Rev. Estud. Ling., Belo Horizonte, v. 27, n. 1, p. 191-212, 2019

Corpus CEFALA-1: Base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia

Corpus CEFALA-1: Audiovisual Database of Speakers for Biometric, Phonetic and Phonology Studies

Arlindo Follador Neto
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais / Brasil
Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, Minas Gerais / Brasil
arlindo.neto@ict.ufvjm.edu.br

Adelino Pinheiro Silva
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais / Brasil
Centro Universitário Newton Paiva, Belo Horizonte, Minas Gerais / Brasil
Instituto de Criminalística de Minas Gerais, Belo Horizonte, Minas Gerais / Brasil
adelinocpp@yahoo.com

Hani Camille Yehia
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais / Brasil
hani@cpdee.ufmg.br

Resumo: A fala humana tem sido estudada em diferentes áreas do conhecimento, as quais incluem desde biometria até fonética e fonologia. Nas pesquisas realizadas em tais áreas, amostras da fala são recursos necessários para a obtenção de resultados e validação de hipóteses. Para isso, amostras de diferentes locutores e conteúdos são armazenadas em arquivos de áudio e organizadas em bases de dados. Tais bases de dados permitem a continuidade, praticidade e confiabilidade de pesquisas, eliminando a difícil e demorada etapa de coleta de dados. Além disso, permitem comparações consistentes entre estudos diferentes. Entretanto, bases de acesso livre na língua portuguesa ou gravadas em ambiente controlado são raramente encontradas. Dessa forma, o objetivo deste trabalho foi construir uma base de dados pública e gratuita do português brasileiro, nomeada Corpus CEFALA-1. A base de dados reúne 104 locutores orientados por um protocolo específico para coleta de amostras audiovisuais de fala gravadas em estúdio.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.27.1.191-212

Este trabalho apresenta as metodologias de processamento, segmentação e organização às quais as amostras de fala foram submetidas, além de análises estatísticas, aplicação à verificação biométrica e análises fonético-fonológicas preliminares do corpus.

Palavras-chave: corpus de locutores; biometria; fonética e fonologia; base de dados audiovisual.

Abstract: Human speech has been studied in different areas of knowledge, which range from biometry to phonetics and phonology. In research conducted in such areas, speech samples are necessary resources for obtaining results and validating hypotheses. For this, samples of different speakers and contents are stored in audio files and organized into databases. Such databases allow the continuity, practicality and reliability of studies, eliminating the difficult and time consuming step of data collection. Moreover, they allow consistent comparisons between different studies. However, free access databases in the Portuguese language or recorded in controlled environments are rarely found. The objective of this paper is to construct a free and public database of Brazilian Portuguese, named Corpus CEFALA-1. The database comprises 104 speakers guided by a specific protocol for the collection of audiovisual speech samples recorded in a studio. The paper presents the methodologies for processing, segmentation and organization of speech samples, statistical analysis, application to biometric verification and preliminary phonetic-phonological analyses.

Keywords: corpus of speakers; biometry; phonetics and phonology; audiovisual database.

Recebido em 10 de abril de 2018

Aceito em 12 de setembro de 2018

1 Introdução

A utilização de informações biométricas de indivíduos é cada vez mais comum em diversas áreas de estudo. Na área de segurança, informações biométricas como impressão digital, íris, geometria da face e voz são empregadas em sistemas que visam verificar a identidade de seus usuários (e.g. bancos, dispositivos pessoais, acessos restritos, sejam eles físicos ou lógicos, urnas eletrônicas, controle de fronteira, etc). A voz, em especial, com sua grande entropia e baixo consumo de recursos para obtenção e processamento, tem sido utilizada em trabalhos inovadores e recentes como fonte de extração de parâmetros biométricos para sistemas de segurança (TRESADERN *et al.*, 2012; WU *et al.*, 2015).

Diferentemente de seu uso para identificação biométrica, outras áreas do conhecimento, como a fonologia e a fonética, exploram a voz há muito mais tempo para outros fins. Nessas áreas as amostras da voz também são utilizadas para entender como os diferentes sons se organizam dentro da fala (SILVA, 1999; SILVA, 2016).

Além da informação acústica, imagens do locutor durante a fala são fontes de estudo de expressões faciais. Tais estudos permitem compreender a dinâmica orofacial e consequentemente enriquecer a animação de desenhos, conhecer distúrbios, detectar emoções, dentre outras aplicações (HORNAK; ROLLS; WADE, 1996; ALEKSIC; KATSAGGELOS, 2006).

As diferentes áreas de estudo que fazem uso de informações de locutores têm em comum a necessidade de utilização de bases de dados. Essas bases são recursos fundamentais para a aquisição de amostras, extração de características e obtenção de resultados. O emprego de tais bases de dados na pesquisa torna possível a validação dos métodos propostos. Entretanto, o acesso atual às bases de dados existentes é restrito, em sua grande maioria é pago e quase sempre em língua estrangeira.

Ainda que a língua não seja um fator crucial para algumas áreas (e.g. verificação de identidade por biometria), a forma com que os dados das bases existentes foram obtidos (e.g. interceptações de ligações telefônicas, registros em ambientes ruidosos, etc) muitas vezes dificulta a execução da pesquisa, que depende da obtenção de amostras biométricas de indivíduos cooperativos em ambiente controlado.

Nesse sentido, tendo em vista a dificuldade de se encontrar uma base de dados de locutores pública e gratuita na língua portuguesa e gravada em um ambiente profissionalmente controlado, o objetivo deste trabalho foi construir a base de dados nomeada Corpus CEFALA-1, por se tratar do primeiro corpus[1] divulgado pelo Centro de Estudos da Fala, Acústica, Linguagem e músicA (CEFALA).

O Corpus CEFALA-1 contou com a participação de 104 indivíduos, dentre eles 49 do sexo feminino e 55 do sexo masculino. Cada indivíduo seguiu um protocolo de gravação que se dividiu em três diferentes etapas de locução: monólogo curto, leitura de parágrafo e leitura

[1] Coleção de amostras de registros de fala, que ocorrem naturalmente ou avocadas, organizadas sistematicamente para emular áreas de uso da língua (BIBER, 1999; HARRINGTON, 2010).

de sentenças. Essas diferentes etapas foram gravadas simultaneamente por cinco microfones, sendo eles um microfone de um *smartphone*, uma câmera de vídeo, um microfone sem fio, um microfone de lapela e um microfone condensador. A base de dados foi gravada em um estúdio profissional de gravação, com o objetivo de reduzir ao máximo o ruído e a reverberação ambientes durante todo o processo.

Os detalhes a respeito do desenvolvimento do Corpus CEFALA-1 são apresentados na Seção 2. Na Seção 3, são apresentados e discutidos um conjunto de experimentos realizados a partir do corpus e na última seção são apresentadas as conclusões deste trabalho.

2 Metodologia

O desenvolvimento da base de dados Corpus CEFALA-1 dividiu-se basicamente em três etapas: i) preparação do estúdio utilizado para aquisição; ii) desenvolvimento de um protocolo de coleta de dados; e iii) processamento e organização dos dados coletados. As três fases do desenvolvimento da base ocorreram durante o primeiro semestre de 2017, sendo que os locutores foram convidados a comparecer para as gravações entre os meses de março e junho. Essas etapas do processo de desenvolvimento serão detalhadas adiante.

2.1 Estúdio de aquisição

A utilização de um ambiente controlado para coleta de áudio permite a construção de uma base de dados de alta qualidade, tornando possível atender aos pré-requisitos do Corpus CEFALA-1, entre eles gravações com menor nível de ruído e reverberação reduzida. Dessa forma, os dados foram colhidos no estúdio do laboratório CEFALA (Centro de Estudos da Fala, Acústica, Linguagem e músicA), localizado na Escola de Engenharia da Universidade Federal de Minas Gerais (UFMG). Esse estúdio possui medidas físicas, em metros, de 2,8 (Largura), 2,9 (Comprimento) e 2,2 (Altura) e conta com revestimento especial para isolamento acústico e redução de reverberação. Utilizando um decibelímetro digital Polimed PM 1900, verificou-se que o ruído de fundo no estúdio era de 34 dB SPL. A Figura 1 exibe o interior do estúdio durante o processo de coleta da base de dados.

O processo de coleta de dados de voz e imagem realizado no estúdio do CEFALA contou com a seguinte estrutura:

- Um decibelímetro digital Polimed PM 1900.
- Uma TV LED Ultra HD 49" LG Modelo 49UB8550. A TV foi utilizada tanto na orientação do locutor quanto no acompanhamento do processo de coleta de dados.
- Um computador Mac Mini, com sistema operacional Mac OS X 10.9.5, processador Intel Core i5 2.5 GHz, 16 GB de RAM e 500 GB de espaço em disco rígido. O computador foi utilizado para processamento e armazenamento dos dados coletados. A opção por esse modelo específico foi em virtude do menor nível de emissão de ruídos, 12 dBA segundo especificação do fabricante.
- Uma mesa de aquisição de áudio M-Audio FireWire 1814, utilizada na digitalização simultânea dos microfones.

FIGURA 1 – Posicionamento dos equipamentos no interior do estúdio de gravação do laboratório CEFALA

Fonte: elaborado pelos autores

A base de dados contou com a utilização de cinco microfones disponíveis no CEFALA e estão nomeados e especificados abaixo para utilização em todo trabalho:

M1 Microfone de lapela da marca DYLAN, modelo DL-09 foi posicionado no peitoral do locutor a aproximadamente vinte centímetros dos lábios.

M2 Microfone sem fio da marca STANER, modelo SW-481, posicionado a aproximadamente dois metros do locutor para coleta do áudio ambiente.

M3 Microfone condensador da marca Brüel & Kjær, modelo 1065, posicionado à altura dos lábios a uma distância de 15 centímetros com 45 graus à direita do plano sagital.

M4 *Smartphone* Samsung Galaxy S2 Lite GT-i9070, utilizado na captação do áudio ambiente, simulando uma aquisição a partir de um dispositivo móvel que utiliza microfone com tecnologia MEMS (*Micro-ElectroMechanical Systems*). O *smartphone* foi posicionado à frente do locutor a uma distância de um metro e elevado a setenta centímetros do piso.

M5 Câmera GoPro Hero 3+ Black Edition, utilizada para captura de vídeo com o *codec* H264 no formato 1280x720 em 60 FPS (*Frames per Second*) e áudio com *codec* AAC em 48 kHz com 32 bits/amostra. Essa câmera foi posicionada frontalmente ao locutor a uma distância de um metro, um exemplo das imagens obtidas por ela é apresentado na Figura 2.

Apesar da construção da base de dados ter utilizado uma câmera de vídeo, descrita no microfone **M5** da lista anterior, o material de vídeo adquirido será de uso restrito do Centro de Estudos da Fala, Acústica, Linguagem e músicA (CEFALA), conforme previamente esclarecido para os voluntários através da autorização de coleta, a qual faz parte do protocolo descrito na próxima seção.

Embora a especificação dos microfones não tenha abrangido detalhadamente suas características técnicas, estas podem ser encontradas nos respectivos manuais fornecidos pelos fabricantes. A influência da utilização desses diferentes microfones não foi amplamente abordada neste trabalho, e está reservado para trabalhos futuros, no entanto a Seção 3.1 apresentará um comparativo da resposta em frequência, em módulo, dos microfones utilizados.

FIGURA 2 – Posicionamento dos locutores em frente à câmera para aquisição do conteúdo audiovisual

Fonte: elaborado pelos autores

2.2 Protocolo de coleta

A utilização de um protocolo de coleta visa orientar os voluntários que tiveram sua fala registrada, tanto no sentido legal quanto no processo de coleta da base de dados. Dessa forma, o protocolo aqui adotado dividiu-se em quatro etapas fundamentais:

I. Autorização de coleta: antes de iniciar o estudo foi apresentada aos participantes uma autorização de coleta de material sonoro a qual esclarecia, dentre outros termos, a finalidade pública dos dados de áudio da base. Tendo concordado com os termos apresentados, o participante assinava a autorização e seguia para o estúdio de coleta. O processo de coleta de dados consistiu em aproximadamente 5 minutos de leitura de texto em voz alta e fala espontânea, as quais são atividades corriqueiramente realizadas em sala de aula. Aprovação por comitê de ética não foi necessária, pois tais atividades não acarretam riscos maiores do que os existentes na vida cotidiana e a identidade dos voluntários foi preservada.[2]

II. Fala espontânea: uma vez no estúdio de gravação, o locutor foi orientado a discorrer a respeito de um assunto de seu interesse (e.g. relato pessoal ou não pessoal, comentário sobre qualquer assunto, descrição de alguma atividade rotineira) por aproximadamente 2 minutos. O intuito desta primeira etapa de gravação foi que o voluntário atingisse o estado de fala espontânea.

III. Leitura de texto: nesta etapa todos os locutores foram orientados a ler um mesmo trecho, com 153 palavras, do livro *A vida de Galileu: o contemplador de estrelas* (HARSANYI, 1957). Esse texto foi escolhido de forma a abranger um número maior de fonemas e também por misturar elementos da fala formal e informal, contendo palavras comuns do português brasileiro e também incluindo uma frase em língua estrangeira (i.e. Italiano).

IV. Leitura de frases: nesta última etapa os locutores foram igualmente submetidos à leitura de vinte frases. A lista de frases apresenta uma média de sete palavras por frase, sendo 3 o mínimo e 19 o máximo de palavras por frase. A seleção das frases levou em consideração

[2] De acordo com o item 8.0.5 da American Psychological Association (2002) e do Artigo 1º da Resolução n. 510 de 7 de abril de 2016 do Conselho Nacional de Saúde.

diferentes tamanhos, frases afirmativas, interrogativas, expressões populares e aliterações. A lista de frases pode ser obtida no protocolo de coleta disponível em http://www.cefala.org.

O intuito das duas últimas etapas do protocolo de coleta foi a obtenção de fala a partir de texto pré-fixado, sendo possível posteriormente análises a respeito do comportamento do locutor durante cada elocução.

2.3 Organização da base de dados

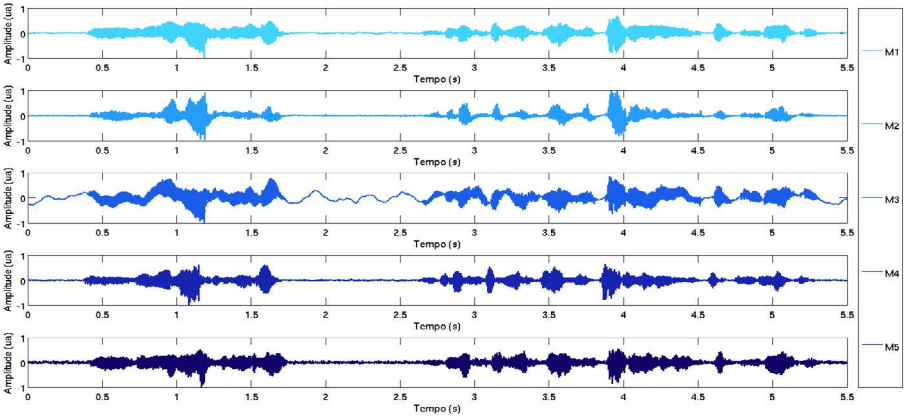
Após a coleta do material audiovisual de todos os voluntários da base, a tarefa subsequente foi o processamento e organização dos dados obtidos. Dessa forma, o primeiro passo consistiu em extrair o conteúdo de áudio dos arquivos em formato MP4, obtidos pela câmera (i.e. M5). O conteúdo de vídeo foi armazenado em seu formato original, enquanto o conteúdo de áudio, obtido no formato ACC, foi extraído e convertido para o formato padrão da base de dados: formato WAV com taxa de amostragem de 44,1 kHz e 16 bits por amostra.

Uma vez organizados os arquivos de áudio de todos os microfones no mesmo formato, a tarefa seguinte foi alinhar os arquivos de todos os locutores, ou seja, fazer com que o áudio de um microfone de um locutor em um ponto específico fosse exatamente o mesmo para os demais microfones. Dessa forma, ainda na etapa de aquisição, antes mesmo do o locutor iniciar o protocolo de coleta, foi solicitado que este aguardasse a execução de um pulso de sincronismo utilizado para marcar o início da gravação em cada um dos microfones. Para o alinhamento foi utilizado o *toolbox* VOICEBOX[3] para MATLAB®.[4] Um exemplo do resultado final do processo é exibido na Figura 3.

[3] VOICEBOX é uma toolbox para processamento de fala com rotinas desenvolvidas para MATLAB®.

[4] MATLAB® é um software interativo de alta performance aplicado ao cálculo numérico.

FIGURA 3 – Zoom em um trecho de áudio de 5,5 segundos para exemplificar o resultado após procedimento de alinhamento dos microfones



Fonte: elaborado pelos autores

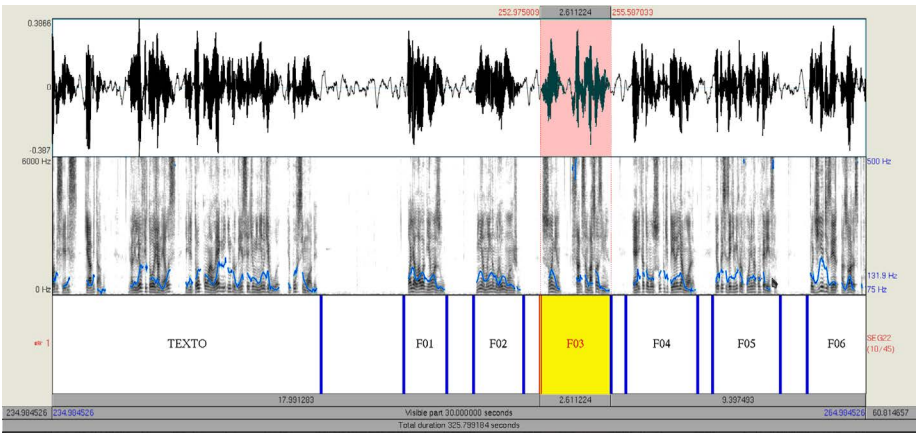
O alinhamento dos arquivos de áudio simplifica a etapa de organização posterior, que é a de segmentação. Uma vez alinhados, a segmentação que for realizada em um arquivo de áudio de um determinado microfone servirá para os demais arquivos dos microfones usados para gravar uma mesma elocução. Para a segmentação, o *software* de processamento e análise de fala *Praat*[5] foi utilizado na criação de *labels* (rótulos) para cada segmento. Essa metodologia descarta a necessidade de fragmentar a base em pequenos arquivos de áudio para cada segmento. Dessa forma, os arquivos de áudio alinhados são preservados, e um novo arquivo contendo os *labels* da segmentação é criado para cada locutor. Essa metodologia permite que existam infinitas formas de segmentação para a mesma base de dados sem alterar os arquivos de áudio alinhados. A segmentação da base realizada neste trabalho foi nomeada SEG22, a qual dividiu as amostras em 22 trechos baseados no protocolo de coleta, sendo eles: o trecho de fala espontânea (i.e. *label* FALAESP); o trecho da leitura de texto (i.e. *label* TEXTO); e a sequência de vinte frases (i.e. *labels* de F01 a F20). A delimitação de cada trecho utiliza o formato *TextGrid* do *Praat* e tem a precisão da ordem de picosegundos (i.e. 10^{-12} segundos). Um exemplo de um arquivo de segmentação é apresentado na Figura 4.

[5] O Praat é um software aplicado na análise e síntese da fala desenvolvido na Universidade de Amsterdã.

A organização final dos arquivos da base de dados foi orientada por uma estrutura de diretórios com nomenclatura padronizada. Os diretórios de locutores são definidos por "Locutor_0XXX?", onde "0XXX" é um número de sequência e "?" identifica o sexo do locutor, sendo M para Masculino e F para Feminino. Dentro de cada diretório de locutor estão presentes os arquivos de áudio com a nomenclatura "Locutor_0XXX?_MY_S.wav", onde "MY" identifica o microfone, a letra "S" indica que o arquivo de áudio foi submetido ao processo de alinhamento e ".wav" que o formato do arquivo é WAV PCM (P*ulse-Code Modulation*). Finalmente, os arquivos no formato TextGrid, que também estão contidos dentro de cada diretório de locutor, seguem a nomenclatura Locutor_0XXX?_NOME.TextGrid, onde as diferenças são a palavra-chave "NOME" que representará o título da segmentação realizada e ".TextGrid" representa a extensão do arquivo, indicando que é o formato reconhecido pela aplicação *Praat*.

Após a segmentação dos arquivos de áudio, foi estimada a quantidade de material sonoro presente no corpus. A Figura 5 a seguir apresenta o gráfico RDI (*Raw-data Description and Inference*) com a duração em segundos dos segmentos de fala espontânea, leitura de texto e leitura de frases. Nessa figura, os pontos representam a duração individual de cada amostra, as curvas laterais são a distribuição de probabilidade empírica, a linha preta horizontal a média e o retângulo escuro é o intervalo de confiança da média para $\alpha = 0{,}05$. A Tabela 1 complementa as informações de duração dos segmentos de fala.

FIGURA 4 – Exemplo de utilização do *software Praat* durante o processo de segmentação (i.e. SEG22) dos trechos de áudio da base de dados



Fonte: elaborado pelos autores

3 Análises preliminares do corpus e exemplo de aplicação

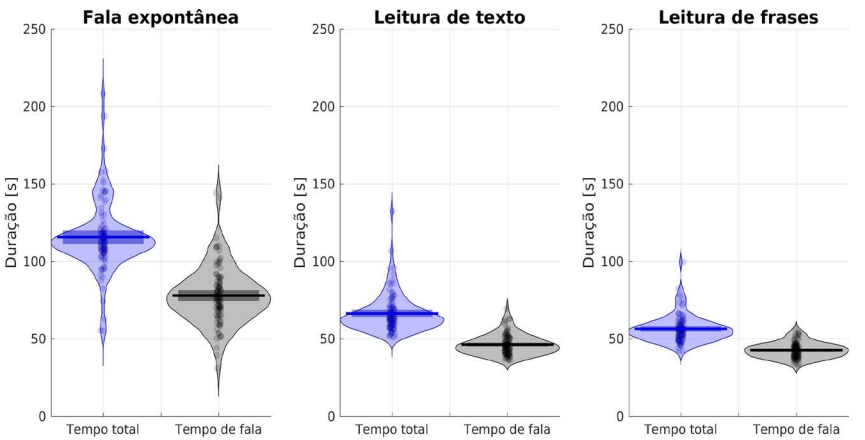
De posse da base de dados adquirida, processada, organizada e segmentada, experimentos foram desenvolvidos no sentido de se obter uma caracterização preliminar do seu conteúdo. Os resultados desses experimentos servem não apenas para identificar melhor os componentes da base de dados, mas também para exemplificar sua utilização em pesquisas voltadas para as áreas de biometria, fonética e fonologia.

TABELA 1 – Duração em segundos com os valores mínimo, médio e máximo dos segmentos (i.e. SEG22) para o tempo total de áudio e para o tempo de fala (i.e. excluindo pausas)

	mínimo		médio		máximo	
	Total	Fala	Total	Fala	Total	Fala
Áudio completo	202	112	273	171	412	251
Fala espontânea	55	31	116	78	208	144
Leitura de texto	51	36	66	46	132	70
Leitura de frases	42	33	56	43	99	56

Fonte: elaborado pelos autores

FIGURA 5 – Gráfico RDI com a duração (em segundos) das etapas de cada amostra do Corpus CEFALA-1



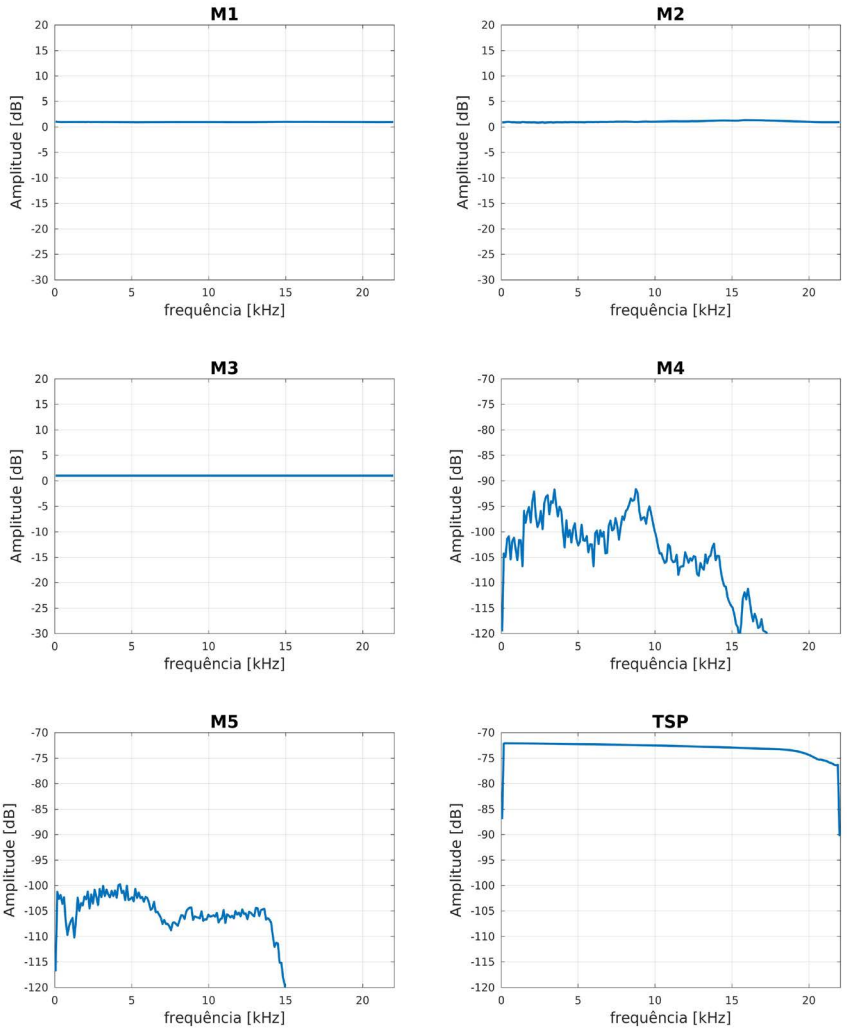
Fonte: elaborado pelos autores

3.1 Resposta em frequência dos microfones no ambiente de gravação

O primeiro experimento tem como objetivo verificar a resposta em frequência dos microfones usados na gravação. Para esse fim foi usado um pulso de varredura do tipo TSP (*Time-Stretched Pulse*) (SUZUKI *et al*; 1995). Para a reprodução do TSP foi utilizado um alto-falante da marca JBL by HARMAN do modelo Micro Wireless, que segundo dados do fabricante possui resposta em frequência de 150 Hz a 20 kHz com uma relação sinal ruído maior que 80 dB. A caixa de som foi posicionada simulando o local onde estarão os locutores, seu volume ajustado para reproduzir uma pressão sonora semelhante àquela observada durante a fala e a reprodução do TSP foi repetida seis vezes para uma maior aquisição de amostras para análise. Durante a reprodução do TSP foi preservada a montagem do estúdio que inclui a posição dos microfones conforme Figura 1. A análise da resposta em frequência de cada microfone foi realizada utilizando o software MATLAB® para aplicação do algoritmo FFT (*Fast Fourier Transform*). Este algoritmo calcula a transformada de *Fourier* a qual converte os dados do domínio do tempo para o domínio da frequência, tornando possível a análise de resposta em frequência em todo o espectro adquirido.

Basicamente, dois fatores são responsáveis por influenciar a resposta em frequência dos microfones. O primeiro está relacionado às características construtivas do microfone, ou seja, sua faixa de passagem intrínseca. O segundo é o posicionamento do microfone no ambiente, uma vez que sons agudos são mais diretivos que os graves. A combinação desses fatores com a resposta em frequência do alto-falante resulta nas curvas apresentadas na Figura 6. Essa análise de resposta em frequência é importante no planejamento de futuros experimentos que venham utilizar amostras de microfones específicos deste corpus.

FIGURA 6 – Resposta em frequência apresentada por cada microfone na execução do TSP



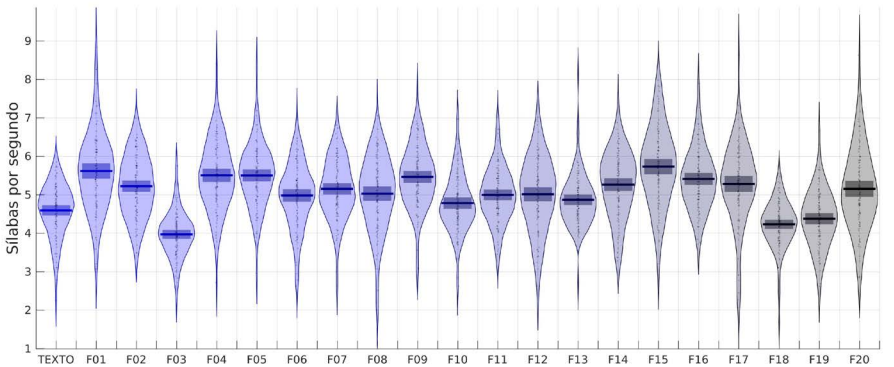
Fonte: elaborado pelos autores

3.2 Desempenho de locutores em leitura de texto e frases

Os experimentos seguintes direcionaram o foco para análises fonéticas e fonológicas. Inicialmente, buscou-se medir as taxas de elocução e de articulação presentes no corpus (GONÇALVES; 2013). O experimento teve como objetivo estimar duas medidas de execução: i) média de geração de sílabas por segundo; ii) tempo médio de pausas durante a fala. Estas estimativas foram obtidas a partir da contagem do número de sílabas e pausas presentes nas etapas de leitura de texto e frases. O processo de segmentação da base de dados e a técnica de VAD (*Voice Activity Detection*) permitiram que esta contagem fosse realizada com precisão, uma vez que sendo aplicado ele separará perfeitamente os trechos contendo fala e pausas dos arquivos de áudio.

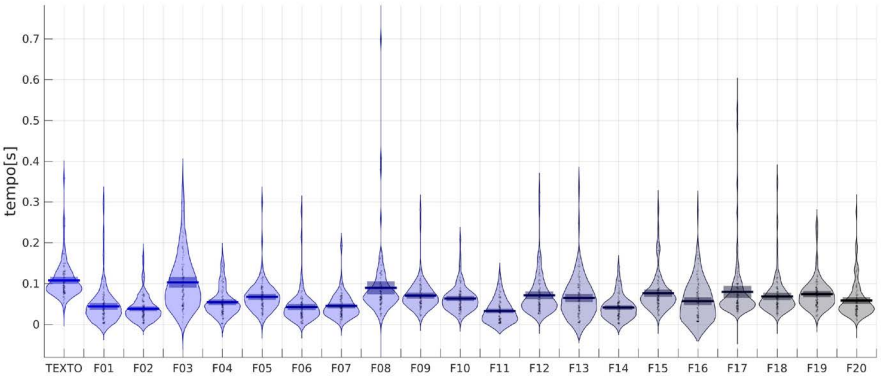
Em relação aos resultados, a taxa média de produção de sílabas oscila em torno de 5 sílabas por segundo, enquanto as pausas (ou tempo sem voz) têm duração média de pouco menos de 100 ms.

FIGURA 7 – Gráfico RDI apresentado a taxa de sílabas por segundo dos falantes



Fonte: elaborado pelos autores

FIGURA 8 – Gráfico RDI apresentado o tempo de pausa dos falantes



Fonte: elaborado pelos autores

3.3 Estatísticas de frequência fundamental e formantes

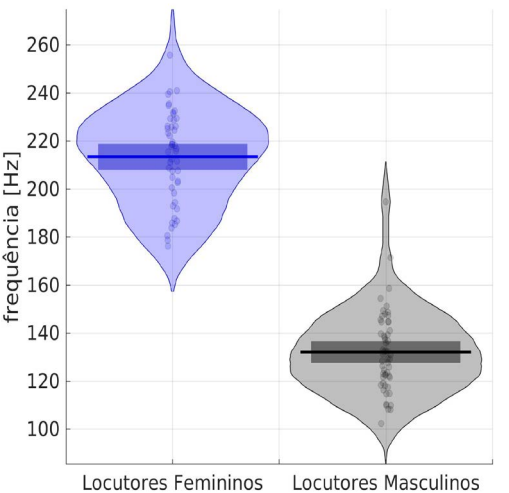
O Corpus CEFALA-1 também pode ser usado para analisar a distribuição dos valores da frequência fundamental (i.e. F_0) e dos formantes (e.g. F_1, F_2, F_3, ...) entre os locutores e entre elocuções. Como exemplo, foram realizadas análises dos formantes separados por sexo. Dessa forma, para os grupos masculino e feminino as análises espectrográficas foram desenvolvidas utilizando o método LPC (*Linear Predictive Coding*). O LPC é uma técnica de análise da fala que modela o sinal como um processo autorregressivo cuja ordem é suficiente para representar os pares de polos referentes aos formantes e inclinação espectral. A partir do LPC é possível encontrar os formantes, frequências em que o trato vocal ressoa em trechos de fala vozeados. Os formantes indicam a frequência, amplitude e largura de banda dessa ressonância do trato vocal. Levando em consideração o modelo do trato vocal como um tubo ressonante (FLANAGAN, 2013), é possível estimar a presença de um formante a cada 1 kHz de banda do sinal de voz. Assim, para o sinal de fala que foi subamostrado a 8 kHz, resultando na banda de 4 kHz, é possível estimar quatro formantes, os quais são excitados pelos harmônicos da frequência fundamental de vibração das pregas vocais.

Para cada locutor foi extraída a frequência fundamental (F_0) por quadro de 25 ms de voz. O algoritmo utilizado foi o *Yet Another Algorithm for Pitch Tracking* que foi publicado nos trabalhos de Kasi e Zahorian

(2002) e Zahorian e Hu (2008). Esse algoritmo retorna os valores da F_0 ao longo dos quadros identificados como vozeados.

Como o valor da F_0 é variável ao longo da fala, cada locutor foi representado pela sua média. A Figura 9 apresenta os valores médios obtidos no gráfico RDI que separa a F_0 entre os locutores do sexo masculino e feminino. Os valores de F_0 para mulheres são aproximadamente o dobro dos observados para homens. Isso acontece porque as cordas vocais femininas são mais curtas e mais leves que as masculinas (TITZE, 1994).

FIGURA 9 – Dispersão da F_0 dos falantes divididos por sexo



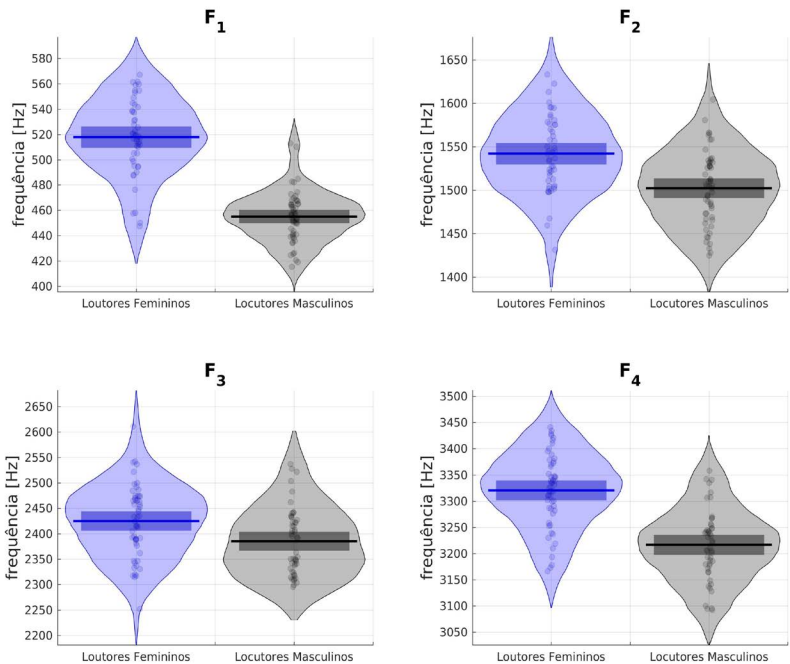
Fonte: elaborado pelos autores

Os formantes foram obtidos por meio de análise LPC com refinamento dos polos, como descrito por Kim, Seo e Sung (2006). A Figura 10 mostra o gráfico RDI do valor médio dos formantes F_1 a F_4 de todas as amostras. O recorte entre locutores de diferentes sexos mostra que os locutores do sexo feminino apresentaram valores médios superiores aos locutores do sexo masculino para os primeiros formantes. Tal diferença ocorre devido a diferença anatômica entre os tratos vocais de homens, cujo trato vocal é mais longo, e de mulheres, cujo trato vocal é mais curto.

3.4 Taxas de identificação de locutores

O último experimento é focado na área de biometria e consiste na avaliação de desempenho da comparação automática de locutores utilizando a metodologia UBM-GMM (*Universal Background Model – Gaussian Mixture Model*). A metodologia UBM-GMM foi proposta por Reynolds *et al.* (2000) e baseia-se na separação de amostras de voz em pequenos quadros. Para cada quadro são calculados os coeficientes ceptrais na escala mel de frequências (MFCC – *Mel Frequency Cepstral Coefficients*) e suas variações temporais de primeira e de segunda ordens, denominadas Δ-cepstrum e ΔΔ-cepstrum. Inspirado na percepção humana para frequências sonoras, o MFCC é uma técnica de extração de parâmetros que utiliza a escala *Mel* para mapear o espectro da voz e que busca reproduzir a resolução espectral da cóclea, tubo ósseo em forma de caracol no qual regiões diferentes são sensíveis a frequências diferentes (PICONE, 1993).

FIGURA 10 – Dispersão das frequências dos formantes dos falantes divididos por sexo



Fonte: elaborado pelos autores

A partir dos valores MFCC, representados por x = {x(1), x(2), …, x(T)}, é possível construir uma função densidade de probabilidade que é aproximada por uma soma ponderada de gaussianas. Dessa forma, o modelo UBM consiste em uma mistura de gaussianas que se ajusta à dispersão dos valores MFCC de todos os locutores da base de dados, enquanto o modelo GMM é uma adaptação do modelo UBM para a dispersão dos valores MFCC de um locutor específico.

Assim, a partir dos modelos UBM-GMM e de uma amostra de áudio é possível verificar a qual locutor essa amostra pertence. Nesse modelo a pontuação (*score*) é obtida em um processo que consiste na comparação de uma amostra de teste com todos os modelos GMM. Caso essa pontuação seja superior a um determinado limiar, a amostra é aceita como sendo do respectivo locutor que originou o modelo GMM. Essa pontuação é definida como o logaritmo da razão de verossimilhança (LLR - *Log-Likelihood Ratio*) e pode ser calculada como

$$LLR(x) = \frac{p(H_0 \mid x, \lambda_{GMM}, \lambda_{UBM})}{p(H_1 \mid x, \lambda_{GMM}, \lambda_{UBM})} = \frac{1}{T}\sum_{t=1}^{T}\left[log\left(p(x(t) \mid \lambda_{GMM})\right) - log\left(p(x(t) \mid \lambda_{UBM})\right)\right],$$

onde, $p(x(t) \mid \lambda_{GMM})$ e $p(x(t) \mid \lambda_{UBM})$ são as avaliações do t-ésimo MFCC dos modelos GMM e UBM. As hipóteses H_0 e H_1 da comparação são definidas como

$$H: \begin{cases} H_0 : x \text{ e } \lambda_{GMM} \text{ são provenientes do mesmo locutor para } LLR(x) > \eta_0 \\ H_1 : x \text{ e } \lambda_{GMM} \text{ são provenientes de locutores diferentes para } LLR(x) \leq \eta_0. \end{cases}$$

Neste experimento a base de dados foi separada em dois grupos, sendo os dados divididos em conjuntos de treinamento e de teste. O conjunto de treinamento é utilizado na parametrização dos modelos UBM-GMM, enquanto o conjunto de teste é utilizado na obtenção dos resultados de verificação. Dessa forma, o conjunto de treinamento consistiu na concatenação dos dados da etapa completa de leitura de texto com 66% da etapa de fala espontânea, enquanto o conjunto de teste consistiu na concatenação dos 34% restantes da etapa de fala espontânea com a etapa completa de leitura de frases. Todo o experimento foi repetido para cada um dos cinco microfones, sendo que em todos os experimentos as amostras foram subamostradas para 8 kHz e restritas a faixa entre 300 e 3500 Hz (i.e. condições de comunicação por telefone).

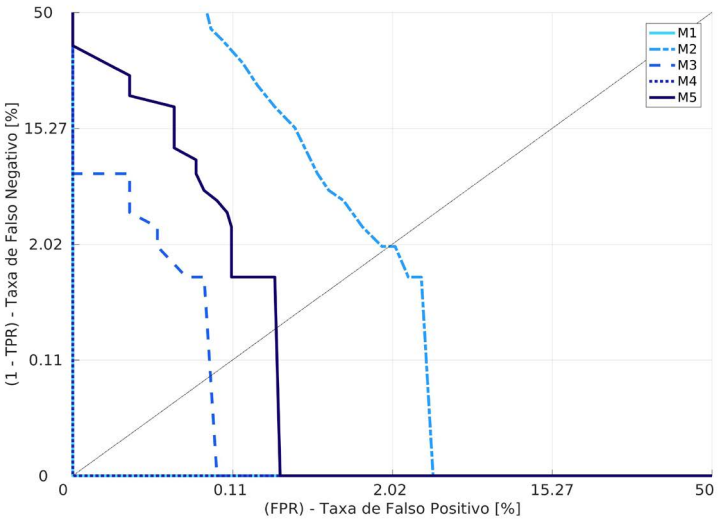
O procedimento de identificação de locutores mostrou-se sensível ao microfone utilizado. A Figura 11 a seguir apresenta a curva DET (*Detection Error Tradeoff*) para a identificação de locutores. Nesta simulação, apenas os microfones M1 e M4 foram capazes de identificar os locutores sem nenhum erro.

TABELA 2 – Limiares de LLR onde obtém-se a menor taxa de mesmo erro e C_{LLR}

	Limiar LLR [nepers]	Taxa de Mesmo Erro [%]	Limiar LLR [nepers]	Mínimo C_{LLR} [nepers]
Microfone M1	0,81	0	0,20	0,25
Microfone M2	1,68	1,92	1,43	0,31
Microfone M3	1,26	0,06	0,91	0,27
Microfone M4	0,65	0	0,14	0,31
Microfone M5	0,32	0,28	0,10	0,35

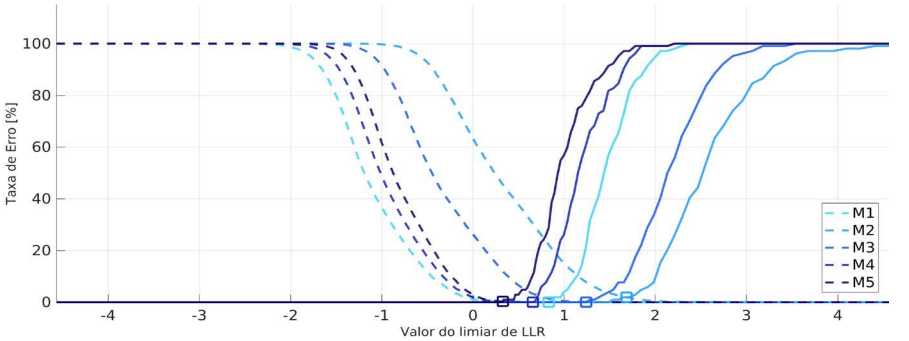
Fonte: elaborado pelos autores

FIGURA 11 – Curva DET para identificação de locutores com a variação do limiar de decisão



Fonte: elaborado pelos autores

FIGURA 12 – Taxa de mesmo erro apresentada no experimento pela variação do limiar de decisão (i.e. valor do logaritmo da razão de verossimilhança). A linha pontilhada representa a taxa de falso negativo enquanto a linha contínua a taxa de falso positivo



Fonte: elaborado pelos autores

As curvas das figuras 11 e 12 mostram o desempenho dos diferentes microfones para a tarefa de verificação de locutor. Utilizando os microfones M1 e M4, foi possível verificar os locutores com uma acurácia de 100%, enquanto erros de verificação para os demais microfones foram observados. A Figura 12 e a Tabela 2 mostram os limiares de separação para cada um dos cinco microfones utilizados.

4 Conclusão

Este artigo apresentou o Corpus CEFALA-1 desenvolvido como ferramenta de suporte para análise de locutores, seja em experimentos de biometria ou em análise fonético-fonológicas. Neste momento o corpus está sendo utilizado em duas teses de doutorado: a primeira na área de autenticação de locutores e a segunda na comparação forense de locutores.

O processo de construção do corpus foi descrito tanto no que se refere à montagem utilizada do estúdio de coleta de dados quanto ao conteúdo e ao processamento das amostras coletadas.

A utilização do corpus foi demonstrada através do levantamento das distribuições de valores da frequência fundamental, dos primeiros formantes, taxas de sílabas, tempo de pausa e de testes de verificação de locutor. Essas análises preliminares são úteis para apresentar o *baseline* do Corpus CEFALA-1, que servirá de baliza para trabalhos futuros.

O objetivo principal deste artigo, entretanto, é difundir o Corpus CEFALA-1, que será disponibilizado sob a licença Creative Common BY-NC-ND mediante cadastro eletrônico em http://www.cefala.org. Espera-se que o Corpus CEFALA-1 torne-se um recurso público e gratuito que contribua tanto com o ensino e a pesquisa da fala em geral quanto do português brasileiro em particular.

Referências

ALEKSIC, P. S.; KATSAGGELOS, A. K. Automatic Facial Expression Recognition Using Facial Animation Parameters and Multistream HMMs. *IEEE Transactions on Information Forensics and Security*, [S.L.], v. 1, n. 1, p. 3-11, 2006.

AMERICAN PSYCHOLOGICAL ASSOCIATION. *Ethical Principles of Psychologists and Code of Conduct*. US: American Psychological Association, 2002.

BIBER, D.; CONRAD, S.; REPPEN, R.; LEECH, G. Corpus linguistics: Investigating language structure and use. *International Journal of Corpus Linguistics*, [s.l.], v. 4, n. 1, p. 185-188, 1999.

CONSELHO NACIONAL DE SAÚDE. *Resolução n. 510* de 7 de abril de 2016. Dispõe sobre as especificidades éticas das pesquisas nas ciências humanas e sociais e de outras que utilizam metodologias próprias dessas áreas. Brasília, 2016.

FLANAGAN, J. L. Speech analysis synthesis and perception. New York: Springer Science & Business Media, 2013.

GONÇALVES, C. S. *Taxa de elocução e de articulação em corpus forense do português brasileiro.* 2013. 192f. Tese (Doutorado) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2013.

HARRINGTON, J. *Phonetic analysis of speech corpora*. Oxford: John Wiley & Sons, 2010.

HARSANYI, Z. *A vida de Galileu*: (o contemplador de estrelas). Rio de Janeiro: Editora José Olympio, 1957.

HORNAK, J.; ROLLS, E.; WADE, D. Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia*, [s.l.], v. 34, n. 4, p. 247-261, 1996.

KASI, K.; ZAHORIAN, S. A. Yet another algorithm for pitch tracking. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP), 2002, Orlando. *Proceedings*… Orlando: IEEE. p. I-361-I-364.

KIM, C.; SEO, K.-D.; SUNG, W. A robust formant extraction algorithm combining spectral peak picking and root polishing. *EURASIP Journal on Applied Signal Processing*, New York, v. 2006, p. 33-33, 2006.

PICONE, J. W. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, [s.l.], v. 81, n. 9, p. 1215-1247, 1993.

REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, [s.l.], v. 10, n. 1-3, p. 19-41, 2000.

SILVA, A. H. P. *Língua Portuguesa I*: fonética e fonologia. Curitiba: IESDE Brasil, 2016.

SILVA, T. C. *Fonética e fonologia do português*: roteiro de estudos e guia de exercícios. São Paulo: Contexto, 1999.

SUZUKI, Y.; ASANO, F.; KIM, H. Y.; SONE, T. An optimum computer generated pulse signal suitable for the measurement of very long impulse responses. *The Journal of the Acoustical Society of America*, [s.l.], v. 97, n. 2, p. 1119-1123, 1995.

TITZE, I. R. *Principles of voice production*. Englewood Cliffs: Prentice Hall, 1994.

TRESADERN, P.; MCCOOL, C.; POH, N.; MATEJKA, P.; HADID, A.; LEVY, C.; MARCEL, S. Mobile biometrics (mobio): Joint face and voice verification for a mobile platform. *IEEE pervasive computing*, p. 79-87, 2012.

WU, Z.; EVANS, N.; KINNUNEN, T.; YAMAGISHI, J.; ALEGRE, F.; LI, H. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication*, [s.l.], v. 66, p. 130-153, 2015.

ZAHORIAN, S. A.; HU, H. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, [s.l.], v. 123, n. 6, p. 4559-4571, 2008.