos os registos de voz envolvidos encontram-se afetados por um canal do tipo telefónico (*i.e.* não é HQ) ou GSM_TL, ou VOI_TL.

Tal como anteriormente também, apresentam-se na Tabela 25 e para referência, resultados da comparação entre bases de dados com registos de voz afetados por uma só condição de canal.

Tabela 25 - Resultados de desempenho EER (em %) obtidos com base na comparação da base de dados 5 (GSM_TL_1h_VOI_TL_2h), em que os registos de voz incluem duas condições de canal, com a base de dados GSM_TL_2h e também VOI_TL_1h, cujo conteúdo de fala é mutuamente exclusivo. Apresentam-se os resultados obtidos para os oradores do género FEM, MAL e FEM & MAL. Para facilitar a discussão, a tabela à direita, sombreada a cor verde, explicita valores relevantes previamente obtidos na comparação de bases de dados com uma só condição de canal.

FEM					
	GSM_TL_2h	VOI_TL_1h		GSM_TL_2h	VOI_TL_1h
BD 5	14,87	5,00	GSM_TL_1h	0,13	
			VOI_TL_2h		0,39
MAL					
	GSM_TL_2h	VOI_TL_1h		GSM_TL_2h	VOI_TL_1h
BD 5	5,70	11,76	GSM_TL_1h	0,00	
			VOI_TL_2h		0,55
FEM & MAL					
	GSM_TL_2h	VOI_TL_1h		GSM_TL_2h	VOI_TL_1h
BD 5	7,88	5,44	GSM_TL_1h	0,04	
			VOI_TL_2h		0,49

Os resultados nesta tabela denotam que, em todos os casos, a utilização de registos de voz com duas condições de canal é sempre prejudicial em relação à utilização de registos com uma só condição de sinal. Este resultado é surpreendente pela negativa. Confirma, porém, a hipótese de que, no *software iVocalise*, a modelização de orador só é eficaz e a discriminação eficiente quando os registos de voz refletem exatamente as mesmas condições de canal.

6.9 Conclusão

Este capítulo apresentou um conjunto de estudos que realizámos com base no *software* de reconhecimento automático de orador, o *iVocalise*. Dada a atualidade e importância prática desta abordagem de reconhecimento de orador em trabalhos de peritagem e investigação, tipicamente conduzida por laboratórios de polícia científica, fornecemos, na Secção 6.1, um breve enquadramento histórico do desenvolvimento na área do reconhecimento de orador até apresentarmos um diagrama de blocos funcional do *software iVocalise*. Seguidamente, na Secção 6.2, destacámos a importância do módulo de VAD no pré-condicionamento dos dados. Na Secção 6.3, caracterizamos resultados consolidados de investigação anteriormente obtidos no reconhecimento automático de orador, através do *software iVocalise*, com a nossa base de dados, e que representam o ponto de partida para a nova investigação reportada nesta dissertação nas secções seguintes. Em especial, estes resultados evidenciam que o melhor desempenho consegue-se usando registos de voz amostrados a 8 kHz, usando VAD e removendo manualmente as regiões de sinal contendo voz do segundo orador. Na Secção 6.4 detalhámos a segmentação que foi realizada sobre os registos da base de dados original, de modo a construirmos novas bases de dados que permitam estudar o efeito no desempenho da duração dos registos, da mútua exclusividade entre os registos comparados quanto ao seu conteúdo de fala, e da assimetria nas condições de canal entre estes registos. Na Secção 6.5, realizámos testes que permitissem concluir sobre o impacto no desempenho que decorre da utilização de registos com metade da duração dos inicialmente usados (e que produziu os valores de referência de desempenho). Exceto para alguns casos pontuais, confirmou-se que o impacto não era expressivo, o que permitiu validar as condições de investigação nas secções seguintes e que fazem uso dos registos com metade da duração dos iniciais. Na Secção 6.6, foram repetidos os testes descritos na Secção 6.5, mas em que os registos implicados nas comparações foram usados de modo a forçar a condição de mútua exclusividade quanto ao conteúdo da fala. Concluiu-se que o seu impacto no desempenho na identificação de orador é no sentido de alguma degradação mas que, apesar de exibir alguma flutuação, não é muito expressiva. Na Secção 6.7, realizámos estudos de comparação de registos de voz

contendo duas condições de canal e concluímos que os resultados de desempenho alcançados (na identificação de orador) eram iguais ou melhores do que os alcançados com registos contendo uma só condição de canal. Estes resultados apontaram para a conclusão de que a modelização de orador é mais eficaz e o reconhecimento de orador é mais eficiente se os registos de voz encerrarem mais do que uma condição de canal de comunicação. Na Secção 6.8, investigámos o impacto no desempenho da identificação de orador quando existe assimetria nos registos de voz que são comparados, designadamente quanto às condições de canal neles refletidas. Concluímos em particular que, exceto em certos casos pontuais, a assimetria é prejudicial para o desempenho, o que coloca mais uma vez em evidência o que parece ser um facto agora confirmado: o *software iVocalise* realiza uma eficaz modelização de orador e uma eficiente identificação desde as condições de canal dos registos de voz comparados sejam exatamente as mesmas, ou seja, simétricas, sendo que uma diversidade destas condições de canal é benéfica para o desempenho em relação às comparações de registos com base numa só condição de canal.

Capítulo 7 – Caso de estudo

7.1 Introdução

O presente capítulo introduz um caso real de peritagem em que estivemos envolvidos e que requereu análise de parâmetros acústicos e seu tratamento estatístico. Assim, explora os métodos, processos e materiais decorrentes de todo o progresso do estudo. Apresenta, também, os resultados a nível estatístico e os resultados relacionados com a extração de parâmetros de voz, como a frequência fundamental. Por fim, justifica-se o facto de o orador em questão ser ou não o falante de registos de voz desconhecidas e alvo de suspeição, com o apoio dos rácios de verosimilhança.

7.2 Trabalho de peritagem

No decorrer da nossa investigação principal e reportada nos capítulos precedentes, colaborámos numa peritagem que implicou conduzir e gerir diálogos de modo a se obterem os parâmetros acústicos de fones específicos que tinham sido produzidos no âmbito dos registos de voz cujo orador era desconhecido. O objetivo da peritagem focou-se em emitir uma apreciação técnica fundamentada sobre vozes que constam de várias interceções telefónicas legais referidas num processo judicial, de modo a avaliar a plausibilidade de algum dos oradores suspeitos corresponder, ou não, ao orador desconhecido.

Neste contexto, recolheram-se, através da gravação de diálogo simulado por via telefónica, amostras de voz especificamente de um indivíduo suspeito, doravante designado de OS (Orador Suspeito), e que participou nessas gravações de forma voluntária. As condições foram idênticas às descritas no Capítulo 3, ou seja, obtivemos a versão de voz telefónica em condições idênticas às das interceções telefónicas, e também a versão de alta-qualidade sem interferência de qualquer canal telefónico de

comunicação.

Para a peritagem, como referido, foi necessário recolher registos de voz de OS para confrontar com umas já existentes gravadas através de interceção telefónica. Estes registos decorreram de duas conversas telefónicas simuladas, sendo gravadas em tempo real, com dois suportes diferentes. O primeiro encontrava-se junto do OS, num local remoto, e o outro encontrava-se na Faculdade de Engenharia da Universidade do Porto.

No local junto ao OS, utilizaram-se os seguintes meios de interface de obtenção de sinal de voz em alta qualidade:

- Sennheiser head microphone (Ear Set1-ew3);
- Adaptador Sennheiser (Modelo MZA 900 com *phantom power*);
- *Interface* áudio de digitalização A/D USB (Cakewalk UA-25EX by Roland);
- Computador pessoal portátil para gravação direta (HP Compaq 6910p com sistema operativo Windows XP Professional) e usando o editor áudio *Adobe Audition* (versão 1.5).

A qualidade deste registo de áudio é definida por uma frequência de amostragem de 48 kHz e 32 bits de resolução por amostra. Para a comunicação celular GSM, o orador OS usou um equipamento iPhone 5S. A MEO foi a operadora de telecomunicações (indicativo 96).

No que diz respeito ao registo realizado na Faculdade de Engenharia, estes foram os meios tecnológicos:

- Um equipamento Nokia 108 e o respetivo conjunto original de *headphone* microfone, sendo a Vodafone a operadora de telecomunicações (indicativo 91).

O sinal de voz remoto do orador OS foi extraído deste último equipamento e fornecido a um computador pessoal (Sony Vaio, modelo PCG-31311M e com sistema operativo Windows 7) para gravação direta através do editor *Audacity* (versão 1.2.6). A qualidade deste registo de áudio é definida por uma frequência de amostragem de 22.05 kHz e 16 bits de resolução por amostra.

Para a investigação subjacente à presente peritagem foi utilizado o *software* que já foi referido na dissertação:

- i) *iVocalise* (versão 2.4.0.1532) – sistema forense de reconhecimento automático de orador, e *Biometrics* (versão 1.5.0.637) sistema associado de representação gráfica de resultados;
- ii) *Adobe Audition* (versão 6.0) – editor de áudio profissional;
- iii) *Audacity* (versão 1.2.6) – editor de áudio;
- iv) *Praat* (versão 6-0-23) – ambiente académico de análise acústica da voz;
- v) *Excel* (versão 15.30) – folha de cálculo para realização de parte da análise estatística.

As interceções telefónicas haviam sido obtidas em qualidade telefónica em registo GSM, com uma frequência de amostragem de 8000 Hz e uma resolução de amostra de 16 bits. Em particular, obtivemos acesso a dois registos das interceções que serão designados por XX01_GSM_TL e XX02_GSM_TL e foram produzidos por um orador desconhecido.

Os quatro registos telefónicos recolhidos por nós e com a voz de OS, têm a designação de OS01_GSM_HQ (conversação número um – alta qualidade), OS01_GSM_TL (conversação número um – qualidade telefónica), OS2_GSM_HQ (conversação número dois – alta qualidade) e OS2_GSM_TL (conversação número dois – qualidade telefónica). Após uma simples conversão de dois deles, estes registos possuem frequência de amostragem de 22050 Hz e resolução de amostra de 16 bits.

As quatro recolhas de voz foram segmentadas manualmente para remover a voz do interlocutor e de ruídos. Seguidamente, foram convertidas para uma frequência de amostragem de 8000 Hz para se encontrarem em condições semelhantes às dos registos da interceção telefónica. Na Tabela 26, figuram as durações das recolhas autorizadas por OS, com silêncio e sem silêncio.

Tabela 26 - Durações das amostras do OS, com e sem silêncio, em minutos e segundos.

Registo	Duração com silêncio (min:seg)	Duração sem silêncio (min:seg)
OS01_GSM_HQ	03:12	>01:57
OS01_GSM_TL	02:36	01:57
OS2_GSM_HQ	02:05	>01:01
OS2_GSM_TL	01:35	01:01

A duração dos registos sem regiões de silêncio é designada de duração de fala útil, sendo fulcral para a construção de modelos de orador, uma vez que são, posteriormente, confrontados no *software iVocalise*, para originar *scores* de distância entre registos de voz.

As duas recolhas realizadas através da interceção também foram objeto de intervenção, na medida em que foi fulcral recorrer à segmentação manual para eliminar registos de voz do orador que conduziu o diálogo, ficando, apenas, os registos de voz de OS. Depois destas tarefas, observou-se que todas as amostras excediam os 15 segundos, o que é exigido para o *software* construir um modelo de orador (15 a 20 segundos). Podemos verificar, na Tabela 27, a duração dos registos áudio de interseções telefónicas com silêncio e sem silêncio, das amostras pertencentes ao orador desconhecido.

Tabela 27 - Durações das amostras do orador desconhecido, com e sem silêncio, em minutos e segundos.

Registo	Duração com silêncio (min:seg)	Duração sem silêncio (min:seg)
XX01_GSM_TL	01:59	00:54
XX02_GSM_TL	01:27	00:38

Para efetuarmos os estudos de comparação de voz, e porque os registos de voz foram recolhidos exatamente nas mesmas condições, recorreremos à base de dados que

constituímos, previamente, e que foi descrita no Capítulo 3, à qual chamamos, neste trabalho de peritagem, “população relevante”.

Além da nossa intervenção, participaram dois investigadores para auxiliar nas tarefas de segmentação de todos os ficheiros áudio, com o objetivo de isolar a voz de cada orador, tanto das interceções telefónicas como das novas gravações. Além disso, tiveram um papel relevante nas tarefas de isolamento de palavras e de fones específicos para análise. Realizámos as tarefas de extração de parâmetros acústicos a partir de fones selecionados. Por fim, procedeu-se à comparação forense de oradores utilizando o *software iVocalise*.

Para ser exequível a análise dos resultados, considerámos duas hipóteses, a saber:

- H_0 : os registos telefónicos que constituem a evidência (E) foram produzidos pelo indivíduo suspeito – esta é a tese da acusação;
- H_1 : os registos telefónicos que constituem a evidência (E) foram produzidos não pelo indivíduo suspeito, mas antes por um indivíduo indiferenciado da ‘população relevante’ – esta é a tese da defesa.

Tendo em conta que possuímos três conjuntos de registos, nomeadamente, obtidos a partir das interceções telefónicas com um orador desconhecido, registados a partir das gravações de OS e, também, os respeitantes à população relevante, foram executadas comparações com o objetivo de estabelecer as distâncias de pontuações entre os registos.

Primeiramente, foram comparados os registos das interceções telefónicas do orador desconhecido com os registos adquiridos a partir da conversação gravada do OS. As pontuações foram obtidas através do *software iVocalise*, com recurso à opção VAD (*Voice Activity Detection*), a sessão S e o classificador ivector-PLDA, e constam na Tabela 28.

Tabela 28 - Scores obtidos a partir da comparação entre amostras do orador OS e entre eles próprios e com os registos das interceções telefónicas do orador desconhecido.

	OS01_GSM_HQ	OS01_GSM_TL	OS02_GSM_HQ	OS02_GSM_TL
OS01_GSM_TL	5,00	127,49	-3,98	87,30
OS02_GSM_TL	-29,29	87,30	-3,42	130,88
XX01_GSM_TL	-89,29	-81,20	-85,06	-88,63
XX02_GSM_TL	-93,35	-91,26	-98,50	-106,19

É observável que os registos do OS comparados entre si dispõem de uma gama de valores que está, maioritariamente, acima de zero ou próxima desse valor. Contudo, quando se estabelecem as comparações dos registos entre o OS e o orador desconhecido, verifica-se que os resultados se distanciam completamente de zero, estando muito abaixo deste valor. Conclui-se, ainda que esta questão requer um estudo mais aprofundado, para apurar se o orador desconhecido poderá não ser o orador OS.

De seguida, foi confrontado o registo XX01_GSM_TL de interceção telefónica do orador desconhecido com as restantes amostras de qualidade telefónica, incluindo OS01_GSM_TL, OS02_GSM_TL e XX02_GSM_TL. Os *scores* obtidos foram os que são ilustrados na Figura 64.

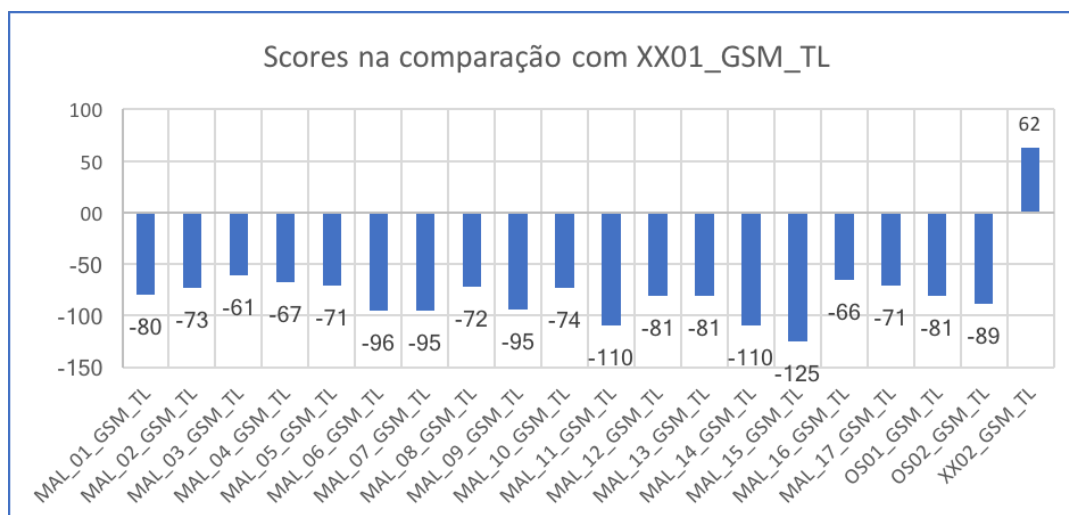


Figura 64 - Scores obtidos entre o registo XX01_GSM_TL e todas as amostras restantes de índole telefónica.

Concluimos com base nesta figura que o resultado obtido para XX02_GSM_TL é positivo e distancia-se substancialmente dos de todos os outros registos. Este facto é sugestivo de que o orador desta amostra e o de XX01_GSM_TL são o mesmo. Demonstra-se, do mesmo modo, que os scores da comparação entre XX01_GSM_TL e OS01_GSM_TL e entre XX01_GSM_TL e OS_02_GSM_TL não se destacam dos resultados obtidos de outro qualquer orador da população relevante, significando que os registos de voz do orador OS são tão distintos dos do orador desconhecido assim como os da população relevante.

Podemos confirmar estas conclusões na comparação de XX02_GSM_TL com as restantes amostras na Figura 65.

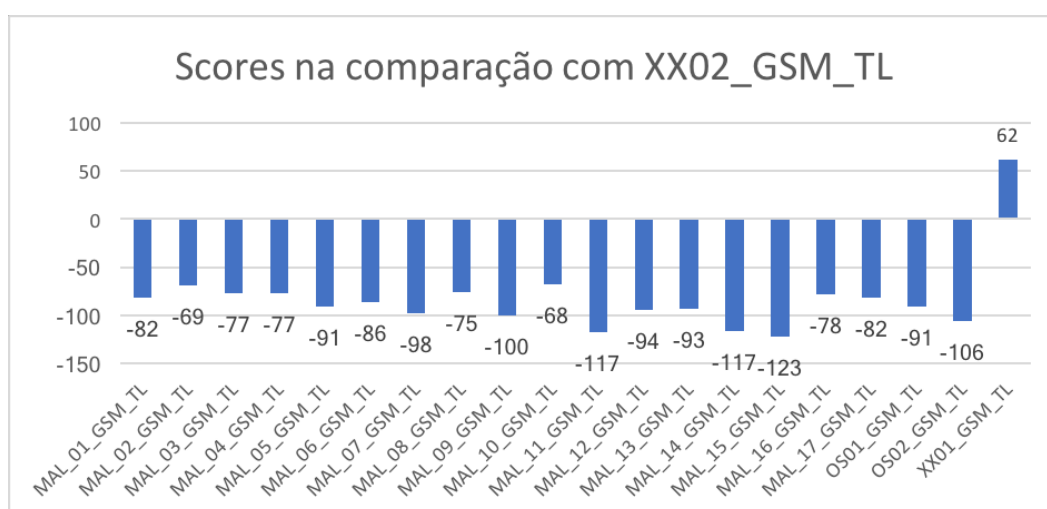


Figura 65 - Scores obtidos entre o registo XX02_GSM_TL e todas as amostras restantes de índole telefónica [52].

Outra perspetiva de análise destes resultados prende-se pelo estudo dos cálculos LR que estão relacionados com as hipóteses H_0 e H_1 anteriormente referidas. Através de curvas de densidade de probabilidade, compararam-se os registos HQ com os registos TL da população relevante, ou seja, obtemos uma matriz de pontuações de 17 oradores por 17 oradores. O passo seguinte foi o de acrescentar a esta base de dados os quatro

registos das conversas telefónicas do orador OS (duas HQ e duas TL), ficando, assim, 19 oradores por 19 oradores. Nos gráficos da Figura 66, observam-se os resultados.

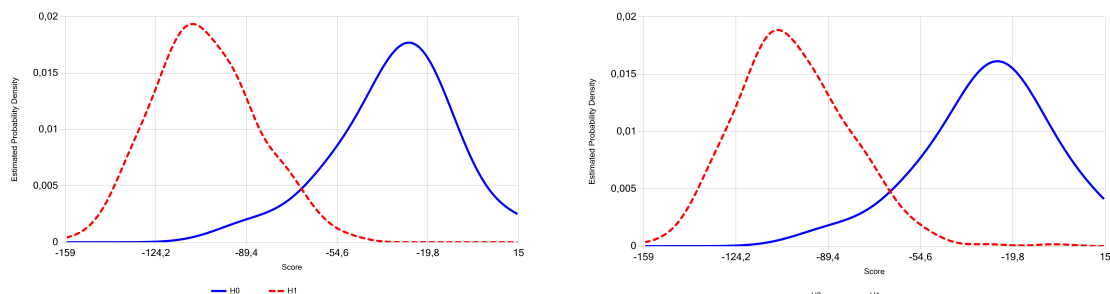


Figura 66 - Curvas de densidade de probabilidade para as hipóteses H_0 e H_1 , tendo em conta o conjunto de 17 oradores da população relevante (esquerda) e a base de dados alargada de 19 oradores (direita) [52].

O parâmetro EER (*Equal Error Rate*) é calculado a partir das curvas de densidade de probabilidade. Indica a igualdade entre a taxa de erro na aceitação incorreta de correspondência entre registos que respeitam a oradores distintos e na rejeição incorreta de correspondência entre registos que são do mesmo orador. Assim, caracteriza-se por ser uma medida que representa a qualidade do reconhecimento de orador que decorre das curvas de densidade de probabilidade para as hipóteses H_0 e H_1 . Os valores desejáveis para o EER são menores do que 7%. Na Figura 67, são patentes os resultados para os testes anteriores.

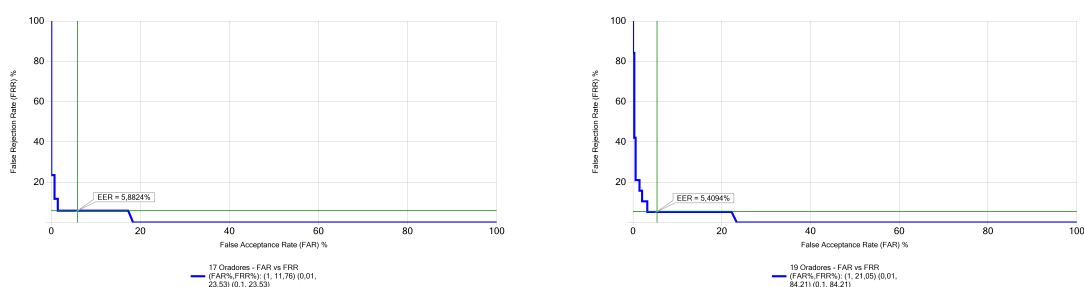


Figura 67 - Valores do desempenho EER tendo em conta o conjunto de 17 oradores da “população relevante” (esquerda) e tendo em conta a base de dados alargada de 19 oradores (direita).

Podemos verificar que a base de dados que inclui os 19 oradores é a que obtém melhores resultados (5,41%) e é, portanto, a mais prometedora para ser utilizada. Os *scores* das distâncias entre os registos telefónicos, obtidos através das opções VAD e características MFCC da voz, uso de coeficientes *delta* e *delta-delta*, projecção para vetor *i-vector*, uso de projecção TV (*Total Variability*) e uso de critério de discriminação PLDA figuram na Tabela 29.

Tabela 29 - *Scores da comparação entre os registos de intercepções telefónicas e de amostras telefónicas das conversações.*

	OS01_GSM_TL	OS02_GSM_TL
XX01_GSM_TL	-81,20	-88,63
XX02_GSM_TL	-91,26	-106,19

Com base nos conceitos introduzidos na Secção 4.5, registam-se na Tabela 30 os rácios LR para a comparação entre o registo XX01_GSM_TL (considerada a evidência E) e os dois registos OS01_GSM_TL e OS02_GSM_TL.

Tabela 30 - *Resultados LR derivados da comparação entre as amostras das intercepções telefónicas e as amostras TL das conversações telefónicas simuladas com o orador OS.*

	OS01_GSM_TL	OS02_GSM_TL
XX01_GSM_TL	0,24	0,15
XX02_GSM_TL	0,12	0,03

Através desta análise com base em procedimentos automáticos e estatísticos de reconhecimento de orador, confirma-se que todos os valores apresentados na tabela anterior são menores do que um e, portanto, a probabilidade de a evidência pertencer à hipótese H_1 é maior do que pertencer à H_0 . Por estas mesmas razões, concluiu-se, na peritagem, que o mais provável é que as amostras do orador OS não correspondam às do orador desconhecido.

As características fonético-acústicas da voz também foram exploradas nesta peritagem. De facto, como referido no Capítulo 2, a frequência fundamental, por exemplo, expressa características idiossincráticas que podem ajudar a identificar o orador [83], mesmo em contexto de registo telefónico, como comprovam estudos recentes [100] [101] e como concluímos no âmbito dos estudos reportados no Capítulo 5. Além disso, as particularidades fonético-acústicas são bastante utilizadas em amostras curtas, uma vez que a curta duração inviabiliza a abordagem de análise estatística.

Procedemos à segmentação manual de regiões sub-silábicas que correspondiam a vogais articuladas em regime fonatório quase-estacionário. As que suscitaram interesse pertinente foram, também, as regiões de hesitação silábica, na medida em que são ricas neste regime fonatório e descrevem melhor os atributos da vibração das cordas vocais [77], como referimos nas Secções 2.3.1 e 5.2.2.

A tarefa seguinte à da segmentação foi analisar cada segmento no *software Praat*. A Figura 68 ilustra os resultados.

Este gráfico da Figura 68 revela que os valores da frequência fundamental encontrados para o orador OS estão abaixo dos 100 Hz e os intervalos de confiança estão sobrepostos, além de que se distanciam do orador desconhecido cuja voz pertence ao falante das amostras das interceções telefónicas. Os intervalos de confiança do orador desconhecido são mais dilatados e possuem uma clara sobreposição entre si. É observável, do mesmo modo, que os intervalos de confiança das médias da frequência fundamental de ambos os oradores não têm qualquer sobreposição. Assim, podemos inferir, a partir da leitura destes dados, que o mais provável é os oradores OS e desconhecido não serem o mesmo indivíduo.

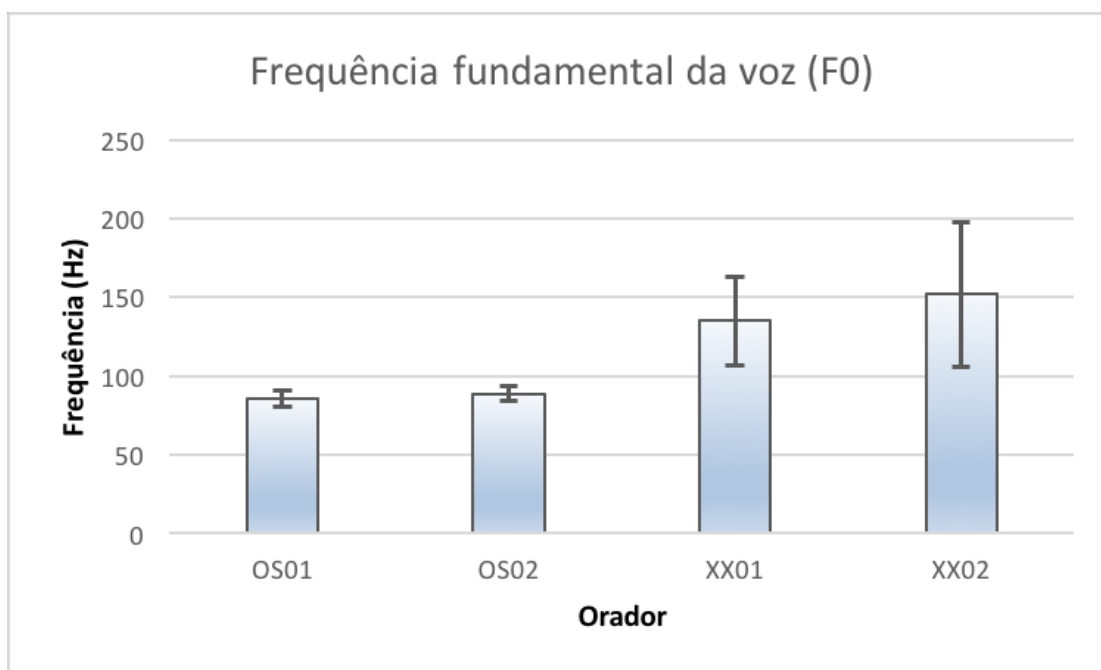


Figura 68 - Valores estimados da frequência fundamental em ambos os registos telefónicos do orador conhecido (OS01, OS02) e do orador desconhecido (XX01, XX02), juntamente com a representação do intervalo de confiança a 95% para o valor médio encontrado.

Os espectrogramas, enquanto abordagem clássica nos estudos forenses, representaram outro teste para consolidar os resultados acima referidos. Assim, com base nos dois registos disponíveis de OS e do orador desconhecido, procuraram-se vocábulos que fossem comuns a ambos e que expusessem características particulares de cada um dos indivíduos. Depois de uma pesquisa exaustiva, as palavras escolhidas foram “carro” e “sete”.

A Figura 69 revela os espectrogramas para uma só ocorrência da palavra “carro” produzida no registo OS02_GSM_TL (orador OS) e uma só ocorrência do mesmo vocábulo em XX01_GSM_TL (orador desconhecido):

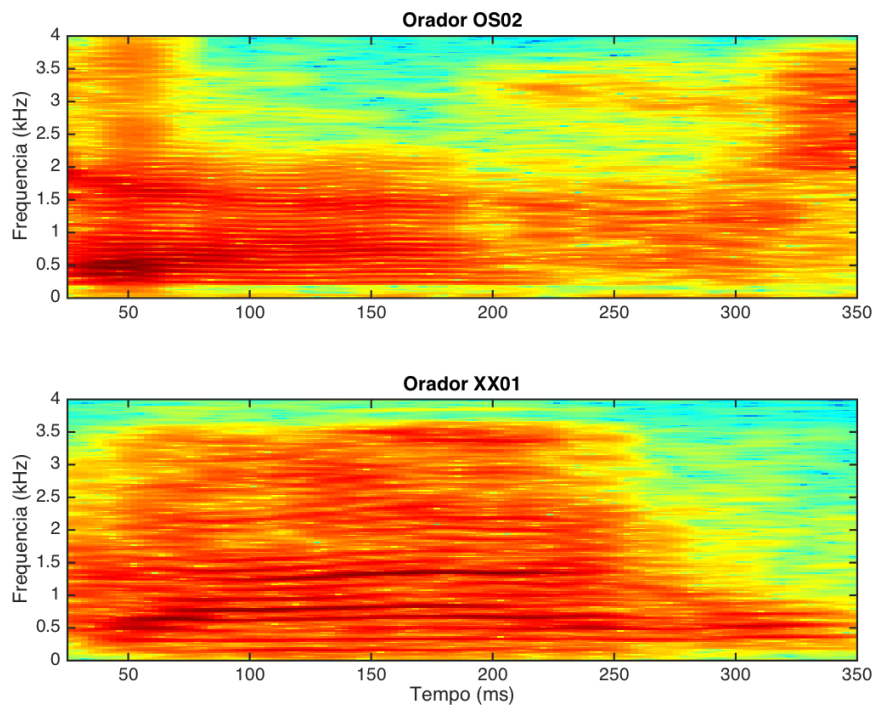


Figura 69 - Espectrogramas correspondentes a uma ocorrência da palavra “carro” no registo de voz telefónica OS02_GSM_TL (painel superior) e no registo de voz telefónica XX01_GSM_TL (painel inferior).

O foco destes espectrogramas é a primeira sílaba de carro, a saber, /ka/, que desperta particular interesse, na medida em que estamos perante o ataque da sílaba e a vogal aberta /a/. Apesar de tudo, o nível de ruído na amostra XX01_GSM_TL é substancial para nos ser possível extrair conclusões firmes quanto ao ataque. Contudo, podemos medir as frequências formantes, como se comprova na Tabela 31.

Tabela 31 - Frequências da formante F1 e da formante F2 do fone /a/ em /ka/, em Hz, para as amostras OS2 e XX01, em qualidade telefónica de GSM.

	OS02_GSM_TL	XX01_GSM_TL
F1 (Hz)	~800	~780
F2 (Hz)	~1450	~1350

Através destes resultados, podemos verificar que as dissemelhanças entre as formantes homólogas não são suficientemente expressivas para serem consideradas. Todavia, a frequência fundamental é notoriamente superior na amostra do orador desconhecido em relação à amostra do orador OS, na medida em que as estrias horizontais do espectrograma do orador desconhecido revelam uma separação entre si significativamente superior à divisão entre as estrias que se verificam no espectrograma do registo do segundo orador, o OS. Estes resultados reforçam as conclusões do teste anterior (ver Figura 69).

Os quatro espectrogramas representados na Figura 70 traduzem a ocorrência da palavra “sete” encontrada em cada um dos quatro registos telefónicos, concretamente, dois do orador OS (OS01_GSM_TL e OS02_GSM_TL) e do orador desconhecido (XX01_GSM_TL e XX02_GSM_TL).

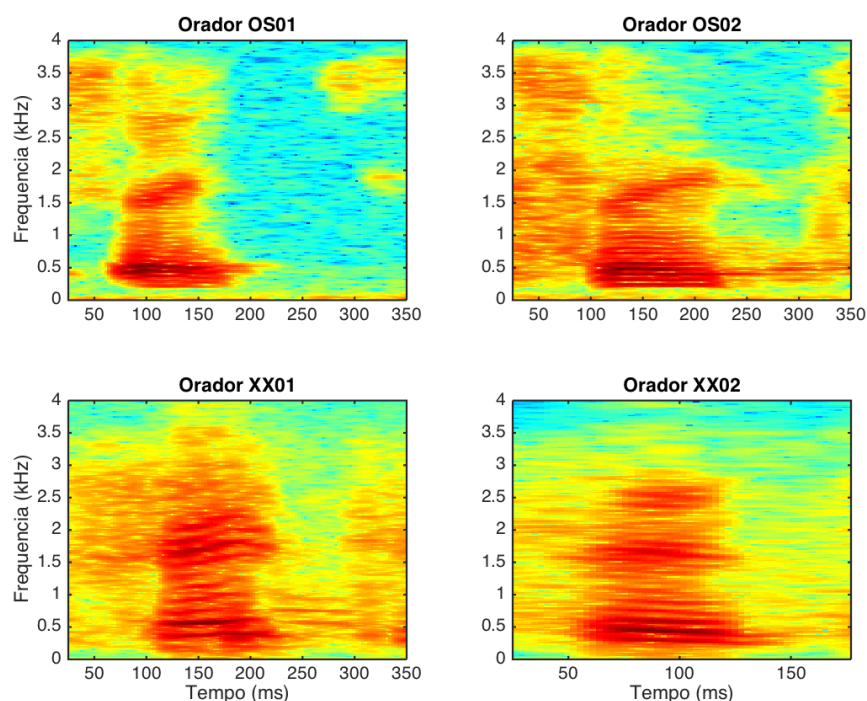


Figura 70 - Espectrogramas correspondentes a uma só ocorrência da palavra “sete” em cada um dos dois registos telefónicos do orador OS (painéis superiores) e em cada um dos dois registos telefónicos do orador XX (painéis inferiores).

A Tabela 32 esquematiza os valores aproximados das frequências das formantes F1 e F2 da vogal semiaberta /e/, pertencente a uma só ocorrência da palavra “sete”, em cada um dos quatro registos de voz, e obtidas por medição direta.

Tabela 32 - Valores aproximados das frequências das formantes extraídas da vogal semiaberta /e/, pertencente a uma ocorrência da palavra “sete”, em cada um dos quatro registos de voz.

	F1 (HZ)	F2 (HZ)
OS01_GSM_TL	~500	~1550
OS02_GSM_TL	~450	~1600
XX01_GSM_TL	~600	~1700
XX02_GSM_TL	~461	~1670

A partir da tabela anterior e dos espectrogramas da Figura 70, concluímos que os valores das formantes homólogas são tendencialmente mais elevadas no orador desconhecido (XX) do que no OS, particularmente no que diz respeito à formante F2, o que sustenta uma demarcação entre ambos os oradores. Esta constatação é consistente com o princípio geral de que um valor mais alto da frequência fundamental potencia valores mais elevados também das frequências das formantes, algo que para um mesmo orador se associa à elevação da laringe.

A diferença entre os valores das formantes de ambos os oradores e o facto do valor da frequência fundamental de OS ser mais baixo do que a do orador desconhecido expressa, possivelmente, que as dimensões do trato vocal e da cavidade nasal do orador OS são maiores do que as do orador desconhecido. Este conhecimento, que sugere que a estatura física do orador OS é maior do que a do orador desconhecido, está estabelecido na literatura [42] [44].

É de salientar também o trajeto da formante F2 ao longo da articulação da vogal aberta /e/. No caso do orador OS, observa-se um trajeto ascendente que é consistente nas palavras “sete”, o que traduz um hábito articulatorio que se demarca dos automatismos articulatorios do orador desconhecido para esta vogal particular.

7.3 Conclusão

Este trabalho de peritagem teve como principal objetivo averiguar as hipóteses de correspondência ou não correspondência entre as amostras de voz telefônica decorrentes de interceções telefônicas legais a um orador desconhecido e a voz de um orador suspeito e conhecido. Para isso, foram necessários meios técnicos e humanos que foram descritos ao longo deste capítulo e que se revestiram de grande importância na preparação dos dados e no ambiente de investigação.

Depois de se recorrer a análise estatística e a análises alternativas de comparação de vozes, como as que se baseiam em características acústico-fonéticas da voz, verifica-se que a probabilidade de as amostras desconhecidas terem sido produzidas pelo orador suspeito são mínimas, tendo em conta os rácios de verosimilhança (*Likelihood Ratios* – LR). Utilizou-se, também, a análise de comparação de voz por meio de espectrogramas, o que confirmou a fraca plausibilidade.

Capítulo 8 - Conclusões

8.1 Síntese dos estudos realizados

O trabalho realizado ao longo da dissertação insere-se na área da Fonética Forense, com especial foco na comparação de vozes e no reconhecimento automático de orador, através de características objetivas extraídas a partir do sinal de voz. Teve como principal objetivo avaliar o nível de degradação que perturba as gravações telefónicas em resultado da influência devida aos canais de comunicação GSM e VoIP, assim como avaliar o seu impacto em tarefas de comparação de vozes através de métodos não-automáticos (Capítulo 5) e métodos automáticos (Capítulo 6).

Um primeiro e decisivo objetivo, como se descreve no Capítulo 3, consistiu na preparação de uma base de dados com registos de voz telefónica que pudessem ser representativos de condições reais de comunicação de sinal de voz telefónica GSM e VoIP, para além de incluir também versões de referência desses sinais de voz libertas de qualquer distorção ou interferência. Depois de construída uma base de dados com amostras de voz de 37 oradores voluntários e em que cerca de metade são indivíduos gêmeos, procedeu-se à segmentação dos correspondentes ficheiros a fim de se remover conteúdo de sinal do segundo orador – o entrevistador –, além das regiões de silêncio. De seguida, procedeu-se à subdivisão das gravações, criando-se uma nova base de dados que seria o ponto de partida para os diferentes testes reportados no Capítulo 6 e recorrendo ao *software* de reconhecimento automático de orador, o *iVocalise*.

Cada gravação incluiu três exercícios de voz: uma conversa simulada, um exercício de leitura e a produção de vogais sustentadas. Após segmentação manual, estas foram objeto de estudo, tal como se descreve no Capítulo 5, com base em parâmetros acústicos selecionados, como a frequência fundamental, o desvio padrão da frequência fundamental, o *jitter*, o *shimmer*, o HNR e as frequências formantes. Foi possível apurar que os mais interessantes, tendo em conta o contexto de análise forense, foram a média da frequência fundamental e o HNR. De facto, estes parâmetros

revelaram maior resistência às influências do canal de comunicação, o que foi possível concluir através de correlação mais elevada e estatisticamente significativa, quando comparados com os outros parâmetros acústicos, tendo por referência os registos de voz de alta qualidade que representam o nosso ‘ground-truth’. Num segundo patamar de relevância, destacaram-se o desvio padrão da frequência fundamental e o *jitter* PPQ5, embora possam ser afetados por fatores pontuais e situacionais, como o stress, e sejam vulneráveis aos efeitos dos canais de comunicação. Por seu turno, o *shimmer* APQ5 mostrou ser um parâmetro irrelevante no que diz respeito à análise de amostras de voz em contexto forense. Quando comparados estes parâmetros acústicos em ambos os *software*, *Praat* e *VoiceStudio*, a frequência fundamental continuou a ser a mais relevante e fidedigna, uma vez que revelou grande imunidade às degradações introduzidas pelo canal telefónico, nos pares GSM e VoIP.

Ainda no âmbito do Capítulo 5 e em relação ao estudo das frequências formantes, concluiu-se que a formante F1 é especialmente sobrestimada em cerca de 11% nas gravações GSM e as formantes F3 e F4 estão subestimadas em cerca de 12% em ambas as gravações GSM e VoIP. Concluiu-se também que as vozes femininas têm propensão a ser mais afetadas pelos erros de estimação da formante F4 do que as vozes masculinas. Os resultados de análise com formantes reportados no Capítulo 5 recomendam também precaução quanto ao seu uso em contexto forense posto que o desvio das formantes induzido pelos próprios canais de comunicação pode exceder a normal variação intrafalante. Decorre que é essencial, em contexto forense, a combinação da análise com base nas formantes com outras formas de análise usando parâmetros alternativos.

No âmbito do Capítulo 6, clarificou-se que em resultado de trabalho anteriormente desenvolvido, foi possível concluir que a remoção do segundo orador é um procedimento útil, pois melhora o desempenho no reconhecimento automático de orador, avaliado através do parâmetro EER, assim como também outras formas mais simples de pré-processamento como a conversão da frequência de amostragem para 8kHz e a utilização de VAD. No que diz respeito aos testes do impacto no desempenho da identificação de orador que resultam da utilização de registos com metade da duração

dos inicialmente constituídos, aferiu-se que esse impacto não era significativo. Quando repetidos estes testes, mas com a condição mútua de exclusividade quanto ao conteúdo da fala, verificou-se que o impacto na identificação do orador está relacionado com alguma degradação, apesar de não ser muito expressiva. Por outro lado, depois de realizarmos estudos de comparação de vozes com duas condições de canal, concluímos que os resultados de desempenho alcançados eram iguais ou superiores àqueles em que os testes haviam sido realizados apenas com uma condição de canal. Este resultado suporta a conclusão de que a modelização de orador é mais eficaz, e o reconhecimento mais eficiente, se os registos de voz comportarem mais do que uma condição canal. Por fim, verificou-se o impacto no desempenho da identificação de orador quando existe assimetria nas condições de canal entre os registos de voz em comparação, sendo que concluímos que, no geral, a assimetria é nociva para o desempenho. A relevante conclusão que se pode inferir, depois de concretizados todos estes testes, é que o *iVocalise* executa uma eficaz modelização de orador e eficiente identificação quando as condições de canal dos registos de voz são simétricas. Contudo, os resultados recomendam que sejam realizados mais testes dos realizados na Secção 6.7 em que registos de voz reflitam diversidade de condições de canal, mas não reflitam simetria destas condições.

Por último, no âmbito do Capítulo 7, descrevemos um caso real de peritagem em que estivemos envolvidos e que constituiu uma oportunidade para aplicarmos abordagens e tecnologias descritas nos Capítulos 5 e 6.

8.2 Principais contribuições

Desta dissertação resultaram algumas apresentações e publicações, como as apresentadas na 3rd *European Conference IAFL (International Association of Forensic Linguistics)* – “*Forensic Linguistics: bridging the gap(s) between language and law*”, *AES International Conference on Audio Forensics*, 13^a *Conferência Bienal IAFL* e 22nd *International Conference on Digital Signal Processing*. Contribuímos, inclusivamente, para um caso de peritagem, descrito no Capítulo 7.

Pretendemos, com a dissertação e as suas principais contribuições, evidenciar a importância da análise de voz em contexto forense, em Portugal e no mundo, e estimular que as teorias e abordagens basilares e fulcrais desta área continuem, pela via da investigação, a conduzir a melhores e novos métodos e ferramentas úteis para a sociedade.

8.3 Trabalho futuro

É nossa percepção que os tópicos abrangidos por esta dissertação são muito diversificados e alguns deles, por exemplo, as tecnologias abordadas no Capítulo 6, são motivo de inúmeras dissertações de doutoramento dedicadas só ao tema aí tratado. No entanto, talvez por isso, compreendemos que avenidas importantes de trabalho futuro podem incluir as seguintes:

- completar o conjunto de testes identificado na Secção 6.7 para investigar, em complemento às conclusões da Secção 6.8, o impacto das condições de diversidade e assimetria do canal de comunicação e preparar um artigo de fundo com o conjunto de todas as conclusões do Capítulo 6;
- desenvolver abordagens que, utilizando as conclusões principais do Capítulo 5 em relação à valia dos parâmetros relacionados com a frequência fundamental e o HNR, permitam a utilização de parâmetros acústicos mais especializados, ou holísticos, em procedimentos automáticos de análise, quer ao nível segmental e dirigidos, por exemplo, às microvariações da frequência fundamental ou à integridade da estrutura harmónica em regiões vozeadas, quer ao nível supra-segmental e dirigido, por exemplo, aos padrões de nasalidade e de co-articulação, à análise da prosódia ou da cadência silábica;
- desenvolver novas abordagens de análise de sinal orientadas à segmentação fonética, como referido no final da Secção 6.2, de modo a permitir a identificação de orador com base numa diversidade de

categorias fonéticas, de que resultariam vários *scores* que seriam depois fundidos num *score* único final. Acreditamos que há fundamento para que esta abordagem de “dividir para reinar” seja realmente a utilizada pelo sistema auditivo humano.

Referências bibliográficas

- [1] LAVER, J. (1994). *Principles of Phonetics*. Cambridge: Cambridge University Press.
- [2] LOAKES, D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. Tese de Doutorado de Humanidades, School of Languages da Universidade de Melbourne.
- [3] NOLAN, F. (1994). “Auditory and Acoustic Analysis in Speaker Recognition. In J. Gibbons (ed.) *Language and the Law*”. London: Longman, 319-345 in LOAKES, D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. Tese de doutoramento apresentada à School of Languages da Universidade de Melbourne.
- [4] NOLAN, F. (1983). *The Phonetic Bases of Speaker Recognition* Cambridge: Cambridge University Press. LOAKES, D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. Tese de Doutorado de Philosophy, School of Languages da Universidade de Melbourne.
- [5] ROSE, P. (2006). “Technical forensic speaker recognition: Evaluation, types and testing of evidence”. *Computer Speech and Language* 20, pp. 159-191. Recuperado em 2016, agosto 24, de <http://www.sciencedirect.com/science/article/pii/S0885230805000355>
- [6] MORRISON, G. S. (2009). “Forensic voice comparison and the paradigm shift”. *Science & Justice*, pp. 298-308. Recuperado em 2012, novembro 6, de <http://www.sciencedirect.com/science/article/pii/S135503060900135X>
- [7] HANSEN, J H. L. & HASAN, T. (2015). “Speaker recognition by machines and

humans”. *IEEE Signal Processing Magazine*, Vol. 32, No. 6, pp. 136–145.

[8] KÜNZEL, H. J. (2010). “Automatic speaker recognition of identical twins”. *Journal of Speech, Language and the Law*, Vol 17, No. 2, 251-277. Recuperado em 2016, outubro 9, de <https://journals.equinoxpub.com/index.php/IJSLL/article/view/7829>

[9] WHITESIDE, S. P. & RIXON E. (2003). “Speech characteristics of monozygotic twin and a same-sex sibling: and acoustic case study of coarticulation patterns in read speech”. *Phonetica*, 60 (4), 273-297. Recuperado em 2012, outubro 30, de <https://www.ncbi.nlm.nih.gov/pubmed/15004495>

[10] BELINCHÓN, M. *et al.* (1994). *Psicología del Lenguaje: Investigación y teoría*. Madrid: Trotta.

[11] ROSE, P. (2002). *Forensic Speaker Identification*. Canberra: Taylor & Francis.

[12] PULAKKA, H. (2005). *Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottography*. Dissertação de Mestrado em Ciência na Tecnologia, University of Technology de Helsínquia.

[13] SOCHOROVÁ, H *et al.* (2003) “In reference to phonation larynx fixation: computer graphic record”. Recuperado em 2017, setembro 7, de <https://www.ncbi.nlm.nih.gov/pubmed/15034610>

[14] BALL, M. J. & RAHILLY, J. (1999). *Phonetics: the science of speech*. London: Arnold.

[15] MENDES, D. (2011). *Reconhecimento de Orador em Dois Segundos*. Dissertação de Mestrado Integrado em Engenharia Eletrotécnica e de Computadores, Faculdade de Engenharia da Universidade do Porto.

[16] Referência sitiográfica disponível em <http://thevoicenotes.com/wp->

[content/uploads/2013/05/abadduction.png](#) e recuperado em 2017, setembro 7

[17] Referência sitiográfica disponível em <http://anatomiaonline.com/wp-content/uploads/2015/10/cortesagital.jpg> e recuperado em 2017, setembro 7

[18] Ana P. Mendes (Editor), "Vocologia do Fado" (em revisão), capítulo IV 'Análise acústica, perceptiva e visual da voz', por Aníbal Ferreira (2016).

[19] BARBOSA, J. M. (1994). *Introdução ao estudo da Fonologia e Morfologia do Português*. Coimbra: Livraria Almedina.

[20] FITCH, W. T (s/d). "Comparative Vocal Production and the Evolution of Speech: Reinterpreting the Descent of the Larynx". Recuperado em 2013, fevereiro 24, de <http://www.st-andrews.ac.uk/~wtsf/downloads/Fitch2002Wray.pdf>

[21] JR., A. P. (s/d). "Evolução humana e natureza da linguagem". Recuperado em 2013, fevereiro 23, de http://www2.ibb.unesp.br/departamentos/Educacao/dep_to_home/Alfredo/documentos/Evolucao_e_Linguagem_Pereira.pdf

[22] CELDRÁN, E. M. (1984). *Fonética*. Barcelona: Teide, 3ª Edição.

[23] FADIGA, L. *et al.* (s/d). "Broca's Area in Language, Action, and Music". Recuperado em 2016, abril 25, de http://cel.huji.ac.il/courses/structureandprocesses/Bibliography/Fadiga_Annals_2009.pdf

[24] HEISER, M. *et al.* (s/d). "The Essential Role of Broca's Area in Imitation". Recuperado em 2016, abril 25, de http://iacoboni.bol.ucla.edu/pdfs/EurJNeurosci_Heiser_v17p1123.pdf

[25] MUSSO, M. *et al.* (2003). “Broca’s area and the language instinct”. Recuperado em 2016, abril 25, de

http://gondabrain.ls.biu.ac.il/neuroling/courses/503/papers/musso_nn2003.pdf

[26] NOVICK, J. M. *et al.* (2010). “Broca’s Area and Language Processing: Evidence for the Cognitive Control Connection”. Recuperado em 2016, abril 25, de

http://www.psych.upenn.edu/stslab/assets/pdf/NovickTrueswellThompsonSchill_2010_BrocasLanguageCognitiveControlConnection.pdf

[27] *Pour la science*, dossiers hors-série: “Les langues du monde” (1997), p.26.

[28] Referência sitiográfica disponível em [https://2.bp.blogspot.com/-](https://2.bp.blogspot.com/-zeyMo8MITqM/WPenDsrpOqI/AAAAAAAAAD6w/i8afFydIf30Ql3Ay16-w2J2onKh7WKzXACLeB/s1600/infant-vs-adult-larynx.jpg)

[zeyMo8MITqM/WPenDsrpOqI/AAAAAAAAAD6w/i8afFydIf30Ql3Ay16-w2J2onKh7WKzXACLeB/s1600/infant-vs-adult-larynx.jpg](https://2.bp.blogspot.com/-zeyMo8MITqM/WPenDsrpOqI/AAAAAAAAAD6w/i8afFydIf30Ql3Ay16-w2J2onKh7WKzXACLeB/s1600/infant-vs-adult-larynx.jpg), acedido em 2017, dezembro 5

[29] SOUSA, R. (2011). *Metodologias de Avaliação Percetiva e Acústica do Sinal de Voz em Aplicações de Ensino do Canto e Diagnóstico/Reabilitação da Fala*. Tese de Doutoramento em Engenharia Eletrotécnica e de Computadores, Faculdade de Engenharia da Universidade do Porto.

[30] KENT, R. & READ C. (2002). *Acoustic analysis of speech*. San Diego: Singular Publishing Group, 2ª edição.

[31] ALMEIDA, V. (2012). *Relação entre características objetivas da voz cantada e seus atributos artísticos e estéticos*. Tese de Mestrado Integrado em Engenharia Eletrotécnica e de Computadores apresentada à Faculdade de Engenharia da Universidade do Porto.

- [32] SEARA, I. C. *et al.* (2011). *Fonética e Fonologia do Português Brasileiro: 2º período*. Florianópolis: LLV/CCE/UFSC.
- [33] DUBOIS, J. *et al.* (1973). *Dicionário de Lingüística*. São Paulo: Cultrix.
- [34] RUA, S. M. (2006). *Estudo Morfológico-dinâmico do tracto vocal humano*. Tese de Mestrado em Engenharia Biomédica, Faculdade de Engenharia da Universidade do Porto.
- [35] BEIGUEILMAN, B. (2008). *O estudo de gêmeos*. E-book recuperado em 2012, fevereiro 16, de <http://www.desvirtual.com/bbeiguel/gemeos01.prn.pdf>
- [36] FERNÁNDEZ, E. S. S. (2014). *Forensic Speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics*. Tese de doutoramento apresentada à Universidade Internacional Menéndez Pelayo.
- [37] FERNÁNDEZ, E. S. S. (2013). “A Phonetic Corpus of Spanish Male Twins and Siblings: Corpus Design and Forensic Application”. *Procedia – Social and Behavioral Sciences*, Vol. 95, 59 – 67. Recuperado em 2017, agosto 14, de <http://www.sciencedirect.com/science/article/pii/S1877042813041426>
- [38] MONTANARI, T. (2013). *Embriologia: texto, atlas e roteiro de aulas práticas*. Porto Alegre: Ed. da autora. Recuperado em 2017, setembro 7, de <http://www.ufrgs.br/livrodeembrio/ppts/5.desenvhumano.pdf>
- [39] GUNNAR, F. (1960). *Acoustic theory of speech production*. The Hague: Mouton.
- [40] LIERDE, K. M. *et al.* (2005). “Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach”. Recuperado em 2012, setembro 27, de <http://www.ncbi.nlm.nih.gov/pubmed/16301097>

[41] SEBASTIAN S. *et al.* (2003). “An investigation into the voice of identical twins”. *Otolaryngology Online Journal* (2013), Vol 3, No. 2. Recuperado em 2017, outubro 7, de <http://www.alliedacademies.org/abstract/an-investigation-into-the-voice-of-identical-twins-4038.html>

[42] SATALOFF, R. T. (1995). “Genetics of voice”. *Journal of Voice*, Vol. 9, No. 1, 16-19. Recuperado em 2017, setembro 3, de http://www.sciencedirect.com/science/article/pii/S0892199705802188?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb

[43] WEINRICH, M. (2010). “Articulatory and acoustic inter-speaker variability in the production of vowels”. Recuperado em 2014, junho 28, de http://www.sprechwissenschaft.uni-jena.de/germsprach_multimedia/Downloads/weirich/Weirich_inter_speaker+variability+in+vowels.pdf

[44] ZUO, D. & MOK P. P. K. (2015). “Formant dynamics of bilingual identical twins”. *Journal of Phonetics*, Vol. 52, pp. 1-12. Recuperado em 2017, setembro 9, de <http://www.sciencedirect.com/science/article/pii/S0095447015000182>

[45] DEBRUYNE, F. *et al.* (2002). “Speaking Fundamental Frequency in monozygotic and dizygotic twins”. Recuperado em 2012, setembro 25, de <http://www.ncbi.nlm.nih.gov/pubmed/12512633>

[46] Whiteside, S. P., & Rixon, E. (2001). Speech patterns of monozygotic twins: An acoustic case study of monosyllabic words. *The Phonetician*, 82, 9–22. Recuperado em 2017, novembro 21, de https://www.researchgate.net/publication/284462890_Speech_patterns_of_monozygotic_twins_An_acoustic_case_study_of_monosyllabic_words

[47] WEINRICH, M. (2011). *The influence of NATURE and NURTURE on speaker-*

specific parameters in twins speech: Articulation, acoustics and perception. Tese de doutoramento apresentada à Humboldt-Universität zu Berlin.

[48] NOLAN, F. (1997). “Speaker recognition and forensic phonetics”. In W. J. Hardcastle, & J. Laver (Eds.), *The handbook of phonetic sciences*, pp. 744–767. Oxford: Blackwell.

[49] CIELO, C. A. *et al.* (2010). “Características vocais de gêmeos monozigóticos”. Recuperado em 2017, novembro 2, de <http://www.scielo.br/pdf/rcefac/v14n6/26-09.pdf>

[50] WEINRICH, M. & LANCIA, L. (2011). “Perceived auditory similarity and its acoustic correlates in twins and unrelated speakers”. Recuperado em 2012, setembro 29, de http://www.sprachwissenschaft.uni-jena.de/germsprach_multimedia/Downloads/weirich/Weirich_PERCEIVED+AUDITORY+SIMILARITY+AND+ACOUSTIC+CORRELATES.pdf

[51] HOLLIEN, H. & SCHWARTZ, R. (2001). “Speaker identification utilizing non contemporary speech”. *Journal of Forensic Science*, Vol. 46, No. 1, 63-67. Recuperado em 2017, novembro 21, de https://www.astm.org/DIGITAL_LIBRARY/JOURNALS/FORENSIC/PAGES/JFS14912J.htm

[52] PINTO, D. (2017). *Identificação de vozes em contexto de análise forense*. Tese de Mestrado Integrado em Engenharia Eletrotécnica e de Computadores, Faculdade de Engenharia da Universidade do Porto.

[53] CHAMPOD, C., MEUWLY, D. (2000). “The inference of identity in forensic speaker recognition”. *Speech Communication*, Vol. 31, pp. 193-203. Recuperado em 2017, setembro 1, de <http://www.sciencedirect.com/science/article/pii/S0167639399000783>

[54] HANSEN, J. H. L. & HASAN, T. (2015). “Speaker recognition by machines and humans,” *IEEE Signal Processing Magazine*, Vol. 32, No. 6, 136–145. Recuperado em

2016, dezembro 6, de <http://ieeexplore.ieee.org/document/7298570/>

[55] KREIMAN, J. & SIDTIS, D. (2013). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Malden: Wiley-Blackwell.

[56] GHIURCAU M. *et al.* (2010). “A study of the effect of emotional state upon the variation of fundamental frequency of a speaker”. *Journal of Applied Computer Science & Mathematics*, Vol. 4, pp. 79-82. Recuperado em 2017, novembro 21, de https://www.researchgate.net/publication/41392010_A_Study_of_the_Effect_of_Emotional_State_upon_the_Variation_of_the_Fundamental_Frequency_of_a_Speaker

[57] GILLIER, R. (2011). *O Disfarce da Voz em Fonética Forense*. Tese de Mestrado em Linguística, Faculdade de Letras da Universidade de Lisboa.

[58] FERREIRA, A. *et al.* (2009). *Comunicações Audiovisuais – Tecnologias, Normas e Aplicações*. Lisboa: IST Press.

[59] LAVER, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

[60] LINDH, J. & MORRISON, G. S. (2011). “Human versus machine: Forensic voice comparison on a small database of Swedish voice recordings”. Recuperado em 2017, novembro 21, de <https://pdfs.semanticscholar.org/67d2/502ddb50910babc3babbad1517a772a45105.pdf>

[61] FRENCH, P. (1994) “An Overview Of Forensic Phonetics With Particular Reference to Speaker Identification”. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, Vol. 1, No. 2, pp. 169-181. Recuperado em 2017, novembro 21, de <https://journals.equinoxpub.com/index.php/IJSLL/article/view/16593>

[62] NOLAN, F. (2002). “Intonation in Speaker Identification: an experiment on pitch alignment features”. *Forensic Linguistics: The International Journal of Speech,*

Language and the Law, Vol. 9, No. 1, pp. 1-21. Recuperado em 2017, novembro 21, de <https://journals.equinoxpub.com/index.php/IJSL/article/view/2006>

[63] MORRISON, G. S. (2009). “Forensic voice comparison and the paradigm shift”. *Science & Justice*, Vol. 49, No. 4, pp. 298-208. Recuperado em 2017, novembro 11, em <http://www.sciencedirect.com/science/article/pii/S135503060900135X>

[64] DRYGAJLO, A. *et al.* (2015). *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises*. Recuperado em 2017, setembro 9, de http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsar_0.pdf

[65] *iVOCALISE - ivector Voice Comparison and Analysis of the Likelihood of Speech Evidence*, Manual do utilizador, Oxford Wave Research (2016).

[66] FRÜHHOLZ, S. *et al.* (2016). “The effect of narrow-band transmissions on recognition of paralinguistic information from human vocalization”. *IEEE Access*, Vol. 4. Recuperado em 2017, setembro 9, de <http://ieeexplore.ieee.org/document/7555372/>

[67] ROBERTS, L. (2011). “Acoustic effects of authentic and acted distress on fundamental frequency and vowel quality”. Recuperado em 2017, setembro 11, de <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Roberts/Roberts.pdf>

[68] NOLAN, F. (1991). “Forensic phonetics”. *Journal of Linguistics*, Vol. 27, no. 2, 483-493. Recuperado em 2016, janeiro 3, de https://www.jstor.org/stable/4176127?seq=1#page_scan_tab_contents

[69] MARTINS, F. *et al.* (2014). “A Fonética Forense na produção de prova do ordenamento jurídico português: o parâmetro do pré-vozeamento”. *ReVEL*, Vol. 12, No. 23. Recuperado em 2017, outubro 1, de <http://www.revel.inf.br/files/5df7d1a0b1e2bc54db33bba8c89111e5.pdf>

[70] BRAZ, J. (2013). *Investigação criminal: a organização, o método e a prova. Os desafios da nova criminalidade*. Coimbra: Almedina, 3ª Edição.

[71] SAN SEGUNDO, E. & GÓMEZ-VILDA, P. (s/d). “Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings”. Recuperado em 2017, agosto 29, de

http://oa.upm.es/40911/1/INVE_MEM_2015_223834.pdf

[72] SAN SEGUNDO, E. (2015). “Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics”. *International Journal of Speech Language and the Law* Vol. 22, No. 2, pp. 249-253, Recuperado em 2017, julho 7, de

https://www.researchgate.net/publication/283906614_Forensic_speaker_comparison_of_Spanish_twins_and_non-twin_siblings_a_phonetic-acoustic_analysis_of_formant_trajectories_in_vocalic_sequences_glottal_source_parameters_and_cepstral_characteristics

[73] MARTINS, M. R. D. (1998). *Ouvir Falar – Introdução à Fonética do Português*. Lisboa: Editorial Caminho. 3ª Edição.

[74] MALMBERG, B. (1998). *A Fonética*. Lisboa: Edição «Livros do Brasil». Tradução de Oliveira Figueiredo.

[75] PATRÍCIO, P. (s/d). “From the Shepard tone to the Perceptual Melody Auditory Illusion”. Recuperado em 2017, dezembro 4, de

<http://smcnetwork.org/system/files/smc2012-145.pdf>

[76] ZHANG, Z. (2016). “Mechanics of human voice production and control”. *The Journal of the Acoustical Society of America* no. 140, pp. 2614-2635. Recuperado em 2017, dezembro 5, de <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5412481/>

- [77] SEGUNDO, E. S. *et al.* (2017). “Euclidian distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics,” *Forensic Science International*, No. 270, pp. 25–38, 2017.
- [78] FREITAS, S. (2012). *Avaliação Acústica e Áudio Percetiva na Caracterização da Voz Humana*. Dissertação de Doutoramento em Engenharia Biomédica, Faculdade de Engenharia da Universidade do Porto.
- [79] FREITAS, S. *et al.* (2014). “Audio-Perceptual Evaluation of Portuguese Voice Disorders—An Inter- and Intrajudge Reliability Study”. *Journal of Voice*, No. 28(2), pp. 210-215. Recuperado em 2017, dezembro 9, de <https://www.ncbi.nlm.nih.gov/pubmed/24050822>
- [80] CASTILHO, S. I. P. (2014). *Estudo da frequência fundamental relativa na tensão laríngea*. Dissertação de Mestrado em Ciências da Fala e da Audição, Universidade de Aveiro.
- [81] FERNÁNDEZ, E. S. S. (s/d). “Glottal source parameters for forensic voice comparison: An approach to voice quality in twins voices”. 21st Conference of the International Association for Forensic Phonetics and Acoustics, Santander, Espanha. Recuperado em 2012, setembro 25, de http://www.iafpa2012.com/AbstractsPDF/SAN_SEGUNDO_EUGENIA.pdf
- [82] LOAKES, D. (2002). “Individuality of the voice: an analysis of telephone recorded twin pairs”. Proceedings of the 9th Australian International Conference on Speech Science & Technology Melbourne. Recuperado em 2012, outubro 30, de <http://assta.org/sst/sst2002/Papers/Loakes062.pdf>
- [83] WEINRICH, M. & FUCHS, S. (s/d). “Vocal tract morphology can influence speaker specific realizations of phonemic contrasts”. ISSP. Recuperado em 2014, junho 28, de <http://www.sprachwissenschaft.uni->

jena.de/germsprach_multimedia/Downloads/weirich/Weirich_Fuchs_Vocal+tract+morphology.pdf

[84] JOHNSON, K. & AZARA, M. (s/d). “The perception of personal identity in speech: evidence from the perception of twins’ speech”. Manuscrito não publicado. Recuperado em 2014, julho 3, de <http://linguistics.berkeley.edu/~kjohnson/papers/twinPerc.pdf>

[85] ARIYAEINIA, A. *et al.* (2008). “A test of the effectiveness of speaker verification for differentiating between identical twins”. *Science & Justice*, Vol. 48, No. 4, pp. 182-186. Recuperado em 2017, novembro 19, de <http://www.sciencedirect.com/science/article/pii/S1355030608000312>

[86] Praat, Manual de utilizador. http://www.fon.hum.uva.nl/rob/Courses/SpeechRecAndSyn/Literature/praat_manual.pdf, recuperado em 2017, novembro 2.

[87] ALEXANDER, A. *et al.* (2016). “VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features”. *The Speaker and Language Recognition Workshop*: Oxford Wave Research Ltd, 2016.

[88] Bio-Metrics, Manual de utilizador, Oxford Wave Research (2016).

[89] CLEAVER – Oxford Wave Research. Oxford Wave Research. Recuperado de <http://www.oxfordwaveresearch.com/products/cleaver/>, recuperado em 2017, dezembro 31.

[90] Página de suporte do *Biometrics* disponível em <http://www.oxfordwaveresearch.com/products/bio-metrics/> e recuperado em 2018, janeiro 1.

[91] Página de suporte do *Biometrics* disponível em <http://www.oxfordwaveresearch.com/products/vocalise/> e recuperado em 2018, janeiro 8.

[92] GERRATT, B. R. *et al.* (2016). “Comparing Measures of Voice Quality From Sustained Phonation and Continuous Speech”. *Journal of Speech, Language, and Hearing Research*, Vol. 59, pp. 994-1001. Recuperado em 2017, dezembro 6, de <http://jslhr.pubs.asha.org/article.aspx?articleid=2553570>

[93] KURTZ, L. O. (2010). *Tempos máximos de fonação de vogais e capacidade vital em mulheres adultas com nódulos vocais*. Dissertação de Mestrado do Programa de Pós-Graduação em Distúrbios da Comunicação Humana, Área de Concentração em Fonoaudiologia e Comunicação Humana – Clínica e Promoção, Universidade Federal de Santa Maria, Brasil. Recuperado em 2017, dezembro 8, de http://cascavel.ufsm.br/tede/tde_arquivos/26/TDE-2010-05-06T115519Z-2601/Publico/KURTZ,%20LAURA%20OLIVEIRA.pdf

[94] ESCUDERO, P. *et al.* (2009). “A cross-dialect acoustic description of vowels: Brazilian and European Portuguese”. *Journal Acoustic Society of America*, Vol. 3, No. 1236, pp. 1379 – 1393.

[95] BYRNE C. & FOULKES. P (2004). “The mobile phone effect on vowel formants”, *International Journal of Speech, Language and the Law*, Vol. 11, No. 1, pp. 83-102.

[96] BITZER J. *et al.* (2014). “Evaluation results of speaker verification for VoIP transmission with packet loss”. 54th AES International Conference on Audio Forensics.

[97] STORY, B. H. (2016). “The Vocal Tract in Singing”. *The Oxford Handbook of Singing*, G.Welch, D.M. Howard & J. Nix, Eds.

- [98] HARRISON, P. (2004). *Variability of formant measurements*. Tese de Mestrado em Fonética e Fonologia, Universidade de York.
- [99] LAL, A. *et al.* (2009). “Study of landline and mobile phone effect on vowel formants”. XX AII India Forensic Science Conference, Jaipur, Rajastão. Recuperado em 2017, dezembro 12, de https://www.researchgate.net/publication/308918242_STUDY_OF_LANDLINE_AND_MOBILE_PHONE_EFFECT_ON_VOWEL_FORMANTS
- [100] FERNANDES, V. & FERREIRA, A. (2017). “On the relevance of F0, Jitter, Shimmer and HNR acoustic parameters in forensic voice comparisons using GSM, VOIP and contemporaneous high-quality voice recordings”. Atas de conferência de AES Conference on Audio Forensics, EUA, junho de 2017.
- [101] FERREIRA, A. & FERNANDES, V. (2017). “Consistency of the F0, Jitter, Shimmer and HNR voice parameters in GSM and VOIP communication”. Publicação submetida para a conferência DSP 2017, Londres, agosto de 2017.
- [102] REGMAN, R. *et al.* (2015). “Auditory Scale Analysis and Evaluation of Phonemes in MISING Language”. *International Journal of Computer Applications*, Vol. 113, No. 15. Recuperado em 2018, janeiro 27, de https://www.researchgate.net/profile/Rizwan_Rehman/publication/273695128_Auditory_Scale_Analysis_and_Evaluation_of_Phonemes_in_MISING_Language/links/5559a55308ae6fd2d8270772/Auditory-Scale-Analysis-and-Evaluation-of-Phonemes-in-MISING-Language.pdf
- [103] SHANNON, B. J. & PALIWAL K. K. (2003). “A comparative study of filter bank spacing for speech recognition”. Microelectronic Engineering Research Conference. Recuperado em 2018, janeiro 27, de

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.1426&rep=rep1&type=pdf>

[104] ROSENBERG, A. E. *et al.* (2008). “Overview of Speaker Recognition”. *Springer Handbook of Speech Processing*, Benesty, Sondhi and Huand (Editores), Capítulo 36, pp. 725-741.

[105] FERREIRA, A. (2018). “First experiments on speaker identification combining a new shift-invariant phase-related feature (NRD), MFCCs and F0 information”. Artigo submetido para a conferência SIGMAP2018.

[106] RAYNOLDS, D. A. (1997). “Comparison of background normalization methods for text-independent speaker verification”. In *Proceedings of Eurospeech*, pp. 963–966.

[107] RAMACHANDRAN, R. P. *et al.* (2002). Speaker recognition - general classifier approaches and data fusion methods. *Pattern Recognition*, 35(12), pp. 2801–2821.

[108] KINNUNEN, T., HAIZHOU L. (2010). “An overview of text-independent speaker recognition: from features to supervectors”. *Speech Communication Vol. 52*, No. 1, pp. 12-40, 2017.

[109] REYNOLDS, D.A. & CAMPBELL, W. M. (2008). “Text-independent speaker recognition”. *Springer Handbook of Speech Processing*, Benesty, Sondhi and Huand (Editores), Springer. Capítulo 38, pp. 763-781.

[110] RABINER, L. & JUANG, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Inc.

[111] DAVIS, S. B. & MERMELSTEIN, P. (1980). “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”.

IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 28, No. 4, pp. 357–366.

[112] KENNY P. (2006) “Joint factor analysis of speaker and session variability: theory and algorithms”. Technical report CRIM.08/08-14.

[113] DEHAK, N. *et al.* (2011) “Front-end factor analysis for speaker verification”. *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 28, No. 4, pp. 788-798.

[114] PRINCE S. & ELDER J. (2007). “Probabilistic linear discriminant analysis for interferences about identity”. Atas de conferência do *IEEE 11th International Conference on Computer Vision*, pp. 1-8, Rio de Janeiro.

[115] MENDES D. & FERREIRA A. (2012). “Speaker identification using phonetic segmentation and normalized relative delays of source harmonics”. *46th AES International Conference on Audio Forensics*, pp. 215–222. Artigo 9-1.

Anexos

Anexo 1

Apelo à participação de voluntários num estudo linguístico - Informantes gémeos

Boa tarde,

Apela-se à participação de indivíduos gémeos, maiores e do mesmo género, numa investigação a realizar no âmbito do programa doutoral Tecnologias da Linguagem Humana da U.Porto, sobre o impacto da comunicação telefónica no reconhecimento de vozes.

Esta investigação, que agrega investigadores da Faculdade de Letras e Faculdade de Engenharia da U.Porto, além de uma entidade portuguesa na área da investigação forense, pretende realizar um conjunto de gravações planeadas de voz telefónica, em diálogo simulado e usando telefonia fixa e móvel. Pretende-se caracterizar as alterações vocais que têm lugar em resultado das tecnologias de comunicação e gravação, e avaliar o seu impacto na comparação e capacidade de identificação de vozes, quer por humanos, quer por computador.

Para tal, apela à manifestação de interesse em participar, de indivíduos gémeos voluntários. Para além dos dois indivíduos gémeos, maiores e do mesmo género, será também importante a participação adicional de um indivíduo da mesma família e género. Apela-se a resposta para Vânia Fernandes (vaniascfernandes@hotmail.com).

Os resultados serão objeto de tratamento estatístico que omitirá a identidade dos participantes. A investigação decorrerá segundo padrões de ética aplicáveis, designadamente ao nível de sigilo e proteção de dados pessoais.

O trabalho será realizado pela doutoranda Vânia Fernandes e contará com a supervisão do professor Aníbal Ferreira (Fac. Eng.).

Anexo 2

Apelo à participação de voluntários num estudo linguístico - Informantes gémeos

Apela-se à participação de indivíduos gémeos, maiores e do mesmo género, numa investigação sobre o impacto da comunicação telefónica no reconhecimento de vozes, a realizar no âmbito do programa doutoral 3º Ciclo de Estudos em Ciências da Linguagem da U.Porto.

Esta investigação, que reúne investigadores da Faculdade de Letras, Faculdade de Psicologia e Ciências da Educação e da Faculdade de Engenharia da U.Porto, pretende realizar um conjunto de gravações planeadas de voz telefónica, em diálogo simulado e usando telefonia fixa e móvel. Pretende-se caracterizar as alterações vocais que têm lugar em resultado das tecnologias de comunicação e gravação, e avaliar o seu impacto na comparação e capacidade de identificação de vozes, quer por humanos, quer por computador.

Para tal, apela-se à manifestação de interesse em participar de indivíduos gémeos. Para além dos dois indivíduos gémeos, maiores e do mesmo género, será também importante a participação adicional de um indivíduo da mesma família e género, de preferência na mesma faixa etária. Apela-se a resposta para Vânia Fernandes (vaniascfernandes@hotmail.com).

A investigação decorrerá segundo padrões aplicáveis de ética, designadamente ao nível de sigilo e protecção de dados pessoais.

O trabalho será realizado pela doutoranda Vânia Fernandes e contará com a supervisão do Professor Doutor Aníbal Ferreira, Prof. Associado da Fac. Engenharia (U.Porto).

Anexo 3

Diálogo Orientado – adultos

E – Boa tarde/ Bom dia. Como vai?

O - Boa tarde/ Bom dia/Olá. Tudo bem, obrigado(a). Com o(a) senhor(a)?

E – Tudo ótimo, obrigado(a). Gostaria que me dissesse, por favor, o seu nome completo, data e local de nascimento.

O – XXXXXXXXXX, XXXXXXXXXX, XXXXXX.

E – Que idade tem?

O – Tenho XXXXXXXXXXXX anos.

E – Como se chamam o seu pai e a sua mãe?

O – O meu pai chama-se XXXXXXXXXXXX e a minha mãe XXXXXXXXXXXXXXXX.

E – Muito bem. E o seu irmão gémeo e o familiar próximo que faz parte do estudo?

O – Chamam-se XXXXXXXXXXXXXXXX e XXXXXXXXXXXXXXXX.

E – Que dia é hoje?

O – Hoje é XXXXXX de XXXXXX de 2014 / (dia da semana).

E – Que horas são?

O – São XXX horas / XXXXX horas e XXXXXX / Faltam XXXXX para as XXXXXX.

E – Hoje está sol / chuva / ... Qual é o seu clima favorito e o clima que detesta?

O – O meu clima favorito é XXXXXXXX e o que detesto é XXXXXXXXXXXXXXXXXXXX.

E – Podia dizer-me o seu estado civil?

O – Sou solteiro(a) / casado(a) / viúvo(a) / ...

E – Por favor, defina-me as palavras “namorado” e “vizinho”.

O – Um namorado é XXXXXXXXXXXXXXXXXXXX e um vizinho é XXXXXXXXXXXXXXXXXXXX...

E – Em que escola andou? Gostava de estudar? Sim ou não?

O – Andei na Escola XXXXXXXXXXXXX. Sim, porque XXXXXXXXXXXXX. / Não, porque XXXXXXXXXXXXX.

E – Qual é a sua profissão?

O – Sou XXXXXXXXXXXXXXXX.

E – E o(a) seu(sua) irmão(ã) e familiar próximo?

O – O(A) meu(minha) irmão(ã) é XXXXXXXXXXXXXXXX e o familiar é XXXXXXXXXXXXXXXX.

E – Como define um bom profissional e um mau profissional?

O – Considero que o mau profissional é XXXXXXXXXXXXXXXXXXXXXXXX e o bom é XXXXXXXXXXXXXXXX.

E – Gosta de chocolate?

O – Sim/ Não.

E – E de beterraba?

O – Sim/ Não.

E – Quase a terminarmos, gostaria que dissesse “aaaaaaa”, “eeeeee”, “iiiiiii”, “ooooooo”, “uuuuuu”, com a duração de três segundos cada.

O – Aaaaaaaaaaaaaa, eeeeeeeeeeee, iiiiiiiiiiiiii, ooooooooooooo, uuuuuuuuuu.

E – Obrigado(a). Gostava que lesse o texto que lhe demos, por favor.

(Neste momento, o orador lê o texto).

O vento norte e o sol discutiam qual dos dois era o mais forte, quando sucedeu passar um viajante envolto numa capa. Ao vê-lo, põem-se de acordo em como aquele que primeiro conseguisse obrigar o viajante a tirar a capa seria considerado o mais forte. O vento norte começou a soprar com muita fúria, mas quanto mais soprava, mais o viajante se aconchegava à sua capa, até que o vento norte desistiu. O sol brilhou, então, com todo o esplendor, e, imediatamente, o viajante tirou a capa. O vento norte teve assim de reconhecer a superioridade do sol.

E – Muito obrigado(a) pela participação e até à próxima. Boa tarde.

Anexo 4

Diálogo Orientado – Público jovem

E – Olá! Tudo bem?

O - Boa tarde/ Bom dia/Olá. Tudo bem, obrigado(a).

E – Ótimo! Gostaria que me dissesse, por favor, o teu nome completo, data e local de nascimento.

O – XXXXXXXXXXX, XXXXXXXX, XXXXXX.

E – Que idade tens?

O – Tenho XXXXXXXXXXXX anos.

E – Como se chamam o teu pai e a tua mãe?

O – O meu pai chama-se XXXXXXXXXXX e a minha mãe XXXXXXXXXXXXXXXX.

E – Muito bem. E o teu irmão gémeo e o familiar próximo que faz parte do estudo?

O – Chamam-se XXXXXXXXXXXXXXXX e XXXXXXXXXXXXXXXX.

E – Que dia é hoje?

O – Hoje é XXXXXX de XXXXXX de 2015 / (dia da semana).

E – Que horas são?

O – São XXX horas / XXXXX horas e XXXXXX / Faltam XXXXX para as XXXXXX.

E – Qual é o teu passatempo favorito?

O – O meu passatempo favorito é XXXXXXXXXX.

E – Para onde gostavas de viajar?

O – Para XXXXXXXX.

E – Em que escola andas? Gostas de estudar? Sim ou não?

O – Ando na Escola XXXXXXXXXXXX. Sim, porque XXXXXXXXXXXX. / Não, porque XXXXXXXXXXXX.

E – O que gostavas de ser?

O – Gostava de ser XXXXXXXX.

E – E o teu (tua) familiar próximo(a) que veio contigo e pais? O que fazem?

O – O meu pai é XXXXXXXXXXXX, a minha mãe é XXXXX e o familiar é XXXXXXXXXXXX.

E – Gostas de chocolate?

O – Sim/ Não.

E – E de beterraba?

O – Sim/ Não.

E – Quase a terminarmos, gostaria que disseses “aaaaaaa”, “eeeeee”, “iiiiiii”, “oooooo”, “uuuuuu”, com a duração de três segundos cada.

O – Aaaaaaaaaaaaaa, eeeeeeeeeeee, iiiiiiiiiiiiii, ooooooooooooo, uuuuuuuuuu.

E – Obrigado(a). Gostava que lesse o texto que te demos, por favor.

(Neste momento, o orador lê o texto).

O vento norte e o sol discutiam qual dos dois era o mais forte, quando sucedeu passar um viajante envolto numa capa. Ao vê-lo, pôem-se de acordo em como aquele que primeiro conseguisse obrigar o viajante a tirar a capa seria considerado o mais forte. O vento norte começou a soprar com muita fúria, mas quanto mais soprava, mais o viajante se aconchegava à sua capa, até que o vento norte desistiu. O sol brilhou, então, com todo o esplendor, e, imediatamente, o viajante tirou a capa. O vento norte teve assim de reconhecer a superioridade do sol.

E – Muito obrigado(a) pela participação e até à próxima. Boa tarde.

Anexo 5

Recolha de diálogo – gémeos + familiar próximo

Sujeito 1:

Sujeito 2:

Familiar próximo:

Grau de parentesco entre os gémeos e o familiar

próximo: _____

Sexo: M ☐ F ☐

Data de nascimento gémeos: ____/____/____

Local de nascimento gémeos:

Data de nascimento familiar próximo: ____/____/____

Local de nascimento familiar próximo:

Morada (Freguesia, Concelho, Distrito):

Indivíduo 1: _____

Indivíduo 2: _____

Familiar próximo: _____

Habilitações:

Indivíduo 1: _____

Indivíduo 2: _____

Familiar próximo: _____

L1 dos sujeitos: _____

L1 do

L1 da

pai: _____

mãe: _____

L1 dos familiar próximo: _____

L1 do

L1 da

pai: _____

mãe: _____

Patologias

<i>Sujeito 1</i>	<i>Sujeito 2</i>

Familiar próximo

--

Particularidades

<i>Sujeito 1</i>	<i>Sujeito 2</i>

<i>Familiar próximo</i>

Observações:

Nome dos ficheiros:

- *Sujeito 1 –*
- *Sujeito 2 –*
- *Familiar próximo –*

Materiais utilizados:

Data de recolha: ____/____/____

Anexo 6

Declaração de consentimento informado – adultos

Declaração de consentimento informado

Eu, _____,

declaro que autorizo a gravação de uma conversação telefónica utilizando canais GSM e VoIP, com objetivo puramente académico de investigar correspondências estatísticas entre as vozes de indivíduos gémeos e de familiares próximos do mesmo género, sobretudo com base em parâmetros acústicos e perceptivos de avaliação qualitativa da voz, analisados pela estudante de Doutoramento em Ciências da Linguagem, variante das Tecnologias da Linguagem Humana, Vânia Sofia Costa Fernandes, da Faculdade de Letras da Universidade do Porto. É coordenador deste estudo o Professor Doutor Aníbal Ferreira, da Faculdade de Engenharia da Universidade do Porto. Os responsáveis por este estudo garantem plena confidencialidade dos dados obtidos e asseguram que qualquer publicação que resulte do estudo apresentará somente resultados tratados estatisticamente e impossibilitando a identificação dos oradores abrangidos.

Porto, de março de 2016

Anexo 7

Declaração de consentimento informado – menores

Declaração de consentimento informado

Eu, _____,
encarregado(a) de educação de _____,

e de _____, declaro que autorizo a gravação de uma conversa  o telef  nica utilizando canais GSM e VoIP, com objetivo puramente acad  mico de investigar correspond  ncias estat  sticas entre as vozes de indiv  duos g  meos, sobretudo com base em par  metros ac  sticos e percetivos de avalia  o qualitativa da voz, analisados pela estudante de Doutoramento em Ci  ncias da Linguagem, variante das Tecnologias da Linguagem Humana, V  nia Sofia Costa Fernandes, da Faculdade de Letras da Universidade do Porto.    coordenador deste estudo o Professor Doutor An  bal Ferreira, da Faculdade de Engenharia da Universidade do Porto. Os respons  veis por este estudo garantem plena confidencialidade dos dados obtidos e asseguram que qualquer publica  o que resulte do estudo apresentar   somente resultados tratados estatisticamente e impossibilitando a identifica  o dos oradores abrangidos.

Porto, ____ de mar  o de 2016

(Encarregado de Educa  o)