Pós-Graduação em Ciência da Computação

“Segmentação de voz em ambientes ruidosos utilizando análise da imagem do espectrograma”

Por

Gilliard Alan de Melo Lopes

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, AGOSTO/2013

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Gilliard Alan de Melo Lopes

Segmentação de voz em ambientes ruidosos utilizando análise de imagem do espectrograma

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO.

ORIENTADOR: PROF. DR. CARLOS ALEXANDRE BARROS DE MELLO

RECIFE, AGOSTO/2013

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

Lopes, Gilliard Alan de Melo

Segmentação de voz em ambientes ruidosos utilizando análise de imagem do espectrograma/ Gilliard Alan de Melo Lopes. - Recife: O Autor, 2013.

xviii, 60 p.: il., fig., tab.

Orientador: Carlos Alexandre Barros de Mello.

Dissertação (mestrado) - Universidade Federal de Pernambuco. CIn, Ciência da Computação, 2013.

Inclui referências.

1. Ciência da Computação. 2. Processamento de voz. I. Mello, Carlos Alexandre Barros de (orientador). II. Título.

004 CDD (23. ed.) MEI2013 – 134

Dissertação de Mestrado apresentada por **Gilliard Alan de Melo Lopes** à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título "**Segmentação de Voz em Ambientes Ruidosos Utilizando Análise da Imagem do Espectrograma**", orientada pelo **Prof. Carlos Alexandre Barros Mello** e aprovada pela Banca Examinadora formada pelos professores:

_____

Prof. George Darmiton da Cunha Cavalcanti
Centro de Informática / UFPE

_____

Prof. Francisco Madeiro Bernardino Junior
Escola Politécnica de Pernambuco / UPE

_____

Prof. Carlos Alexandre Barros de Mello
Centro de Informática / UFPE

Visto e permitida a impressão.
Recife, 13 de agosto de 2013

_____

Profa. Edna Natividade da Silva Barros
Coordenadora da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Agradecimentos

Em primeiro lugar, agradeço a Deus pelo dom da vida. Agradeço a minha família pelo apoio e força em tempo integral. Agradeço a minha namorada Andreza Gouveia pelo carinho e companheirismo nos momentos mais turbulentos.

Agradeço fortemente ao meu orientador, Dr. Carlos Alexandre, pela excelente orientação, dedicação, compreensão e, sobretudo, paciência para me guiar no decorrer do mestrado apesar de todos os problemas enfrentados.

Um agradecimento especial ao amigo doutorando Diogo Costa, pelos momentos de estudo e pesquisa ao longo do meu mestrado.

E por último, porém não menos importante, agradeço a FACEPE pelo crédito no projeto aqui descrito.

Resumo

Esta dissertação apresenta um novo algoritmo para segmentação de sinais de voz baseado em técnicas de processamento de imagem, tais como análise de espectrograma, morfologia matemática, componentes conectados, análise de projeção e binarização. O algoritmo proposto opera em dois ciclos: o primeiro age separando o sinal de voz do fundo (silêncio ou ruído). O segundo utiliza esse sinal de voz segmentado para realizar a segmentação de sílabas fonéticas (agrupamento de fonemas). A base de dados de áudio *MIT* (*MIT Mobile Device Speaker Verification Corpus)* e a *TIMIT* (*Texas Instruments/Massachussets Institute of Technology*) foram utilizadas para validação do algoritmo proposto. Os sinais de voz escolhidos variam desde o gênero do locutor, a regionalidade (sotaque), tipos de fonemas e ruídos de fundo, como: ruídos de apito, chuva, vento e de um cruzamento de ruas com tráfego intenso. A técnica proposta mostrou eficiência na segmentação, no que diz respeito aos segmentos fonéticos, em ambientes com ausência e presença de ruídos, utilizando os mesmos parâmetros em ambas as situações.

Palavras-chave: Segmentação de Voz, Segmentação de Fonemas, Espectrograma, Análise de Imagens.

Abstract

This dissertation presents a novel speech segmentation technique based on image processing features like spectrogram analysis, mathematical morphology, connected components, projection profile and thresholding. The proposed algorithm works in two loops: the first segments the sound in search for the speech signal. After, the segmented speech returns to the algorithm for phoneme segmentation. For evaluation and validation, the algorithm was applied to MIT Mobile Device Speaker Verification Corpus (MIT-MDSVC) and TIMIT (Texas Instruments/ Massachussets Institute of Technology) audio databases. We choose audio signals, both male and female voices, with different phonemes, and with increasing noise difficulties, like whistle, rain, wind and road traffic noises. The proposed technique showed to be robust, in relation to the phonetic segments, under environments with and without noise, using the same parameters settings for both situations with satisfactory results.

Keywords: *Speech Segmentation, Phoneme Segmentation, Spectrogram, Image Analysis.*

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ASI	*Automatic Speaker Identification* (Identificação Automática de Locutor)
ASM	*Automatic Segmentation Machine* (Máquina de Segmentação Automática)
ASR	*Automatic Speech Recognition* (Reconhecimento Automático de Fala)
DFT	*Discrete Fourier Transform* (Transformada Discreta de Fourier)
DIP	*Digital Image Processing* (Processamento Digital de Imagens)
DSP	*Digital Speech Processing* (Processamento Digital de Voz)
DWT	*Discrete Wavelet Transform* (Transformada Discreta Wavelet)
HMM	*Hidden Markov Models* (Modelos Ocultos de Markov)
MFCC	*Mel-Frequency Cepstrum Coefficients* (Coeficientes Cepstrais da Frequência-Mel)
MIT – MDSVC	*MIT Mobile Device Speaker Verification Corpus*
MLP	*Multi-Layer Perceptron* (Perceptron de multicamadas)
SR	*Speech Recognition* (Reconhecimento de Voz)
STFT	*Short-Time Fourier Transform* (Transformada de Fourier de Tempo Curto)
TTS	*Text-to-speech Synthesis* (Síntese de Voz)
VAD	*Voice Activity Detection*

1 Introdução

É comum dizer que os seres humanos são criaturas visuais. Nós percebemos a importância da nossa visão em situações bem simples como fechar os olhos ou andar a noite em uma estrada com ausência de iluminação. Nesses dois casos, nossa percepção sobre o ambiente em que nos encontramos é parcial ou totalmente perdida, porém, o nosso sistema auditivo tem um bom desempenho, permitindo-nos escutar com a mesma perfeição em ambientes escuros [2]. Nós somos capazes de escutar sons em todas as direções (mesmo sem estarmos de frente para a fonte do som), também podemos ouvir através de obstáculos como portas e até mesmo paredes (onde a luz não pode penetrar). Portanto, é evidente a importância da nossa audição, principalmente para compor uma das habilidades mais fundamentais do ser humano: comunicação (através da fala, por exemplo).

Os seres humanos são capazes de entender, separar, compreender e até mesmo antecipar uma sentença pronunciada em sua linguagem nativa com muita naturalidade e facilidade. Mesmo em ambientes ruidosos, o ser humano possui a habilidade de focar e extrair o som da elocução que lhe é de interesse. Um dos principais objetivos das pesquisas na área de reconhecimento de fala é a construção de uma máquina capaz de transcrever, de maneira eficaz e em tempo real, o que lhe é pronunciado. Entretanto, mesmo após mais de 50 anos de pesquisa na área de Reconhecimento de Voz (*Speech Recognition - SR*), a capacidade de reconhecimento dos sistemas ainda está muito longe de se equiparar com a do ser humano [1]. O sinal de voz é complexo e multivariado e as soluções propostas até hoje não conseguem atender todas as possibilidades da percepção humana. Porém, algumas aplicações específicas da área de reconhecimento de fala já começaram a ser utilizadas nos últimos anos, aumentando a interação homem-máquina através da comunicação por voz.

Uma das subáreas do reconhecimento de voz que ainda precisa de muitas melhorias é a segmentação de voz. A segmentação é o processo de identificação dos limites entre as palavras, sílabas, ou fonemas em uma linguagem natural. Encontrar os limites de um texto escrito é uma tarefa simples, em geral, no entanto, a mesma tarefa para uma pronúncia em voz requer muitos cuidados pelo fato de não existir o “espaço em branco” entre as palavras pronunciadas (fazendo uma analogia com o texto escrito). Além de ser uma das etapas mais importantes de um sistema de reconhecimento de voz, é também uma das tarefas mais complexas nessa área de pesquisa [3]. Realizar a segmentação de voz de modo confiável é fundamental para a construção de um reconhecedor automático da fala (*Automatic Speech Recognition - ASR*) [4]. A segmentação também é um requisito para trabalhos em indexação de documentos enunciados (*speech document indexing*) [5], análise fonética da voz (*phonetic analysis of speech*) [6], classificação

de conteúdo de áudio (*audio content classification*) [7] e reconhecimento de palavras (*word recognition*) [8]. De acordo com Rabiner e Sambur, a tarefa de segmentação automática da fala por meio de uma máquina é inerentemente um problema difícil [9]. Um dos fatores que atrapalham a segmentação é devido à falta de uma analogia acústica confiável entre as palavras pronunciadas assim como existem os espaços em branco entre as palavras escritas em um texto [10][11].

Outro problema que torna a segmentação uma tarefa tão complicada é devido ao efeito de co-articulação, o qual não causa uma descontinuidade aparente no sinal de voz [12][13]. O efeito da co-articulação se manifesta pela alteração do padrão articulatório de um determinado segmento sonoro pela influência de outro adjacente, ou próximo, na elocução completa [14]. Os efeitos da co-articulação fazem com que, por exemplo, o fonema "p" da palavra "paro" seja distinto do fonema "p" da palavra "puro". Neste último caso, o movimento articulatório necessário à produção do "u" resulta em uma mudança na pronúncia do fonema anterior (o "p"). Os fonemas são estruturas muito flexíveis, eles podem ser facilmente modificados ou suprimidos durante a pronúncia de uma elocução e, portanto, precisam de uma grande atenção na segmentação [13]. Existem técnicas que utilizam uma segmentação fixa de quadro (*frame*) de maneira explícita, baseada nos modelos ocultos de Markov (*Hidden Markov Models - HMM*), mas que acabam sofrendo com problemas específicos, sobretudo com o problema de seleção dos limites do fonema (*phoneme boundary selection*) [15]. Uma abordagem alternativa é utilizar uma segmentação implícita, sem fixação dos quadros, o que evita o problema de seleção dos limites do fonema.

1.1 **Processamento Digital da Voz**

O sinal de voz é uma onda acústica que carrega informação. Ele é contínuo no tempo e para ser chamado de 'sinal digital', ele precisa ser discretizado, e geralmente, é submetido a dois procedimentos: amostragem e quantização [16]. A amostragem é a obtenção de uma sequência de amostras provenientes de um sinal contínuo, em instantes de tempo igualmente espaçados. O processo de quantização consiste em atribuir um determinado número de níveis discretos em amplitude e fazer a comparação entre o sinal amostrado e o nível discreto mais próximo. O processamento de sinais [16] possui um enfoque na representação, transformação e manipulação dos sinais e da informação contida neles. O Processamento Digital de Voz (*Digital Speech Processing - DSP*) é uma subárea de processamento de sinais onde o sinal a ser estudado é o sinal de voz e a informação contida nele é um som produzido por um ser humano ou por um computador simulando um ser humano. A Figura 1.1 representa um sistema onde um sinal de voz é processado por técnicas de *DSP*. De acordo com Rabiner e Schafer [17], os sistemas de

processamento de voz estão divididos em quatro aplicações (a última não é mencionada diretamente, mas possui destaque nesta área):

1. Codificação da fala (*speech coding*);
2. Síntese da fala (*speech synthesis or text-to-speech synthesis - TTS*);
3. Reconhecimento automático da fala (*ASR*);
4. Supressão de ruído.

A primeira aplicação diz respeito aos processos cuja finalidade é obter uma representação compacta do sinal de voz, e, por esse motivo, também podemos chamá-la de "compressão do sinal de voz". As técnicas de codificação do sinal de voz são usadas tanto para transmitir como para armazenar os sinais de voz de maneira mais compacta que o sinal original.

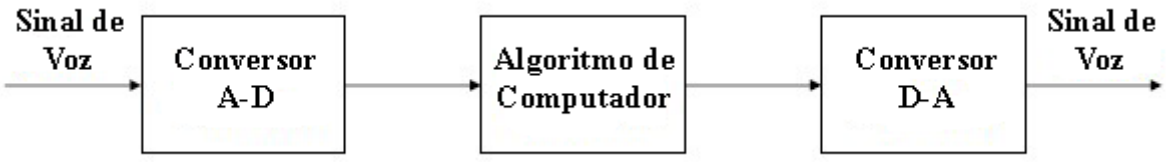


Figura 1.1: Diagrama de blocos padrão para aplicações em DSP para um sinal de voz (traduzido de [17]).

A síntese da fala é o ramo da *DSP* que estuda a síntese da voz humana através de computadores, ou seja, pesquisa a simulação digital da produção de voz do ser humano. Sendo ainda mais específico, essa área estuda a geração de sons da voz humana a partir do texto escrito, ou seja, conversão de texto em voz, por essa razão é comumente encontrado na literatura como *text-to-speech - TTS*.

O sistema de reconhecimento automático da fala possui como finalidade o reconhecimento da voz do ser humano. Em outras palavras, o objetivo de um sistema *ASR* é transcrever as frases que um ser humano pronuncia a ele. A Figura 1.2 ilustra a gama de aplicações baseadas em processamento digital de voz.

A supressão de ruídos em sinais de voz degradados visa o processamento e restauração de sinais de voz que sofreram distorções devido à introdução ou presença de ruído. O objetivo é reduzir, ou até mesmo eliminar o ruído presente nos sinais. As técnicas de filtragem são as mais comuns na utilização de supressão de ruídos em sinais de voz.

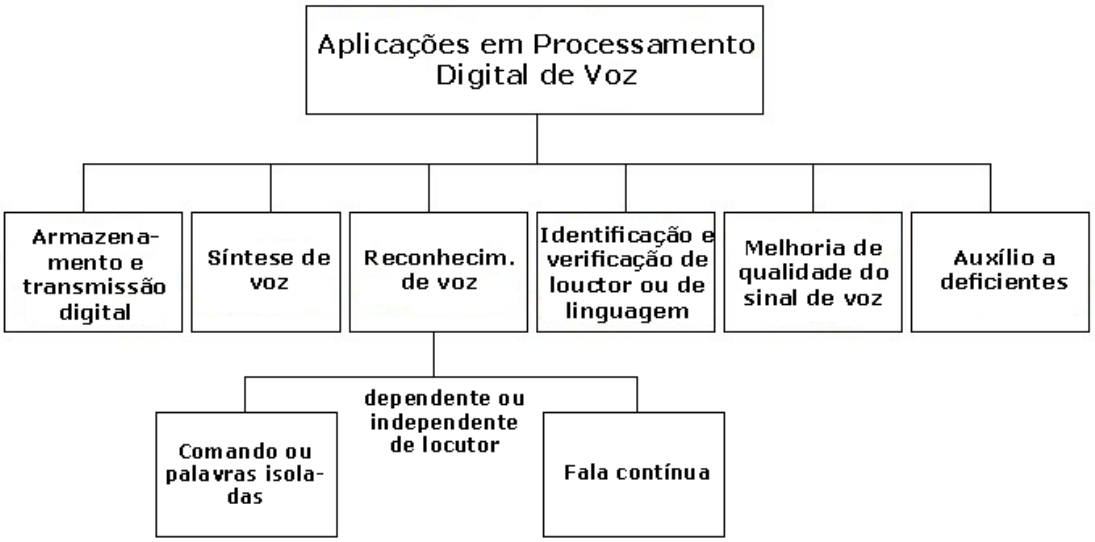


Figura 1.2: A extensa gama de aplicações em processamento digital de voz (adaptado de [21])

Uma das grandes dificuldades de se construir um sistema *ASR* robusto e eficiente é sua natureza interdisciplinar. Podemos citar algumas áreas de conhecimento que se aplicam ao problema de reconhecimento da fala: processamento de sinais, física acústica, reconhecimento de padrões, comunicação e teoria da informação, linguística, fisiologia, ciências da computação e psicologia. É uma gama de conhecimento tão ampla que é improvável que uma única pessoa possa dominar completamente esse tema [1].

Um sistema *ASR* que tenha o reconhecimento realizado com sucesso pode tomar decisões como ativar/desativar dispositivos ou mesmo digitar as palavras que lhe foram pronunciadas em um texto de *e-mail*. Existe ainda uma subdivisão no campo de reconhecimento automático da fala: Identificação do locutor (identificar a pessoa que está falando, também chamado de biometria da voz), Identificação da linguagem (o idioma que está sendo pronunciado por uma pessoa) e Reconhecimento de palavras (que pode ser o reconhecimento de uma palavra isolada ou o reconhecimento de fala contínua em uma conversa entre pessoas). Esses dois tipos de reconhecimento de fala podem ser realizados em um modo dependente do locutor, ou seja, apenas identifica comandos ou frases pronunciados por uma única pessoa ou podem ser realizados em um modo independente de locutor, que é o mesmo conceito do anterior, porém aplicado a diversas pessoas. Para alcançar esse objetivo final, os sistemas *ASR* possuem várias etapas, como pode ser observado na Figura 1.3. No próximo tópico, os sistemas *ASR* são discutidos com maiores detalhes. Este trabalho visa apresentar um novo algoritmo para o processo de segmentação da voz, representado na primeira etapa (de Pré-processamento) de um *ASR*.

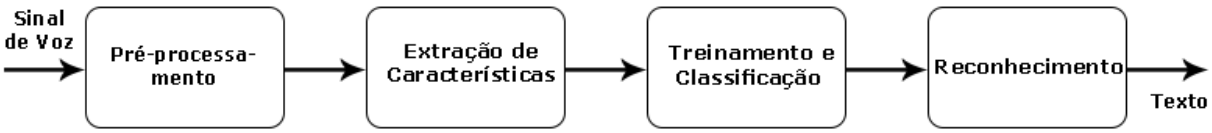


Figura 1.3: Etapas de um sistema *ASR*, adaptado de [17].

1.1.1 Sistema *ASR*

Neste tópico, são explanadas as etapas ilustradas na Figura 1.3. Existem vários modelos diferentes para representar um sistema *ASR*, porém todos apresentam o mesmo *layout* básico. Na Figura 1.3, pode-se visualizar um modelo de quatro etapas, que poderia ser de cinco, caso considerássemos a etapa de "aquisição do sinal de voz" como sendo a predecessora do "sinal de voz" apresentado na figura. Essa etapa de "aquisição do sinal de voz" é responsável pela obtenção do sinal de voz analógico e em seguida pela discretização em um sinal digital (como já foi explicado no início desta seção), para que seja possível a aplicação das técnicas de *DSP*. O que é mostrado na Figura 1.3 é um sinal de voz já digitalizado e, portanto, pronto para o processamento.

A etapa de "pré-processamento" é composta por vários outros processos de transformação do sinal, tais como: filtragem (remoção de ruídos), normalização, segmentação, entre outras. Apesar de ser "apenas" um processo dentro da etapa de pré-processamento do sinal de voz, a segmentação é uma das tarefas mais complexas em um *ASR* [3].

A etapa de "extração de características" de um sinal de voz é responsável por converter o sinal de voz em um vetor de características. Essa conversão é geralmente baseada em análises em pequenas janelas do sinal no domínio do tempo, que são capazes de representar com eficiência as características de um sinal sobre esse domínio. Não existe um conjunto padrão de características para o reconhecimento de voz. Ao invés disso, combinações de características acústicas, articulatórias e auditivas são utilizadas em vários sistemas de reconhecimento de voz. O descritor de características acústicas mais popular é, sem dúvida, o *MFCC (Mel-Frequency Cepstrum Coefficients)* e seus derivados [18].

Na verdade, a etapa de "treinamento e classificação" é, de fato, caracterizada por duas etapas, pois o treinamento é realizado antes da classificação. Após o sistema estar devidamente

treinado, os sinais de entrada são direcionados para a classificação. Caso sejam inseridos novos padrões de voz, será necessário um novo treinamento do sistema. Como essa etapa não é o foco do trabalho, e por motivos de simplicidade, convém juntar essas duas etapas em apenas uma. Dessa forma, a etapa de treinamento serve para que o sistema aprenda a reconhecer o conjunto de vetor de características do sinal e reconheça o padrão de voz pronunciado. Quando o sistema está treinado, a classificação realiza a escolha do padrão de voz mais adequado (de acordo com sua lógica de implementação) para o sinal de voz de entrada. O classificador mais utilizado em sistemas *ASR* é o *HMM (Hidden Markov Models)* [19].

1.2 **Motivação**

O processo de comunicação entre os humanos é baseado, principalmente, na habilidade de reconhecer e compreender os sinais de voz que são pronunciados entre eles. Portanto, os sinais de voz definem uma parte essencial da interação entre os humanos, e é, sem dúvida, a maneira mais eficiente e eficaz para a troca de informações. Por conta disso, o reconhecimento automático de fala e também a identificação de locutor (*Automatic Speaker Identification – ASI)* surgem como grandes atrativos de uma tecnologia repleta de possibilidades para novos produtos e serviços de comunicação. Na busca por soluções que alcancem esses exemplos, diversos modelos e algoritmos de reconhecimento de voz vêm sendo desenvolvidos e aperfeiçoados na última década, sempre procurando atingir melhores resultados com um menor custo computacional.

Para alcançar os objetivos mencionados, é imprescindível que alguns subcampos da área de reconhecimento de voz avancem em suas pesquisas e produzam métodos mais robustos. Em muitas abordagens do reconhecimento de voz, os sinais de voz precisam ser divididos em segmentos antes de o reconhecimento ser efetuado de fato. A segmentação de voz é uma importante tarefa para a construção de um sistema *ASR* [4]; ela é fundamental para o treinamento dos modelos acústicos de um *ASR* e para o desenvolvimento de sistemas do tipo *TTS*. À medida que existe um crescente desenvolvimento de sistemas que usam fala para uma interface homem-máquina, a demanda por uma segmentação automática confiável também cresce, visto que a segmentação é um processo importante e necessário em diversas áreas do processamento de fala.

Uma vez que a segmentação manual é inviável devido ao tamanho das bases de fala atuais, além de ser um processo tedioso, e que gera diferenças, já que é um processo subjetivo, este trabalho tem como objetivo principal desenvolver um sistema de segmentação automática de fala que possa produzir segmentos acústicos a partir de uma locução, evitando dessa forma as incoerências e a demora causada pela segmentação manual.

A motivação para combinar as técnicas de processamento digital de imagens com técnicas de processamento do sinal de voz é proveniente, principalmente, da clareza com que as informações são fornecidas pela imagem do espectrograma.

1.3 Objetivos

Neste trabalho, é proposta uma nova técnica de processamento de voz, baseada no sinal de voz e em características do processamento digital de imagens (*digital image processing - DIP)* como análise de espectrograma, morfologia matemática, componentes conectados, análise de projeção e binarização. A técnica opera em dois passos e é capaz de segmentar voz e sílabas fonéticas mesmo em ambientes ruidosos. A sílaba fonética pode ser considerada um agrupamento de fonemas e ela é mais bem destacada na imagem do espectrograma, em comparação ao fonema isolado.

Outros objetivos também podem ser citados:

- Expandir o estado-da-arte de uma área ainda sem solução ótima e de grande aplicabilidade;
- Avaliar a combinação de técnicas de processamento de imagens com técnicas que operam sobre os sinais de voz;
- Contribuir para a disseminação e pesquisa de segmentação de voz no Brasil, ajudando na formação de capital humano nacional;
- Geração de material didático na área de segmentação, ajudando na exemplificação de aplicações práticas para as disciplinas de processamento de voz e sinais.

1.4 Estrutura da Dissertação

Este trabalho encontra-se estruturado em seis capítulos.

No Capítulo 1, é introduzida uma visão geral sobre o processamento de sinais de voz, segmentação da voz e reconhecimento automático de fala, além da motivação e importância da área de pesquisa. Por fim, os objetivos que devem ser alcançados no trabalho também são listados.

No Capítulo 2, explicam-se os conceitos básicos relativos aos procedimentos e processos abordados na pesquisa. Estes processos são divididos em técnicas de processamento digital de voz e técnicas de processamento digital de imagem. O capítulo inicia com um esboço do processo de produção de voz pelos seres humanos.

No Capítulo 3, é apresentado o estado da arte de algoritmos de segmentação de voz e explica os algoritmos que contribuem para o desenvolvimento do algoritmo proposto neste trabalho.

No Capítulo 4, o novo algoritmo de segmentação de voz proposto neste trabalho é apresentado e explicado em detalhes.

No Capítulo 5, são relatados os testes e experimentos, além dos resultados obtidos e análises sobre os mesmos.

O Capítulo 6 finaliza a dissertação com uma conclusão sobre o trabalho, contribuições que ele ofereceu e sugestões para trabalhos futuros.

2 Conceitos Básicos

Este capítulo visa a demonstrar os principais conceitos relacionados às áreas abordadas no trabalho. Desde já se reporta que o objetivo não é fornecer um *background* completo das áreas relacionadas, mas sim explicar os pontos necessários para o entendimento da presente dissertação, isto é, as questões envolvidas nos algoritmos ou processos utilizados na pesquisa.

2.1 Processo de produção de voz pelos seres humanos

Segundo Levelt [22], o processo de produção de sons (no caso dos seres humanos, a voz) se inicia com a construção de uma mensagem na mente, estágio conhecido como conceitualização. Devido à complexidade que o nosso cérebro possui, ainda se sabe pouco sobre esse estágio. Logo em seguida, o locutor realiza uma etapa de formulação, onde ele converte a mensagem para um padrão linguístico, levando em consideração as regras léxicas (seleção das palavras apropriadas) e sintáticas (ordenação correta das palavras) da língua em questão. Por fim, o locutor precisa realizar os movimentos motores necessários para externalização e consequente transmissão da mensagem, essa etapa é a mais relacionada com o objeto de estudo desta dissertação e é descrita na próxima seção, ela é chamada de etapa de articulação. A Figura 2.1 ilustra os três estágios explicados em uma abordagem de alto nível para produção da voz.

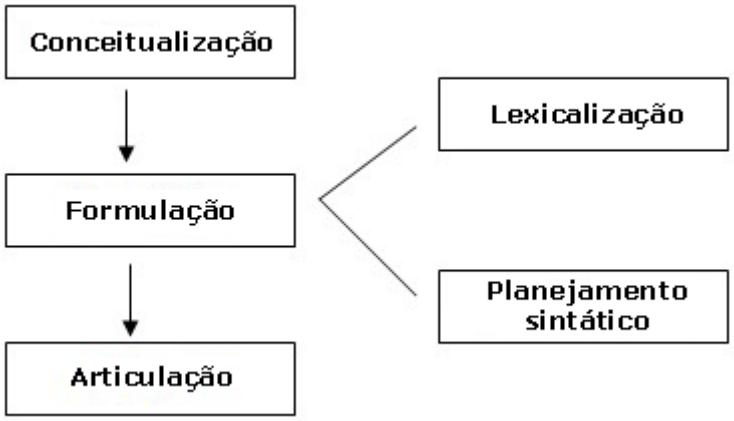


Figura 2.1: Esquema em alto nível do processo de produção da voz (traduzido de [22])

2.1.1 Articulação da voz

Após a conceitualização e formulação da mensagem, nós precisamos realizar a produção do som que representa a mensagem, é nesse momento que o estágio da articulação atua. Esse estágio geralmente é subdividido em quatro processos: iniciação, fonação, *oro-nasal* e a articulação propriamente dita [23].

A iniciação, como diz o nome, é o primeiro processo, ocorre quando o ar é expelido dos pulmões e entra pela traqueia, onde inicia o segundo processo.

A traqueia é a "porta de entrada" da laringe (Figura 2.2 (c)). O processo de fonação ocorre ao longo da laringe, principalmente onde estão dois tecidos horizontais chamados de "pregas vocais". O espaço entre as pregas vocais é chamado de glótis, e esta tem um papel fundamental na produção do som. A glótis pode permanecer fechada, como na Figura 2.2 (a), e, portanto, nenhum ar passa, ou pode estar estreitamente aberta, possibilitando a passagem do ar pelas pregas vocais, fazendo-as vibrar e consequentemente produzir sons audíveis (*voiced sounds)*. A glótis também pode estar totalmente aberta, que é a situação em que estamos respirando normalmente, dessa forma a vibração das pregas vocais é reduzida, produzindo os sons inaudíveis (*voiceless sounds)* [23]. As pregas vocais vibram muito rapidamente. Nos homens, esse número de ciclos vibratórios fica em torno de 125 vezes em um segundo. Na mulher, que tem voz geralmente mais aguda, o número aumenta para 250 vezes por segundo. A essa característica damos o nome de frequência fundamental.

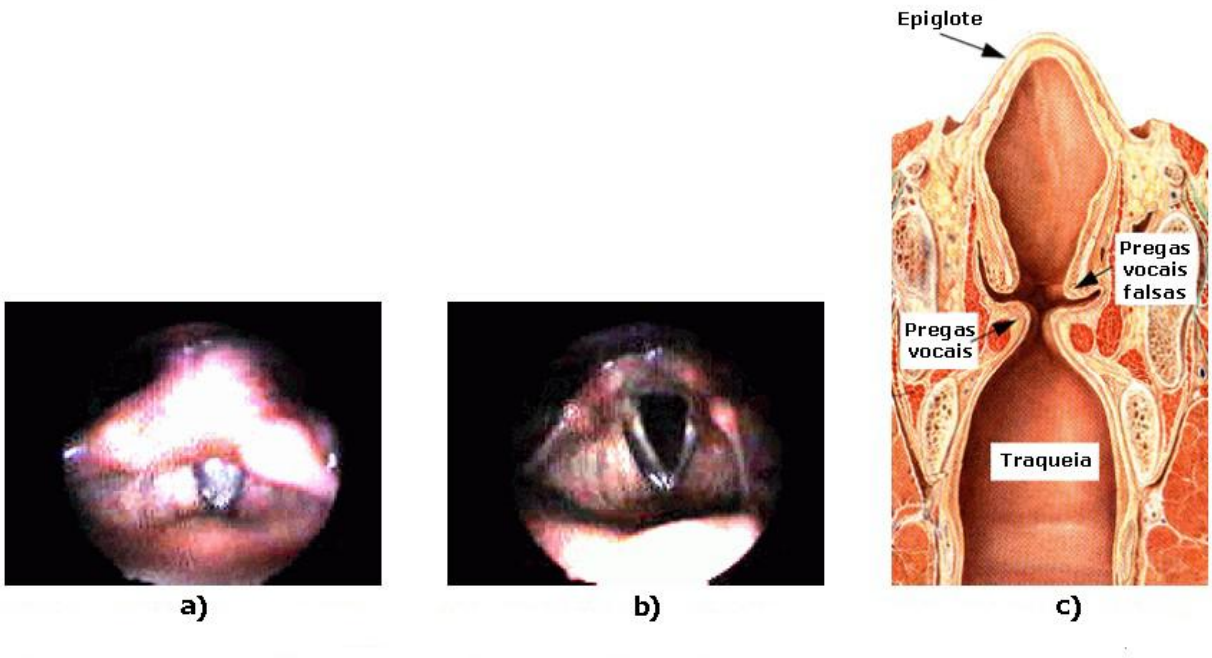


Figura 2.2: Componentes do processo de fonação: (a) glótis fechada, (b) glótis aberta e (c) ilustração básica da laringe, adaptado de [24].

Após passar pela laringe, o fluxo de ar é "cortado" em pulsos quase-periódicos os quais são modulados na frequência enquanto passam através da faringe (cavidade da garganta) e da cavidade oral ou nasal [21]. O véu-palatino é o responsável por selecionar se o fluxo de ar vai passar pela cavidade oral ou nasal (Figura 2.3). Por meio do processo *oro-nasal*, podemos diferenciar as consonantes nasais (/m/, /n/) dos outros sons.

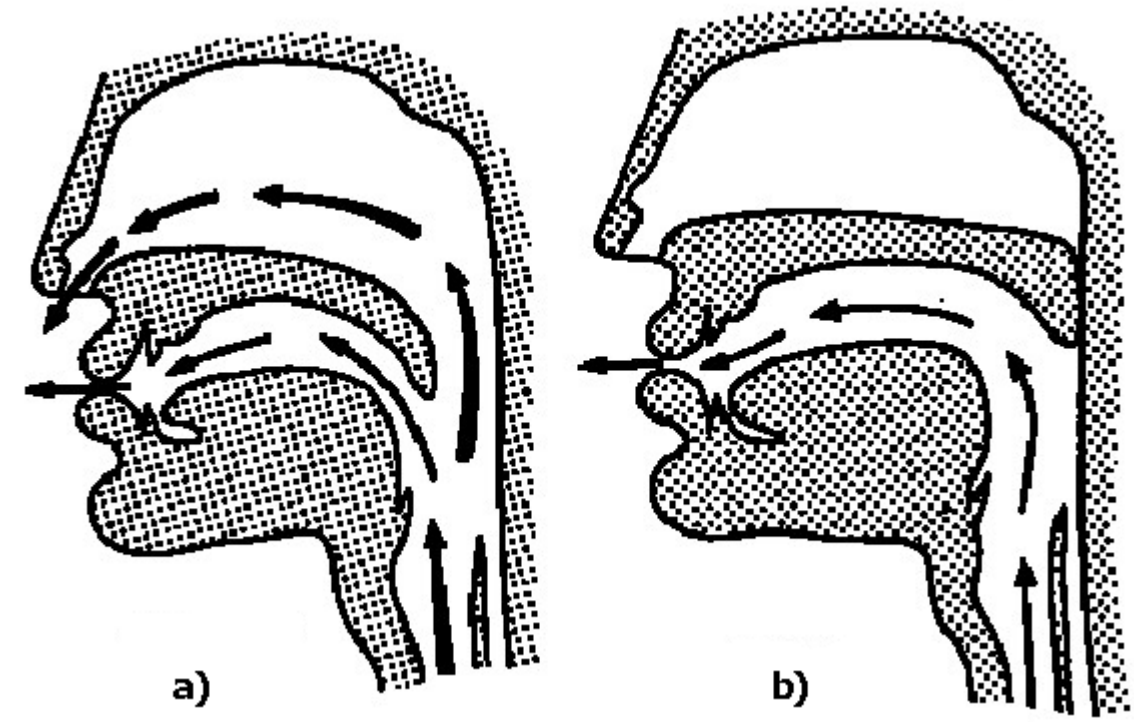


Figura 2.3: O processo *oro-nasal*: (a) articulação nasal – o fluxo de ar passa pelas duas cavidades, (b) articulação oral – o fluxo de ar passa apenas pela cavidade oral, adaptado de [24].

Por último, o processo de articulação é efetuado na boca e é o responsável por diferenciar e "moldar" os sons para aquilo que reconhecemos como sons de fala. Dentro da boca, a cavidade oral funciona como um ressoador do som e os articuladores (lábios, língua, dentes, palato duro, véu palatar e mandíbula) trabalham sozinhos ou em conjunto para, de fato, produzir os diferentes tipos de sons que percebemos no nosso dia-a-dia (vogais, consoantes, ditongos etc).

A Figura 2.4 apresenta um diagrama de blocos para o processo de produção dos sons pelo ser humano.

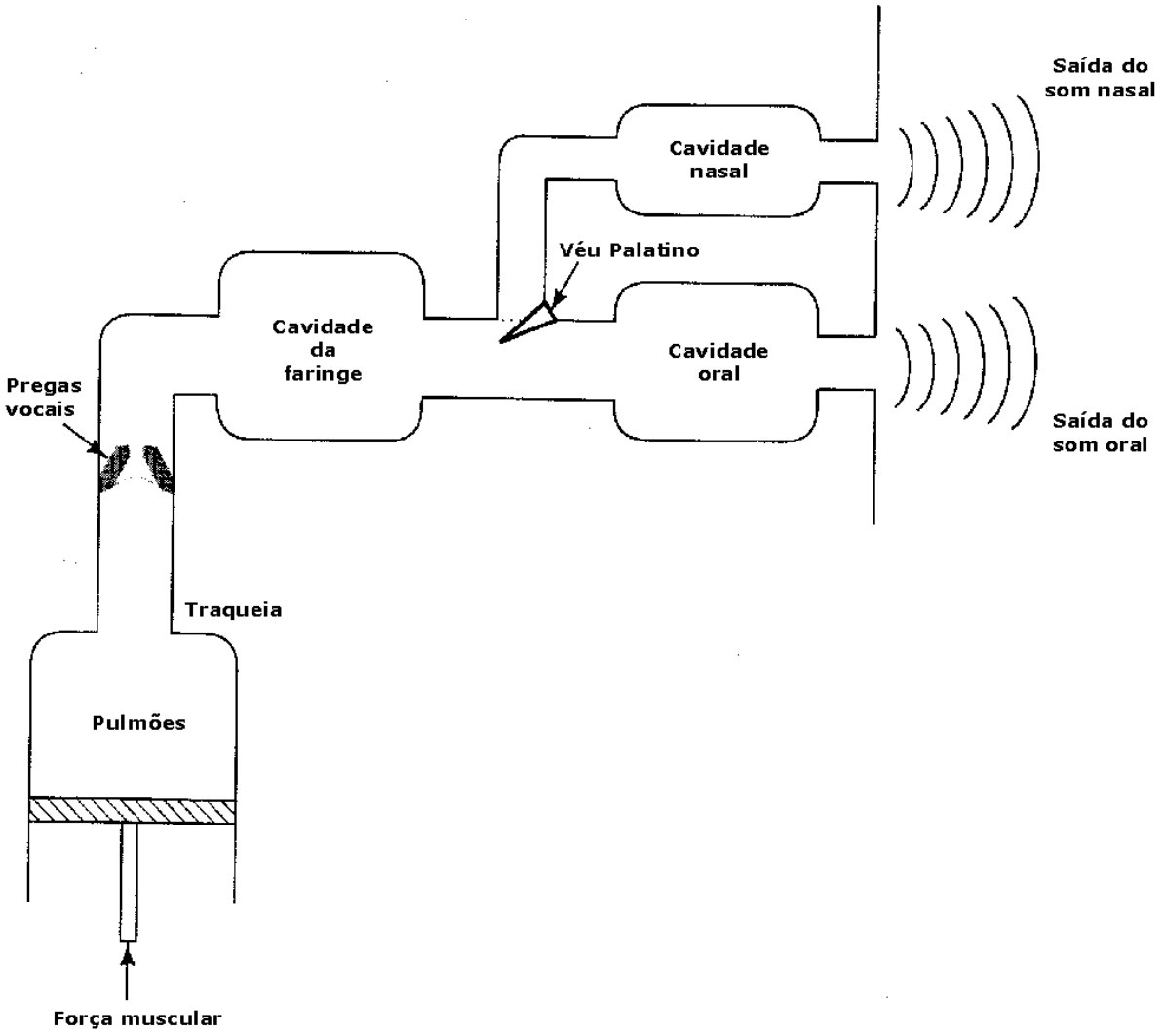


Figura 2.4: Diagrama de blocos para produção da voz humana, adaptado de [1].

2.2 Processo de percepção de som pelos seres humanos

Os sons são ondas mecânicas, criadas quando corpos materiais vibram. Essas vibrações causam mudanças na pressão de um determinado meio (ou ambiente). O meio em que vivemos é a atmosfera terrestre e, por isso, os sons que ouvimos causam uma variação na pressão do ar para chegar até nossos ouvidos. As características básicas de uma onda sonora são a amplitude e a frequência. A amplitude é a diferença entre a maior e a menor área de pressão deslocada. A amplitude está relacionada com a sonoridade, ou seja, com a intensidade do nível de som (forte ou fraco). Já a frequência é o número de vezes por segundo que um padrão de variação de pressão se repete, e está relacionada com a percepção de grave e agudo [2].

Nos seres humanos, a função auditiva é realizada por um órgão que chamamos de ouvido, ele é o responsável pela captação das vibrações no ar (ondas sonoras) e pela transformação delas

em impulsos nervosos. No entanto, muitos pesquisadores defendem que grande parte do processo de audição depende do processamento de dados que é realizado no sistema nervoso central [25].

2.2.1 Anatomia e fisiologia do ouvido

A anatomia ou fisiologia do ouvido é geralmente dividida em três partes: ouvido externo, ouvido médio e ouvido interno. Nesta dissertação, vamos abordar de maneira resumida cada uma dessas partes.

- Ouvido externo: Formado pelo pavilhão auricular e pelo canal auditivo, o qual é encerrado pelo tímpano. A função do ouvido externo é contribuir para determinação da direção da origem dos sons. Além disso, o canal auditivo funciona como um ressoador, ele é capaz de amplificar os sons com frequências de aproximadamente 3000 Hz. As ondas sonoras que entram pelo canal auditivo são então conduzidas até o tímpano.
- Ouvido médio: Consiste de três pequenos ossos, também chamados de ossículos, cujos nomes são: martelo, bigorna e estribo. Eles são conectados ao ouvido externo através do tímpano. Esses ossículos atuam como uma alavanca, o que altera a pressão exercida por uma onda sonora ao tímpano, numa maior pressão sobre a janela oval no ouvido interno. Além disso, a diferença nas áreas da janela oval e do tímpano resulta na amplificação do sinal.
- Ouvido interno: Consiste de canais semicirculares, do nervo auditivo e de uma coclea. Os canais semicirculares são detectores do equilíbrio do corpo humano, porém contribuem muito pouco para o sistema auditivo. A coclea contém todos os mecanismos responsáveis pela transformação das variações de pressão em impulsos nervosos propriamente ditos, que passam ao longo do nervo auditivo para o cérebro. Em um proceso que ainda não é compreendido inteiramente, o cérebro é capaz de interpretar as qualidades do som pela reação dos impulsos nervosos. O ouvido interno é conectado com o ouvido médio através da janela oval.

2.2.2 Percepção do som

O sistema auditivo, de uma maneira, geral, trabalha de uma forma não linear. O comportamento fisiológico dos nossos ouvidos em resposta a simples tons é relativamente simples. No entanto, a maior parte dos sons é variante no tempo e possui diversos componentes espectrais. O resultado disso é que a percepção de energia sonora em uma determinada frequência é dependente da distribuição do som em outras frequências e do intervalo de tempo da energia antes e depois do som. Aquilo que uma pessoa ouve em resposta a um determinado som

é uma questão bastante complexa [2]. Por conta disso, os pesquisadores buscam compreender os mecanismos de percepção da voz nos seres humanos para entender o relacionamento complexo entre o sinal acústico e o som que de fato escutamos. Existem duas razões principais para a complexidade: o problema de segmentação e a variabilidade do sinal acústico. A falta de solução para esses dois problemas dificulta bastante a construção de máquinas capazes de reconhecer discursos contínuos de voz.

2.2.3 O problema da segmentação

Quando escutamos alguém falar, nós conseguimos segmentar facilmente as frases em palavras isoladas. Se analisarmos esse sinal acústico, nós vamos constatar que ele não é perfeitamente separado em palavras ou fonemas. O sinal acústico é contínuo e não há, necessariamente, pausas específicas no sinal, pelo contrário, pode haver pausas que não correspondem às pausas que percebemos naturalmente entre as palavras ou fonemas.

Em alguns casos, o contexto também é necessário para a segmentação ocorrer corretamente, pois duas palavras diferentes podem ter uma estrutura fonética muito semelhante, ou até igual, no caso das palavras homófonas, que possuem pronúncias iguais e significados diferentes. No português brasileiro podemos citar exemplos como: “espere a mente raciocinar” e “experimente isso”; “Ela tinha algo” e “Pega a latinha!”; “Ele foi ao concerto” e “Ele realizou o conserto da peça”. No inglês, podemos citar exemplos como: “*I scream*” e “*Ice cream*”; “*Ruth*” e “*Roof*”; “*Them all*” e “*The mall*”. Nesses exemplos citados, devem ser realizadas diferentes segmentações de acordo com o significado da sentença em que cada palavra se encontra.

A segmentação fonética do sinal de voz é essencialmente uma tarefa de determinar a instância de tempo em que um fonema termina e outro inicia [10][11]. Contudo, ainda não foi descoberto nenhum evento físico que acontece na produção da voz, no momento em que ocorre essa transição de fonemas, apenas sabe-se que os órgãos vocais movem-se lentamente de uma posição para outra. Apesar de vários anos de pesquisa, determinar onde um fonema termina e o outro inicia continua sendo uma difícil tarefa nessa área [10][11].

2.2.4 O problema da variabilidade

O fonema pode ter diferentes formas que são determinadas de acordo com a variedade de fontes do som. Este fenômeno costuma ser chamado de “problema da não invariância fonético-acústica” [26]. No contexto de fonemas, espera-se que eles possuam alguma característica acústica que sirva para diferenciar um fonema de outro. Apesar disso, os pesquisadores ainda não encontraram provas da real existência dessa invariância fonética para todos os fonemas [26].

2.2.4.1 Variabilidade de acordo com o contexto do fonema

De acordo com o contexto da frase, as propriedades acústicas do sinal associado com os fonemas podem mudar. Esse efeito é resultante da maneira como a voz é produzida. Quando falamos, os articuladores estão em constante movimento, portanto a forma do trato vocal para um determinado fonema é influenciada pelas formas dos fonemas antecedentes e posteriores. Esse fenômeno é também conhecido por co-articulação e foi exemplificado no Capítulo 1.

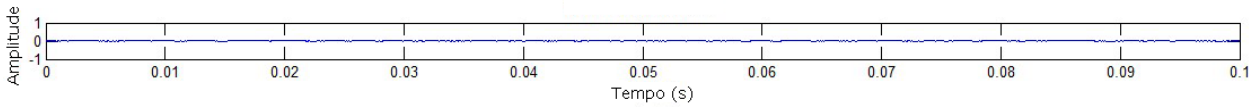
2.2.4.2 Variabilidade de acordo com o locutor

Diferentes locutores podem produzir sinais acústicos muito distintos para um mesmo fonema. As pessoas pronunciam os mesmos fonemas e palavras com sotaques, timbres e velocidades diferentes. Além disso, a estrutura do trato vocal pode mudar ligeiramente de acordo com cada pessoa. A maneira "desleixada" (ou informal) que muitas pessoas pronunciam as palavras também introduz variabilidade no sinal acústico.

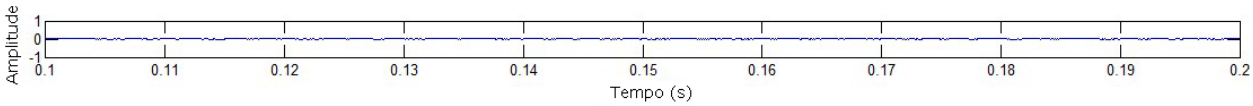
A variabilidade do sinal acústico, causada pelas razões explicadas nos parágrafos anteriores, gera diversos problemas para o ouvinte. Os sinais de voz variam bastante e geralmente são transformados pelo nosso cérebro em palavras familiares. Por conta do problema de segmentação e o problema da variabilidade, tem sido difícil a concepção de máquinas que possam reconhecer a fala contínua [1].

2.3 Representação do sinal de voz

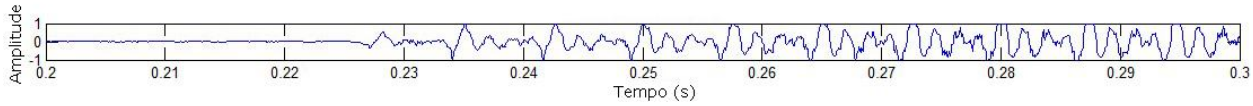
O sinal de voz é um tipo de onda que varia lentamente com o tempo, de maneira que, se uma análise for realizada em períodos curtos de tempo (entre 5 e 20 milisegundos), as características resultantes serão praticamente estacionárias. A ilustração desse efeito é apresentada na Figura 2.5. Vale ressaltar que, se períodos de tempo acima de 100ms forem considerados, as características do sinal podem ser diferentes e consequentemente podem refletir as variações dos sons de falas que estão sendo pronunciados naquele instante [1].



(a) Sinal de 0 a 100ms



(b) Sinal de 100 a 200ms



(c) Sinal de 200 a 300ms

Figura 2.5: Gráfico em forma de onda do sinal de voz da parte inicial da pronúncia "É agora".

O sinal praticamente não varia nos primeiros 200ms (primeiro e segundo gráficos da Figura 2.5), o que corresponde ao silêncio de fundo e, portanto, possui uma amplitude baixa. A terceira linha (de 200 a 300ms) inicia com um breve período de silêncio, acompanhado por um pequeno aumento na amplitude (som inaudível, próximo aos 0,23s) e em seguida um aumento maior na amplitude e uma mudança no desenho e regularidade da forma de onda, transformando-se em uma onda quase-periódica.

A forma mais simples e direta de se classificar eventos em sinais de voz é a partir do estado das cordas vocais. Seguindo essa ideia, a convenção padrão utiliza uma representação de três estados: silêncio (S - *Silence*), onde nenhum som está sendo pronunciado; som inaudível (U – *unvoiced*), onde as cordas vocais não estão vibrando, resultando em um sinal aperiódico, de natureza aleatória (ruído); e som audível, onde as cordas vocais estão vibrando periodicamente e o sinal resultante é considerado quase-periódico. A aplicação desse tipo de classificação no terceiro gráfico (sinal de 200 a 300ms) da Figura 2.5 resulta na Figura 2.6.

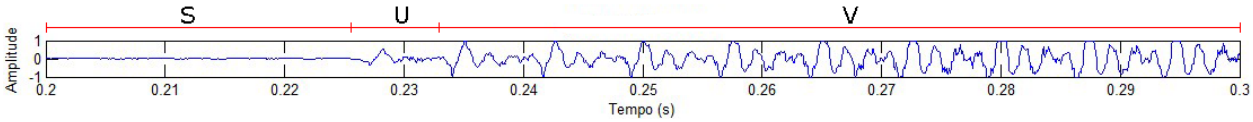


Figura 2.6: Classificação do sinal de voz a partir do estado das cordas vocais.

No trecho inicial, antes do usuário falar, a forma de onda é classificada como siêncio, indicado pelo 'S', e é caracterizado pela amplitude praticamente constante e próxima de zero. No segundo trecho, indicado pelo 'U', há um som inaudível, que resulta em uma pequena modificação na forma de onda, especificamente nesse caso, é o momento em que o elocutor aspira o ar antes de iniciar a pronúncia da elocução. Uma voz baixa de fundo, um sussuro, alguns tipos de ruídos e alguns fonemas também podem ser caracterizados como sons inaudíveis. O

último trecho da figura corresponde à voz do elocutor de fato, e nesse caso, se refere ao ínicio da frase "É agora", mais especificamente, esse trecho corresponde ao fonema /é/. Apesar de simples, esse tipo de classificação não é muito preciso e, por vezes, é difícil distinguir entre os sons inaudíveis fracos (como /f/), os sons audíveis fracos (como /v/ ou /m/) e o silêncio de fundo. O gráfico completo da frase pronunciada nos exemplos acima pode ser visualizado na Figura 2.7.

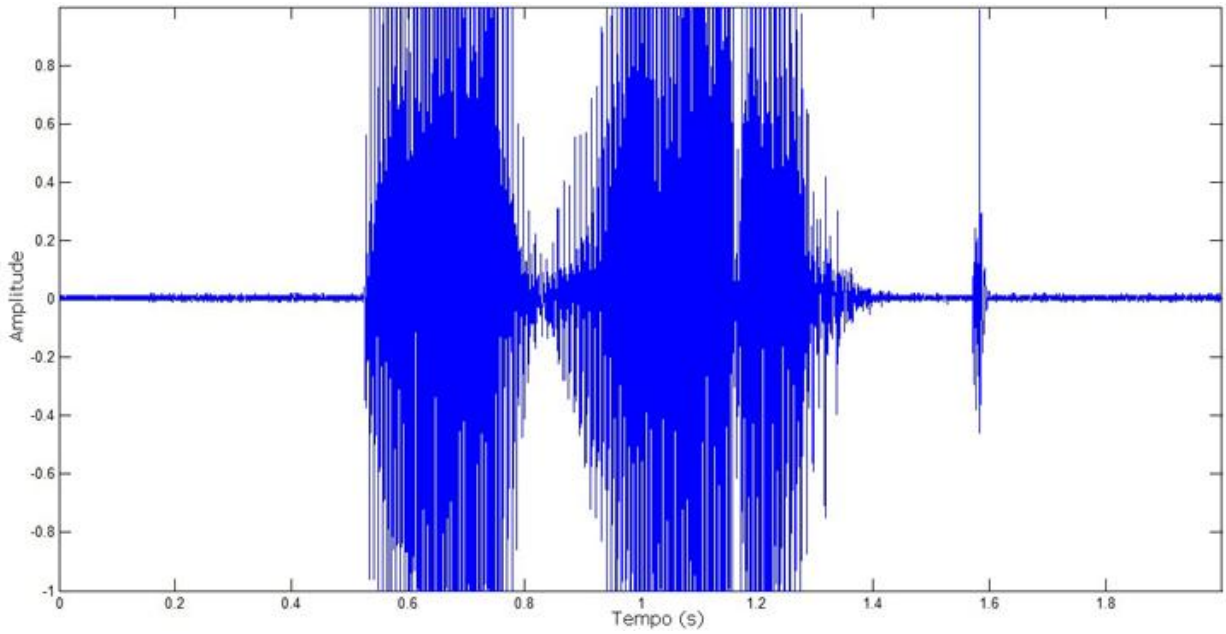


Figura 2.7: Gráfico em forma de onda do sinal de voz da pronúncia "É agora".

Uma maneira alternativa de caracterizar o sinal de voz e apresentar a informação associada aos sons é através da representação espectral. No nosso contexto, o espectrograma de um sinal de voz é uma representação em três dimensões da intensidade da voz, em diferentes bandas de frequência, sobre o tempo. Geralmente, utilizamos apenas os valores absolutos para a amplitude do sinal de voz, removendo a parte negativa dele. O espectrograma é calculado através da Transformada de Fourier de Tempo Curto (*short-time Fourier transform – STFT*). Para realçar o gráfico do espectrograma, normalmente calcula-se a energia do sinal, elevando os valores do sinal ao quadrado. A Figura 2.8 ilustra o gráfico do espectrograma para a energia do sinal de voz da pronúncia "É agora". Os pontos mais vivos (laranjas e vermelhos) representam uma maior intensidade do sinal de voz de acordo com o tempo e a frequência em que ele se encontra. A escala de cores utilizada no espectrograma da Figura 2.8 é a mesma para todas as imagens de espectrograma deste trabalho.

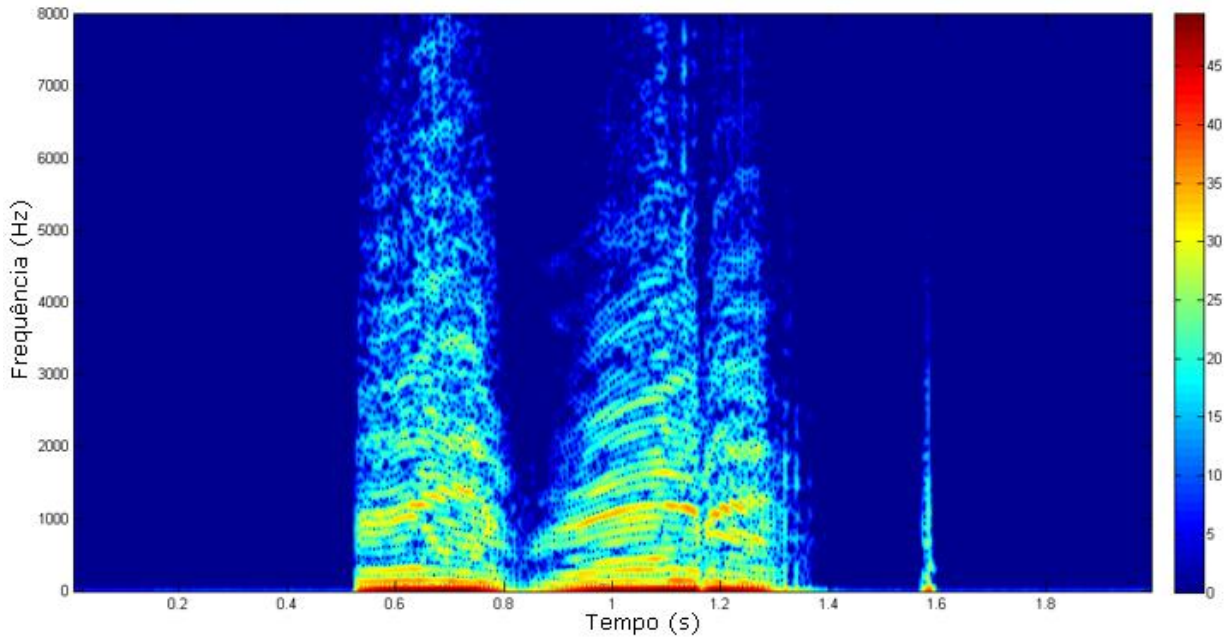


Figura 2.8: Espectrograma da energia do sinal de voz da pronúncia "É agora". Os pontos mais vivos (laranjas e vermelhos) representam uma maior intensidade do sinal de voz.

2.4 Estudo de técnicas de processamento digital de imagens

Ao longo dos últimos anos, o interesse em técnicas de processamento digital de imagens (PDI) vem aumentando bastante [27]. Diversas aplicações utilizam algoritmos associados ao processamento digital de imagens, seja para segurança, saúde, automação, entretenimento etc [28]. Normalmente, as técnicas de PDI são utilizadas para melhorar a qualidade das imagens, tornando mais compreensível o reconhecimento por um humano (ou por uma máquina), e/ou para fazer a compressão das imagens para transmissão e/ou armazenamento.

Os sistemas de processamento de imagens são sistemas que recebem uma imagem de entrada, realizam algum tipo de processo sobre a imagem e devolvem uma imagem de saída. Neste trabalho, foram aplicadas algumas técnicas de processamento digital de imagens (explicadas a seguir) na imagem resultante do espectrograma de energia de um dado sinal acústico.

2.4.1 Morfologia matemática

A morfologia matemática, elaborada inicialmente por Georges Matheron e Jean Serra [29], concentra seus esforços no estudo da estrutura geométrica das entidades presentes em uma imagem. A morfologia matemática pode ser aplicada em várias áreas de processamento e análise de imagens, como: realce, filtragem, segmentação, detecção de bordas, esqueletização, afinamento, entre outras. A ideia geral da teoria consiste em extrair as informações relativas à forma de um conjunto desconhecido (uma imagem), pela transformação através de outro

conjunto completamente definido, chamado elemento estruturante. Portanto, a base da morfologia matemática é a teoria de conjuntos. A morfologia matemática é baseada em duas operações básicas, a saber: dilatação e erosão. Nos próximos parágrafos, serão apresentadas algumas definições úteis sobre a teoria dos conjuntos, no intuito de facilitar o entendimento sobre as operações de dilatação e erosão.

Sejam A e B conjuntos em Z^2, cujos componentes são $a = (a_1, a_2)$ e $b = (b_1, b_2)$, respectivamente. A translação de A por $x = (x_1, x_2)$, denotada $(A)_x$, é definida na equação

$$(A)_x = \{ c | c = a + x, \ para \ \ a \in A \}. \quad \textbf{(1)}$$

A reflexão de B, denotada $\hat{B}$, é definida como

$$\hat{B} = \{ x | x = -b, \ para \ \ b \in B \}. \quad \textbf{(2)}$$

O complemento do conjunto A é

$$A^c = \{ x | x \notin A \}. \quad \textbf{(3)}$$

Por último, a diferença entre dois conjuntos A e B, representado por $A - B$ é

$$A - B = \{ x | x \in A, x \notin B \} = A \cap B^c. \quad \textbf{(4)}$$

Como já dito anteriormente, as operações fundamentais da morfologia matemática são a dilatação e a erosão. A dilatação, de uma maneira simples, torna os objetos mais largos e de maneira oposta, a erosão torna os objetos mais estreitos. Sejam A e B conjuntos no espaço Z^2 e seja $\emptyset$ o conjunto vazio. A dilatação de A por B, denotada $A \oplus B$, é definida como

$$A \oplus B = \{ x | (\hat{B})_x \cap A \neq \emptyset \}. \quad \textbf{(5)}$$

Logo, a operação de dilatação consiste em obter a reflexão de B sobre sua origem e depois deslocar esta reflexão de x. A dilatação de A por B é, então, o conjunto de todos os x deslocamentos para os quais a interseção de $(\hat{B})_x$ e A inclui, pelo menos, um elemento diferente de zero. Com base nisso, a equação anterior pode ser reescrita como

$$A \oplus B = \{ x | [(\hat{B})_x \cap A] \subseteq A. \quad \textbf{(6)}$$

O conjunto *B* é normalmente chamado de elemento estruturante.

Sejam *A* e *B* conjuntos no espaço Z^2, a erosão de *A* por *B*, denotada A $\ominus$ B, é definida como

$$A \ominus B = \{ x | (B)_x \subseteq A \}. \quad \textbf{(7)}$$

O que significa que a erosão de *A* por *B* resulta no conjunto de pontos *x* tais que *B*, transladado de *x*, está contido em *A*. Em outras palavras, o elemento estruturante (*B*) irá percorrer toda a imagem e, a cada iteração, verificará se todos os *pixels* ativos na vizinhança do *pixel* de origem estão sobrepostos por *pixels* ativos na imagem, em caso positivo o *pixel* de saída fica ativo, em caso negativo fica inativo. A dilatação e a erosão são operações duais entre si com respeito à complementação e reflexão, de tal forma que

$$(A \ominus B)^c = A^c \oplus B^c. \quad \textbf{(8)}$$

2.4.2 Análise de projeção (vertical ou horizontal)

A projeção horizontal corresponde à soma dos valores perpendiculares à coordenada horizontal (normalmente a coordenada de eixo *x*), enquanto que a projeção vertical corresponde à soma dos valores perpendiculares à coordenada vertical (normalmente a coordenada de eixo *y*) [30]. Essa técnica é comumente utilizada para realizar a segmentação de linhas em imagens de documentos [31][32][33], onde ela realiza a contagem do número de *pixels* pretos em cada linha do documento. As linhas que tiverem um número de *pixels* acima de certo limiar são consideradas texto, e as restantes são *background* (fundo). As projeções horizontais e verticais são definidas, respectivamente, por P_h de tamanho *m* e P_v de tamanho *n*, sendo

$$P_h(i) = \sum_{j=1}^{n} Im(i,j), \qquad i = 1, 2, 3, \dots, m\ , \quad \textbf{(9)}$$

$$P_v(j) = \sum_{i=1}^{m} Im(i,j), \qquad j = 1, 2, 3, \dots, n\ , \quad \textbf{(10)}$$

em que *Im* é uma imagem de *i* linhas e *j* colunas.

2.4.3 Binarização

A binarização (ou limiarização) consiste em separar as regiões de uma imagem quando esta apresenta duas classes (o fundo e o objeto). A binarização é uma das abordagens mais importantes da segmentação de imagens. Como já diz o nome, ela tem como finalidade produzir uma imagem binária com duas cores: preto e branco. Segundo Gonzalez [27], matematicamente, a limiarização pode ser descrita como uma operação que envolve testes de uma função *T*, sendo

$$T = f(\, x, y, p(x, y)\,). \tag{11}$$

em que *f(x, y)* é o nível de cinza do ponto *(x, y)* e *p(x, y)* denota alguma propriedade local desse ponto. Uma imagem limiarizada *g(x, y)* é definida como

$$g(x, y) = \begin{cases} 1 \; se \; f(x, y) > T \\ 0 \; se \; f(x, y) \leq T \end{cases} \tag{12}$$

Os *pixels* rotulados como ‘1’ são correspondentes ao fundo (ou *background*) da imagem, e os *pixels* rotulados como ‘0’ correspondem ao objeto, ou ao texto se tomarmos como referência um texto preto escrito sobre fundo branco.

3 Algoritmos de Segmentação de Voz

O objetivo deste capítulo é apresentar os principais conceitos e desafios na área de segmentação de voz além de trabalhos importantes de seu estado da arte. Segundo Ostendorf *et al.* em [34], a segmentação de fala pode ser dividida em: "diarização" de áudio (*audio diarization*) ou segmentação estrutural (*structural segmentation*). A diarização do áudio busca diferenciar a fala contínua da música através do agrupamento de regiões acusticamente homogêneas. Já a segmentação estrutural, trabalha com o lado acústico e com o lado léxico dos sinais de voz. Existem várias abordagens que procuram identificar os limites dos fonemas utilizando as características dos sinais de voz. Os limites identificados são utilizados para determinar os blocos de segmentação. Várias técnicas usam a abordagem mencionada [35][36][37], além de utilizar também características derivadas do conhecimento acústico dos fonemas. Alguns métodos baseados em: reconhecimento de padrões [38], modelos ocultos de Markov [39], análise fractal [40][41], transformada *wavelets*, e outras técnicas, também já foram propostos.

Vostermans *et al.* [43] desenvolveram um sistema para segmentação automática e classificação de falas. Inicialmente, o sistema realiza uma segmentação identificando as maiores alterações (chamadas de "marcos de referência") no sinal acústico, obtidas através do modelo auditivo. O sistema possui três etapas: identificação, geração e eliminação. Até quatro segmentos iniciais consecutivos são incorporados para a construção de um conjunto de segmentos fonéticos candidatos. Uma rede neural do tipo *MLP (Multi-Layer Perceptron)* é utilizada na etapa de segmentação fonética para calcular a probabilidade de um limite ser, de fato, um limite fonético, de acordo com as evidências de um limite fonético precedente e também das características acústicas. A classificação fonética também é realizada por meio de uma rede *MLP*, que classifica o vetor acústico em uma classe dentre cinco classes fonéticas. Um procedimento de busca, utilizando o algoritmo de Viterbi [44], alinha o sinal de voz com o modelo de transição de estado, derivado da transcrição da pronúncia, levando em conta as saídas das duas redes *MLPs,* nas etapas de segmentação e classificação fonética. O resultado da busca Viterbi é um conjunto de marcos (limites) e rótulos (*labels*), que maximiza a probabilidade combinada de limites fonéticos e sequências de fonemas, de acordo com as características fonéticas e as transcrições fornecidas. Segundo o autor, o sistema é bastante adaptável de um idioma para outro e não requer um conhecimento acústico extensivo e nem um grande tempo de treinamento para a rede neural. Uma taxa de acerto de 76% do posicionamento dos marcos (com uma margem de 20ms da posição ideal) foi alcançada no sistema, que foi executado utilizando a base de dados *TIMIT*.

Um método de segmentação de voz baseado em máquinas de segmentação múltipla automática (*ASM – Automatic Segmentation Machines*) é proposto em [42]. A ideia geral é realizar uma segmentação final utilizando os marcadores de segmentação provenientes de cada *ASM* (que funciona como variação de um *HMM*). Os parâmetros de peso e *bias* são ditos como referências, mas esses valores são definidos *a priori* através de um processo de segmentação manual. O método proposto foi testado em uma base de dados com sentenças pronunciadas por um narrador profissional, situado em uma sala de isolamento acústico. Obviamente, esse tipo de abordagem é bem restrito e não é adequado para aplicações no mundo real.

Um algoritmo que combina codificação de entropia, análise de multi-resolução e mapas auto-organizáveis (*SOM – Self-Organizing Maps*) [66] é proposto para a tarefa de segmentação de voz em [45]. O algoritmo de segmentação, chamado de CME + SOM, é utilizado para síntese de voz. Após a amostragem e filtragem por janelamento, uma transformada contínua *wavelet* é aplicada ao sinal. Os autores sugerem uma função *wavelet* conhecida por 'chapéu mexicano' (ou Hermitian *wavelet*) [46], devido a certas propriedades de localização temporal, inerentes a ela. A entropia de Shannon [47] é calculada em cada janela, de cada escala. As matrizes resultantes são concatenadas e normalizadas e cada coluna é utilizada como um vetor de entrada para o treinamento da rede *SOM*. Este método é aplicado a uma base de dados de sílabas espanholas e também não é recomendado para sinais com ruído.

Recentemente, algumas abordagens para processamento de voz, através da utilização de técnicas de processamento de imagens, vêm sendo utilizadas. O trabalho de Hines e Harte em [56] verifica a inteligibilidade de uma pronúncia, ou seja, o entendimento da pronúncia pela percepção acústica, usando um índice de similaridade entre imagens [57]. Um índice de similaridade calcula o quanto duas imagens (uma imagem de referência e uma imagem alvo) são semelhantes. Hines e Harte utilizam essa medida para comparar estruturas chamadas neurogramas e para estimar a degradação fonética. Os mesmos autores propuseram uma nova técnica de predição de inteligibilidade da fala, baseado em outro índice de similaridade, chamado índice de similaridade de neurogramas (*NSIM – Neurogram Similarity Index*) [58], que é uma adaptação do trabalho de Wang *et al*. SSIM (*Structural Similarity Index*) [57]. A identificação de locutor através da utilização de técnicas de processamento de imagem sobre o espectrograma do sinal de voz é o foco do trabalho de Ajmera *et al*. em [59]. A transformada discreta do cosseno é aplicada nas projeções da transformada Radon [27] para criar um vetor de características no intuito de uma identificação posterior do elocutor da pronúncia.

A análise espectral é um método bastante eficiente para extrair informações de sinais de voz. A transformada discreta *wavelet* (*DWT – Discrete Wavelet Transform*) vem sendo utilizada com sucesso em várias aplicações de voz [20][50][51][52][53] para análise espectral dos sinais. Inclusive, alguns experimentos apontam que os métodos baseados em *DWT* são superiores aos métodos clássicos, baseados em *MFCC* [20][50][51]. Um método de segmentação de voz que realiza análises do sinal no domínio da frequência, utilizando decomposição de funções *wavelet*, é encontrado em [48][49]. Basicamente, o método consiste de oito etapas: normalização do sinal, decomposição *wavelet* em seis níveis (os autores sugerem a utilização da *wavelet* de família Meyer), cálculo da soma da energia das amostras em todas as sub-bandas de frequência, cálculo do envelope da energia para cada sub-banda, cálculo da primeira derivada da energia, definição e agrupamento dos índices dos candidatos a segmentação e cálculo do índice médio de cada grupo. Uma convolução utilizando a máscara [1, 2, -2, -1] também é realizada nas sub-bandas de energia visando obter um gráfico suavizado. Existem pequenas diferenças entre os algoritmos apresentados nos dois trabalhos referenciados. Este método foi completamente reproduzido e utilizado como comparativo do algoritmo proposto neste trabalho, por conta disso, ele é explanado em detalhes nas próximas seções.

3.1 Algoritmo de Ziólko *et al.* [49]

A ideia geral do algoritmo é encontrar os candidatos em cada nível de decomposição e usar essa informação para encontrar o ponto correto da segmentação. A seguir, abordamos cada etapa do algoritmo.

3.1.1 Etapa 1: Normalização

A normalização do sinal de voz é realizada através da divisão de cada valor da amostra original pelo maior valor entre as amostras. O valor normalizado é calculado de acordo com

$$v = \frac{x_i}{x_{max}}, \tag{13}$$

em que v é o valor normalizado; x_i é o valor original da amostra e x_{max} é o valor máximo entre as amostras do sinal de voz.

3.1.2 Etapa 2: Decomposição *DWT*

As transformadas *wavelets* podem ser entendidas como mecanismos para decompor ou quebrar sinais em suas partes constituintes, permitindo a análise dos dados em diferentes domínios de frequências com a resolução de cada componente vinculada à sua escala. O autor utiliza a seguinte equação para obtenção da *DWT* e seus coeficientes

$$s(t) = \sum_{i} c_{m+1,i}\phi_{m+1,i(t).} \quad \textbf{(14)}$$

em que $\phi_{m+1,i}$ é a *i*-ésima função *wavelet* no nível de resolução (m+1). As conhecidas equações [52][54] também são calculadas:

$$c_{m,n} = \sum_{i} h_{i-2n}c_{m+1,i\,,} \quad \textbf{(15)}$$

$$d_{m,n} = \sum_{i} g_{i-2n}c_{m+1,i.} \quad \textbf{(16)}$$

em que *m* é o nível de resolução, *n* é a escala, e *h* e *g* são coeficientes constantes que dependem do par de funções: escala (ϕ) e *wavelet mãe* (ψ). As equações (15) e (16) são utilizadas para decomposição do sinal através da filtragem digital dos coeficientes *wavelets*. Os elementos da *DWT*, para um nível específico, podem ser coletados em um vetor, por exemplo: $d_m = (d_{m,1},\ d_{m,2},\ \ldots)^T$. Os coeficientes de outros níveis de resolução são calculados de maneira recursiva aplicando as equações (15) e (16). Sendo assim, os valores dos M+1 níveis da *DWT* são obtidos no seguinte formato:

$$DWT(s) = \{d_M, d_{M-1}, \ldots, d_1, c_1\} \quad \textbf{(17)}$$

A transformada *wavelet* pode ser vista como uma árvore. O nó raiz consiste dos coeficientes da série *wavelet* do sinal de voz original. O próximo nível da árvore é o resultado do primeiro passo da *DWT*. Os níveis subsequentes na árvore são construídos aplicando recursivamente a transformada *wavelet* para dividir o sinal em partes baixas (aproximação) e altas (detalhes). Um exemplo de ilustração da árvore de decomposição da transformada *wavelet*

pode ser observado na Figura 3.1, onde S é o sinal original, cada cA_i e cD_i é a componente de aproximação e detalhe, respectivamente, no nível *i*. De acordo com os experimentos de Ziólko *et al* [48][49], seis níveis de decomposição foram adotados, com a justificativa que esse número é capaz de cobrir todas as bandas de frequências da voz humana. A família de *wavelet* Meyer (Figura 3.2) foi escolhida como base para *DWT* por conta da sua simetria no domínio do tempo e do seu suporte compacto no domínio da frequência.

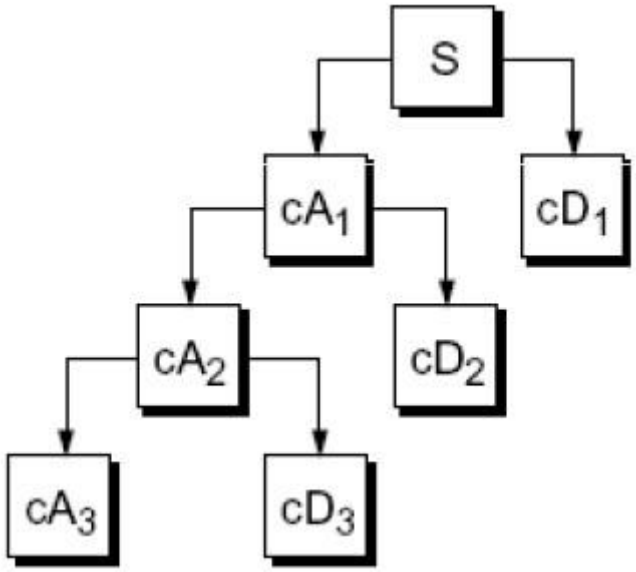


Figura 3.1: Decomposição de três níveis da transformada *wavelet*, retirado de [60].

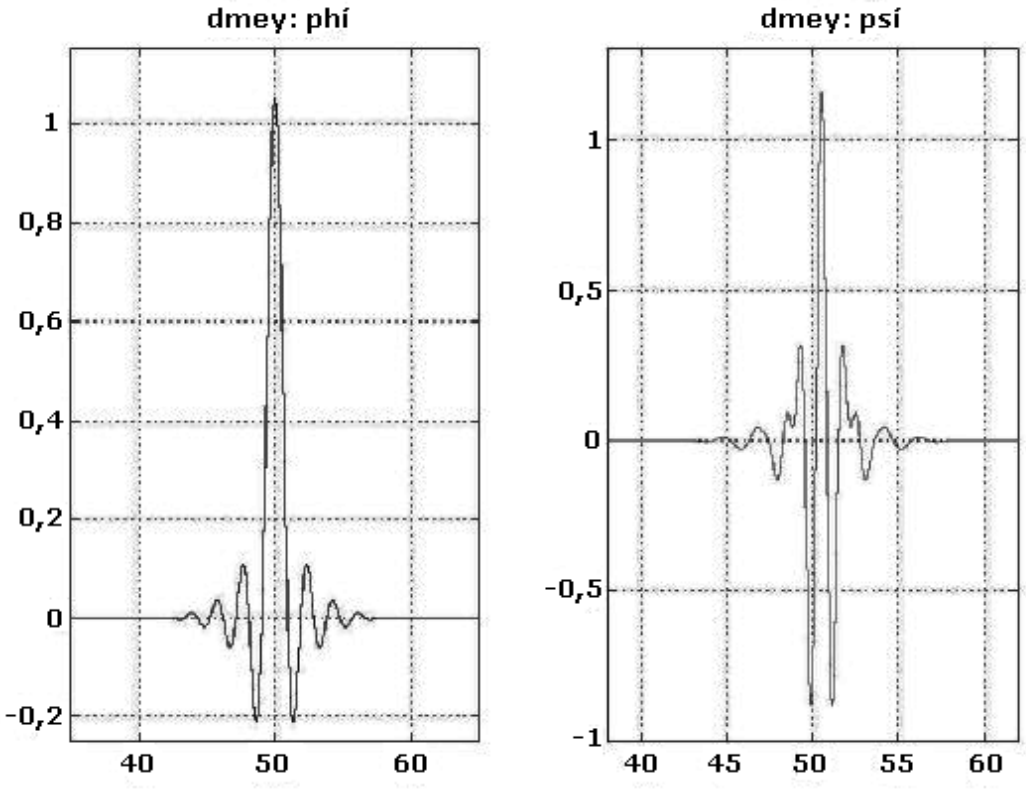


Figura 3.2: Wavelet Meyer discreta, retirado de [49].

3.1.3 Etapa 3: Cálculo da Soma da Energia das Amostras

O cálculo do somatório da energia de todas as amostras de cada sub-banda de frequência (p_n) é realizado de acordo com

$$p_n(i) = \sum_{j=1}^{2^{n-1}} d^2_{n,j+2^{n-1}i}\,, \text{onde } i = 0, \dots, 2^{-M}N - 1\,. \quad \textbf{(18)}$$

O valor $2^{-M+n-1}N$ de amostras de espectro *wavelets* no level *n* (onde *n* = 1,...,M) depende do tamanho *N* do sinal de voz no domínio do tempo, assumindo *N* como uma potência de 2. Caso *N* não seja uma potência de 2, as amostras restantes são descartadas. A Tabela 1 apresenta esse número em cada nível, relativo ao menor nível de resolução. Para cada nível de decomposição *n*, a energia é calculada de uma maneira diferente para obter a equidade no número de amostras de energia.

Tabela 1: Características dos níveis da *DWT* e de seus envelopes

Nível da *DWT*	Banda de Frequência (Hz)	Nº de amostras em comparação ao nível 1	Tamanho da janela ω
6	2756-5512	32	3
5	1378-2756	16	3
4	689-1378	8	3
3	345-689	4	5
2	172-345	2	5
1	86-172	1	5

3.1.4 Etapa 4: Cálculo dos Envelopes

O cálculo dos envelopes p'_n para flutuações de energia em cada sub-banda de frequência é realizado através da escolha dos maiores valores de p_n em uma janela de tamanho ω (Tabela 1). A Figura 3.3 ilustra um exemplo de segmentação da palavra de origem polonesa 'Andrzej' onde pode ser observado o envelope para cada nível da *DWT*.

3.1.5 Etapa 5: Cálculo da Função de Taxa de Variação (*rate-of-change*)

O cálculo da função de taxa de variação $r_n(i)$ é realizado através da filtragem de $p_n(i)$ pela máscara [1, 2, -2, -1].

3.1.6 Etapa 6: Cálculo dos Candidatos (limites de segmentação)

Dado um limiar p de distância entre $r_n(i)$ e p'_n e outro limiar p_{min} do valor mínimo de p_n, esta etapa do algoritmo deve encontrar os índices nos quais

$$\left|\beta|r_n(i)| - p'_n(i)\right| < p \; AND \begin{pmatrix} \left|\beta|r_n(i+1)| - p'_n(i+1)\right| > p \; OR \\ \left|\beta|r_n(i-1)| - p'_n(i-1)\right| > p \end{pmatrix} AND \; p'_n(i) > p_{min},$$

onde $\beta = 1$.

Estes índices são escritos em um único vetor (representado pelos asteriscos vermelhos na Figura 3.3). O valor de $p_{min} = 0{,}003$ foi escolhido experimentalmente em [48] e é o mesmo adotado neste trabalho.

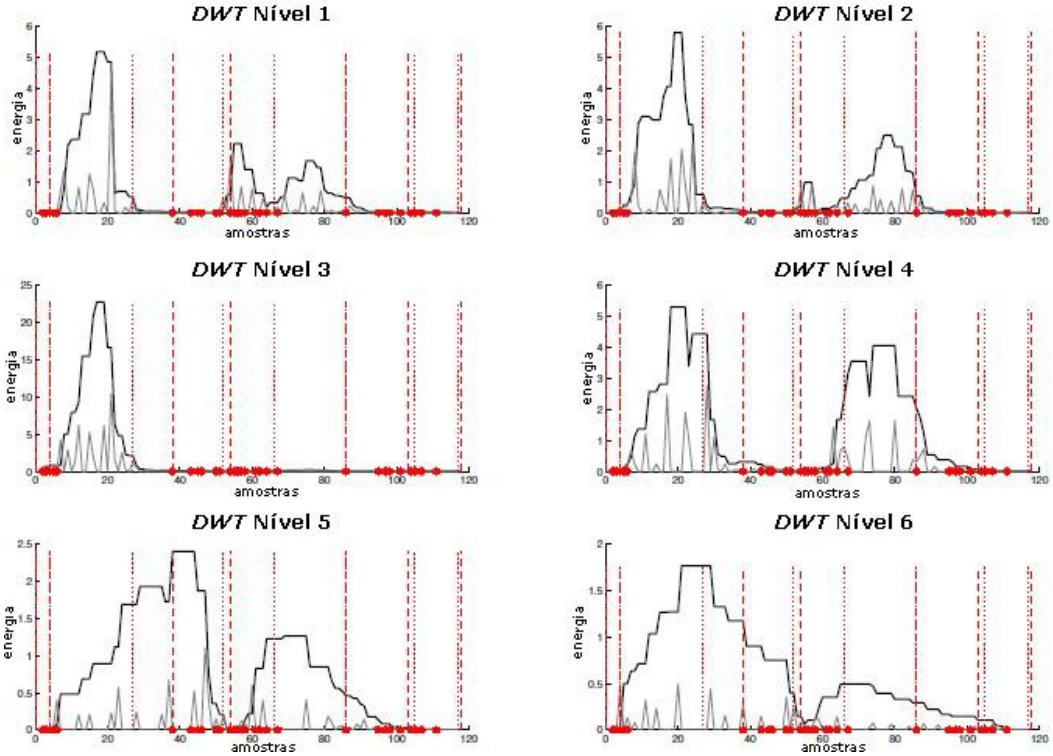


Figura 3.3: Exemplo de segmentação do nome 'Andrzej' /Λ:ndʒɛɪ/. As linhas pontilhadas significam limites da segmentação manual; As linhas tracejadas significam os limites da segmentação automática; As linhas grossas (cinza escuro) são os envelopes e as linhas finas (cinza claro) são as funções de taxa de variação. Os asteriscos vermelhos sobre o eixo *x* são os candidatos a limites nos quais $r_n(i)$ e $p'_n(i)$ são muitos próximos, traduzido de [48]

3.1.7 Etapa 7: Agrupamento dos Candidatos

Esta etapa é responsável por encontrar e agrupar os índices candidatos a segmentação que estejam próximos uns dos outros. O algoritmo calcula que os índices, cuja distância seja menor que o atributo α, devem ser agrupados. O autor sugere o valor cinco para o atributo α, que é um valor discreto de energia que representa 29ms (o menor tamanho de um fonema).

3.1.8 Etapa 8: Cálculo do Representante de cada Grupo de Candidatos

Esta etapa é responsável por calcular o valor do índice médio (obviamente arredondado para um número inteiro) para cada grupo encontrado na etapa 7. Esse índice médio de cada grupo é chamado de representante do grupo.

3.1.9 Resultados e Conclusões sobre o Algoritmo de Ziólko

O autor utilizou uma base de dados com 50 palavras polonesas e segmentou todas elas manualmente para fazer a comparação com a segmentação automática gerada pelo seu algoritmo. De uma maneira geral, o algoritmo se sai bem, ele encontra os limites de segmentação bem próximos dos limites manuais, porém o método trabalha apenas considerando o sinal de voz com ausência de ruídos, e se aplicado em sinais do mundo real, pode encontrar vários candidatos em lugares que não possuem informação acústica de voz, e sim de ruído.

4 Algoritmo Proposto

Este capítulo tem como meta apresentar um novo algoritmo de segmentação de voz que foi desenvolvido nesta pesquisa. Dessa forma, explica-se detalhadamente cada passo do algoritmo, mostrando sua importância no processo, e também ilustrando as ideias e assunções tomadas no desenvolvimento do mesmo.

O algoritmo proposto neste trabalho é baseado no sinal de voz e em características do processamento digital de imagens, como análise de espectrograma, operações de matemática morfológica, análise de projeção etc. O algoritmo funciona basicamente em dois ciclos. O primeiro ciclo age segmentando o sinal de voz do fundo (silêncio ou ruído). O segundo ciclo realiza a segmentação de sílabas fonéticas. Existem diferentes definições para as sílabas fonéticas, talvez a que melhor se encaixe para este trabalho é a do famoso linguista Roman Jakobson, em [67]: "*os fonemas são traços distintivos que se reúnem em feixes simultâneos e que se concatenam em sequência. O padrão elementar sotoposto a um dado grupo de fonemas é a sílaba fonética*". Em outras palavras, a sílaba fonética pode ser considerada um agrupamento de fonemas e ela é mais bem destacada na imagem do espectrograma, em comparação ao fonema isolado. Ela é mais destacada, principalmente pelo fato de ser possível visualizar seu início e término na imagem do espectrograma.

4.1 Ciclo 1: Segmentação da Voz

Cada ciclo do algoritmo executa uma sub-rotina de cinco passos, que foi desenvolvida para este trabalho, chamada *stepFive,* baseada no sinal e em técnicas de processamento de imagem, cujos passos são:

1. Pré-processamento e estimativa da energia;
2. Análise de espectrograma do sinal de voz;
3. Realce da imagem do espectrograma e análise de projeção;
4. Limiarização baseada na mediana, análise de componentes conectados e operações de matemática morfológica;
5. Segmentação do sinal de voz em relação ao fundo (*background*).

Para a execução da sub-rotina é necessário a utilização de quatro parâmetros ilustrados na Tabela 2.

Tabela 2: Parâmetros utilizados no ciclo 1 da sub-rotina *stepFive*.

Parâmetro	Descrição	Valor
S	Sinal de voz a ser processado	*a[n]*
ρ	Constante experimental para multiplicação da mediana	4
se_w	Largura do elemento estruturante utilizado na dilatação morfológica	60
r_s	Constante experimental para auxílio na remoção de segmentos	10

Nas próximas subseções, são explicados os cinco passos da sub-rotina *stepFive*, em seguida é explicado o segundo ciclo que realiza a segmentação das sílabas fonéticas.

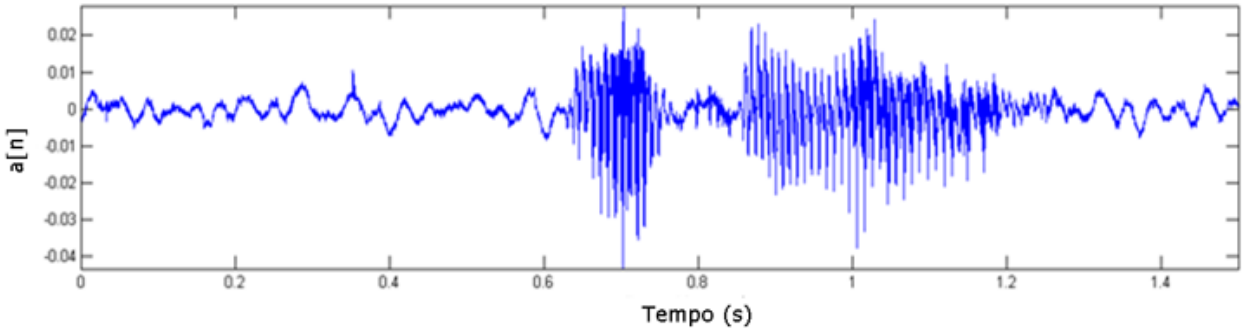
4.1.1 1º passo: Pré-processamento e Cálculo da Energia do Sinal

A função discreta (ou vetor), *a[n]*, representa a redução de um sinal contínuo de voz, *s(t)*, para um sinal discreto e define o valor de amplitude do sinal para cada amostra, *n*, no tempo, *t*, em um processo de amostragem de frequência *fs*. Para reduzir o tempo de execução do algoritmo, este primeiro passo processa e converte o sinal de áudio para 16 KHz, caso ele tenha sido amostrado em uma frequência superior a essa. A densidade de energia do sinal é calculada elevando ao quadrado cada elemento do vetor que contém o sinal de voz, como mostra a equação

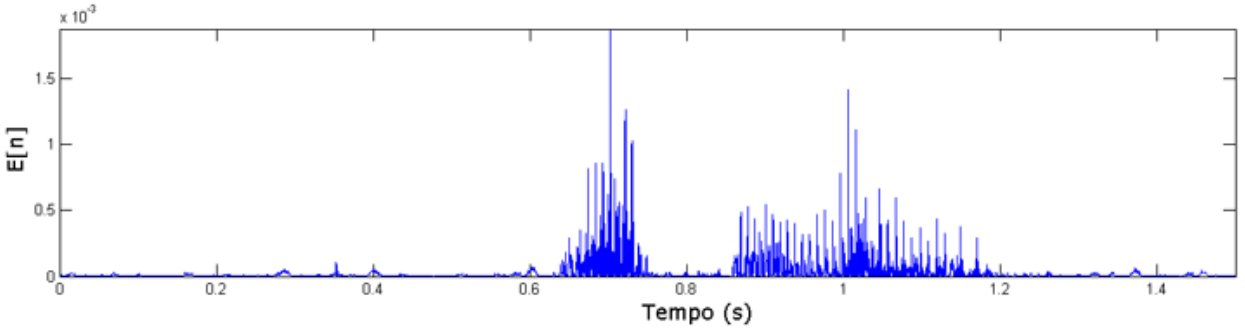
$$E[n] = |a[n]|^2 , \quad \textbf{(19)}$$

em que *E[n]* é o vetor de densidade de energia do sinal.

A Figura 4.1 ilustra um exemplo de sinal de áudio original (discreto), *a[n]*, e a densidade de energia, *E[n]*, para a pronúncia (por voz masculina) do nome "Rocky Road", retirado da base MIT Mobile Device Speaker Verification Corpus (MIT - MDSVC) [55].



(a)



(b)

Figura 4.1: (a) Sinal discreto de áudio, *a[n]* e (b) densidade de energia, *E[n]*, da pronúncia "Rocky Road".

4.1.2 2º Passo: Análise de Espectrograma do Sinal de Voz

O gráfico de um sinal de voz, no domínio do tempo, dificilmente fornece uma visão geral das principais características que ele contém. É possível observar certas propriedades, por exemplo, na Figura 4.1, podemos visualizar que a voz se concentra no intervalo de tempo que vai de 0,6 até 1,2 segundos. Poderíamos até inferir que no intervalo de 0,6 até 0,8 segundos o elocutor pronunciou a palavra "Rocky", e no intervalo restante até 1,2 segundos ele pronunciou "Road". No entanto, contar apenas com esse tipo de informação não é suficiente para analisar a grande parte de sinais de voz do mundo real. Nesse contexto, é interessante utilizar uma importante ferramenta matemática: a Transformada de Fourier [61]. A Transformada é um procedimento matemático que tem como principal objetivo o mapeamento de um conjunto de coordenadas para outro conjunto definido, inclusive, possibilitando a operação inversa. Podemos utilizar as transformadas para observar características de um sinal que não são visíveis no domínio padrão em que ele se encontra (no nosso exemplo, o domínio padrão para o sinal de voz é o tempo).

A Transformada de Fourier permite a mudança de um sinal do domínio do tempo para o domínio da frequência. Além de apresentar certas propriedades que são visíveis apenas no domínio da frequência, ele possui outra vantagem: a frequência pode ser facilmente controlada, ou modificada (ao contrário do tempo). A Transformada clássica de Fourier opera sobre sinais contínuos e inclui implicitamente uma hipótese sobre a estacionaridade deles [61]. Como estamos trabalhando com sinais digitais e não estacionários, a Transformada de Fourier utilizada neste trabalho é a Transformada Discreta de Fourier (*DFT – Discrete Fourier Transform*), que é a Transformada de Fourier para sequências de duração finita. Mais especificamente, iremos utilizar a Transformada Discreta de Fourier de Tempo Curto (*Short-Time Fourier Transform - STFT*), que analisa o sinal dentro de uma pequena "janela", onde este permanece aproximadamente estacionário (e consequentemente passível de ser utilizado na Transformada de Fourier).

O espectrograma da energia de densidade do sinal de voz é calculado utilizando a Transformada de Fourier de Tempo Curto, definida em (20) [62]

$$STFT\{a[n]\}(m,\omega) = \sum_{n=-\infty}^{\infty} a[n]w[n-m]e^{-j\omega n}\ , \tag{20}$$

onde, *a[n]* é o sinal de áudio e *w[n-m]* é a função janela centrada em *n*. Neste caso, *m* é um valor discreto e ω é contínuo, representando a frequência. A função de janelamento escolhida para o trabalho foi a de *Hamming*.

A imagem de espectrograma, *f[x, y]*, é calculada com amplitude 50 dB (decibéis) abaixo do máximo e pode ser visualizada na Figura 4.2. A imagem *f[x, y]* também foi normalizada na escala de 0 a 1, para processamento posterior, utilizando a (21).

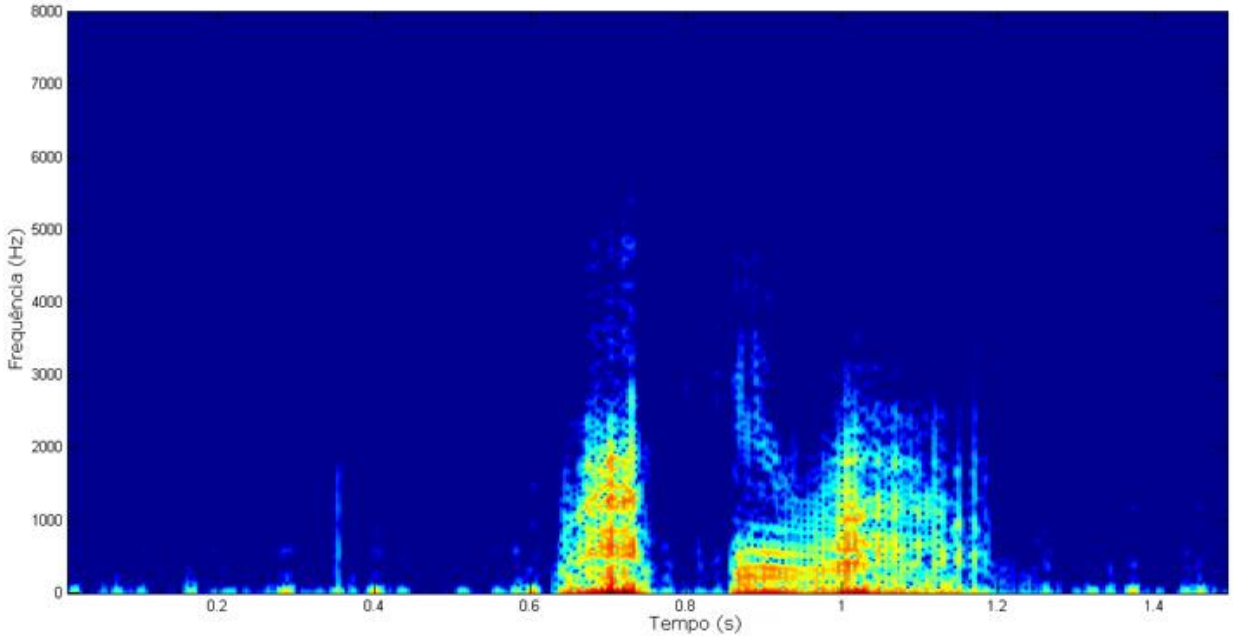


Figura 4.2: A imagem do espectrograma da densidade de energia, *f[x,y]*, do sinal de voz da pronúncia "Rocky Road".

$$v = \left(\frac{x_i - x_{mín}}{x_{máx} - x_{mín}}\right) \tag{21}$$

Na (21), *v* representa o valor normalizado, x_i é o valor a ser normalizado e $x_{mín}$ e $x_{máx}$ são os valores mínimo e máximo, respectivamente, da matriz que representa a imagem *f[x, y]*.

4.1.3 3º Passo: Realce de *f[x, y]* e Análise de Projeção

A média de cada linha, *Mean[x]*, da imagem do espectrograma, *f[x. y]*, é calculada de acordo com a equação (22). Em seguida, cada pixel de *f[x, y]* é subtraído pelo valor de *Mean[x]* encontrado, e pode ser visualizada na equação (23).

$$Mean[x] = \frac{1}{f_w}\sum_{y=1}^{f_w} f[x,y] \tag{22}$$

$$g[x,y] = f[x,y] - Mean[x] \tag{23}$$

onde, f_w é a largura da imagem, *x* é o eixo das linhas, *y* é o eixo das colunas e *g[x, y]* é a imagem realçada do espectrograma.

Este procedimento diminui a frequência de ruído no sinal e é comumente utilizado quando o ruído é aproximadamente constante no domínio do tempo [63]. Apesar de nem todos os ruídos encontrados nas bases de dados de áudio permanecer constantes (no tempo), esse procedimento é útil para os demais passos. A análise de projeção horizontal, *h[x]*, (24), da imagem do espectrograma, mostra uma alta densidade do sinal nas baixas frequências, uma região rica em ruído, o que significa que esta região possui pouca contribuição para a informação do sinal, como pode ser observado na Figura 4.3. A subtração pela média horizontal contribui para destacar as outras regiões de menor frequência. A Figura 4.4 ilustra a imagem realçada do espectrograma, *g[x, y]*.

$$h[x] = P_h(f[x,y]) = \sum_{y=1}^{f_w} f[x,y] \tag{24}$$

onde, *h[x]* ou $P_h(f[x,y])$ representa a análise de projeção horizontal da imagem do espectrograma e f_w é a largura da imagem.

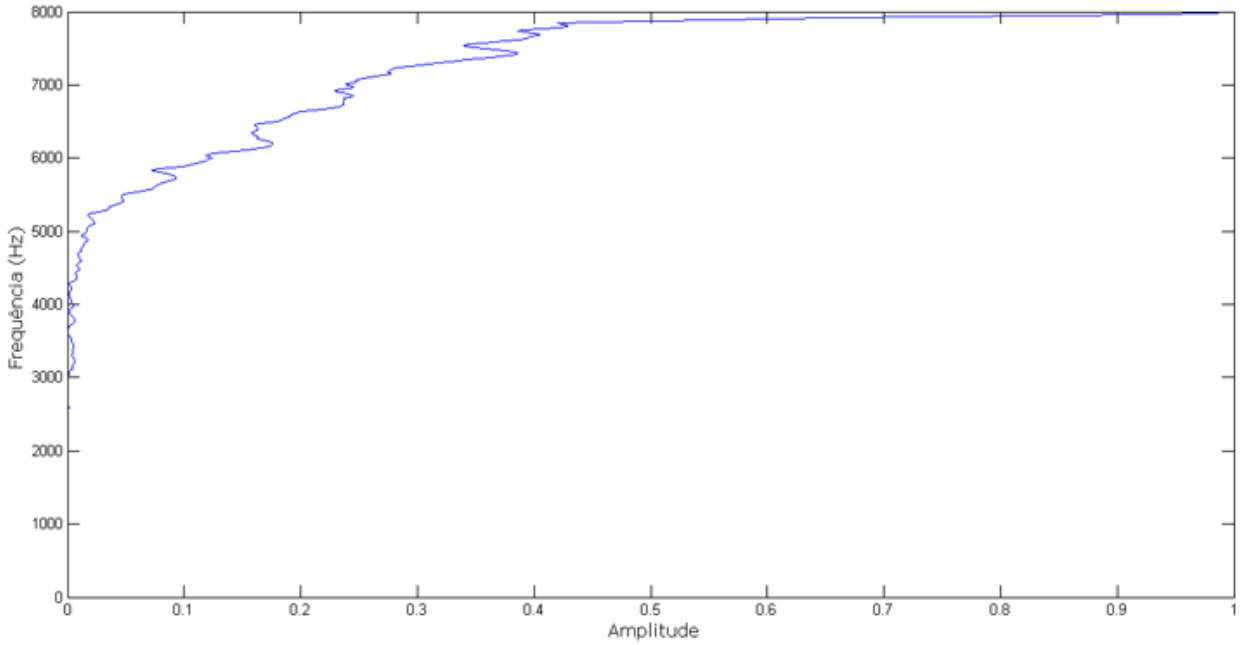


Figura 4.3: Projeção horizontal do espectrograma, *h[x]*, para a Figura 4.2.

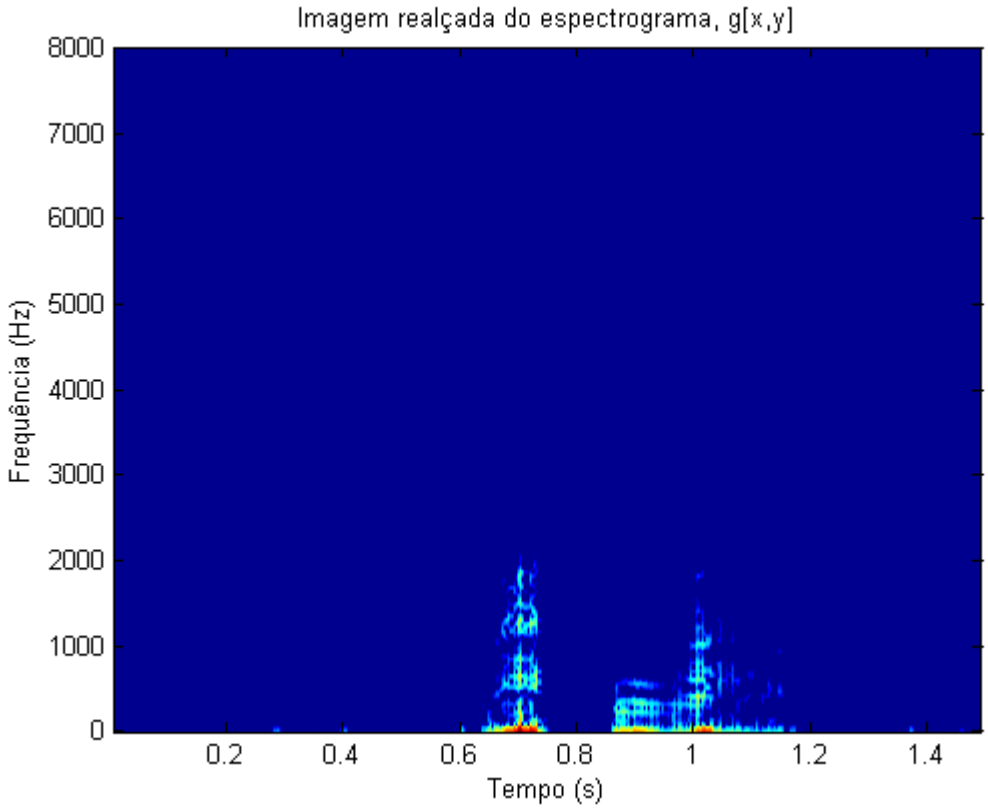
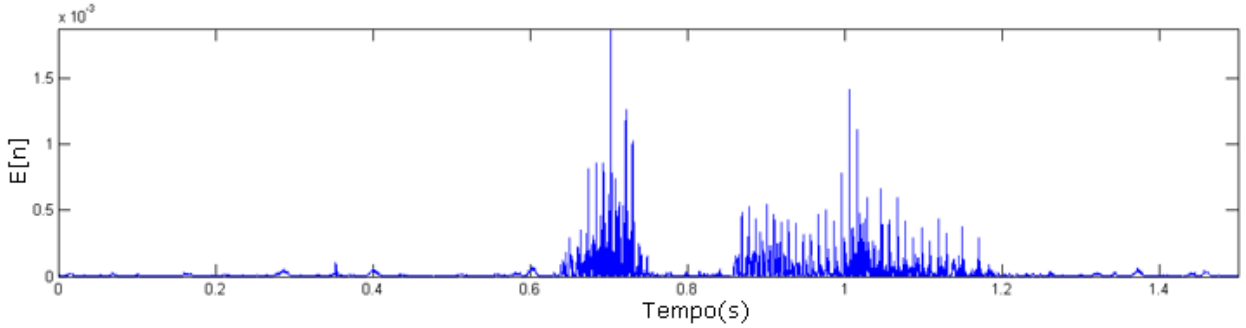


Figura 4.4: Imagem realçada do espectrograma, *g[x, y]*.

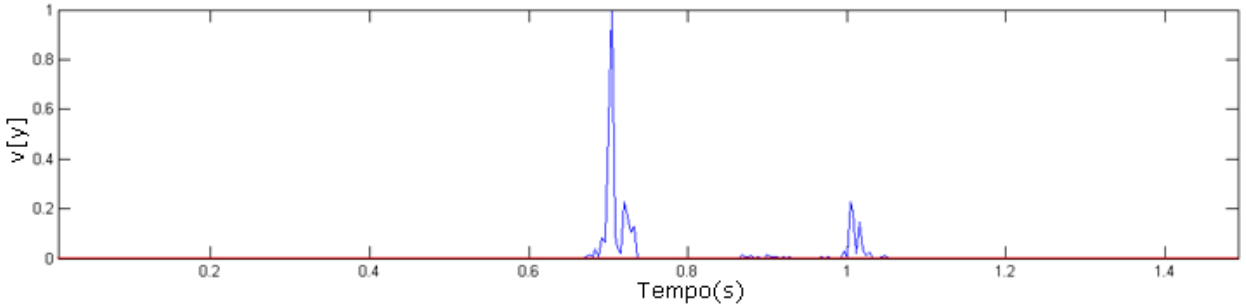
A projeção vertical, (25), da imagem realçada do espectrograma, *g[x, y]*, é calculada e elevada à terceira potência para um aumento do sinal em relação ao ruído. A Figura 4.5 ilustra uma comparação entre a densidade de energia do sinal original e a projeção vertical do sinal. É possível visualizar que a projeção vertical fornece um sinal mais limpo do que àquele fornecido somente pela densidade de energia.

$$v[y] = \left(P_v(f[x,y])\right)^3 = \left(\sum_{x=1}^{f_h} f[x,y]\right)^3 \quad \textbf{(25)}$$

onde, *v[y]* ou $P_v(f[x,y])$ representa a análise de projeção vertical da imagem do espectrograma e f_h é a altura da imagem.



(a)



(b)

Figura 4.5: (a) Densidade de energia do sinal original, *E[n]*; (b) Projeção vertical do espectrograma do sinal, *v[y]*.

4.1.4 4º Passo: Limiarização baseada na mediana, análise de componentes conectados e operações de matemática morfológica

Uma limiarização baseada na mediana é realizada sobre a projeção vertical do espectrograma, *v[y]*. Ela é alcançada seguindo as equações (26) e (27).

$$b[y] = Th(v[y]) = \begin{cases} 1, se\ v[y] \geq T \\ 0, se\ v[y] < T \end{cases} \quad \textbf{(26)}$$

$$T = \rho \times Median(v[y]) \quad \textbf{(27)}$$

Na Figura 4.5 (b), *T* é a linha horizontal vermelha; A Figura 4.6 (b) ilustra os segmentos binários, *b[y]*. O valor de *T* é delimitado pelo intervalo [0, 1]. O valor do parâmetro ρ é 4, ele foi definido de forma empírica e pode ser observado na Tabela 2.

Portanto, uma análise de componente conectado das regiões binárias, *b[y]*, (voz vs. fundo) é realizada. Todos os segmentos de voz, menores que o valor do parâmetro r_s, são removidos. Por último, uma dilatação morfológica com um elemento estruturante de largura, se_w, é executada nos componentes restantes, como podemos observar, na Figura 4.6 (c). Os valores dos parâmetros r_s e se_w, necessários para produzir $b_r[n]$ e $b_d[n]$, na Figura 4.6 (d), estão na Tabela 2.

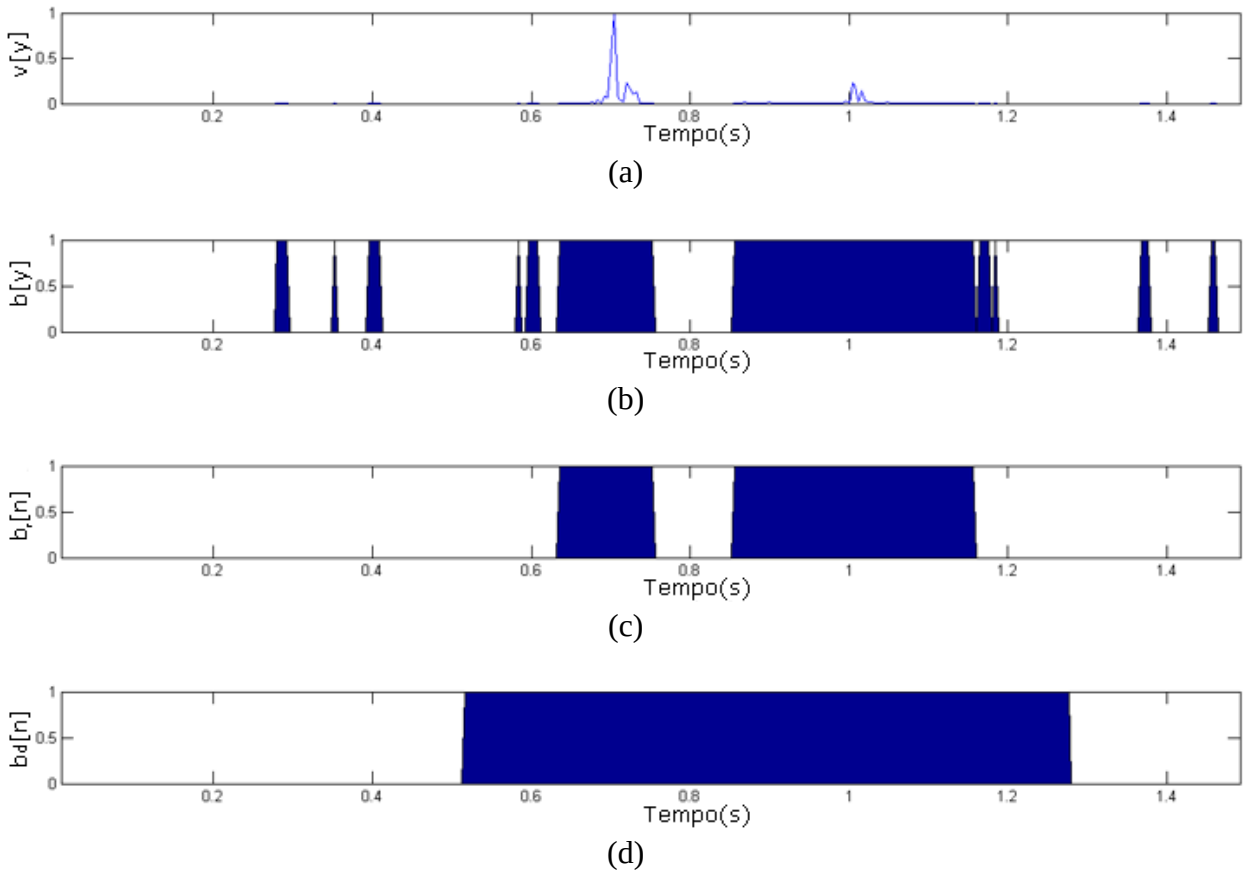
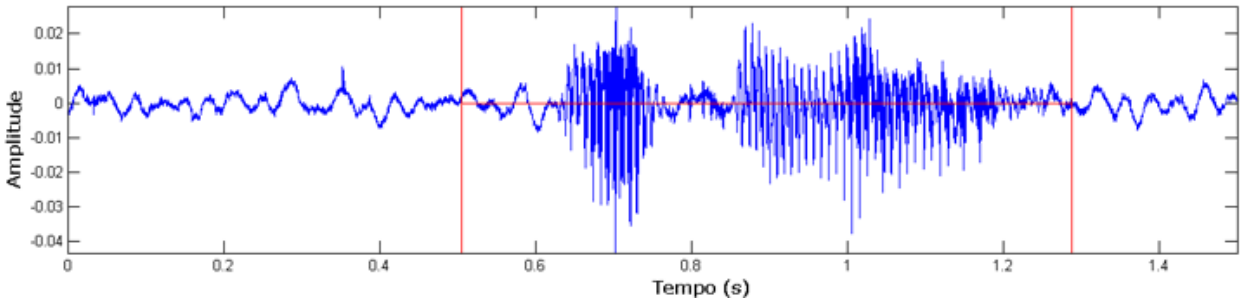


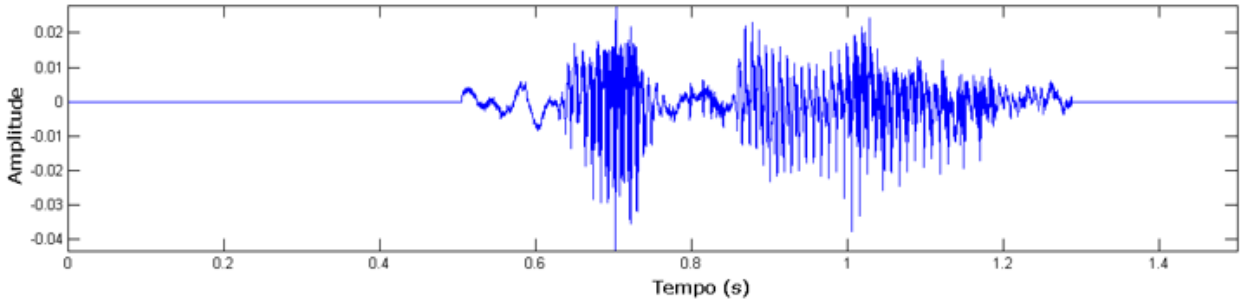
Figura 4.6: (a) Projeção vertical do espectrograma, *v[y]*; (b) Os segmentos binários gerados, *b[y]*; (c) Remoção dos segmentos de tamanho menor que r_s, produzindo $b_r[n]$; (d) Dilatação dos segmentos restantes, produzindo $b_d[n]$.

4.1.5 5º Passo: Segmentação do sinal de voz em relação ao *background*

O sinal de áudio original, *a[n]*, é cortado de acordo com as regiões de voz encontradas pelo algoritmo. Esse passo funciona de modo semelhante a uma binarização de imagem, onde as regiões que possuem valor 1 são indicadas como objeto (ou *foreground*, no nosso caso, a voz) e as regiões que possuem valor 0 são indicadas como fundo (ou *background*). A Figura 4.7 (a) apresenta o sinal original e região de segmentação de voz, enquanto a Figura 4.7 (b) mostra o áudio final segmentado, *Seg[n]*, deixando a voz e removendo todo o ruído ou silêncio de fundo.



(a)



(b)

Figura 4.7: (a) áudio original e região de segmentação de voz; (b) áudio final segmentado, *Seg[n]*.

4.2 Ciclo 2: Segmentação de Sílabas Fonéticas

Após a execução do primeiro ciclo, o áudio segmentado resultante, *Seg[n]*, será processado em um novo ciclo, utilizando a mesma sub-rotina *stepFive*, porém com um ajuste no valor dos parâmetros, como pode ser visto na Tabela 3. Vale ressaltar que, apenas a região que contém voz será processada neste ciclo. A Figura 4.8 apresenta o áudio segmentado, contendo apenas a voz e as regiões de segmentação das sílabas fonéticas. A Figura 4.9 ilustra o corte final do áudio, onde apenas permanecem as sílabas fonéticas, *Pho[n]*. As sílabas fonéticas foram corretamente segmentadas, de acordo com a segmentação manual do nome "Rocky Road", resultando em "\rok\"\ro\"\ad\".

Tabela 3: Parâmetros utilizados no ciclo 2 da sub-rotina *stepFive*

Parâmetro	Descrição	Valor
S	Sinal de voz a ser processado	*Seg[n]*
ρ	Constante experimental para multiplicação da mediana	1
se_w	Largura do elemento estruturante utilizado na dilatação morfológica	1
r_s	Constante experimental para auxílio na remoção de segmentos	5

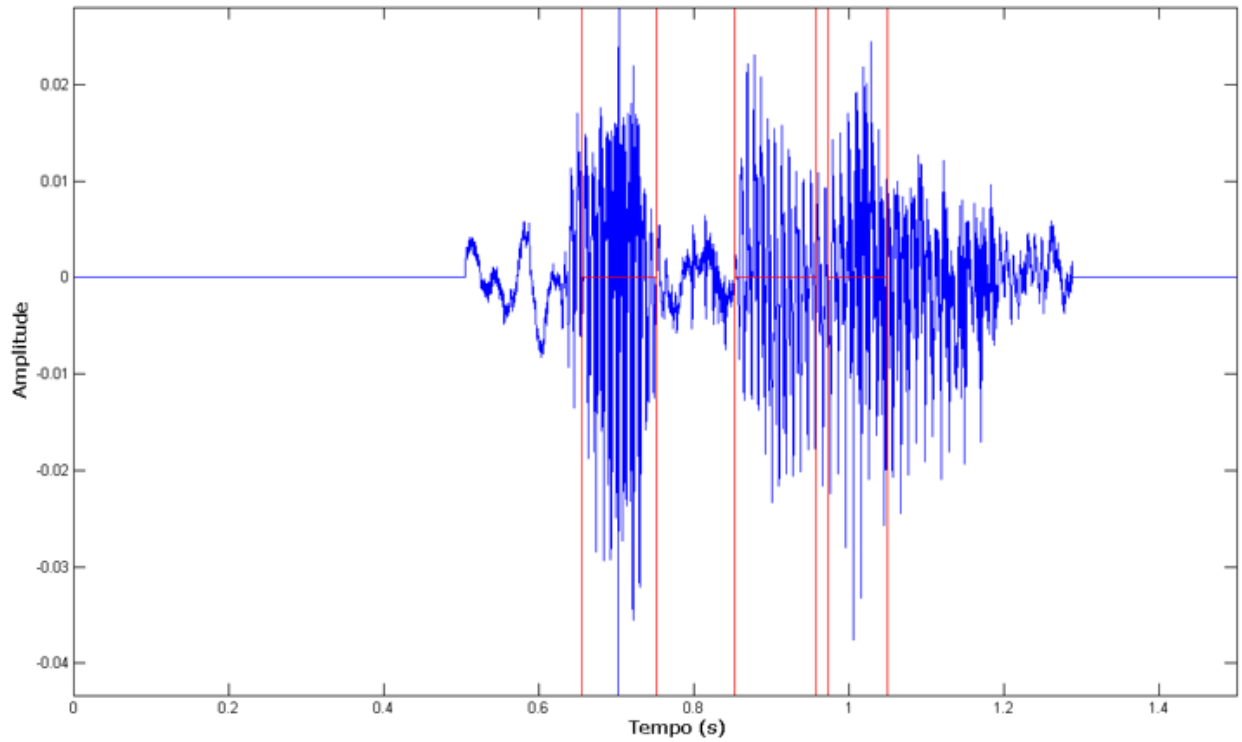


Figura 4.8: Áudio segmentado com as regiões de sílabas fonéticas (em vermelho).

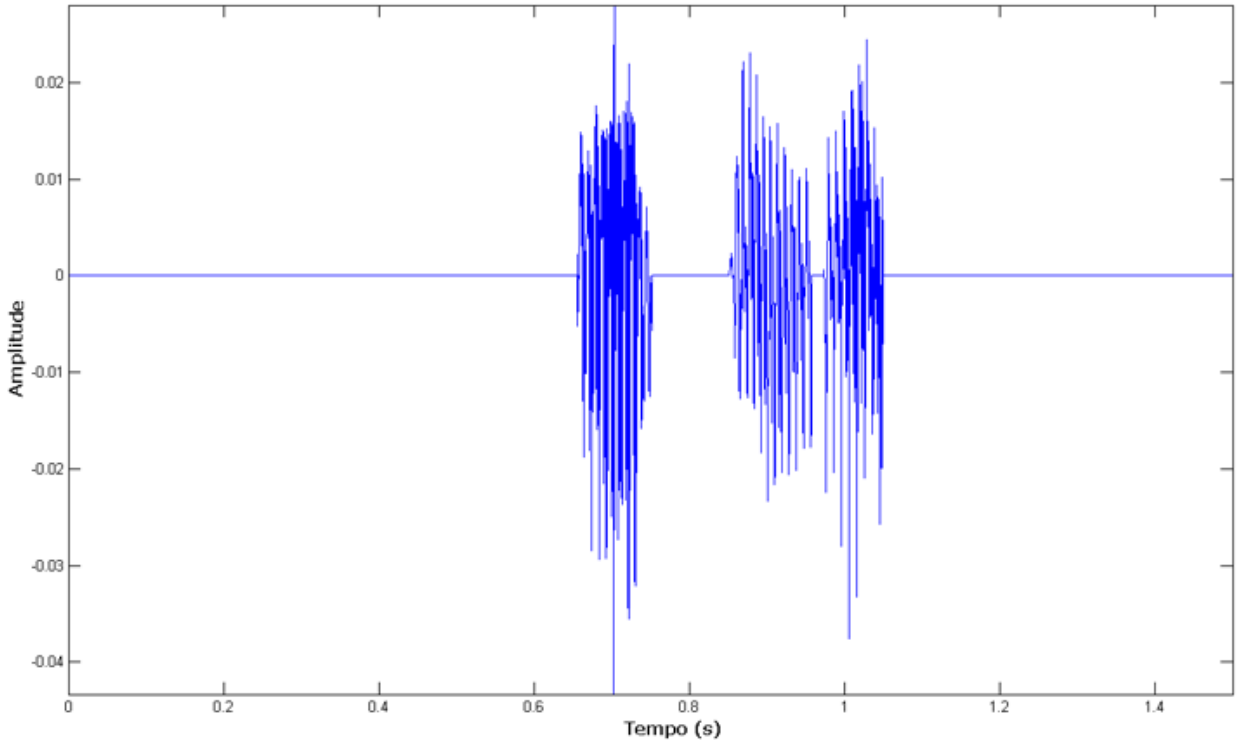


Figura 4.9: Áudio final, contendo apenas as sílabas fonéticas, *Pho[n]*.

O esquema completo do algoritmo proposto neste trabalho pode ser visto na Figura 4.10.

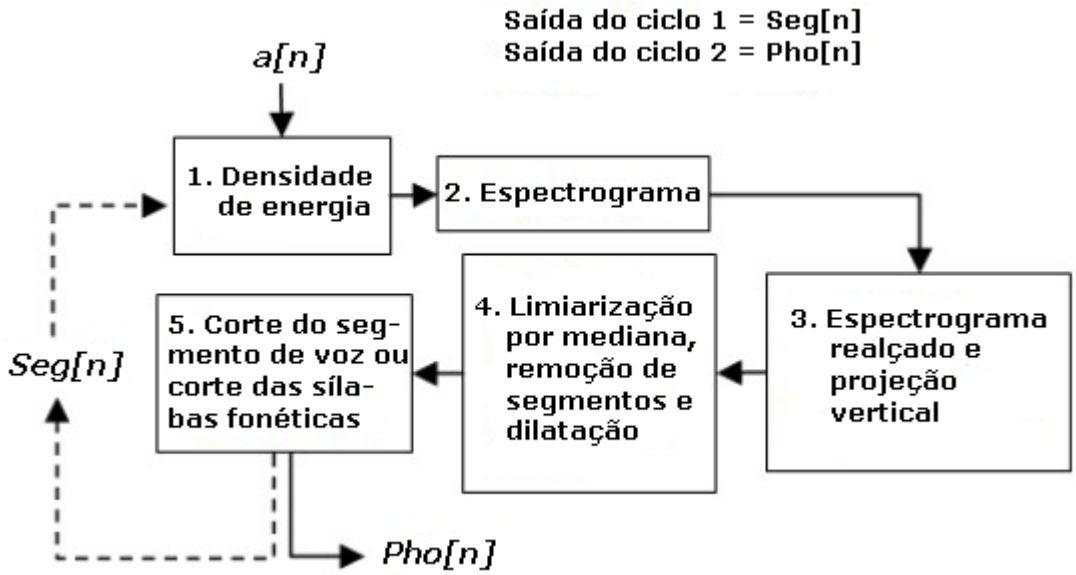


Figura 4.10: Diagrama de blocos do algoritmo de segmentação proposto neste trabalho, adaptado de [64].

A Figura 4.11 apresenta um exemplo completo da execução do algoritmo proposto sobre o sinal de voz da pronúncia “Barbara Taylor”.

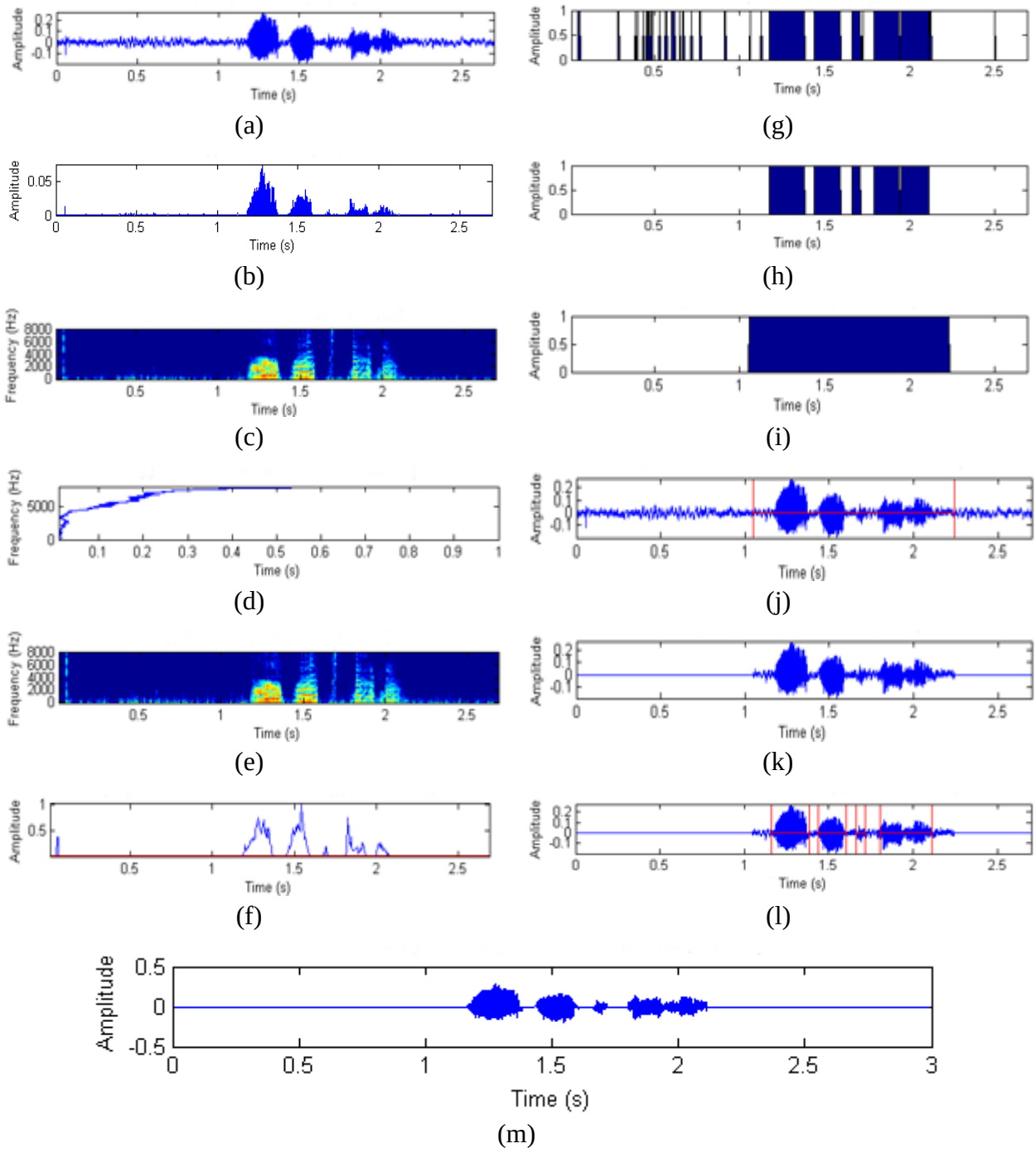


Figura 4.11: Saída passo-a-passo da execução do algoritmo de segmentação proposto sobre o sinal de voz da pronúncia "Barbara Taylor": (a) sinal de áudio original, *a[n]*; (b) densidade de energia, *E[n]*; (c) espectrograma da energia; (d) projeção horizontal do espectrograma, *h[x]*; (e) imagem realçada do espectrograma, *g[x, y]*; (f) projeção vertical do espectrograma, *v[y]*; (g) segmentos binários, *b[y]*; (h) remoção de pequenos segmentos, $b_r[n]$; (i) dilatação dos segmentos restantes, $b_d[n]$; (j) região de segmentação sobre o sinal original; (k) apenas o sinal de voz segmentado, *Seg[n]*; (l) sílabas fonéticas segmentadas sobre o sinal original; (m) apenas as sílabas fonéticas segmentadas, *Pho[n]*.

5 Experimentos e Análise de Resultados

Este Capítulo apresenta os experimentos realizados nesta pesquisa, a fim de validar e verificar a eficácia do novo algoritmo de segmentação exposto no Capítulo 4.

5.1 Base de Dados

Os pesquisadores utilizam medidas e base de dados diferentes nas comparações encontradas em trabalhos da literatura. Levando isso em conta, duas bases de dados populares foram utilizadas para este trabalho: a MIT *Mobile Device Speaker Verification Corpus* (MIT – MDSVC), que é uma base para trabalhos relacionados ao tema de reconhecimento de voz [43] e a TIMIT (*Texas Instruments/Massachussets Institute of Technology*) [68], que é uma base de fala independente de locutor, provavelmente a mais utilizada na área de reconhecimento de fala. Ela foi desenvolvida inicialmente com a finalidade de desenvolver e testar sistemas de reconhecimento de fala, porém, devido ao aumento crescente de pesquisas em diversas áreas de processamento digital da fala, a TIMIT também passou a ser utilizada para testes em segmentação.

A base de dados MIT-MDSVC [55] foi construída para fornecer sinais de voz para a aquisição de conhecimento fonético-acústico. Ela também é muito utilizada para desenvolvimento de algoritmos de verificação de locutor e avaliação de sistemas automáticos de reconhecimento de voz. A base de dados foi transcrita nos laboratórios do departamento de ciências da computação do Instituto de Tecnologia de Massachussetts (*Massachussetts Institute of Technology – MIT).*

Além disso, a base de dados é dividida em dois conjuntos: um de usuários legítimos e outro de impostores. Para o conjunto de usuários legítimos, os dados de voz foram coletados em duas diferentes sessões, de vinte minutos cada, que ocorreram em dias diferentes. Cada sessão resultou em 54 pronúncias por usuário, originando 2.592 amostras de voz, totalizando, portanto, 5.184 amostras do conjunto de usuários legítimos. Para o conjunto de impostores, os usuários participaram em uma única sessão de vinte minutos. O conjunto de amostras de impostores contém 2.700 amostras, porém como possui aplicação tradicionalmente em sistemas de identificação/verficação de locutor, não faz muito sentido de ser utilizada neste trabalho. Dentro do conjunto de usuários legítimos, existem 48 locutores, dos quais 22 são mulheres e 26 são homens.

Os dados foram coletados em diferentes ambientes (um escritório silencioso, um saguão razoavelmente barulhento, e um cruzamento de ruas congestionado) e em dois tipos de

microfone (microfone interno de um dispositivo portátil e microfone externo de um fone de ouvido).

Para a confecção da TIMIT foram utilizados 630 locutores, abrangendo os 8 principais dialetos do Inglês Americano. Cada locutor pronunciou 10 locuções, totalizando 6300 locuções foneticamente balanceadas. Desse total, normalmente apenas 5040 são utilizadas para os testes. Esse conjunto é dividido em locuções de treinamento (3696 locuções) e locuções de teste (1344). É muito importante destacar que do total de locuções de teste, 624 são distintas. As locuções da TIMIT foram amostradas a 16 kHz, com uma resolução de 16 bits por amostra. Cada locução do conjunto de treinamento e teste apresenta sua transcrição fonética, a segmentação manual em termos de palavras e também em termos de fones. Na transcrição fonética das locuções é utilizado um conjunto de 64 fones distintos. A documentação da TIMIT sugere que apenas 48 fones sejam utilizados no treinamento dos modelos acústicos e sugere ainda que, na fase de teste esse conjunto seja simplificado para 39.

Esta base de dados vem sendo utilizada como padrão pela comunidade de reconhecimento de voz desde o ano de 1994, logo após a sua confecção, e até hoje é largamente utilizada, tanto para aplicações de identificação de locutor quanto para reconhecimento de voz. Além de fornecer, para cada sentença, a segmentação manual em nível de fonema, ela também é considerada pequena o suficiente para garantir um rápido retorno de tempo na execução de experimentos completos, e grande o suficiente para demonstrar as capacidades de um sistema.

5.2 **Metodologia dos Experimentos**

Devido à grande quantidade de amostras existentes na base de dados e o pouco tempo para realizar a segmentação manual das sentenças, surge a dúvida em relação a quantos e quais exemplos devem ser utilizados. Para a base de dados MIT-MDSVC, que possui em geral a pronúncia de nomes curtos, foi estabelecido que o conjunto de sentenças utilizado no trabalho tivesse no mínimo duas, e no máximo quatro palavras por sentença, porém, vale ressaltar que não foram utilizadas todas as sentenças da base que se encaixam nessa regra. As sentenças selecionadas da base TIMIT possuem no mínimo cinco palavras por sentença. A maior sentença escolhida possui quatorze palavras. Dessa maneira, o trabalho de segmentação manual é reduzido e podemos verificar o comportamento do algoritmo antes de elaborar testes com sentenças mais complexas. O conjunto de sentenças utilizado no trabalho pode ser visto na Tabela 4 e na Tabela 5. As sentenças mais curtas são pronunciadas em cerca de dois segundos e as mais longas em cerca de quatro segundos. O primeiro conjunto foi empregado na publicação deste algoritmo proposto, em [64].

Tabela 4: Lista de sentenças curtas utilizadas no trabalho.

Sentenças da base MIT – MDSVC utilizadas neste trabalho
“Carol Owens”
“Chocolate fudge”
“Jessica Brown”
“Lee Hetherigton”
“Peppermint stick”
“Pralines and cream”
“Rocky Road”
“Sangita Sharma”
“Stephanie Seneff”
“Thomas Cronin”
“Adriana Girton”
“Barbara Taylor”
“Mint chocolate chip”
“Patricia Wilson”
“Peter Adamson”
“Justin Rattner”

Tabela 5: Lista de sentenças longas utilizadas no trabalho

Sentenças da base TIMIT utilizadas neste trabalho
“She had your dark suit in greasy wash water all year” (homem)
“She had your dark suit in greasy wash water all year” (mulher)
“Don't ask me to carry an oily rag like that”
“Maybe they will take us”
“Assume, for example, a situation where a farm has a packing shed and fields”
“Fuming, helpless, he watched them pass him”
“What outfit does she drive for?”
“Is a relaxed home atmosphere enough to help her outgrow these traits?”
“Fill small hole in bowl with clay”
“Publicity and notoriety go hand in hand”
“Norwegian sweaters are made of lamb's wool”
“Brush fires are common in the dry underbrush of Nevada”
“Those answers will be straightforward if you think them through carefully first”
“I just saw Jim near the new archeological museum”
“We'll serve rhubarb pie after Rachel's talk”
“Biblical scholars argue history”

5.3 **Resultados e Análise**

5.3.1 Base MIT-MDSVC

As sentenças da base MIT-MDSVC possuem locutores masculinos e femininos, uma grande variabilidade de fonemas e um aumento incremental no ruído de fundo. Uma comparação com o algoritmo de Ziólko *et al.* [49] é apresentada na Tabela 6.

A execução do algoritmo proposto depende apenas da *STFT* e da análise de componentes conectados, resultando em um custo computacional de $O(n^2)$. O algoritmo comportou-se bem em ambientes com ou sem ruído. Os parâmetros utilizados nos experimentos foram iguais aqueles utilizados no Capítulo 4, mostrando que a técnica proposta realiza a segmentação em várias situações, e com ambientes diferentes. Nos testes efetuados, a região de voz foi corretamente segmentada para todas as sentenças de fala. Os picos de ruído provenientes da ação de ligar o microfone foram removidos com sucesso (é possível visualizar um desses picos no início da Figura 4.11). Graças à subtração pela média horizontal, em (23), é possível segmentar sílabas fonéticas como a \ens\ da palavra "Owens" e também outros sons nasais como \n\ ou \m\. O uso da dilatação admite a inclusão de alguns fonemas nos segmentos resultantes, por exemplo, \s\, \j\ e \c\. Em algumas sentenças, o algoritmo agrupou dois ou mais fonemas em uma mesma região de segmentação, ou perdeu a parte inicial ou final de alguns fonemas (ver Tabela 7). A parte central dos sinais de áudio testados foi corretamente segmentada em todos os casos, mesmo com a presença de ruído.

O algoritmo de Ziólko *et al.* funciona bem em sinais de voz com ausência de ruído, porém produz resultados ruins quando processado em ambientes ruidosos (Tabela 6). Por esse motivo, para melhorar o seu desempenho, um pré-processamento foi realizado com o primeiro ciclo do algoritmo proposto, objetivando a remoção do ruído de fundo, e deixando apenas o segmento de voz. Este procedimento melhorou o desempenho do algoritmo de Ziólko *et al.* como pode ser visto na Tabela 6. Os parâmetros utilizados para o algoritmo de Ziólko *et al.* foram: $\rho = 0{,}02$, $\rho_{min} = 0{,}003$, $\beta = 1$, $\alpha = 5$, como podem ser vistos em [49].

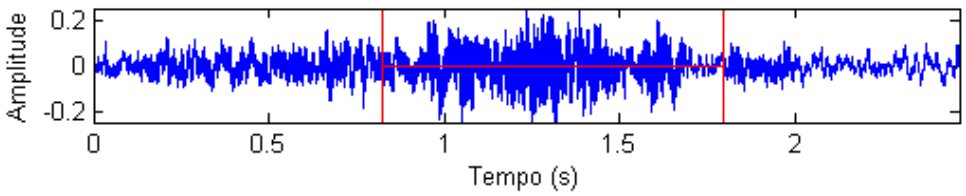
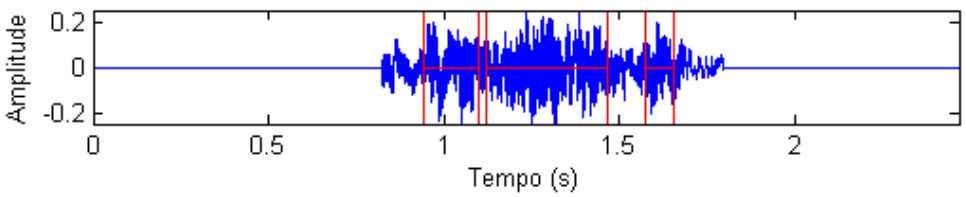
Tabela 6: Número de sílabas fonéticas segmentadas do algoritmo proposto em comparação ao algoritmo de Ziólko *et al.* aplicados em sentenças da base MIT-MDSVC

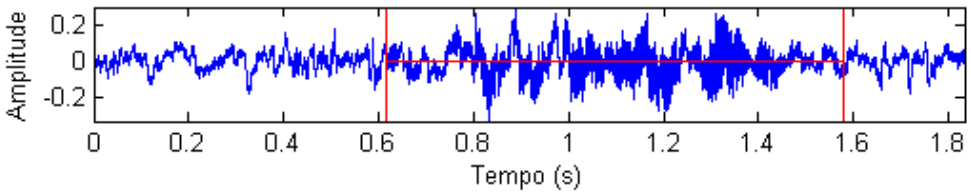
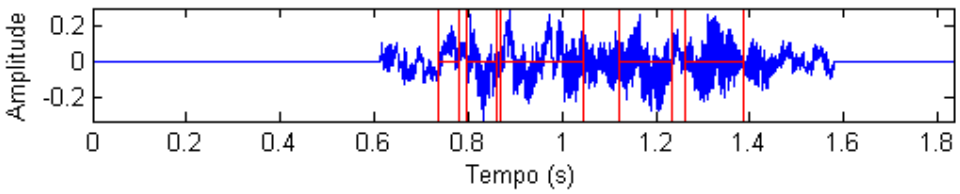
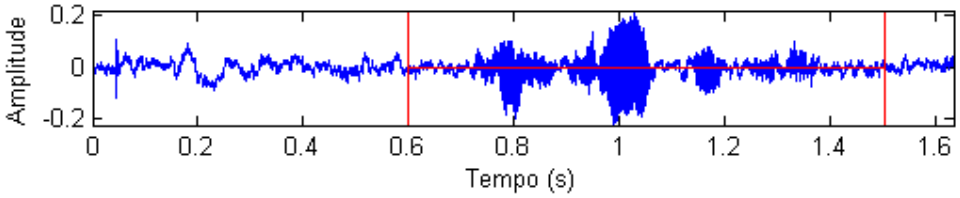
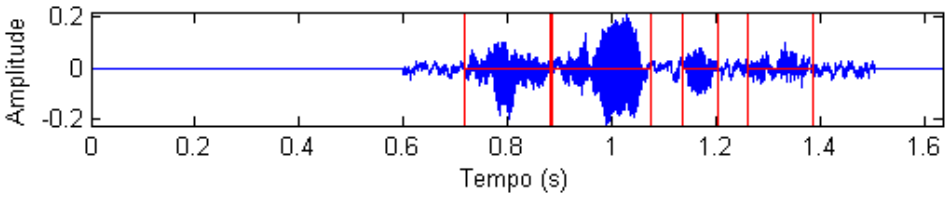
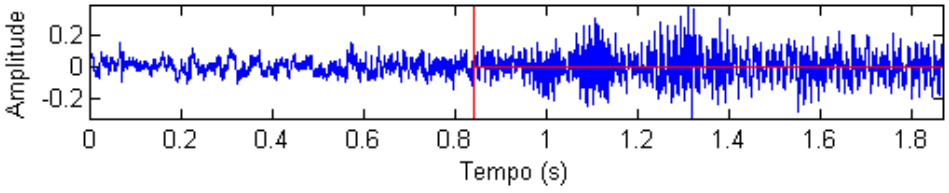
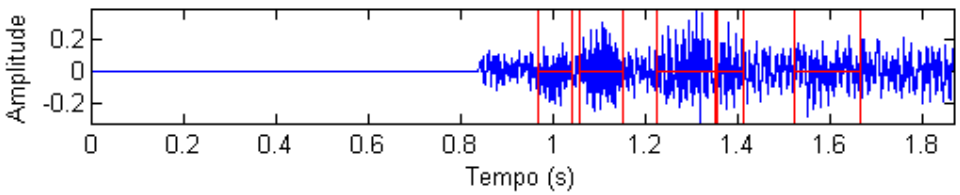
Número de Segmentos de Sílabas Fonéticas			
Ideal (*ground truth*)	Algoritmo proposto	Ziólko *et al.*	Algoritmo proposto + Ziólko *et al.*
"Carol Owens" (mulher pronuncia) – arquivo: f00-s1-L3a-phrase50_16k			
04	03	28	13
"Chocolate fudge" (mulher pronuncia) – arquivo: f14-s1-L2a-phrase56_16k			
06	03	30	16

“Jessica Brown” (mulher pronuncia) – arquivo: f16-s1-L3a-phrase54_16k			
04	04	39	14
“Lee Hetherington” (homem pronuncia) – arquivo: m09-s2-L2b-phrase03_16k			
05	05	23	14
“Peppermint stick” (mulher pronuncia) – arquivo: f13-s1-L1a-phrase58_16k			
04	04	17	07
“Pralines and cream” (mulher pronuncia) – arquivo: f21-s1-L3a-phrase52_16k			
05	03	19	08
“Rocky Road” (homem pronuncia) – arquivo: m16-s2-L1b-phrase06_16k			
03	03	30	13
“Sangita Sharma” (mulher pronuncia) – arquivo: f16-s1-L4a-phrase23_16k			
05	04	46	22
“Stephanie Seneff” (mulher pronuncia) – arquivo: f14-s2-L2b-phrase13_16k			
05	04	33	20
“Thomas Cronin” (homem pronuncia) – arquivo: m01-s1-L1a-phrase35_16k			
04	04	27	10

Na Tabela 7, o algoritmo de segmentação proposto é executado em quatro sinais de voz ruidosos. O primeiro experimento tem um ruído de apito, o segundo tem um ruído de chuva, o terceiro tem um ruído originado pelo vento, e o quarto tem um ruído do tráfego de uma rua movimentada. A segmentação do sinal original bem como das sílabas fonéticas são apresentadas. Mais uma vez, o algoritmo segmentou corretamente todas as sentenças.

Tabela 7: Resultados da segmentação de voz e sílabas fonéticas do algoritmo proposto.

Experimento	**Segmentação de voz / Segmentação de sílabas fonéticas**
1º	(mulher pronuncia) - arquivo: f00-s1-L3a-phrase51_16k
Pronúncia: “Adriana Girton”	
Segmentos: “ad”, “riana”, “girt”	
2º	(mulher pronuncia) – arquivo: f19-s1-L4a-phrase53_16k

Pronúncia: “Patricia Wilson”	
Segmentos: “a”, “tri”, “cia”, “wil”, “so”	
3º	(mulher pronuncia) - arquivo: f19-s1-L4a-phrase53_16k
Pronúncia: “Mint chocolate chip”	
Segmentos: “mint”, “choco”, “la”, “techip”	
4º	(mulher pronuncia) – arquivo: f19-s1-L4a-phrase53_16k
Pronúncia: “Peter Adamson”	
Segmentos: “pe”, “ter”, “a”, “dam”, “son”	

A Figura 5.1 (a) apresenta o sinal original da sentença “Justin Rattner”, pronunciada por um homem (arquivo: m02-s2-L2b-phrase57_16k). A Figura 5.1 (b) apresenta a segmentação de sílabas fonéticas, alcançada pelo algoritmo proposto (aplicado após o ciclo de segmentação de voz). A Figura 5.1 (c) ilustra os resultados alcançados pelo algoritmo de Ziólko *et al.* O método proposto encontrou três segmentos para as sílabas fonéticas “jus”tin”ratt” (a sílaba fonética “ner” é perdida por conta do ruído), enquanto o algoritmo de Ziólko *et al.* encontrou 45 fonemas.

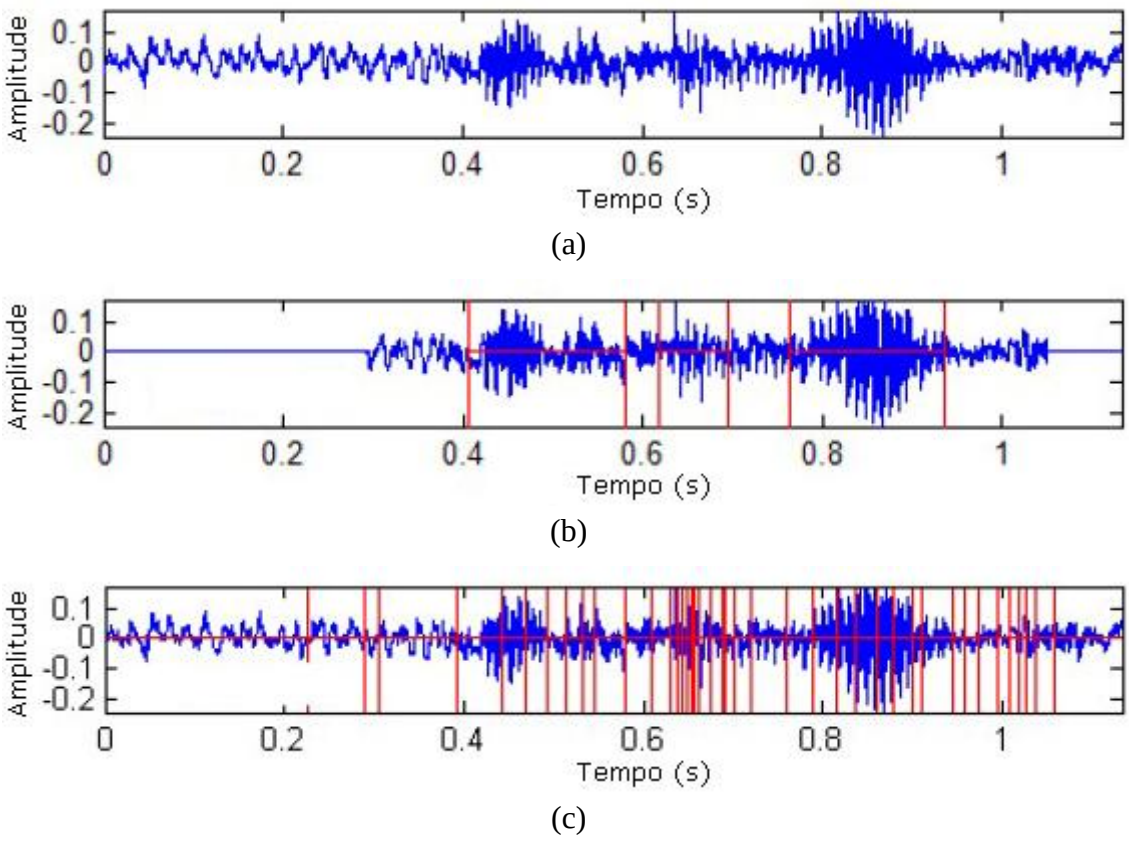


Figura 5.1: Comparação da saída da segmentação de sílabas fonéticas para o algoritmo proposto e Ziólko *et al.* (a) áudio original; (b) segmentação de fonemas pelo algoritmo proposto (após ciclo de segmentação de voz); (c) segmentação de fonemas pelo algoritmo de Ziólko *et al.*

5.3.2 Base TIMIT

Algumas sentenças da base de áudio TIMIT foram utilizadas para verificar o comportamento do algoritmo em relação a sentenças longas. As sentenças desta base não possuem ruídos, já que foram gravadas em ambiente controlado. Os parâmetros utilizados nos experimentos foram iguais aqueles utilizados no Capítulo 4. No geral, o algoritmo comportou-se de maneira similar a sua utilização no primeiro conjunto de dados com sentenças curtas. O algoritmo de Ziólko *et al.* alcança um número de segmentos mais próximo do ideal, já que as sentenças da base TIMIT foram gravadas em ambiente controlado (ausência de ruído). A região

de voz também foi corretamente segmentada para todas as sentenças de fala. A Tabela 8 ilustra os resultados da execução do algoritmo no conjunto de sentenças selecionadas da base TIMIT.

Tabela 8: Número de sílabas fonéticas segmentadas do algoritmo proposto em comparação ao algoritmo de Ziólko *et al.* aplicados em sentenças da base TIMIT

Número de Segmentos de Sílabas Fonéticas			
Ideal (*ground truth*)	Algoritmo proposto	Ziólko *et al.*	Algoritmo proposto + Ziólko *et al.*
"She had your dark suit in greasy wash water all year" (homem pronuncia) – arquivo: dr1 - mcpm0\sa1.wav			
13	13	29	20
"She had your dark suit in greasy wash water all year" (mulher pronuncia) – arquivo: dr2 - faem0\sa1.wav			
13	14	30	22
"Don't ask me to carry an oily rag like that" (homem pronuncia) – arquivo: dr1 - mcpm0\sa2.wav			
11	06	18	14
"Maybe they will take us" (homem pronuncia) – arquivo: dr2 - marc0\si1188.wav			
06	05	09	07
"Assume, for example, a situation where a farm has a packing shed and fields" (mulher pronuncia) – arquivo: dr2 - faem0\si1392.wav			
20	19	41	29
"Fuming, helpless, he watched them pass him" (homem pronuncia) – arquivo: dr2 - marc0\si1818.wav			
09	07	17	13
"What outfit does she drive for" (mulher pronuncia) – arquivo: dr2 - faem0\si2022.wav			
06	05	16	09
"Is a relaxed home atmosphere enough to help her outgrow these traits?" (homem pronuncia) – arquivo: dr2 - marc0\si558.wav			
18	14	37	30
"Fill small hole in bowl with clay" (mulher pronuncia) – arquivo: dr2 - faem0\si762.wav			
07	07	19	11
"Publicity and notoriety go hand in hand" (mulher pronuncia) – arquivo: dr2 - faem0\sx132.wav			
10	08	21	16
"Norwegian sweaters are made of lamb's wool" (homem pronuncia) – arquivo: dr2 - marc0\sx198.wav			
09	07	23	18
"Brush fires are common in the dry underbrush of Nevada" (homem pronuncia) – arquivo: dr2 - marc0\sx288.wav			
12	06	26	16
"Those answers will be straightforward if you think them through carefully first" (mulher pronuncia) – arquivo: dr2 - faem0\sx312.wav			
19	13	32	23

“I just saw Jim near the new archeological museum” (homem pronuncia) – arquivo: dr2 - marc0\sx378.wav			
12	09	25	18
“We'll serve rhubarb pie after Rachel's talk” (mulher pronuncia) – arquivo: dr2 - faem0\sx402.wav			
11	08	24	19
“Biblical scholars argue history” (mulher pronuncia) – arquivo: dr2 - faem0\sx42.wav			
08	07	17	13

A Figura 5.2 apresenta um exemplo completo da execução do algoritmo proposto sobre o sinal de voz da sentença “She had your dark suit in greasy wash water all year”, pronunciada por uma voz masculina.

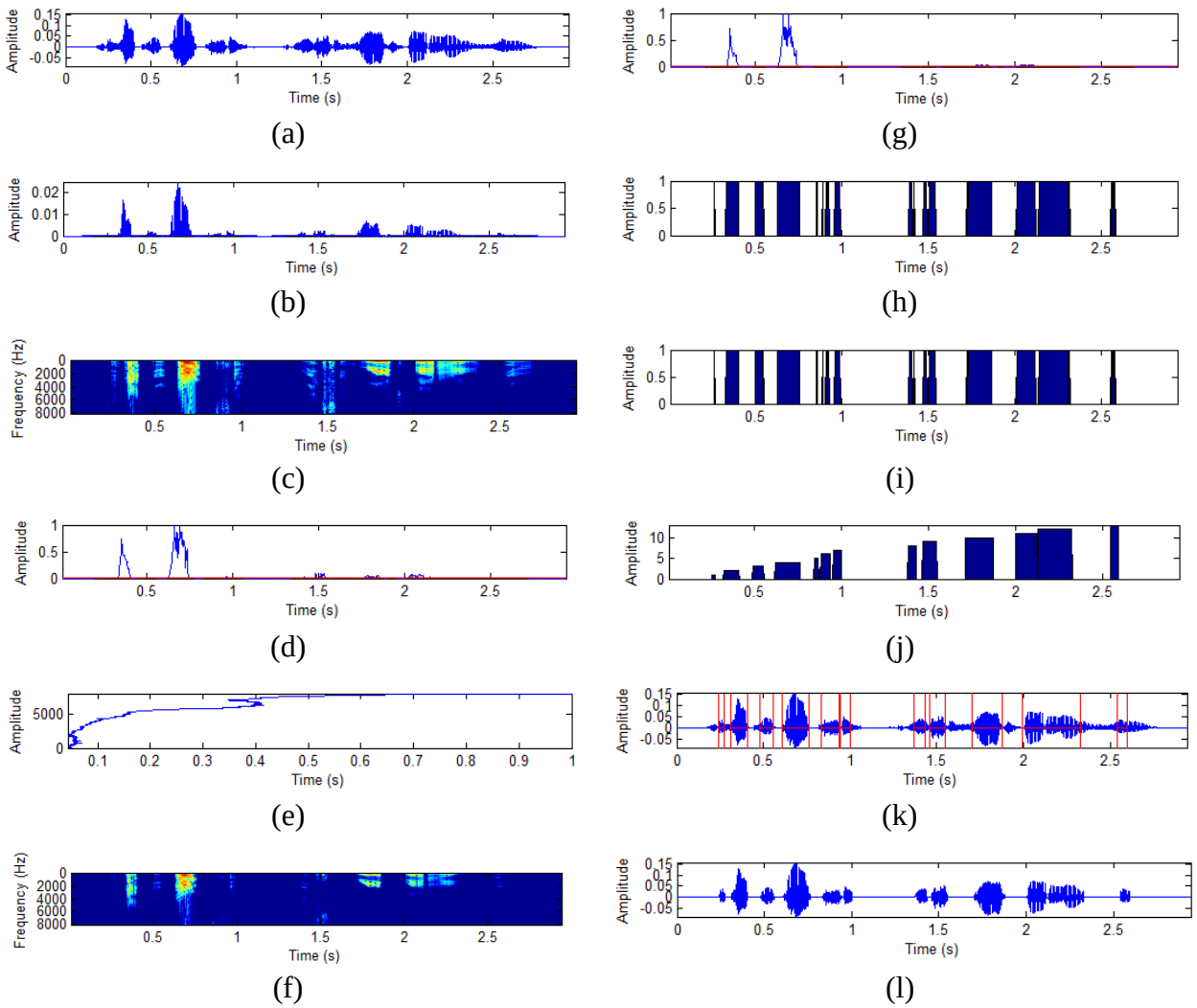


Figura 5.2: Saída passo-a-passo da execução do algoritmo de segmentação proposto sobre o sinal de voz da pronúncia “She had your dark suit in greasy wash water all year”: (a) sinal de áudio original; (b) densidade de energia; (c) espectrograma da energia; (d) projeção vertical do espectrograma; (e) projeção horizontal do espectrograma; (f) imagem realçada do espectrograma; (g) nova projeção vertical do espectrograma; (h) segmentos binários; (i) remoção de pequenos segmentos; (j) dilatação dos segmentos restantes; (k) sílabas fonéticas segmentadas sobre o sinal original; (l) apenas as sílabas fonéticas segmentadas.

6 Conclusões e Trabalhos Futuros

O reconhecimento automático de voz tem sido a meta de pesquisadores da área desde os anos 60. Neste período, ocorreram vários avanços, algumas técnicas tornaram-se padrão para determinadas tarefas e outras perderam a relevância na área. Entretanto, mesmo após mais de 50 anos de pesquisa na área de reconhecimento automático de voz e com a utilização na prática de algumas aplicações simples de reconhecimento de voz, o desempenho da máquina ainda está longe de se equiparar com a do ser humano. Este trabalho tem como objetivo adicionar uma contribuição à literatura da área, apresentando um novo algoritmo de segmentação de voz e fonemas. Como já foi dito, a etapa de segmentação de voz é uma das etapas mais importantes de um sistema de reconhecimento de voz, e é também uma das tarefas mais complexas nessa área de pesquisa [3].

Em resumo, uma nova técnica de processamento de voz, baseada na análise de imagem do espectrograma, e que nos testes efetuados neste trabalho mostrou um bom desempenho na segmentação de voz e fonemas em ambientes ruidosos, foi apresentada. A técnica opera em dois ciclos: primeiro segmentando o sinal de voz e depois utilizando esse sinal segmentado para encontrar as sílabas fonéticas ou os fonemas. A base de dados de áudio *MIT-MDSVC* e a *TIMIT* foram utilizadas para validação do algoritmo. Foram escolhidos sinais com vozes masculina e feminina, com diferentes fonemas, e com aumento incremental de ruído (apito, chuva, vento e tráfego de rua movimentada).

O algoritmo proposto segmentou corretamente grande parte das sentenças utilizadas, em ambientes com ausência e presença de ruído. Nos experimentos, a região de voz (1º ciclo) de todas as amostras utilizadas foi corretamente separada do fundo. O algoritmo também conseguiu segmentar corretamente fonemas nasais de 5 sentenças curtas (oriundas da base *MIT-MDSVC),* de um total de 8 que possuem esse tipo de fonema, por exemplo, \n\, \m\ e \ens\. A parte central do sinal de voz também foi corretamente segmentada em todos os exemplos, mesmo com a presença de ruído. Em algumas amostras com alta taxa de ruído, a parte inicial ou final de determinados fonemas foram perdidas.

Uma comparação com o algoritmo de Ziólko *et al.* [48][49] também foi realizada. O algoritmo mostrou-se eficiente em ambientes com ausência de ruído, porém apresentou uma baixa precisão em ambientes ruidosos, compostos de situações rotineiras no mundo real.

As operações necessárias para o algoritmo proposto são de ampla utilização e adaptáveis a uma implementação via *hardware*. A técnica é determinística, já que não foram utilizadas variáveis probabilísticas ou aleatórias, e a fase de treinamento pode ser efetuada de maneira

opcional, já que o algoritmo mostrou bom desempenho com os parâmetros estabelecidos neste trabalho, tornando-se uma boa alternativa para aplicações *online*.

6.1 Contribuições

Através deste trabalho, verifica-se que é possível realizar, de maneira eficiente, a junção de técnicas de processamento de imagem com técnicas de processamento de voz. Neste caso, especificamente, foram aplicadas algumas operações clássicas de processamento de imagens (dilatação, análise de projeção, entre outras) sobre as imagens de espectrogramas, que foram gerados por sinais de voz, provenientes da base *MIT-MDSVC*. Os experimentos mostram que o algoritmo proposto é bastante promissor, inclusive, obteve para todos os sinais testados, um melhor desempenho em relação ao algoritmo de Ziólko *et al*. [48][49]. Desta forma, este trabalho resulta em uma interessante contribuição para a área de segmentação de voz.

A pesquisa também produziu um artigo publicado na *IEEE International Conference on Systems, Man and Cybernetics* (SMC), realizada em outubro de 2012, em Seul, Coreia do Sul [64].

6.2 Trabalhos Futuros

Existem diversas diretrizes a serem seguidas para os trabalhos futuros. Por exemplo, pode-se desenvolver um procedimento de remoção de ruído, baseado na informação fornecida pela segmentação do *background*, que contém, na maioria dos casos, apenas ruído. Outra iniciativa é a utilização de uma mediana local, ao invés da mediana global utilizada neste trabalho.

Uma alternativa bastante promissora para trabalho futuro é a utilização de imagens de neurograma [56][58], ao invés das imagens de espectrogramas utilizadas neste trabalho. O neurograma é análogo ao espectrograma, ele é uma representação em imagem de um sinal no domínio tempo-frequência, utilizando cores para indicar a intensidade da atividade. Basicamente, o neurograma é formado pelos padrões de descarga produzidos por fibras dos nervos auditivos em resposta a frases pronunciadas ou outros sons complexos. A Figura 6.1 ilustra o exemplo de neurogramas para o sinal de voz da pronúncia "*ship"*.

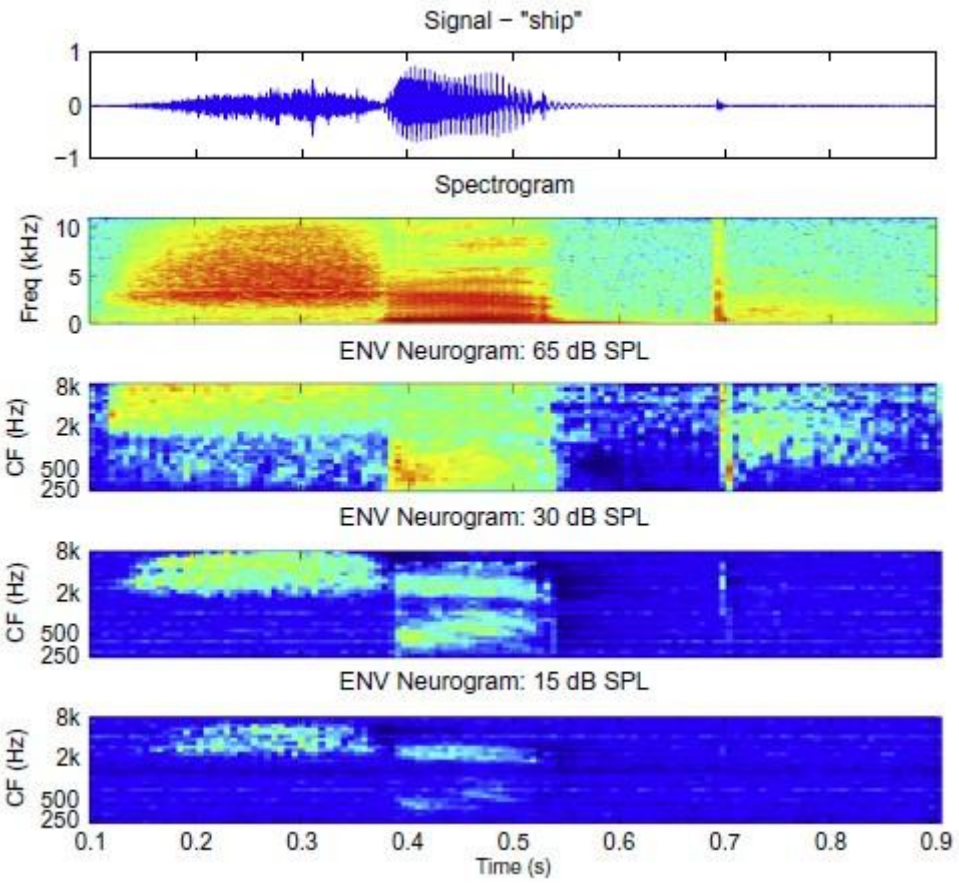


Figura 6.1: O primeiro gráfico mostra os sinal original do áudio da palavra "ship"; Os outros três gráficos ilustram neurogramas derivados do sinal original, com intesidades de 65, 30 e 15dB respectivamente. Retirado de [58].

Por último, pode-se verificar o comportamento do algoritmo na detecção de atividade de voz (*VAD – Voice Activity Detection*). Como o algoritmo segmentou a região de voz (1º ciclo) em todos os sinais testados, incluindo sinais com ruídos inerentes ao mundo real, ele pode ser adaptado para funcionar bem em aplicações desse tipo. Alguns autores vêm trabalhando com técnicas estatísticas e baseadas em computação inteligente [69][70][71], porém a maioria dos algoritmos de VAD falha à medida que o ruído de fundo aumenta.

Referências Bibliográficas

[1] RABINER, L.; JUANG, B.-H. **Fundamentals of Speech Recognition,** New Jersey, EUA, Ed. Prentice Hall, 1993.

[2] WOLFE J. M.; KLUENDER K. R.; LEVI, D. M.; BARTOSHUK, L. M.; HERZ, R. S.; KLATZKY, R. L.; LEDERMAN, S. J.; MERFELD, D. M. **Sensation & Perception**, 2ª edição, Massachusetts, EUA, Ed. Sinauer Associates, 2009.

[3] PAUL, C.; RICHARD, S.; NICK C.; JOE, L. **Bootstrapping Word Boundaries: A Bottom-up Corpus-Based Approach to Speech Segmentation,** Cognitive Psychology**,** Elsevier, vol. 33, pp. 111-153, 1997.

[4] ANNA, E.; GUIDO, A. **Text Independent Methods for Speech Segmentation** in G. Chollet et.al, "Non Linear Speech Modeling", Lecture Notes in Computer Science, Springer Berlin, vol. 3445, pp. 261-290, 2005.

[5] AL-HADDAD, S. A. R. *et al*. **Automatic Segmentation and Labeling for Malay Speech Recognition**, Internacional Conference on Signal Processing, Computational Geometry & Artificial Vision, pp. 217-221, Elounda, Grécia, 2006.

[6] MERMELSTEIN P. **Automatic segmentation of speech into syllabic units**, Journal of Acoustic Society America, vol. 58, p. 880-883, 1975.

[7] ZHANG, T.; KUO, C. **Hierarchical classification of audio data for archiving and retrieving,** Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3001-3004, Arizona, EUA, 1999.

[8] ANTAL M. **Speaker independent phoneme classification in continuous speech**. Studia Univ. Babes-Bolyal, Informatica, vol. 49, pp. 55-64, 2004.

[9] RABINER, L.R.; SAMBUR, M.R. **Some preliminary Experiments in the Recognition of Connected Digits**, IEEE Trans. of Acoustic, Speech and Signal Processing, vol. 24, Issue 2, pp. 170-182, 1976.

[10] MICHAEL, R.B. **Speech Segmentation and Word Discovery: A Computational Perspective**, Trends in Cognitive Science, vol. 3, Issue 8, pp. 294-301, 1999.

[11] ABDULLAH, H. **Buku Linguistik Am**, Kuala Lumpur, Malásia, Ed. PTS Professional Publishing, 2005.

[12] INGRAM, J. C. L. **Neurolinguistics: An Introduction to Spoken Language Processing and its Disorders**, 1st Edition, Reino Unido, Cambridge University Press, 2007.

[13] PRASAD, V. K.; NAGARAJAN, T.; MURTHY, H. A. **Automatic Segmentation of Continuous Speech Using Minimum Delay Phase Group Delay Functions**, Speech Communication Journal, vol. 42, issues 3-4, pp. 429-446, 2004.

[14] ANINDYA, S.; SREENIVAS, T. V. **Automatic Speech Segmentation Using Average Level Crossing Rate Information**, Proceeding of International Conference of Acoustic, Speech and Signal Processing (ICASSP), pp. 397-400, Pennsylavania, EUA, 2005.

[15] SALAM, M. S. **A Fusion of Statistical and Connectionist Approaches for Segmentation of Connected Digit**, PhD Thesis, Universiti Teknologi Malaysia, 2010.

[16] HAYKIN, S.; VEEN, B. V. **Sinais e Sistemas,** Porto Alegre, Bookman, 2001.

[17] RABINER, L.; SCHAFER, R. **Theory and Applications of Digital Speech Processing,** 1st edition, New Jersey, EUA, Ed. Prentice Hall, 2010.

[18] HOSSAN, M. A.; MEMON, S.; GREGORY, M. A. **A novel approach for MFCC feature extraction**, International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-5, 13-15, Queensland, Austrália, 2010.

[19] RABINER, L. R. **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**, Proceedings of the IEEE, vol. 77, Issue 2, pp. 257–286, 1989.

[20] FAROOQ, O.; DATTA, S. **Wavelet based robust subband features for phoneme recognition**, IEEE Proceedings: Vision, Image and Signal Processing, vol. 151, no. 3, pp. 187–193, 2004.

[21] RABINER, L. R.; SCHAFER, R. W. **Introduction to Digital Speech Processing,** Foundations and Trends in Signal Processing, vol. 1, no. 1-2, pp. 1-194, 2007.

[22] LEVELT, W. **Speaking: From intention to articulation,** Cambridge, EUA, MIT Press, 1989.

[23] GIEGERICH, H. J. **English Phonology: An introduction**, Cambridge, EUA, Cambridge University Press, 1992.

[24] MONAGHAN, A. **Phonetics: Processes of Speech Production**, 1998, disponível em: http://www.compapp.dcu.ie/~alex/CA162/PHONETICS/processes.html

[25] ROSSING, T. D. **The Science of Sound**, Michigan University, EUA, Ed. Addison-Wesley, 2nd edition, 1990.

[26] WARREN, R. M. **Auditory Perception**, Cambridge, EUA, Cambridge University Press, 1999.

[27] GONZALEZ, R. C.; WOODS R. E. **Digital Image Processing**, Ed. Pearson Education, 3rd Edition, 2010.

[28] UMBAUGH, S. **Digital Imaging Processing and Analysis: Human and Computer Applications with CVIPtools**, Ed. CRC, 2nd Edition, 2011.

[29] SERRA, J.; NOEL, A. C. C. **Image Analysis and Mathematical Morphology**, Ed. Academic Press, 1982.

[30] CHUANG, C-H.; LI, Z-P.; LIN, G-S.; SHEN, W-C. **A Projection Profile-based Algorithm for Content-aware Image Resizing**, International Conference on Computer Graphics, Imaging and Visualization, (CGIV), pp. 125-130, Singapura, 2011.

[31] ANTONACOPOULOS, A; KARATZAS, D. **Document image analysis for World War II personal records**, Proc. First Intl. Workshop on Document Image Analysis for Libraries (DIAL), pp. 336–341, Palo Alto, EUA, 2004.

[32] ARIVAZHAGAN, M; SRINIVASAN, H; SHIRARI, S. **A statistical approach to line segmentation in handwritten documents**, Proceedings of the International Society for Optics and Photonics (SPIE), vol. 6500, California, EUA, 2007.

[33] BAR-YOSEF, Y.; HAGBI, N.; KEDEM, K. DINSTEIN, I. **Line segmentation for degraded handwritten historical documents**, Proceedings of International Conference on Document Analysis and Recognition (ICDAR), pp. 1161–1165, Barcelona, Espanha, 2009.

[34] OSTENDORF, M. *et al*. **Speech segmentation and spoken document processing**, IEEE Signal Processing Magazine, vol. 25, nº. 3, pp. 59-69, 2008.

[35] GRAYDEN, D. B.; SCORDILIS, M. S. **Phonemic segmentation of fluent speech**, International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 73–76, Adelaide, Austrália, 1994.

[36] WEINSTEIN, C. J.; MCCANDLESS S. S.; MONDSHEIN L. F.; ZUE V. W. **A system for acoustic-phonetic analysis of continuous speech**, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 23, no. 1, pp. 54-67, 1975.

[37] ZUE, V.W. **The use of speech knowledge in automatic speech recognition**, Proceedings of the IEEE, vol. 73, no. 11, pp. 1602–1615, 1985.

[38] SUH, Y.; LEE, Y. **Phoneme segmentation of continuous speech using multi-layer perceptron**, International Conference on Spoken Language Processing, pp. 1297-1300, Philadelphia, EUA, 1996.

[39] OSTENDORF, M.; DIGALAKIS, V. V.; KIMBALL, O. A. **From HMM's to segment models: A unified view of stochastic modeling for speech recognition**, IEEE Transactions on Speech and Audio Processing, vol. 4, no. 5, pp. 360–378, 1996.

[40] FANTINATO, P. C. *et al*. **A fractal-based approach for speech segmentation**, IEEE Internation Symposium on Multimedia (ISM), Califórnia, EUA, pp. 551-555, 2008.

[41] AL-AKAIDI, M.; BLACKLEDGE, J.; FEKKAI, S. **Fractal dimension segmentation: isolated speech recognition**, IEEE Seminar on Speech Coding for Algorithms for Radio Channels, pp. 4/1 – 4/5, Londres, Reino Unido, 2000.

[42] PARK, S. S.; KIM, N. S. **On using multiple models for automatic speech segmentation**, IEEE Transactions on Audio, Speech and Language Processing, vol. 15, n. 8, pp. 2202-2212, 2007.

[43] VOSTERMANS, A., MARTENS, J.-P., COILE, B. V. **Automatic segmentation and labeling of multi-lingual speech data**, Speech Communication, vol.19, pp. 271-293, 1996.

[44] FORNEY, G. D. **The viterbi algorithm**, Proceedings of the IEEE, vol.61, no.3, pp. 268-278, 1973.

[45] TORRES, H. M.; GURLEKIAN, J. A. **Acoustic speech unit segmentation for concatenative synthesis**, Computer Speech and Language, vol. 22, no. 2, pp. 196–206, 2008.

[46] MALLAT, S. **A Wavelet Tour of Signal Processing**, Ed. Academic Press, 3rd Edition, 2009.

[47] SHANNON, C. **A mathematical theory of communication**, Bell System Technical Journal, vol. 3, no. 27, pp. 379–423, 1948.

[48] ZIÓLKO, B.; MANANDAR S.; WILSON, R. C. **Phoneme Segmentation of Speech**, International Conference on Pattern Recognition, pp. 282-285, Hong Kong, 2006.

[49] ZIÓLKO, B.; MANANDAR S.; WILSON, R. C.; ZIÓLKO, M. **Wavelet Method of Speech Segmentation**, European Signal Processing Conference (EUSIPCO), Florence, Itália, 2006.

[50] DEVIREN, M.; DAOUDI, K. **Frequency and wavelet filtering for robust speech recognition**, Joint International Conference on Artificial Neural Networks (ICANN)/International on Neural Information Processing (ICONIP), pp. 452-460, Istanbul, Turquia, 2003.

[51] GOWDY, J. N.; TUFEKCI, Z. **Mel-scaled discrete wavelet coefficients for speech recognition**, Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 1351-1354, Istanbul, Turquia, 2000.

[52] RIOUL, O.; VETTERLI, M. **Wavelets and signal processing**, IEEE Signal Processing Magazine, vol. 8, pp.11–38, 1991.

[53] WANG, D.; NARAYANAN. S. **Piecewise linear stylization of pitch via wavelet analysis**, Proceedings of Interspeech, pp. 3277-3280, Lisboa, Portugal, 2005.

[54] DAUBECHIES, I. **Ten lectures on Wavelets**, Pennsylvania, EUA, Society for Industrial and Applied Mathematics, 1992.

[55] WOO, R. H.; PARK, A.; HAZEN, T. J. **The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments**, IEEE Odyssey Speaker and Language Recognition Workshop, pp. 1-6, San Juan, Porto Rico, 2006.

[56] HINES, A.; HARTE, N. **Speech intelligibility from image processing**, Speech Communication, vol. 52, no. 9, pp. 736–752, 2010.

[57] WANG, Z.; BOVIK, A.; SHEIKH, H.; SIMONCELLI, E. **Image quality assessment: from error visibility to structural similarity**, IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[58] HINES, A.; HARTE, N. **Speech intelligibility prediction using a Neurogram Similarity Index Measure**, Speech Communication, vol. 54, no. 2, pp. 306–320, 2012.

[59] AJMERA, P. K.; JADHAV, D. V.; HOLAMBE, R. S. **Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram**, Pattern Recognition, vol. 44, no. 10, pp. 2749–2759, 2011.

[60] MELLO, C.A.B.M. **Análise Wavelet**, 2013, disponível em: {HYPERLINK "http://www.cin.ufpe.br/~cabm/pds/PDS_Aula09_Wavelet.pdf"}

[61] OPPENHEIM, A. V.; SCHAFER, R. **Discrete Time Signal Processing**, Ed. Prentice Hall, Michigan, EUA, 1989.

[62] BENESTY, J.; CHEN, J.; HABETS, E. A. P. **Speech Enhancement in the STFT Domain**, Springer, 1st Edition, 2011.

[63] DANIELS, J. D. **Ground Penetrating Radar**, Institution of Engineering and Technology, 2nd Edition, Londres, Reino Unido, 2004.

[64] COSTA, D. C.; LOPES, G. A. M.; MELLO, C. A. B.; VIANA, H. O. **Speech and phoneme segmentation under noisy environment through spectrogram image analysis**, IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1017-1022, Seoul, Coreia do Sul, 2012.

[65] LIPPMANN, R. P. **Speech recognition by machines and humans**, Speech Communication, vol. 22, pp. 01-15, 1997.

[66] KOHONEN, T. **The self-organizing map**, Proceedings of the IEEE , vol.78, no.9, pp. 1464-1480, 1990.

[67] JAKOBSON, R. **Verbal Art, Verbal Sign, Verbal Time**, University Of Minnesota Press, EUA, 1985.

[68] GAROFOLO, J. S. *et al*. **TIMIT Acoustic-Phonetic Continuous Speech Corpus**, Linguistic Data Consortium, Philadelphia, 1993.

[69] ELIZALDE, B.; FRIEDLAND, G. **Lost in segmentation: Three approaches for speech/non-speech detection in consumer-produced videos**, IEEE International Conference on Multimedia and Expo (ICME) , pp.1-6, California, EUA, 2013.

[70] CHOI, J.-H.; CHANG, J.-H. **On using spectral gradient in conditional MAP criterion for robust voice activity detection**, IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp.370-374, Pequim, China, 2012.

[71] MOUSAZADEH, S.; COHEN, I. **Voice Activity Detection in Presence of Transient Noise Using Spectral Clustering**, IEEE Transactions on Audio, Speech, and Language Processing, vol.21, no.6, pp.1261-1271, 2013.