Os efeitos da raiva na identificação automatizada de locutores baseada em espectros de longo prazo

Ortega-Rodríguez, M.[1]; **Solís-Sánchez, H.**[1]; **Valverde-Méndez, D.**[1,2]; **Venegas-Li, A.**[1,3]

[1] Escola de Física e Centro de Investigação Geofísica, Universidade da Costa Rica, San José, Costa Rica, {manuel.ortega, hugo.solis}@ucr.ac.cr
[2] Department of Physics, Princeton University, Princeton, Estados Unidos, dsmendez@princeton.edu
[3] Physics Department, University of California at Davis, Davis, Estados Unidos, avenegasli@ucdavis.edu

Resumo

A identificação forense de locutores tem considerado tradicionalmente abordagens ao problema baseadas na análise de espectros a longo prazo (várias dezenas de segundos de duração). Essas abordagens demonstraram ser especialmente robustas, no sentido de que continuam funcionando bem mesmo se as gravações forem curtas; além disso, o método não é sensível a mudanças na intensidade sonora da amostra, e mantém um desempenho adequado na presença de ruído e largura de banda limitada. Por todos esses motivos, constitui uma das técnicas preferidas para a identificação forense, juntamente com a análise de formantes, velocidade da fala e determinação da frequência fundamental. No entanto, verificou-se que o estado de raiva produz uma distorção importante no sinal acústico para efeitos da análise de espectros de fala a longo prazo. Mesmo que o nível de raiva seja apenas moderado, há um desvio dos resultados quantitativos da identificação forense de locutores que representa 33% da distância em direção a um locutor totalmente diferente (no espaço de correlação entre amostras). Portanto, conclui-se que é importante ter cautela ao aplicar este método.

Palavras-chave: identificação forense de locutor e locutora, espectros a longo prazo, acústica forense, distorções emocionais, raiva.

PACS: 43.72.Uv, 43.72.Ar, 43.72.Fx.

The effects of anger on automated long-term-spectra based speaker-identification

Abstract

Forensic speaker identification has traditionally considered approaches based on long-term (a few tens of seconds) spectra analysis as especially robust. This is because they work well for short recordings, are not sensitive to changes in the intensity of the sample, and continue to function in the presence of noise and limited passband. Because of this, the long-term spectra approach is one of the preferred tools for forensic speaker identification, in addition to formant analysis, speed of speech, and determination of the fundamental frequency. However, we find that anger induces a significant distortion of the acoustic signal for long-term spectra analysis purposes. Even moderate anger offsets speaker identification results by 33% in the direction of a different speaker altogether (in the space of sample correlations). Therefore, caution should be exercised when applying this tool.

Keywords: automated speaker identification, long term spectra, forensic acoustics, emotional distortions, anger.

Esta é a versão em português do artigo também publicado em espanhol (10.55753/aev.v37e54.28). DOI: 10.55753/aev.v37e54.193

1. INTRODUÇÃO

O propósito deste artigo é estudar como determinar quantitativamente o efeito das distorções causadas por estados emocionais (em particular, o de raiva) na análise do espectro médio da fala a longo prazo (conhecido em inglês como *Long Term Spectra*, LTS) para fins de *identificação forense de locutores* (IFL). (Exemplos de artigos de especialistas no tema da IFL são fornecidos por Hollien [1] e Hollien [2].) A abordagem desta pesquisa é realizada por meio de uma metodologia cuidadosa e reprodutível, explicada mais adiante. O objetivo do processo da IFL é identificar uma pessoa falante por meio da análise de sua voz, geralmente sob condições que não são ideais. Um dos desafios fundamentais com os quais a IFL tem que lidar é determinar se a variabilidade intra-locutor é menor que a variabilidade interlocutor (o que é claramente desejável), e como essa relação se mantém para diferentes condições (Hollien [3]). Entre as condições mais comuns, podem-se mencionar as distorções tecnológicas devido ao equipamento usado para fazer as gravações, assim como as distorções ambientais causadas por ruído ou sons ásperos de fundo.

Em particular, os próprios falantes podem ser a fonte da distorção, já que podem haver uma variedade de sentimentos, como medo, raiva e ansiedade (uma situação provável, por exemplo, na IFL, quando o falante poderia estar cometendo um crime). Essas emoções desencadeiam uma modificação na produção da fala que se manifesta como uma mudança nos valores dos parâmetros do sinal (como as frequências e a velocidade da fala) (Williams e Stevens [4]; Banse e Scherer [5]; Johnstone [6]). A produção da voz consiste em pulsos de ar causados pela vibração das cordas vocais (sendo depois modificados pelo trato vocal supralaríngeo), de forma que os fatores dominantes de vocalização são os padrões de respiração e a tensão variante dos músculos envolvidos no processo. Como esses fatores têm uma alta correlação com as emoções, é muito provável que tais mudanças dos fatores sejam detectáveis na onda acústica (Scherer [7]).

Este tópico, no entanto, não foi estudado de maneira extensiva para nenhum idioma, em grande parte devido a restrições éticas e metodológicas que tornam complicada a produção controlada de emoções fortes. O consenso (Johnstone [6]) tem sido que condições controladas de laboratório são viáveis apenas para estados emocionais de baixa intensidade, para os quais é mais difícil notar uma mudança na eficácia da IFL. Outra dificuldade reside em como induzir a emoção desejada. Martin [8] oferece uma visão geral de algumas possíveis técnicas usadas para gerar emoções específicas, como música e imagens emotivas, bem como técnicas de autogeração, como o uso da imaginação e memórias. Estes métodos são classificados de acordo com a forma como as emoções são produzidas.

Para os propósitos desta pesquisa, selecionou-se o método de recordações autobiográficas autoinduzidas. Nesta técnica, pede-se aos participantes (todos homens, como explicado mais abaixo) que se lembrem de eventos emotivos com o objetivo de gerar a emoção desejada. Embora esse não seja o único método aplicável ao problema em questão, foi escolhido por sua simplicidade e porque permitia aos participantes ter privacidade enquanto faziam as gravações. Além disso, a qualidade de experimento cego era assegurada, fazendo com que os próprios participantes determinassem seu nível de raiva (em uma escala numérica), como descrito mais adiante.

2. FUNDAMENTOS

Nesta seção, será discutida a lógica do funcionamento do processo de identificação forense de locutores por meio do emprego da análise de *espectro a longo prazo* (LTS), processo esse que será designado doravante pela abreviação IFLLTS. A despeito da existência de muitos marcadores (também conhecidos como "vetores") que auxiliam na distinção de gravações de diferentes falantes, desde o trabalho pioneiro de Hollien e Majewski [9], um dos métodos mais comumente empregados na IFL é a análise LTS (Kinnunen *et al.* [10]; Ortega-Rodríguez *et al.* [11]). A análise LTS quantifica o timbre médio temporal da voz, propriedade acústica que permite a um ouvinte distinguir, por

exemplo, um clarinete de um violino tocando a mesma nota musical com igual intensidade sonora. Obtém-se essa distribuição por meio do cálculo da Transformada de Fourier do sinal em um longo período, como cerca de 30 segundos, permitindo que o falante pronuncie várias vezes quase todos os sons do espanhol, abrangendo assim todo o espaço de fase sonoro.

Este vetor tem sido extensivamente estudado quanto à sua eficiência na identificação do locutor, revelando-se um dos mais confiáveis, especialmente por funcionar mesmo na presença de ruído e largura de banda limitada (Hollien [3]).

Um desafio no uso deste vetor é definir a correlação entre dois espectros. Existem diversas abordagens para esse problema, incluindo a atribuição de um número específico a cada conjunto de dados de LTS conforme algum algoritmo, ou até a inspeção visual de gráficos. Contudo, para os propósitos deste artigo, é necessária uma abordagem mais sofisticada. Foram considerados dois coeficientes de correlação: a Desvio Padrão das Diferenças de Distribuição (SDDD, *Standard Deviation of the Differences Distribution*) (Harmegnies [12]) e o coeficiente de correlação cruzada de Bravais-Pearson, R (Stanton [13]). Experimentos exploratórios realizados pelo nosso grupo, sem o componente de raiva, indicaram que o coeficiente de Bravais-Pearson é mais eficaz para esta pesquisa, sendo, portanto, selecionado. O critério para essa escolha foi o seguinte: um método é considerado superior a outro quando a correlação entre amostras do mesmo falante é mais próxima de 1, e a correlação cruzada (diferentes falantes) é mais distante de 1.

No método de Bravais-Pearson, o espectro da análise LTS é visto como um vetor de dimensão k, com um total de k canais de frequência. O espectro pode ser definido como:

$$S \equiv (S_1, S_2, \ldots, S_i, \ldots, S_k)\,, \qquad (1)$$

em que S_i representa o nível da i-ésima componente de frequência (Harmegnies [12]). Nesse contexto, o coeficiente R mede a relação entre as duas amostras LTS. R é definido por:

$$R_{SS'} \equiv \frac{1}{k} \frac{\sum_{i=1}^{k} (S_i - M_S)\,(S_{i'} - M_{S'})}{\sigma_S\, \sigma_{S'}}\,, \qquad (2)$$

em que M_S e $M_{S'}$ referem-se às médias de cada espectro, enquanto σ_S e $\sigma_{S'}$ são os respectivos desvios padrão. O coeficiente de Bravais-Pearson tem várias vantagens, como uma grande capacidade discriminatória e a independência de diferenças de intensidade relativa entre os dois espectros (Harmegnies [12]), permitindo a comparação de gravações realizadas sob distintas condições ambientais ou de posicionamento do microfone.

A maioria dos estudos sobre a relação entre fala e emoções concentra-se na capacidade de distinguir diferentes estados emocionais do falante por meio da análise do sinal acústico (Williams e Stevens [4]; Fuller [14]; Scherer [7]; Johnstone [6]; Harnsberger *et al.* [15]). Em particular, a análise LTS tem sido utilizada para identificar estados emocionais e de depressão (Pittam [16]), embora a filtragem humana tenha sido por vezes utilizada no processo (Banse e Scherer [5]).

Menos comum é a investigação de como as emoções ou a intenção deliberada de alterar a voz afetam a IFL. Rodman e Powell [17] sugeriram e planejaram estudos sobre os efeitos de mascaramento na IFL, mas não os realizaram. Mais relacionado a este artigo, Hollien e Majewski [9] estudaram o efeito do estresse (induzido por eletrochoques nos sujeitos) na análise LTS do sinal, sem encontrar impactos significativos para a realização adequada da IFL. Para além do exposto, não existem, tanto quanto é do conhecimento dos autores, estudos sobre o efeito das emoções no IFLLTS.

3. MATERIAIS E MÉTODOS

Esta seção descreverá como a população do estudo foi definida e como os respectivos registros foram realizados.

3.1 Gravações dos sujeitos

O problema da IFL possui muitas variáveis. Considere-se, por exemplo, o gênero, a origem geográfica e a idade dos sujeitos (Hertrich e Ziegelmayer [18]; Linville [19]; Hollien e Majewski [9]; Pittam [16]; Yüksel e Gündüz [20]). Por essa razão, escolheram-se sujeitos com características semelhantes. Assim, buscou-se destacar o efeito da emoção sobre outras variáveis, reduzindo a complexidade geral do problema. Os sujeitos atendiam aos seguintes critérios: todos eram homens, com idades entre 18 e 25 anos, e sua origem geográfica era a Grande Área Metropolitana do Vale Central da Costa Rica (essa região é composta pelas quatro maiores cidades do país, localizadas na região central, que possui a maior densidade populacional). Todos os sujeitos cursaram o ensino fundamental nessa região.

3.2 Condições de gravação

As condições de gravação foram homogeneizadas tanto quanto possível, com o objetivo de obter os melhores resultados. Cada amostra de fala foi gravada em um local privado onde os sujeitos se sentissem confortáveis. Além disso, solicitou-se que a fala fosse fluente e espontânea, não recitada ou decorada. Utilizaram-se microfones de *smartphone* da melhor qualidade possível, como os de iPhone e Samsung Galaxy.

Quanto ao estado emocional, houve dois tipos de gravação: gravações normais, nas quais se pedia aos sujeitos que falassem sobre suas vidas cotidianas, a fim de ter o menor grau de resposta emocional possível; e gravações com raiva, nas quais se pedia aos sujeitos que induzissem um estado de raiva descrevendo uma situação pessoal que os tivesse irritado. A entrevistadora deixou o sujeito sozinho nesta parte para evitar que se inibisse.

A duração das amostras foi de 45 e 60 segundos para os casos normal e de raiva, respectivamente (o tempo adicional nos casos de raiva permitia um período de transição).

4. METODOLOGIA

Um total de 32 sujeitos (que satisfaziam os critérios de seleção mencionados anteriormente) foram entrevistados. Esse número foi escolhido para ter uma amostra estatisticamente significativa (National Institute of Standards and Technology [21]) da Grande Área Metropolitana do Vale Central da Costa Rica. Cada entrevista consistiu em três gravações: para 16 dos 32 entrevistados, implementou-se a seguinte ordem de gravação: normal-normal-irritado. Para os outros 16 entrevistados, utilizou-se a ordem normal-irritado-normal. A motivação por trás de fazer dois tipos de ordenamento das gravações é a de reduzir os efeitos de possíveis erros sistemáticos devidos à ordem de amostragem.

Os sujeitos foram levados a um local privado onde um conjunto de instruções foi lido para eles; cada sujeito realizou sua gravação separadamente dos demais. A entrevistadora explicou a cada sujeito que a experiência fazia parte de um projeto de pesquisa da Universidade da Costa Rica e que o conteúdo das gravações não seria usado ou ouvido. Foi enfatizado que apenas as propriedades acústicas das amostras eram de interesse, e que o nível de irritação seria determinado unicamente por meio de autoavaliação própria no final do processo de gravação (tornando o processo um experimento cego). Além disso, a natureza do projeto foi descrita apenas em termos muito gerais para que a produção de voz fosse o mais natural possível.

No caso das gravações normais, os falantes receberam a instrução de falar por 45 segundos sobre suas vidas (por exemplo, seu dia, seu animal de estimação, um evento recente, etc.). No caso das gravações com irritação, pediu-se que falassem durante um minuto sobre algo que os irritasse, como alguém com quem tivessem problemas ou algum evento que os tivesse irritado muito. Indicou-se aos sujeitos que não fingissem a emoção (por exemplo, forçando-a elevando a voz). Eles foram deixados sozinhos com o gravador e instruídos a falar quando estivessem prontos. Após concluir a gravação com irritação, pediu-se a cada participante que classificasse seu nível de irritação em uma escala

numérica de 1 a 5, em que "1" significava "não consegui ficar totalmente irritado", "3" significava "moderadamente irritado" e "5" significava "furioso".

Note-se que, quando se usou o primeiro ordenamento de gravações (normal-normal-irritado), os sujeitos não sabiam sobre a parte de irritação até a última gravação, enquanto que no outro ordenamento (normal-irritado-normal), o falante já estava ciente desse aspecto da pesquisa durante sua última gravação. Como mencionado anteriormente, ambos os ordenamentos foram utilizados e promediados para reduzir qualquer possível viés sistemático.

5. PROCESSAMENTO DE DADOS

O processamento de dados no qual a presente investigação está baseada foi desenvolvido pelo nosso grupo (Ortega-Rodríguez *et al.* [11]) estudando a IFL (sem o componente de raiva) e tem sido extensivamente testado e otimizado para assegurar os melhores resultados de identificação. Por otimização, referimo-nos a testar diferentes tipos de janelamento temporal (por exemplo, retangular, gaussiana, Welch, Hanning, Hamming etc.) e diferentes taxas de amostragem para determinar quais funcionam melhor de acordo com o critério mencionado na Seção 2.

O processamento de dados correspondente a este artigo pode ser resumido da seguinte forma. Segmentos de 30 segundos são obtidos a partir das gravações originais. No caso das gravações com estado de raiva, esse procedimento é particularmente importante para eliminar a parte de transição do estado normal para o estado de raiva. Em seguida, utilizou-se o *software* de processamento de áudio Audacity 2.1.0 (Audacity Team [22]) para obter a Transformada Rápida de Fourier (FFT, do inglês *Fast Fourier Transform*) para cada gravação. Uma janela de Hanning foi usada com uma amostragem de 4096 valores. A utilização desse parâmetro deve-se ao fato de que demonstrou dar os melhores resultados em nossos estudos prévios exploratórios. Cada FFT foi salva como um arquivo de texto para continuar sendo processada. Nas Figuras 1 e 2, podem ser vistas amostras desses espectros para os casos normal e irritado, respectivamente.

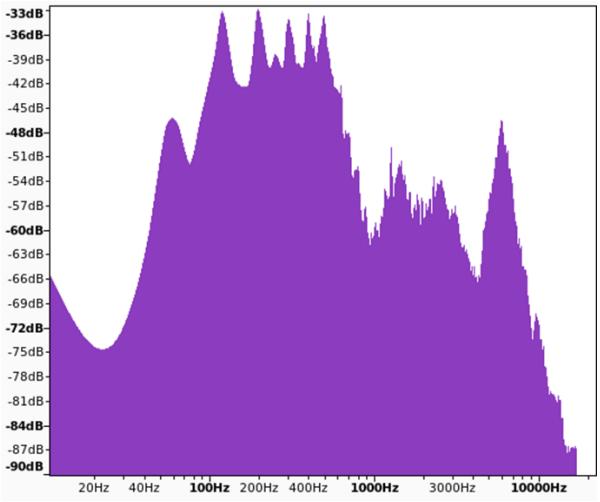

Figura 1: Espectro de uma das gravações correspondentes a fala normal (isto é, sem raiva). A duração da amostra é de 30 segundos, e o espectro foi produzido utilizando uma janela Hanning com 4096 valores.

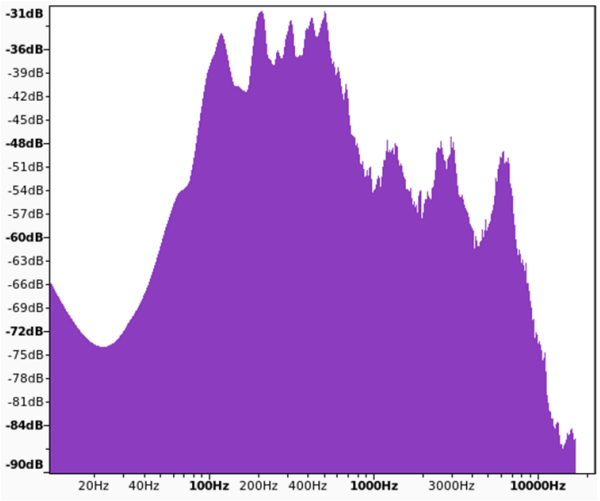

Figura 2: Espectro de uma das gravações correspondentes a fala com raiva de nível 4 (ver a Tabela 2). A duração da amostra é de 30 segundos, e o espectro foi produzido utilizando uma janela Hanning com 4096 valores. O falante é o mesmo para o caso da Figura 1.

Posteriormente, um programa em C++ calculou o coeficiente de correlação de Bravais-Pearson para as amostras. Finalmente, procedeu-se à obtenção da média, do desvio padrão (DP) da amostra e do erro padrão da média (EP) para as medições dos coeficientes de correlação. Foram efetuados os dois casos: normal-normal-irritado e normal-irritado-normal.

6. RESULTADOS E DISCUSSÃO

A Tabela 1 mostra os resultados dos cálculos descritos na seção anterior.

Tabela 1: Resultados estatísticos do coeficiente de correlação de Bravais-Pearson *R* para o caso intra-locutor. Um total de 32 sujeitos participou para o contraste entre os casos normal-normal e normal-irritado.

	Coeficiente de correlação, caso normal-normal	Coeficiente de correlação, caso normal-irritado
Média	0,950	0,934
Desvio padrão	0,028	0,037
Erro padrão da média	0,005	0,005

Após entrevistar os 32 sujeitos, procedeu-se ao cálculo do coeficiente de correlação de Bravais-Pearson *R* para obter a correlação entre fala normal e normal, e a correlação entre fala normal e irritada (intra-locutor). Como se pode observar na tabela, existe uma diferença notável entre os dois valores médios. Para ter uma ideia de quão significativa é essa diferença, é útil compará-la com os resultados do trabalho mencionado do nosso grupo (Ortega-Rodríguez *et al.* [11]), que é uma versão simplificada do processo descrito neste artigo (já que o elemento de raiva não estava presente), embora as condições experimentais fossem as mesmas. Nesse trabalho, o coeficiente de correlação médio *R* entre dois locutores diferentes foi medido em 0,890 com um EP de 0,010, enquanto a correlação entre medições do mesmo locutor teve um *R* médio de 0,955 com um EP de 0,005.

Ao comparar estas três correlações: 0,950 (normal-normal, mesmo falante), 0,934 (normal-irritado, mesmo falante) e 0,890 (falantes diferentes, ambos em modo normal), conclui-se que os efeitos da raiva desviam o sinal em uma significativa proporção de 33% em direção a um falante diferente. Isso é especialmente notável considerando que o nível médio de raiva autorelatado foi de apenas 2,9 numa escala de 1 a 5, indicando que, em média, os sujeitos estavam apenas moderadamente irritados. A distribuição do nível de raiva para os sujeitos é apresentada na Tabela 2, e os respectivos resultados estatísticos encontram-se na Tabela 3.

Tabela 2: Raiva autorelatada pelos sujeitos; 1 significa "não consegui ficar totalmente irritado", 3 significa "moderadamente irritado", e 5 significa "furioso".

Quantidade de sujeitos	Nível de raiva autorelatado
5	4
18	3
9	2

Tabela 3: Resultados estatísticos para os dados da Tabela 2.

	Nível de raiva autorelatado
Média	2,90
Desvio padrão	0,65
Erro padrão da média	0,12

É esperado que níveis mais altos de raiva gerem maiores efeitos na IFLLTS. Para testar essa hipótese, os resultados foram filtrados, deixando apenas os 5 indivíduos mais irritados (aqueles com nível 4 na escala de 1 a 5). A Tabela 4 mostra os resultados obtidos com essa condição. No caso de raiva forte, o desvio da normalidade está próximo de 50% do caminho para outro falante.

Tabela 4: Resultados estatísticos intra-locutor para o coeficiente de correlação de Bravais-Pearson para o caso das cinco gravações com maior raiva. Como esperado, a raiva forte tem um efeito maior na IFLLTS do que a raiva moderada.

	Coeficiente de correlação, caso normal-normal	Coeficiente de correlação, caso normal-irritado
Média	0,950	0,922
Desvio padrão	0,030	0,047
Erro padrão da média	0,013	0,015

Isso demonstra que os efeitos da raiva na IFLLTS realmente aumentam à medida que o nível de raiva se eleva.

7. CONCLUSÕES

Embora alguns autores elogiem o método da IFLLTS por ser robusto diante do estresse do falante (Hollien e Majewski [9]), foi encontrado na presente pesquisa que existe uma distorção significativa na voz humana devido à raiva para efeitos da IFLLTS. Mesmo quando a resposta emocional dos participantes foi moderada, observa-se uma diferença apreciável nos coeficientes de correlação entre os casos de gravações normal-normal e normal-irritado. A raiva moderada desvia os resultados da IFL em 33% na direção de outro falante. É importante ressaltar que esses resultados foram obtidos com um método totalmente automatizável, oferecendo uma abordagem objetiva independente dos erros humanos de percepção. O método também evita avaliar a sinceridade dos participantes, estando, portanto, em conformidade com o código de prática da Associação Internacional de Fonética e Acústica Forense (The International Association for Forensic Phonetics and Acoustics [23]).

Os resultados deste artigo são relevantes para a investigação forense, já que a análise LTS tem sido tradicionalmente considerada um vetor robusto na IFL, especialmente por não ser sensível a mudanças na intensidade sonora da fala, funcionar bem para gravações curtas e continuar operante na presença de ruído e larguras de banda limitadas. No entanto, os resultados deste artigo indicam que se deve ter cuidado ao usar a IFLLTS ao calcular razões de verossimilhança (*likelihood ratios*) em um contexto de raiva, mesmo que essa raiva não seja intensa. Em aplicações forenses, recomenda-se sempre registrar o grau de raiva do falante, tendo sempre em mente os valores de distorção obtidos neste artigo como referência.

Uma vez que outras emoções também podem afetar significativamente a eficácia da IFLLTS, seu estudo em processos de automatização é amplamente justificado e recomendado. O estudo também poderia ser estendido a mulheres ou a falantes de outros idiomas.

AGRADECIMENTOS

Este trabalho recebeu apoio do projeto 805-B2-175 da Vice-Reitoria de Pesquisa da Universidade da Costa Rica, bem como do Centro de Pesquisas Geofísicas da mesma universidade.

CONTRIBUIÇÃO DOS AUTORES E AUTORAS

Todos os autores trabalharam conjuntamente na geração do conceito geral deste artigo e na definição de sua metodologia. Todos discutiram e aprovaram os resultados.

Em particular, D. Valverde-Méndez e A. Venegas-Li foram responsáveis pelas gravações, D. Valverde-Méndez realizou o processamento de dados e fez o primeiro rascunho do artigo (incluindo a busca de antecedentes), enquanto M. Ortega-Rodríguez ficou encarregado da redação final.

CONFLITO DE INTERESSES

Os autores declaram não ter conflito de interesses em relação ao conteúdo deste artigo.

REFERENCIAS

1. HOLLIEN, Harry. Barriers to Progress in Speaker Identification with Comments on the Trayvon Martin Case. *Linguistic Evidence in Security, Law and Intelligence*, University Library System, University of Pittsburgh, v. 1, n. 1, p. 76–98, dez. 2013. ISSN 2327-5596. doi: 10.5195/lesli.2013.3.

2. HOLLIEN, Harry. An Approach to Speaker Identification. *Journal of Forensic Sciences*, Wiley, v. 61, n. 2, p. 334–344, fev. 2016. doi: 10.1111/1556-4029.13034, pMID: 27404606.

3. HOLLIEN, Harry Francis. *Forensic Voice Identification*. Londres, Inglaterra: Academic Press, 2002. ISBN 0123526213.

4. WILLIAMS, Carl E.; STEVENS, Kenneth N. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, Acoustical Society of America (ASA),

v. 52, n. 4B, p. 1238–1250, out. 1972. doi: 10.1121/1.1913238.

5. BANSE, Rainer; SCHERER, Klaus R. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, American Psychological Association (APA), v. 70, n. 3, p. 614–636, 1996. doi: 10.1037/0022-3514.70.3.614.

6. JOHNSTONE, Tom. *The effect of emotion on voice production and speech acoustics*. Tese (PhD) — University of Western Australia & University of Geneva, Perth, Australia, 2001. doi: https://doi.org/10.31237/osf.io/qd6hz.

7. SCHERER, Klaus R. Voice, Stress, and Emotion. In: ____. *Dynamics of Stress: Physiological, Psychological and Social Perspectives*. 1. ed. [S.l.]: Springer US, 1986. p. 157–179. ISBN 978-1-4684-5122-1. doi: 10.1007/978-1-4684-5122-1_9.

8. MARTIN, Maryanne. On the induction of mood. *Clinical Psychology Review*, Elsevier BV, v. 10, n. 6, p. 669–697, jan. 1990. ISSN 1873-7811. doi: 10.1016/0272-7358(90)90075-l.

9. HOLLIEN, Harry; MAJEWSKI, Wojciech. Speaker identification by long-term spectra under normal and distorted speech conditions. *The Journal of the Acoustical Society of America*, Acoustical Society of America (ASA), v. 62, n. 4, p. 975–980, out. 1977. ISSN 1520-8524. doi: 10.1121/1.381592.

10. KINNUNEN, Tomi; HAUTAMAKI, Ville; FRANTI, Pasi. On the Use of Long-Term Average Spectrum in Automatic Speaker Recognition. In: *Proc. International Symposium on Chinese Spoken Language Processing*. [s.n.], 2006. p. 559–567. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d3b4740466aeb1d25831b6329599b615a5bab9b1.

11. ORTEGA-RODRIGUEZ, Manuel. *Relatório final: Articulação de um sistema de identificação de locutor para fins forenses (Original: Informe Final: Articulación de un sistema de identificación de locutor con fines forenses)*. [S.l.], 2016. Acessado em novembro de 2021. Disponível em: https://hdl.handle.net/10669/85190.

12. HARMEGNIES, Bernard. SDDD: A new dissimilarity index for the comparison of speech spectra. *Pattern Recognition Letters*, Elsevier BV, v. 8, n. 3, p. 153–158, out. 1988. ISSN 1872-7344. doi: 10.1016/0167-8655(88)90093-1.

13. STANTON, Jeffrey M. Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, Informa UK Limited, v. 9, n. 3, jan. 2001. ISSN 1069-1898. doi: 10.1080/10691898.2001.11910537.

14. FULLER, Fred H. *Detection of emotional stress by voice analysis final report*. Bethesda, Maryland, USA, 1972. Disponível em: https://www.ojp.gov/ncjrs/virtual-library/abstracts/detection-emotional-stress-voice-analysis-final-report.

15. HARNSBERGER, James D.; HOLLIEN, Harry; MARTIN, Camilo A.; HOLLIEN, Kevin A. Stress and Deception in Speech: Evaluating Layered Voice Analysis. *Journal of Forensic Sciences*, Wiley, v. 54, n. 3, p. 642–650, maio 2009. ISSN 1556-4029. doi: 10.1111/j.1556-4029.2009.01026.x.

16. PITTAM, Jeffery. The Long-Term Spectral Measurement of Voice Quality as a Social and Personality Marker: A Review. *Language and Speech*, SAGE Publications, v. 30, n. 1, p. 1–12, jan. 1987. ISSN 1756-6053. doi: 10.1177/002383098703000101.

17. RODMAN, Robert D.; POWELL, Michael S. Computer Recognition of Speakers Who Disguise Their Voice. In: *The International Conference on Signal Processing Applications and Technology (ICSPAT 2000)*. [s.n.], 2000. Disponível em: https://api.semanticscholar.org/CorpusID:16980245.

18. HERTRICH, I.; ZIEGELMAYER, G. Sexual dimorphism in the long term speech spectrum. *Human Evolution*, Springer Science and

Business Media LLC, v. 2, n. 3, p. 255–262, maio 1987. doi: 10.1007/bf03016110.

19. LINVILLE, Sue Ellen. Source Characteristics of Aged Voice Assessed from Long-Term Average Spectra. *Journal of Voice*, Elsevier BV, v. 16, n. 4, p. 472–479, dez. 2002. doi: 10.1016/s0892-1997(02)00122-4.

20. YÜKSEL, Mustafa; GÜNDÜZ, Bülent. Long term average speech spectra of Turkish. *Logopedics Phoniatrics Vocology*, Informa UK Limited, v. 43, n. 3, p. 101–105, set. 2017. doi: 10.1080/14015439.2017.1377286.

21. National Institute of Standards and Technology. *NIST/SEMATECH e-Handbook of Statistical Methods*. [s.n.], 2012. Acessado em outubro de 2021. Disponível em: https://www.itl.nist.gov/div898/handbook/prc/section2/prc222.htm.

22. Audacity Team. *Audacity (v. 2.1.0), editor e gravador de áudio*. 2015. Disponível em: https://www.audacityteam.org/.

23. The International Association for Forensic Phonetics and Acoustics. *Code of Practice*. [S.l.], 2004. Acessado em janeiro de 2018. Disponível em: https://www.iafpa.net/the-association/code-of-practice/.