Ênio dos Santos Silva

EXTENSÃO ARTIFICIAL DE LARGURA DE BANDA PARA SINAIS DE FALA EM TELEFONIA USANDO CLASSIFICAÇÃO FONÉTICA

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Rui Seara.

Florianópolis
2016

Ficha de identificação da obra elaborada pelo autor
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Silva, Ênio dos Santos
Extensão Artificial de Largura de Banda para Sinais de Fala em Telefonia Usando Classificação Fonética / Ênio dos Santos Silva; orientador, Rui Seara - Florianópolis - SC, 2016.
103p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia Elétrica.

Inclui referências

1. Engenharia Elétrica. 2. Classificação fonética. 3. Codificação de fala. 4. Extensão artificial de largura de banda. 5. Mineração de dados. 6. Mineração de dados. 7. Realce de voz. 8. Reconhecimento automático de fala. 9. Reconhecimento de padrões. 10. Seleção de atributos. I. Seara, Rui. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Elétrica. III. Título.

Ênio dos Santos Silva

EXTENSÃO ARTIFICIAL DE LARGURA DE BANDA PARA SINAIS DE FALA EM TELEFONIA USANDO CLASSIFICAÇÃO FONÉTICA

Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Engenharia Elétrica, Área de Concentração *Comunicações e Processamento de Sinais*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina

Florianópolis, 15 de março de 2016

_____
Prof. Carlos Galup Montoro, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia Elétrica

Banca Examinadora:

_____
Prof. Rui Seara, Dr. - UFSC
Orientador

_____
Prof. Hans Helmut Zürn, Ph.D. - UFSC

_____
Prof. Sidnei Noceti Filho, Dr. - UFSC

_____
Prof. Eduardo Luiz Ortiz Batista, Dr. - UFSC

Dedico este trabalho a minha querida filha Kelline Silva, minha amada noiva Marília Medeiros e aos meus estimados pais, Francisco e Jandira Silva. Dedico ainda à memória de minha amiga e companheira Keyla Melo.

Agradecimentos

Inicialmente, agradeço a Deus pela oportunidade de conhecer e conviver com pessoas maravilhosas durante todos os anos dedicados ao curso de mestrado, e por me confortar, trazendo paz e esperança, mesmo nos momentos mais difíceis.

Agradeço aos meus pais, que, com seu amor incondicional, não mediram esforços pra me apoiar e incentivar nessa árdua caminhada. Gordão (Pai) e dona Janda (mãe), minha gratidão não cabe em palavras, mas expresso dizendo que os amo muito.

Não posso deixar de agradecer quem me incentivou desde o princípio, minha irmã Elaine. Pessoa que, solidária ao irmão, ajudou a cuidar da minha filha e que embarcou (literalmente) comigo na retomada do curso de mestrado. Elaine, muito obrigado!

Também agradeço a Keyla Melo, minha fiel amiga, companheira e mãe da minha filha. Juntos compartilhamos sonhos e momentos de muita alegria. Agradeço também aos pais da Keyla, tia Francy e tio Mozart, que sempre confiaram em mim.

Agradeço a minha filha, Kelline Silva, por todo o carinho, pelo incentivo e pela compressão em dividir nosso precioso tempo juntos, de pai e filha, com o tempo que eu necessitava dedicar aos estudos. Filha, você sempre foi o meu maior incentivo!

Agradeço aos grandes amigos que conquistei nesse período, amigos que se tornaram verdadeiros irmãos: Salomão, Breno, Marcelo Bolinha, Rafael, Yader, André Tahim, Fábio Itturriet, Tales, Emerson Fedechen e Felipe Clemente (Mocotó).

A conclusão dessa jornada não seria possível sem a participação fundamental de minha noiva, Marília Medeiros, com quem voltei a formar uma família e a sonhar com um futuro cada vez mais feliz. Agradeço também aos meus sogros Jorge e Elaine Medeiros, pessoas maravilhosas que me acolheram como um filho e também não mediram esforços pra me apoiar. Serei eternamente grato!

A todos os colegas do Laboratório de Circuitos e Processamento de Sinais - LINSE, pela amizade, pela troca de informação e de conhecimento e pela colaboração dentro do nosso ambiente de trabalho. Aos amigos Barbosa, Becker, Kuhn, Matsuo e Zilli, pela atenção e assessoria remota na reta final da dissertação. Aos amigos Walter e Elton, pela hospitalidade, conselhos e agradáveis conversas.

Ao Prof. Ferando Pacheco, pelas relevantes contribuições ao trabalho e por todos os ensinamentos. Aos membros da banca examinadora pelos valorosos comentários que contribuíram para o aprimoramento deste trabalho.

Por fim, agradeço especialmente ao Prof. Rui Seara, meu orientador, pela paciência, persistência, confiança no meu trabalho, por todas as conversas, pelas palavras de motivação e pelas relevantes contribuições para o desenvolvimento desta dissertação.

Seja você quem for, seja qual for a posição social que você tenha na vida, a mais alta ou a mais baixa, tenha sempre como meta muita força, muita determinação e sempre faça tudo com muito amor e com muita fé em Deus, que um dia você chega lá. De alguma maneira você chega lá.

(Ayrton Senna, 1991)

Resumo

Este trabalho de pesquisa apresenta uma nova estratégia para implementar sistemas de extensão artificial de largura de banda (*artificial bandwidth extension* - ABWE) para sinais de fala aplicados à rede pública (convencional) de telefonia (*public switched telephone network* - PSTN). Especificamente, aqui é proposta uma estratégia baseada em classificação fonética visando representar satisfatoriamente segmentos de fala com energia concentrada em altas frequências, superando outros resultados apresentados na literatura. Para tal, técnicas de seleção de atributos aplicadas a sinais de fala de banda limitada são investigadas, aprimorando a classificação em grupos fonéticos abrangentes (*broad group phonetic* - BGP) com ênfase na discriminação de fonemas pertencentes ao grupo fonético fricativo. Adicionalmente, neste trabalho é discutida a integração do sistema de ABWE proposto em sistemas de reconhecimento automático de fala (*automatic speech recognition* - ASR) para o português brasileiro aplicados à PSTN. Particularmente, visando o aprimoramento de ASR em PSTN, as etapas de extração de atributos do sinal da fala e a etapa de construção do modelo acústico são desenvolvidas baseadas em sinais sintéticos de banda larga (*wideband* - WB) estimados a partir do realce de sinais de banda estreita (*narrowband* - NB) usando ABWE. Os resultados obtidos apresentam realce na qualidade subjetiva dos sinais de fala reconstruídos e ganho no desempenho do ASR, confirmando a eficácia das estratégias propostas neste trabalho de pesquisa.

Palavras-chave: Classificação fonética. Codificação de fala. Extensão artificial de largura de banda. Mineração de dados. Realce de voz. Reconhecimento automático de fala. Reconhecimento de padrões. Seleção de atributos.

Abstract

This research work presents a new strategy for implementing artificial band width extension (ABWE) systems for speech signals applied to the public switched telephone network (PSTN). Specifically, a strategy based on phonetic classification is proposed here aiming to represent speech segments with concentrated energy at high frequencies, outperforming other approaches from the open literature. In this context, feature selection techniques applied to limited band width speech signals are investigated, improving the broad group phonetic (BGP) classification with an emphasis on discrimination of phonemes belonging to the fricative phonetic group. In addition, the integration of the proposed ABWE approach in automatic speech recognition (ASR) systems for Brazilian Portuguese applied to the PSTN is also discussed. Particularly, in order to improve PSTN ASR systems, synthetically estimated wideband (WB) signals, from the narrowband (NB) enhancement by ABWE, are used to obtain more discriminating attributes of speech signals as well as for achieving better performance of acoustic models (AM). The obtained results show an enhancement in the quality of reconstructed speech signals with very good performance in ASR systems, confirming the effectiveness of the proposed strategies in this research work.

Keywords: Phonetic classification. Speech coding. Artificial bandwidth extension. Data mining. Speech enhancement. Automatic speech recognition. Pattern recognition. Attribute selection.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ABWE *artificial bandwidth extension*

ACR *absolute category rating*

ASR *automatic speech recognition*

BGP *broad group phonetic*

CCR *comparison category rating*

CFS *correlation-based feature selection*

CMOS *comparative mean opinion score*

DCT *discrete cosine transform*

GMM *gaussian mixture model*

HMM *hidden Markov model*

HP *high pass*

ISDN *integrated service digital network*

ITU *International Telecommunication Union*

***k*-NN** *k-nearest neighbor*

LP *low pass*

LPC *linear predictive coding*

LSF *line spectral frequency*

MFCC *mel-frequency cepstrum coefficients*

MI *mutual information*

MLP *multilayer perceptron*

MMSE *minimum mean squared error*

MOS *mean opinion score*

NB *narrowband*

NSR *network speech recognition*

PB *português brasileiro*

PDF *probability density function*

PLP *perceptual linear predictive*

PSTN *public switched telephone network*

SF *spectral folding*

SVM *support vector machine*

UP *upband*

VoIP *voice over internet protocol*

WB *wideband*

Sumário

Capítulo 1

Introdução

1.1 Motivação para o Uso de Extensão Artificial de Largura de Banda

A extensão artificial de largura de banda (*artificial bandwidth extension* - ABWE) é uma técnica que possibilita a estimação de componentes de frequência adicionais capazes de ampliar a largura de banda do espectro de um sinal de fala de banda limitada, originalmente transmitido pela rede de telefonia pública (*public switched telephone network* - PSTN). Nos últimos anos, essa técnica ganhou evidência por estar contida em um dos mais difundidos *codecs* de telefonia: o *codec* G.729 [1], em sua Recomendação ITU-T G.729.1 [2]. A necessidade de extensão de largura de banda surgiu devido à baixa qualidade do sinal de fala oriundo da PSTN que, historicamente, por razões econômicas e para evitar interferências entre canais (*cross-talk*), adotou, como padrão para transmissão, sinais de banda estreita (*narrowband* - NB) [3]. Nesse cenário, os sinais são amostrados a 8000 Hz e contêm componentes de frequências limitados entre 300 e 3400 Hz. Tal limitação de largura de banda provoca perda de qualidade nos sinais de fala, tornando-os “abafados", “sem brilho" e com degradação de naturalidade e inteligibilidade [4], [5], [6].

Estudos realizados pela entidade de regulamentação de telecomunicações denominada União Internacional de Telecomunicações (*International Telecommunication Union* - ITU) com usuários de PSTN mostraram que a largura de banda do sinal de fala afeta significativamente sua qualidade perceptual bem como sua inteligibilidade. Dessa forma, sinais de fala de banda larga (*wideband* - WB) são preferidos pelos usuários por utilizarem uma largura de banda mais abrangente, com frequências entre 50 e 7000 Hz, apresentando maior riqueza espectral quando comparado aos sinais de NB [7].

Em [2], [8] e [9], a utilização de codificadores de WB vem sendo discutida e adotada em alguns sistemas de comunicações. Em um futuro próximo, é inevitável a extinção das PSTNs, como já vem ocorrendo por meio da adoção de sistemas de voz sobre IP (*voice over internet protocol* - VoIP) [3]; no entanto, a PSTN ainda é uma das redes mais difundidas em todo o mundo e sua modernização para WB demandaria um esforço enorme, o qual seria,

em curto prazo, economicamente inviável [3]. Portanto, durante os próximos anos, preveem-se apenas migrações graduais para terminais de WB e, por um certo período de transição, redes de telefonia mistas de NB e de WB irão coexistir [4], [6], [10]. Assim, para contornar as limitações da comunicação em NB, a extensão artificial de largura de banda pode ser vista como uma alternativa interessante. Nesse contexto, os componentes de frequência não transmitidos pelos codificadores de NB devem ser sintetizados artificialmente através da estimação de parâmetros do modelo fonte-filtro de produção do sinal da fala, no qual são consideradas duas etapas: estimação do sinal de excitação e estimação do envelope do trato vocal [1], [3], [4]. Essa técnica proporciona melhorias na qualidade do sinal de fala, tornando-o mais próximo de um sinal de WB, requerendo apenas alterações nos terminais receptores (*far-end*), mantendo o sistema compatível com a maior parte das redes de telefonia existentes [3].

Nos últimos anos, diversos sistemas usando diferentes estratégias vêm sendo propostos para a realização de ABWE [9]. Em [2], [8] e [11], a ABWE é obtida através da transmissão de informações extras ao *far-end* (*side information*). Entretanto, a implementação de tais estratégias resulta em aumento na taxa de bits do sinal, bem como em adaptações nos terminais transmissores (*near-end*) e receptores (*far-end*) da rede de comunicação. Para contornar tais problemas, [12] e [13] sugerem estratégias que não necessitam de *side information*. No entanto, tais estratégias levam a um alto índice de ocorrência de ruídos do tipo musical e impulsivo devido à falta de um tratamento adequado dos parâmetros discriminativos do trato vocal.

Apesar da crescente evolução dos sistemas de ABWE, ainda não se dispõe de qualquer procedimento consolidado apresentando desempenho satisfatório na estimação de WB, principalmente nos casos em que o sinal de fala contém energias concentradas em altas frequências, isto é, maiores do que 5000 Hz. Nesses casos, notam-se geralmente artefatos indesejados no sinal de fala reconstituído [6], [10], [14], [15], [16] e [17].

1.2 Objetivos e Contribuições do Trabalho

Este trabalho de pesquisa apresenta uma nova estratégia para implementar sistemas de ABWE para sinais de fala aplicados à telefonia em redes de pacotes. A estratégia para ABWE aqui desenvolvida é baseada nos procedimentos descritos em [9], sendo propostas alterações nas etapas de representação e estimação do trato vocal, para a qual é sugerida uma estratégia baseada em classificação fonética. Dessa forma, objetiva-se melhorar o desempenho desses sistemas, mantendo a compatibilidade com as redes existentes e com

os *codecs* de NB e de WB amplamente adotados no mercado de telefonia, tais como as Recomendações ITU-T G.729 [1] e G.729.1 [2], respectivamente.

A etapa de classificação fonética é responsável por um tratamento específico em diferentes classes de sinais de fala, buscando, dessa forma, uma melhor representação e, consequentemente, uma melhoria na clusterização dos parâmetros discriminativos do trato vocal, que resultem em uma síntese mais "limpa" e agradável dos sinais de WB. Nesse contexto, são investigadas técnicas de seleção de atributos aplicadas a sinais de fala de banda limitada, assim como o aprimoramento da classificação em grupos fonéticos abrangentes (*broad group phonetic* - BGP) com ênfase na discriminação de fonemas pertencentes ao grupo fonético fricativo. Os atributos característicos do sinal de fala são investigados através de métodos baseados em algoritmos de aprendizagem de máquinas (*machine learning*) e métricas de análise de componentes, tais como, informação mútua (*mutual information* - MI) e seleção de parâmetros baseada em correlação (*correlation-based feature selection* - CFS). Dessa forma, torna-se possível a análise acerca da seleção dos melhores atributos a serem considerados na composição dos classificadores fonéticos.

Tendo em vista a versatilidade do sistema de ABWE desenvolvido, também é investigada aqui a integração do sistema de ABWE em sistemas de reconhecimento automático de fala (*automatic speech recognition* - ASR) para o português brasileiro aplicados à PSTN. Nesse cenário, o estado da arte informa que sistemas de ASR que decodificam sinais de NB apresentam desempenho inferior aos sistemas que operam com sinais de WB. Objetivando o aprimoramento de ASR em PSTN, as etapas de extração de atributos do sinal da fala, bem como a etapa de construção do modelo acústico são desenvolvidas baseadas em sinais sintéticos de WB estimados a partir do realce de sinais de NB usando ABWE.

A eficácia das estratégias propostas neste trabalho é verificada através de avaliações subjetivas e objetivas, através de espectrogramas, resultados de medidas de distâncias, avaliações perceptuais da qualidade do sinal de fala [18], [19], [20] e através da análise de desempenho da estratégia de ABWE aplicada a sistemas de ASR.

Ressalta-se que parte das contribuições desta dissertação já foram publicadas no XXXI Simpósio Brasileiro de Telecomunicações - SBrT 2013, realizado na cidade de Fortaleza-CE, em setembro de 2013, na forma de um artigo técnico intitulado "Extensão artificial de largura de banda para sinais de fala usando classificação fonética".

1.3 Estrutura da Dissertação

Este trabalho está organizado como segue. No Capítulo 2, são discutidas importantes definições relativas ao processo de produção da fala, assim como são apresentados uma breve introdução à fonética e fonologia do português brasileiro e o principal método de aproximação do processo de produção da fala para sistemas de telecomunicações. No Capítulo 3, o conceito de extensão de largura de banda, suas aplicações e os principais métodos de extensão, são discutidos. No Capítulo 4, são investigados os grupos fonéticos e são apresentadas as classes fonéticas consideradas neste trabalho. As Seções 4.2, 4.3 e 4.4 apresentam os procedimentos de extração e seleção de atributos do sinal de fala e descrevem a etapa de desenvolvimento da estratégia proposta para o processo de classificação fonética. No Capítulo 5, uma estratégia de implementação do sistema de ABWE é apresentada e a descrição de cada estágio de processamento é discutida. No Capítulo 6, é apresentada a aplicação de ABWE em sistemas de ASR. Nesse cenário, a estratégia baseada em reconhecimento de fala em rede (*network speech recognition* - NSR) é proposta e o desenvolvimento dos sistemas de ABWE e de ASR são descritos. Finalmente, o Capítulo 7 apresenta as conclusões e comentários finais deste trabalho de pesquisa.

Capítulo 2

O Sinal de Fala

A fala é uma característica pertencente exclusivamente aos seres humanos. A linguagem falada é uma modalidade natural de comunicação e consequentemente a mais utilizada entre os seres humanos. A informação transmitida através da fala é intrinsecamente de natureza discreta, ou seja, pode ser representada pela concatenação de elementos a partir de um conjunto finito de símbolos [21], [22]. Esses símbolos representam as sequências de sons que compõem o sinal de fala. O estudo desses sons e suas articulações caracterizam a fonética e fonologia de um determinado idioma, fundamentando a linguagem que é a fonte de comunicação. Nesse contexto, o termo *fonema* é definido como a menor unidade acústica dos sinais de fala [23].

A representação do sinal de fala deve ser tal que o conteúdo da informação possa ser extraído facilmente por um ouvinte humano ou, de forma automática, por uma máquina. Contudo, a fala não transmite apenas informações léxicas. Devido à capacidade de o nosso cérebro interpretar informações complexas, podemos, de forma praticamente inconsciente, reconhecer a identidade do falante, sua posição no espaço físico, seu estado emocional e informações implícitas no tom de voz usado, como por exemplo, ironia, seriedade ou tristeza [24].

O sinal de fala ocupa toda a escala de frequência perceptível pelo sistema auditivo humano [3], cujos limites de audição possuem frequências entre aproximadamente 20 Hz e 20 kHz [21]. A Figura 1 ilustra um exemplo de um sinal de fala adquirido a uma taxa de amostragem de $44,1$ kHz, suas representações no domínio temporal e espectral, e a correspondente representação em símbolos fonéticos (fonemas).

2.1 Processo de Produção do Sinal de Fala

O sinal de fala experimenta diversas transformações durante seu processo de produção. Essas transformações podem ser associadas às singularidades acústicas de cada segmento do sinal de fala. As características de produção da fala correspondem a uma combinação de características físicas e características adquiridas, sendo que estas últimas correspondem aos diferentes hábitos e maneiras de falar adquiridos pelos locutores ao longo do tempo [24].

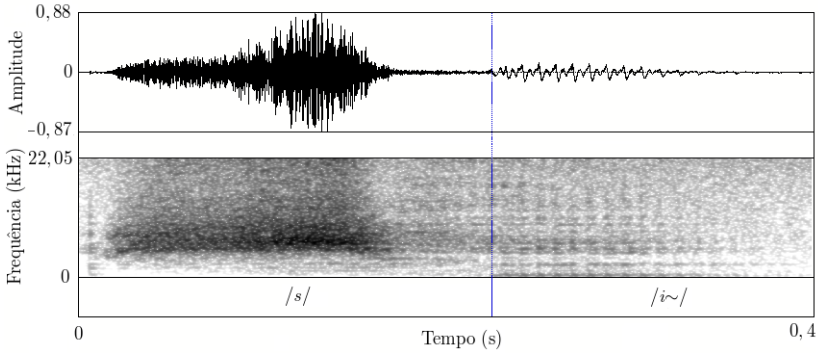
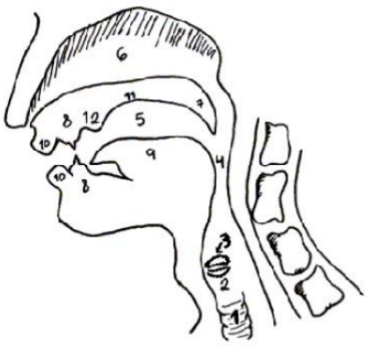


Figura 1 – Exemplo de características de um de sinal de fala e sua representação fonética.

Devido ao seu comportamento dinâmico, o sinal de fala é caracterizado por uma forma de onda apresentando uma certa complexidade. Todavia, o processamento de sinais de fala pode ser melhor compreendido através do conhecimento dos principais mecanismos envolvidos em seu processo de produção. Para tanto, esta seção apresenta uma descrição introdutória do aparelho fonador humano e de seu modelo de produção de fala considerado em diversas aplicações de telecomunicações.

2.1.1 Aparelho Fonador

O processo de produção da fala é um tanto complexo e envolve a participação de diversos órgãos. Como nenhum órgão realiza função específica à fonação, para a área de interesse deste trabalho de pesquisa, utiliza-se o termo *aparelho fonador* para descrever os órgãos, direta ou indiretamente, relacionados à produção de fala. Esses órgãos, envolvidos no aparelho fonador humano são, além dos músculos e nervos, os brônquios, a traquéia, a laringe (com as pregas vocais), a faringe, as cavidades nasais e a boca com a língua (dividida em ápice, dorso e raiz), o palato duro (ou céu da boca), o palato mole (ou véu palatino), os dentes com os alvéolos, e os lábios. Para evidenciar a localização das estruturas que constituem o aparelho fonador humano um diagrama simplificado é ilustrado na Figura 2.



1 – Traquéia
2 – Laringe
3 – Glote (pregas vocais)
4 – Faringe
5 – Cavidade bucal
6 – Cavidade nasal
7 – Véu palatino, palato mole
8 – Maxilares (dentes)
9 – Língua
10 – Lábios
11 – Palato duro
12 – Alvéolos dentais

Figura 2 – Aparelho Fonador.

Dentre as diferentes funções desempenhadas pelas estruturas que constituem o aparelho fonador humano, podemos destacar, no escopo da produção de sinais de fala, os seguintes elementos:

- **Pulmões**. Órgãos respiratórios que fornecem a corrente de ar, matéria-prima da fonação.
- **Brônquios**. Tubos que conectam os pulmões à traquéia.
- **Traquéia**. Faz a conexão da laringe com os brônquios. A cavidade, formada pelos brônquios e a traquéia, atua como um ressonador de baixa frequência.
- **Laringe**. Situa-se entre a traquéia e a língua, sendo constituída por uma série de cartilagens revestidas por uma membrana mucosa que é movimentada pelos músculos da laringe[1]. As dobras da membrana mucosa dão origem às pregas vocais.
- **Faringe**. Interliga as cavidades nasais e oral com a laringe. Suas cavidades[2] funcionam como uma caixa de ressonância que pode mudar de forma e tamanho, influindo na ressonância dos sons vocálicos.
- **Cavidades Nasais**. Cavidades paralelas entre as narinas e a faringe. Também assume papel de ressoador.

[1] Músculos cricoaritenóideos posterior e lateral.
[2] A faringe é formada pela naso, oro e laringofaringe.

- **Boca**. Graças, sobretudo, ao movimento da língua e do maxilar, a boca pode variar de forma e volume, alterando assim o fluxo de ar, resultando, consequentemente, na produção de diferentes sons. Na boca, encontra-se ainda a úvula que, mesmo não possuindo movimentação própria, pode ser vista como um importante articulador, pois se movimenta em conjunto com a língua. A úvula tem a função de impedir ou permitir a passagem de ar pelas cavidades nasais através, respectivamente, de seu levantamento ou abaixamento. Além disso, a úvula também pode vibrar. Por fim, os lábios constituem a terminação do aparelho fonador. O grau de projeção dos lábios para frente como também o grau de abertura da boca aumentam a caixa de ressonância, influindo no timbre dos sons produzidos.

No processo de produção de fala, o ar expelido dos pulmões por via dos brônquios, penetra na traquéia e chega à laringe, onde, ao atravessar a glote, encontra o primeiro obstáculo à sua passagem. Nesse ponto, o fluxo de ar pode encontrar a glote aberta ou fechada. Se estiver aberta, o ar força a passagem através das pregas vocais retesadas, fazendo-as vibrar e produzir o som musical característico dos segmentos vozeados (ou sonoros). Se estiver fechada, com as pregas vocais relaxadas, o ar escapa da laringe sem vibrações, formando assim os segmentos de fala denominados não-vozeados (ou surdos) [21], [24].

As cavidades supraglóticas atuam como ressoadores, impondo importantes alterações ao sinal de excitação. Dentre as características físicas, a forma do trato vocal pode ser considerada como um aspecto de distinção importante. O trato vocal é considerado um ressoador variante no tempo, isto é, sua forma muda com o tempo. O trato vocal corresponde aos órgãos de produção da fala situados acima das pregas vocais, envolvendo a faringe, a boca e a cavidade nasal [21].

As frequências de ressonância do trato vocal são conhecidas como formantes e resultam das diferentes configurações possíveis das cavidades ressoadoras em conjunto com os demais elementos do aparelho fonador. O valor da frequência dos quatro primeiros e principais formantes (F_1, F_2, F_3 e F_4) tem relação com as seguintes configurações:

- O deslocamento da língua no plano vertical basicamente define o valor do primeiro formante (F_1).

- A frequência do segundo formante (F_2) é determinada pelo deslocamento da língua no plano horizontal.

- O grau de obstrução formado entre língua e faringe define o valor do terceiro formante (F_3).

- A frequência do quarto formante (F_4) é função da posição vertical da laringe.

O valor da frequência dos formantes pode também sofrer pequenas alterações em função da posição dos lábios. Quando o palato mole se encontra abaixado, ocorre acoplamento do trato vocal com a cavidade nasal. Esse acoplamento resulta na interação das frequências de ressonâncias. Além disso, é possível ocorrer frequências de anti-ressonâncias ou anti-formantes. Os anti-formantes são frequências de ressonância próximas das frequências dos formantes que causam a redução de amplitude desses formantes devido à perda de energia causada pelo acoplamento entre o trato vocal e a cavidade nasal [21], [24].

O entendimento das possíveis combinações dos articuladores no processo de produção de fala, das limitações fisiológicas do aparelho fonador e da forma como a fonética e fonologia do português brasileiro se organizam são de grande valia no desenvolvimento de sistemas de processamento de sinais de fala. Em função disso, importantes considerações a respeito da produção do sinal de fala e suas correspondentes classes fonéticas serão discutidas no Capítulo 4.

2.1.2 Fonética e Fonologia do Português Brasileiro

Vislumbrando o entendimento dos mecanismos presentes no processo de produção dos sinais de fala e da relação entre os órgãos do sistema fonador, apresentado na Seção 2.1.1, as seções seguintes introduzem conceitos da fonética e fonologia do português brasileiro (doravante PB).

2.1.2.1 Vogais

As vogais se diferem dos demais segmentos de fala pelo fato de o sinal de excitação que as produz não experimentar qualquer obstrução.

O número de vogais do PB varia em função da posição da vogal na palavra. Quando em posição tônica, as vogais são definidas conforme Tabela 1.

2.1.2.2 Semivogais

Semivogais são as vogais assilábicas i̯ e u̯ em encontros vocálicos, formando ditongos decrescentes e ditongos crescentes. Os ditongos decrescentes são formados pela sequência de uma vogal e uma semivogal, podendo ser orais ou nasais. Ditongos crescentes são constituídos de uma semivogal seguida de uma vogal e são sempre orais [23].

Tabela 1 – Vogais em posição tônica

	Não-arredondadas			Arredondadas			
altas	/i/					/u/	
médias		/e/			/o/		(2^o grau)
médias			/ɛ/	/ɔ/			(1^o grau)
baixa			/a/				
	anterior		central	posterior			

2.1.2.3 Consoantes

Em contrapartida ao processo de produção das vogais, durante o processo de produção das consoantes, o sinal de excitação sofre obstrução total ou parcial em um ou mais pontos do sistema fonador. O resultado dessa obstrução é um ruído característico das consoantes, em contraste com os segmentos provenientes das vogais. Esse ruído pode ser caracterizado por um som aperiódico, tanto contínuo quanto plosivo, e apresenta uma energia acústica consideravelmente menor do que as das vogais [24].

Em [23] e [24], as consoantes são definidas como elementos que se combinam com as vogais para formarem as sílabas. De acordo com a posição que ocupam na sílaba, as consoantes podem ser, pré-vocálicas e pós-vocálicas. Nas consoantes pré-vocálicas, a articulação predominante se concentra na fase em que se desfaz a obstrução bucal ao sinal de excitação (corrente expiratória). Enquanto nas consoantes pós-vocálicas, as articulações se concentram na fase de fechamento e a abertura da boca se modifica (fechando) para criar o elemento consonântico de travamento da sílaba.

As consoantes do PB podem ser separadas em labiais, anteriores e posteriores, conforme Tabela 2 [24].

Tabela 2 – Consoantes do português brasileiro

Labiais	/p/	/b/	/f/	/v/	/m/		
Anteriores	/t/	/d/	/s/	/z/	/n/	/l/	/ɾ/
Posteriores	/k/	/g/	/ʃ/	/ʒ/	/ɲ/	/λ/	/ɾ/

As consoantes podem também ser classificadas em função da região e maneira como são articuladas, isto é, o ponto e o modo de articulação. Quando classificadas quanto ao modo de articulação, são divididas em oclu-

sivas e constritivas (fricativas, laterais ou vibrantes). Quando a classificação é realizada de acordo com o ponto de articulação, as consoantes podem ser classificadas como bilabiais, labiodentais, linguodentais, alveolares, pós-alveolares[3], palatais e velares. Ainda, considerando o papel das pregas vocais, as consoantes podem ser denominadas vozeadas ou não-vozeadas. Por fim, as consoantes podem ser orais ou nasais, dependendo do papel das cavidades bucal e nasal [24].

2.1.3 Modelo Para Produção do Sinal de Fala

Em resumo, como apresentado na Seção 2.1.1, os sinais de fala podem ser representados pelos seguintes estados:

- **Fala não-vozeada**. Ocorre quando não há vibração das pregas vocais. A forma de onda resultante é não-periódica.

- **Fala vozeada**. Ocorre quando há vibração das pregas vocais. A forma de onda resultante é periódica.

Ainda como descrito na Seção 2.1.1, o trato vocal pode ser visto como um filtro que modela o fluxo de ar provenientes dos pulmões. Esse processo permite a geração de diferentes sons. Os pulmões propiciam a excitação para o trato vocal e essa excitação pode ser periódica ou aperiódica, dependendo do estado das pregas vocais.

O comportamento dinâmico entre o trato vocal e a fonte de excitação vem fomentando pesquisas sobre o desenvolvimento de modelos matemáticos para processo de produção do sinal de fala. A Figura 3 ilustra um modelo de produção do sinal de fala no qual o trato vocal é representado por um sistema quase-linear excitado por uma fonte periódica ou aperiódica.

Nota-se da Figura 3 que o trato vocal é representado por um filtro $v(n)$ linear invariante no tempo e a sua saída é conectada em série com outro filtro $r(n)$ que modela o efeito da radiação do som nos lábios. Geralmente, um filtro passa-altas de primeira ordem é usado para implementar o efeito da radiação do som pelos lábios. Nota-se também que a fonte de excitação é definida como $g(n)$ para o caso em que a entrada é uma sequência periódica. Enquanto, para o caso em que a fonte é um ruído, como, por exemplo, durante a produção da consoante /s/, uma sequência aleatória $n(n)$ é usada, contendo geralmente espectro plano (ruído branco).

[3]Pós-alveolares (ou palatoalveolares) têm como articulador ativo a parte anterior da língua e como articulador passivo, a região entre o palato duro e o alvéolo dental

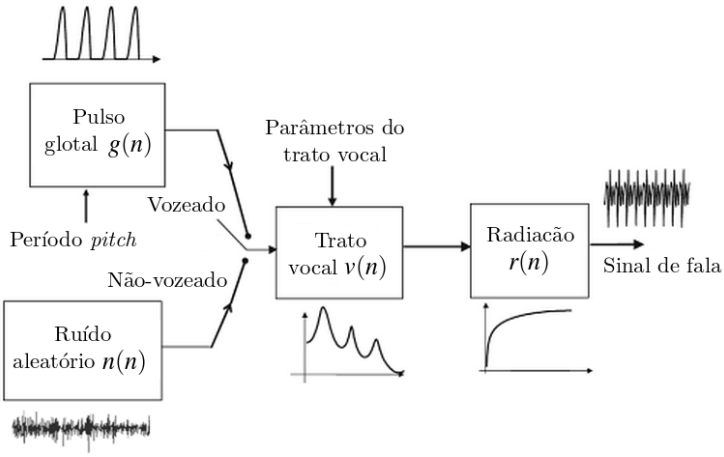


Figura 3 – Diagrama de blocos do modelo para produção do sinal de fala.

O modelo de produção da fala ilustrado na Figura 3, também conhecido como modelo fonte-filtro, é ideal para transmissão de fala em canais de banda estreita por necessitarem de apenas três conjuntos de parâmetros para a transmissão, a saber: a frequência fundamental, o nível de vozeamento e os parâmetros do trato vocal. Por essa razão, esse modelo vem sendo utilizado extensivamente em aplicações de codificação de fala com baixas taxas de bits em sistemas de telecomunicações [1], [4].

O modelo fonte-filtro é uma aproximação do processo de produção da fala, em que é assumido que a fonte e o filtro (caracterizando o trato vocal) são independentes e que a relação entre a pressão e velocidade do volume do fluxo de ar de excitação é linear. Na realidade, o acoplamento existente entre a fonte de excitação e o filtro do trato vocal é bem mais complexo e não linear, existindo contribuições aeroacústicas para a geração de sons do trato vocal que não são levadas em conta pelo modelo fonte-filtro. Todavia, na prática, a aproximação fornecida pelo modelo fonte-filtro para a produção da fala, apesar de suas limitações, satisfaz uma vasta gama de aplicações de processamento de fala utilizados atualmente em sistemas de telecomunicações [3].

2.2 Sinais de Fala de Banda Limitada

Sinais de fala de banda limitada são sinais que originalmente apresentavam componentes de frequência distribuídos em toda a escala de frequên-

cia perceptível pelo ouvido humano (20 a 20 kHz) mas que devido a perdas relacionadas ao seu processo de codificação e/ou transmissão, passaram a apresentar restrições em suas correspondentes larguras de bandas, como, por exemplo, as restrições impostas aos sinais de NB quando transmitidos via PSTN.

Dentre as diversas informações contidas no sinal de fala, o conteúdo da mensagem é o de primordial importância e é a informação mínima necessária para o exercício da comunicação. Com base nesse conceito, para proporcionar uma comunicação básica, a ITU inicialmente adotou a PSTN como base para sistemas de comunicação. Nesse contexto, para que a inteligibilidade do sinal de fala seja preservada, é necessário apenas cerca de 3100 Hz de largura de banda com banda passante de 300 a 3400 Hz [3].

A degradação da qualidade do sinal de fala na PSTN é causada pela introdução de filtros limitadores de banda associados aos amplificadores usados para manter um nível adequado de sinal em chamadas de longas distâncias. Esses filtros apresentam uma banda passante de aproximadamente 300 a 3400 Hz e reduzem a interferência (*cross-talk*) entre canais. A aplicação desses filtros (passa-banda) atenuam consideravelmente uma porção do sinal de fala e afetam diretamente a qualidade subjetiva do sinal. A Figura 4 apresenta o gabarito de ganho do filtro passa-banda padronizado pela ITU em sua Recomendação ITU-T 88b [3].

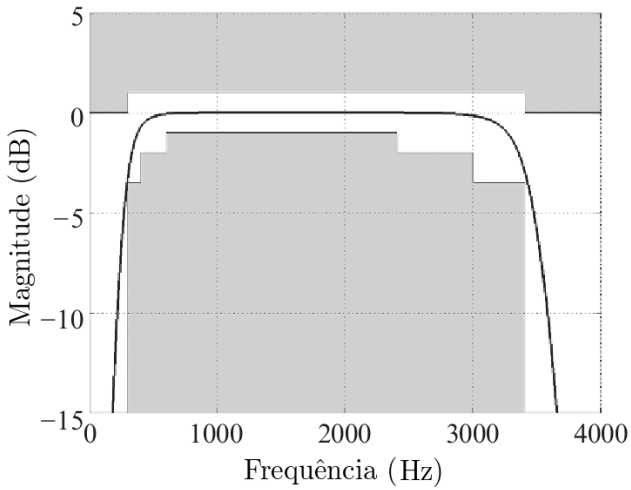


Figura 4 – Filtro passa-banda utilizado na PSTN.

Redes digitais, tais como redes VoIP e a rede digital de serviços integrados (*integrated service digital network* - ISDN), são capazes de transmitir sinais de fala de alta qualidade. Nesses cenários, os componentes de frequência abaixo de 300 Hz assim como os componentes maiores do que 3400 Hz podem ser transmitidos sem atenuações, limitando-se apenas pela frequência de *Nyquist* do sinal [22]. As Figuras 5 e 6 ilustram as características acústicas dos sinais de fala correspondentes à elocução “sim”, sendo transmitidos por um canal operando com frequência de *Nyquist* igual a 8000 Hz (Figura 5) e pela PSTN (Figura 6).

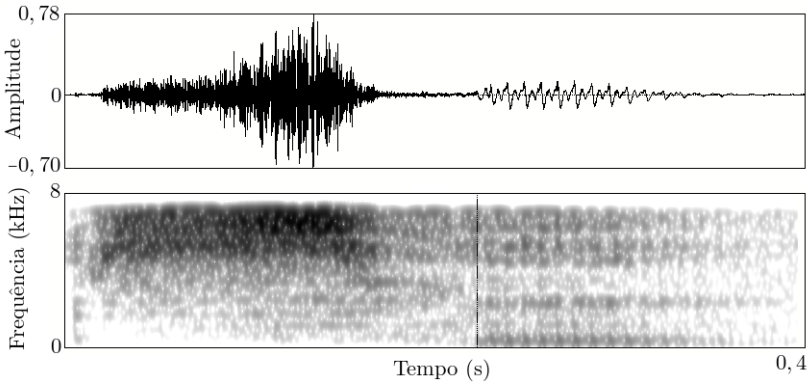


Figura 5 – Exemplo de um sinal de fala amostrado limitado apenas pela frequência de *Nyquist* de 8000 Hz.

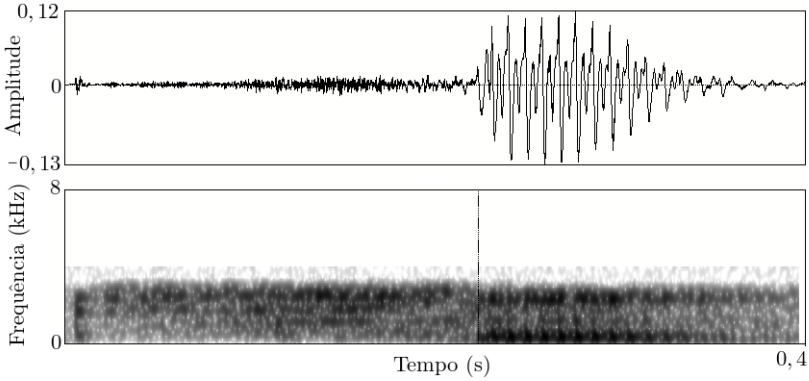


Figura 6 – Exemplo de um sinal de fala transmitido pela PSTN.

Nota-se das Figuras 5 e 6 que o sinal original (veja Figura 1) experimenta diferentes transformações ao ser transmitido pelas redes supramencionadas. No canal correspondente da Figura 5, a limitação de banda é dada apenas pela frequência de *Nyquist*. Enquanto na Figura 6, a limitação de banda é oriunda das restrições da PSTN. Ainda com respeito às Figuras 5 e 6, conforme ilustrado nos segmentos de fala referentes ao fonema /s/, os detalhes de altas frequências, distribuídos entre 4000 e 8000 Hz na Figura 5, são fortemente atenuados no sinal de fala proveniente da PSTN e representado pela Figura 6.

Nos últimos anos, diversas contribuições vêm sendo elencadas com o objetivo de aprimorar a qualidade do sinal de fala em redes de telefonia de pacotes. Atualmente, codificadores de banda larga são capazes de aumentar a largura de banda de sinais para 7000 Hz ou até mesmo para larguras maiores, utilizando apenas uma complexidade moderada [2], [3], [11]. Outras aproximações tentam aumentar a largura de banda através da transmissão de *side information* contendo atributos pertencentes às altas frequências, do sinal de fala, originalmente não transmitidas via PSTN [2], [8]. Todavia, a aplicação dessas aproximações demandaria a alteração da rede existente ou pelo menos, no segundo caso, a utilização de dispositivos que fossem capazes de codificar *side information* utilizando, por exemplo, marcas d'águas acústicas no usuário de origem, sendo capaz de decodificar tais informações no usuário de destino. Outra possibilidade é aumentar artificialmente a banda do sinal na recepção através de uma extensão da largura de banda do correspondente sinal.

2.3 Conclusões

Neste capítulo, importantes características dos sinais de fala foram investigadas. Posteriormente, foram introduzidas definições relativas à fonética e fonologia do português brasileiro, assim como os aspectos fisiológicos do processo de produção da fala. Nesse contexto, o principal método de aproximação do processo de produção da fala para sistemas de telecomunicações foi apresentado. Além disso, foram discutidos os efeitos da limitação de banda em sinais de fala.

Capítulo 3

Extensão Artificial de Largura de Banda

Na literatura, a grande maioria dos algoritmos que implementam sistemas de ABWE são baseados no modelo fonte-filtro de produção da fala [9], [10], [5]. Nesse modelo, um sinal de fala é gerado através de um filtro (caracterizando o trato vocal), o qual é excitado por um sinal (com características específicas) gerado por uma dada fonte. Especificamente, o termo "fonte" refere-se à fonte de excitação, ou sinal de excitação, e o termo "filtro" refere-se ao filtro de síntese, ou envelope espectral do trato vocal. O modelo fonte-filtro é aplicado em diversas áreas de processamento de fala, tais como análise, síntese, codificação, dentre outras [19]. Por exemplo, em sistemas de codificação, o modelo fonte-filtro é usado visando reduzir a quantidade de parâmetros utilizados para a representação do sinal de fala, reduzindo assim, a taxa de bits durante a transmissão do sinal [19], [21]. Então, ao invés da codificação de todas as amostras de um quadro do sinal de fala, faz-se necessária a codificação de apenas um conjunto reduzido de parâmetros da fonte e do filtro. Neste capítulo, são discutidos os modelos para a representação do sinal de excitação (fonte) e do envelope de trato vocal (filtro) do modelo fonte-filtro de produção da fala.

3.1 Representação do Modelo Fonte-Filtro

Para a estimação dos parâmetros do modelo fonte-filtro linear de produção do sinal de fala, assume-se aqui a técnica de análise preditiva linear [21] em que a amostra corrente de um sinal de fala $s(n)$ pode ser obtida aproximadamente através da combinação linear de suas p amostras passadas. Assim,

$$s(n) \cong a_1(n)s(n-1) + a_2(n)s(n-2) + \ldots + a_p(n)s(n-p). \tag{3.1}$$

Os coeficientes a_i correspondem aos coeficientes de um filtro recursivo de ordem p. Definindo $\sigma(n)e(n)$ como um sinal de erro de aproximação proveniente da predição linear, (3.1) pode ser expressa como

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + \sigma(n)e(n), \tag{3.2}$$

onde $e(n)$ denota um sinal de excitação e $\sigma(n)$ um fator de ganho. Aplicando a transformada Z e considerando a_i e $\sigma(n) = \sigma$ constantes dentro de um quadro do sinal de fala, o sinal $s(n)$ no domínio da transformada Z pode ser expresso como

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + \sigma E(z). \tag{3.3}$$

Assim, realizando algumas manipulações algébricas, obtém-se

$$\begin{aligned} \frac{S(z)}{\sigma E(z)} &= \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \\ &= H(z) \\ &= \frac{1}{A(z)} \end{aligned} \tag{3.4}$$

onde a função de transferência $H(z)$ é denominada "filtro de síntese" e sua função inversa, isto é, $[H(z)]^{-1} = A(z)$, é denominada "filtro de análise". A resposta em frequência de $H(z)$ é considerada o envelope espectral do trato vocal de um quadro do sinal $s(n)$, sendo representada pelos coeficientes a_i da codificação preditiva linear (*linear predictive coding* - LPC), enquanto

$$e_\sigma(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{3.5}$$

representa o sinal de excitação do correspondente quadro. Particularmente, os coeficientes LPC a_i são usualmente computados a partir dos algoritmos recursivos de *Dubin-Levison* [18], [21].

A Figura 7(a) apresenta o diagrama de blocos do processo de produção do sinal de fala considerando o modelo linear fonte-filtro, onde $e_\sigma(n) = \sigma(n)e(n)$ e $H(z)$ representa o filtro de síntese composto pelos coeficientes a_i. A Figura 7(b) apresenta o diagrama de blocos para a obtenção do sinal de erro $e_\sigma(n)$, onde $A(z)$ representa o filtro de análise.

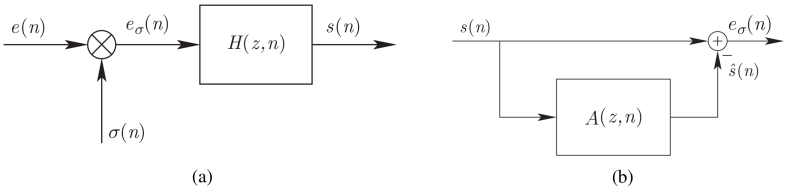


Figura 7 – Diagrama de blocos do modelo linear fonte-filtro do processo de produção da fala, (a) estrutura de síntese e (b) estrutura de análise.

Assim, cada quadro do sinal de fala $s(n)$ é caracterizado pelo sinal de excitação $e_{\sigma(n)}$ e pelos coeficientes LPC a_i. A Figura 8 ilustra o espectro de amplitude de um quadro do sinal de fala e o seu correspondente envelope espectral calculado para $H(z)|_{z=e^{j\omega}}$, considerando coeficientes LPC de ordem $p = 12$.

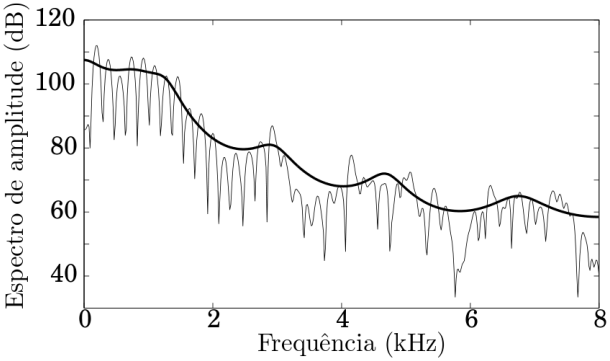


Figura 8 – Representação típica do espectro de amplitude de um quadro do sinal de fala (linha sólida cinza) e do envelope espectral (linha sólida escura) estimado através dos coeficientes LPC de ordem $p = 12$.

A Figura 9 ilustra a estimação do sinal de fala $s(n)$ a partir dos espectros do sinal de excitação $e_\sigma(n)$ e do envelope de trato vocal caracterizado pelos coeficientes LPC a_i.

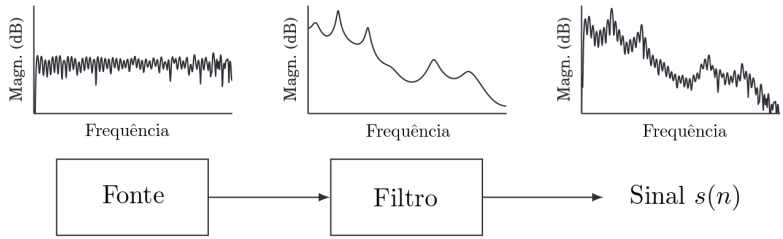


Figura 9 – Estimação do sinal de fala a partir de $E_\sigma(z)$ e $H(z)$.

A partir dos coeficientes LPC, outras representações podem ser obtidas, como os coeficientes denominados linhas espectrais de frequências (*line spectral frequency* - LSF). Esses coeficientes são muito conhecidos e usados

em codificação de fala e em aplicações de extensão de largura de banda [3]. Comparados aos LPC, os coeficientes LSF apresentam maior robustez na quantização dos coeficientes e, consequentemente, na estabilidade dos filtros. Esses coeficientes são computados usando os polinômios $P(z)$ e $Q(z)$ (para detalhes veja [21], [25]), dados a seguir:

$$\begin{aligned} P(z) &= A(z) + z^{p+1}A(z^{-1}) \\ Q(z) &= A(z) - z^{p+1}A(z^{-1}) \end{aligned} \tag{3.6}$$

Dentre as diversas técnicas propostas para a representação do envelope espectral [3], [5], [25], os coeficientes de LPC e LSF são os mais utilizados, sendo, por isso, considerados neste trabalho de pesquisa.

3.2 Estratégias para a Realização de Sistemas de ABWE

A ideia básica de um sistema de ABWE é recriar artificialmente componentes de alta frequência, convertendo um sinal de NB em um sinal de WB, o qual possui maior definição e qualidade. Essa técnica assume que ambos sinais (isto é, de NB e de WB) sejam gerados pelo mesmo modelo fonte-filtro de produção da fala [4].

Assumindo que o sinal de fala seja quase-estacionário dentro de um quadro de curta duração (aproximadamente 20 ms) [18], parâmetros do modelo fonte-filtro relacionados ao sinal de WB, tais como o sinal de excitação e o envelope do trato vocal, são estimados através de informações implícitas contidas no sinal de NB [3].

Considerando o espaço de dados dos parâmetros de banda alta (*up-band* - UP) denotado como S_{UP} e de NB denotado como S_{NB}, a função de mapeamento

$$\begin{aligned} f : S_{\mathrm{NB}} &\rightarrow S_{\mathrm{UP}} \\ x &\mapsto y = f(x) \end{aligned} \tag{3.7}$$

representa a solução ideal para um sistema de ABWE. Nesse contexto, uma função f, que determine os parâmetros de UP, $y \in S_{\mathrm{UP}}$, a partir de parâmetros de NB, $x \in S_{\mathrm{NB}}$, deve ser estimada. Dessa forma, com a combinação de ambos os espaços S_{UP} e S_{NB} é possível uma estimação dos parâmetros de WB $S_{\mathrm{WB}} \supset (S_{\mathrm{NB}} \cup S_{\mathrm{UP}})$.

Na literatura, diversos sistemas usando diferentes estratégias vêm sendo propostos para a implementação do sistema de ABWE [2], [4], [11], [12]. Neste trabalho, é discutida uma estratégia de extensão de largura de banda baseada no diagrama de blocos ilustrado na Figura 10.

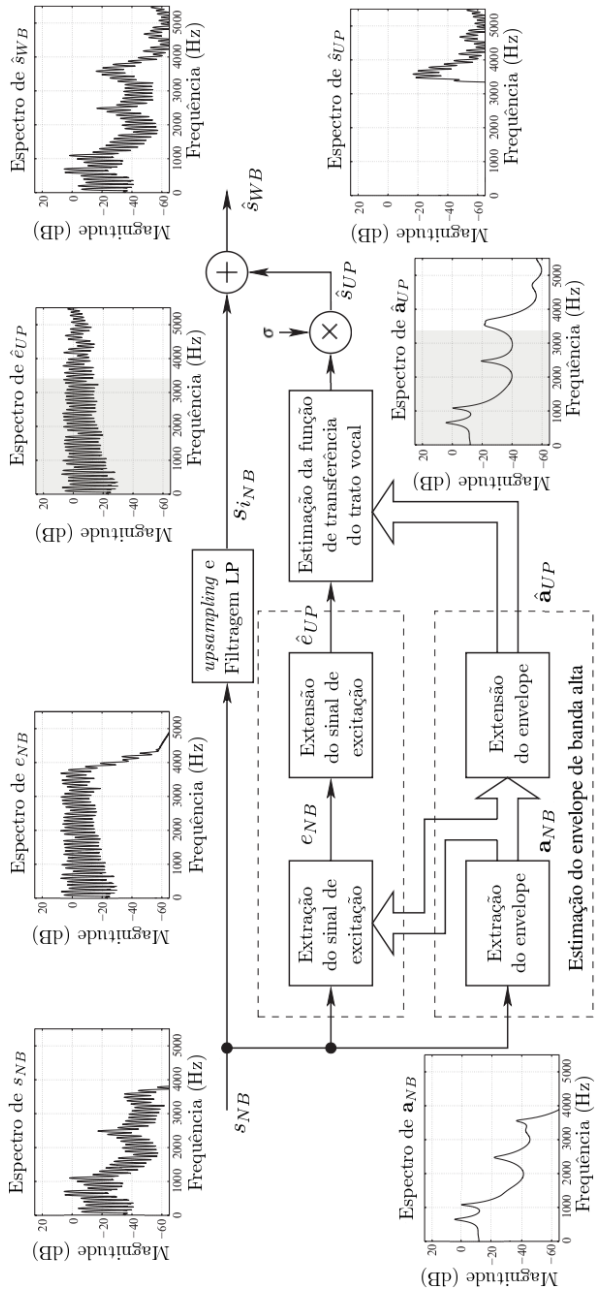


Figura 10 – Diagrama de blocos para a implementação de ABWE.

A estratégia de extensão de largura de banda ilustrada na Figura 10 considera o processo de conversão de um sinal de fala de NB s_{NB} em um sinal estimado de WB $\hat{s}_{\mathrm{WB}}$. O sinal s_{NB} é a entrada do sistema de ABWE e, a partir da análise LPC, os parâmetros originais de NB, $\mathbf{a}_{\mathrm{NB}}$ e e_{NB}, são obtidos. Em seguida, o sinal de excitação de NB e_{NB} é usado para a estimação do sinal de excitação de UP $\hat{e}_{\mathrm{UP}}$. O sinal, espectralmente plano, de excitação estimado $\hat{e}_{\mathrm{UP}}$ é convolvido com o envelope espectral estimado de UP representado pelos coeficientes $\hat{\mathbf{a}}_{\mathrm{UP}}$. Esses coeficientes são estimados a partir do sinal de NB s_{NB} com seus correspondentes coeficientes $\mathbf{a}_{\mathrm{NB}}$. Após a convolução dos sinais estimados de excitação $\hat{e}_{\mathrm{UP}}$ com o envelope espectral $\hat{\mathbf{a}}_{\mathrm{UP}}$, obtém-se o sinal de fala estimado de UP $\hat{s}_{\mathrm{UP}}$. A fim de ajustar as potências de NB e de UP, um fator de correção de ganho σ é aplicado após o filtro de síntese. O sinal resultante de UP $\hat{s}_{\mathrm{UP}}$ é adicionado ao sinal $s_{i_{\mathrm{NB}}}$ interpolado (*upsampling*) e processado por um filtro passa-baixas (*low pass* - LP). Dessa forma o sinal estimado de WB $\hat{s}_{\mathrm{WB}}$ é finalmente obtido.

3.2.1 Extensão do Sinal de Excitação

Conforme discutido na Seção 2.1.3 e ilustrado na Figura 9, o sinal de excitação contém os detalhes da estrutura espectral do sinal de fala $s(n)$ e geralmente apresenta um envelope espectral plano. Os sinais de excitação (utilizados para geração do sinal de fala) são comumente representados pelos sinais de erro $e(n)$ da técnica LPC. Usualmente, a estimação do sinal de excitação de UP $\hat{e}_{\mathrm{UP}}$ é efetuada a partir do seu correspondente sinal de excitação de NB e_{NB}. Na literatura, diversas técnicas são propostas para a extensão do sinal de excitação (isto é, obter $\hat{e}_{\mathrm{UP}}$ através de e_{NB}), como, por exemplo, as técnicas de extensão através da geração de ruídos e/ou senoides [5] e técnicas que usam funções não lineares como funções quadráticas $e^2_{\mathrm{NB}}(n)$, cúbicas $e^3_{\mathrm{NB}}(n)$ e retificação de onda completa $|e_{\mathrm{NB}}(n)|$ [3]. No entanto, a utilização de funções não lineares gera componentes distorcidos harmonicamente sem qualquer associação com a frequência fundamental ou com o período de *pitch* do sinal, resultando em sinais estimados de excitação de UP $\hat{e}_{\mathrm{UP}}$ com espectro variável nas altas frequências, sendo necessários ajustes para torná-lo espectralmente plano [3].

A utilização explícita das informações contidas no sinal de NB vem mostrando-se eficaz para extensão do sinal de excitação. Dessa forma, a fim de manter a estrutura harmônica original do sinal de excitação e_{NB}, as técnicas de translação espectral, escalamento de *pitch* e espelhamento espectral (*spectral folding* - SF) [3], [4], [5] são preferencialmente usadas por apresentarem melhores desempenhos em termos de qualidade de áudio [10]. Neste

trabalho de pesquisa, a técnica de espelhamento espectral é a considerada.

3.2.1.1 Espelhamento Espectral

A técnica de SF é uma maneira eficiente de estender a largura de banda de um sinal através de uma cópia espelhada do espectro de NB na escala de UP [3], [4], [5]. Esse procedimento pode ser implementado facilmente no domínio do tempo através da inserção de amostras nulas entre as amostras originais do sinal[4] ou no domínio da frequência via espelhamento direto dos coeficientes da transformada de *Fourier* [21]. A Figura 11 ilustra os efeitos (no domínio da frequência) do espelhamento espectral de um sinal de excitação de NB e_{NB}. O resultado é uma versão estendida de e_{NB} contendo, nas altas frequências (de 4000 a 8000 Hz), a imagem espelhada do espectro de NB, mantendo dessa forma a estrutura harmônica, bem como a estrutura espectral plana do sinal e_{NB}.

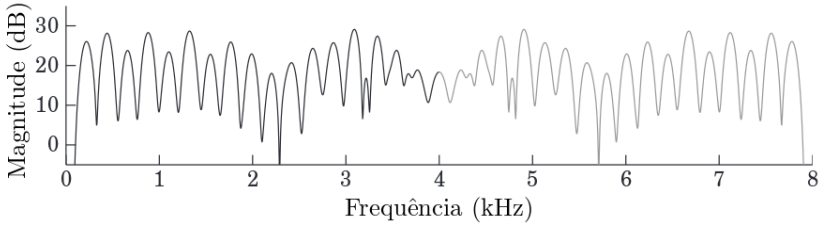


Figura 11 – Exemplo do espectro do sinal de excitação estendido usando SF.

3.2.2 Extensão do Envelope Espectral

Conforme ilustrado na Figura 10, a etapa responsável pela estimação dos coeficientes de UP $\hat{\mathbf{a}}_{\mathrm{UP}}$ refere-se ao bloco de extensão do envelope, mostrado em detalhes na Figura 12. Nota-se da Figura 12 que o envelope de UP é estimado com base em um vetor de parâmetros $\mathbf{x}$ (extraído a cada quadro de entrada do sinal de fala) e também levando em consideração conhecimentos *a priori* da relação entre as propriedades acústicas do sinal e os parâmetros do vetor $\mathbf{x}$. Dessa forma, uma vez estimado o envelope de UP, os seus coeficientes $\hat{a}_{\mathrm{UP}}$ podem ser diretamente obtidos.

[4] Procedimento equivalente ao processo de *upsampling* sem a utilização do filtro anti-recobrimento, fazendo uso do espectro espelhado para a composição da UP estimada.

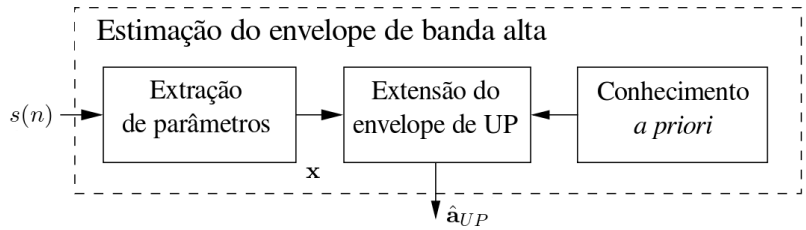


Figura 12 – Diagrama de blocos para extensão do envelope do trato vocal.

Portanto, para que a extensão do envelope do trato vocal seja realizada, é necessário um conhecimento *a priori* sobre a relação espectral entre as frequências de NB e as de UP. A escolha da função para modelar a relação entre NB e UP requer a determinação de um conjunto de parâmetros de NB **x** que seja capaz de discriminar satisfatoriamente as diferentes formas do trato vocal. A partir desses parâmetros é realizada a estimação de um outro conjunto de parâmetros **y** representante do envelope de UP. A qualidade da estimação da UP é fortemente dependente da função de mapeamento implementada e do conjunto de parâmetros discriminativos usados na etapa de treinamento [3], [4], [5].

Na literatura, diversas funções de mapeamento são propostas para a representação da relação entre NB e UP [10], [26], [27]. Em [2] e [11], informações extras (*side information*) são utilizadas para a extensão do envelope, aumentando, no entanto, a taxa de bits para a transmissão dos sinais, o que implica adaptações dos terminais transmissores e receptores da rede de comunicação. A fim de manter a compatibilidade com a rede de comunicação existente, outras funções de mapeamento que não necessitam de *side information* são discutidas em [12], [13], [5], [27].

Em [25] e [28], a extensão do envelope espectral é obtida através do mapeamento linear, em que a matriz $\mathbf{M} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ representa a solução ideal obtida usando a minimização do erro quadrático médio (*minimum mean squared error* - MMSE) entre os parâmetros de NB representados pelos vetores coluna da matriz **X** e os parâmetros de UP representados pelos vetores coluna de **Y**. Nessa técnica, devido à inversão $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, o cálculo da matriz **M** é conveniente apenas para a utilização de um número reduzido de parâmetros de NB **X**. Entretanto, um número reduzido de parâmetros não é capaz de representar satisfatoriamente o comportamento dinâmico do trato vocal, contaminando o sinal de WB estimado com diversos artefatos (*clicks*) que degradam a qualidade do sinal resultante [3], [4].

Em [6] e [26], redes neurais são usadas como funções de mapeamento

entre os espaços de dados S_{NB} e S_{UP}. Tal abordagem, por utilizar mapeamento não linear entre os espaços de dados, não dispõe de uma fácil interpretação das relações espectrais entre a NB e a UP. Em [3], [5], [10] e [13], a extensão do envelope do trato vocal é realizada através de modelagens estatísticas, resultando em relações espectrais entre a NB e a UP que podem, agora, ser melhor investigadas. Em [10] e [13], modelos com misturas de gaussianas (*gaussian mixture model* - GMM) são considerados, enquanto em [3] e [5], são usados modelos ocultos de *Markov* (*hidden Markov model* - HMM). Ambas as técnicas GMM e HMM apresentam desempenhos satisfatórios através de estimativas de histogramas e funções densidades de probabilidade (*probability density function* - PDF) dos espaços de dados S_{NB} e S_{UP}. Particularmente, a HMM apresenta a vantagem de considerar o comportamento dinâmico do sinal de fala através da observação de quadros passados. Entretanto, assim como para a modelagem usando redes neurais, as GMM e HMM necessitam de um grande conjunto de dados de treinamento para a obtenção de desempenho satisfatório [3].

Em [12], [26], [27] e [29], a utilização de mapeamento por *codebooks* é apresentada como uma técnica interessante e eficaz para a modelagem do conhecimento *a priori* das relações entre os envelopes de NB e de UP. Tais abordagens apresentam melhores resultados quando comparadas com as redes neurais [26], proporcionando fácil investigação dos espaços de dados (S_{NB} e S_{UP}) e das propriedades acústicas dos sinais de fala, sem a necessidade de um grande conjunto de dados para treinamento. Assim, neste trabalho de pesquisa, adota-se, como função de estimação do espaço de dados UP, as funções de mapeamento por *codebook*.

3.2.2.1 Mapeamento através de *Codebook*

Em resumo, a função de mapeamento por *codebook* consiste, primeiramente, no procedimento de extração de parâmetros e, posteriormente, no agrupamento (clusterização) desses parâmetros (segundo algum critério de similaridade). Após a etapa de agrupamento, um único conjunto de parâmetros é adotado para representar cada grupo. Nesse contexto, o termo "centroide" é utilizado para associar o centro de gravidade dos parâmetros de um grupo, sendo que cada grupo é representado por sua correspondente centroide, comumente denominada *codeword*. Especificamente, o agrupamento dos parâmetros é realizado através da análise de similaridades entre os quadros dos sinais de fala disponíveis para o treinamento. Na literatura, essa análise é efetuada através de funções-objetivo, tais como distâncias de *Itakura-Saito*, *Mahalanobis*, euclidiana, dentre outras [19], [21]. Neste trabalho, a distân-

cia euclidiana é a considerada. A Figura 13 ilustra um *codebook* durante a etapa de agrupamento de parâmetros ($\{x_1, x_2\}$) e definição das centroides (•) e *codewords* ($\mathbf{y}_i$).

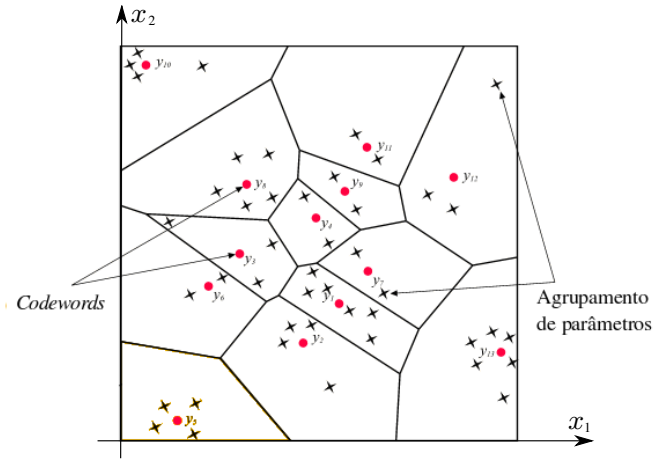


Figura 13 – Ilustração de um *codebook* durante a etapa de agrupamento de parâmetros ($\{x_1, x_2\}$) acústicos e determinação de *codewords*.

Para a estimação do envelope de UP, em [25] e [30], foram propostos *codebooks* distintos para a representação dos quadros do sinal de fala referente aos segmentos vozeados e não-vozeados. Em [12] e [25], além da distinção de *codebooks*, um mapeamento usando combinação linear foi proposto. Nesse contexto, o mapeamento dos espaços de dados S_{NB} e S_{UP} é realizado através de um *codebook* dual composto por um *codebook* de NB $\mathbb{C}_{\text{NB}}$ e um de UP $\mathbb{C}_{\text{UP}}$ treinados simultaneamente. Dessa forma, a cada quadro do sinal de fala $s(n)$, parâmetros do vetor $\mathbf{x}$ são extraídos e as medidas de distância de $\mathbf{x}$ para cada *codewords* de $\mathbb{C}_{\text{NB}}$ são computadas. A partir dessas distâncias, são atribuídos fatores de pesos associados a cada *codeword* de $\mathbb{C}_{\text{UP}}$ e, assim, o espaço de dados de UP S_{UP} pode ser determinado através da combinação linear das k *codewords* de $\mathbb{C}_{\text{UP}}$ com seus correspondentes fatores de pesos w. Em [12], foi proposto o mapeamento de *codebooks* com memórias, implementado através da interpolação do envelope estimado do quadro atual com o envelope estimado do quadro anterior. As abordagens utilizando *codebooks* com combinação linear e com memória resultam em melhorias na qualidade dos sinais estimados de UP. A Figura 14 ilustra o processo de obtenção do envelope espectral de UP a partir de *codebook* dual usando combinação linear.

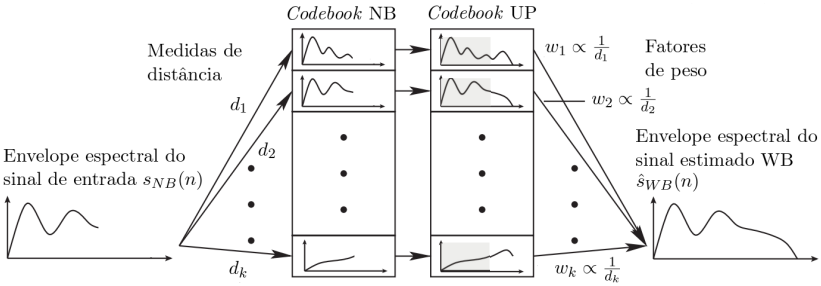


Figura 14 – Ilustração do processo de mapeamento dos *codebooks* de envelopes espectrais de NB e de WB.

3.3 Conclusões

Neste capítulo, foi apresentado um resumo dos principais algoritmos disponíveis na literatura para realização de ABWE usando apenas informações pertencentes ao sinal de NB. Na etapa de extensão do envelope de trato vocal, observa-se a necessidade de conhecimentos *a priori* das relações dos espaços de dados S_{NB} e S_{UP}, bem como das singularidades do sinal de fala em cada quadro de curta duração. Portanto, a análise do sinal de fala e sua classificação em grupos discriminativos é de grande importância para o aprimoramento da qualidade dos sinais de WB estimados.

Capítulo 4

Mineração de Dados e Classificação Fonética de Sinais de Fala de Banda Limitada

A convergência da computação e comunicação vem produzindo uma sociedade que consome, no seu dia-a-dia, cada vez mais informações. No entanto, a maioria das informações está em sua forma bruta de dados. Os dados podem ser caracterizados como fatos registrados e a informação pode ser considerada como o conjunto de padrões, ou expectativas, que fundamentam os dados. A tecnologia atual nos permite capturar e/ou armazenar uma vasta quantidade de dados. Todavia, há uma enorme quantidade de informações que são potencialmente importantes, mas não possuem bancos de dados adequados para a sua representação [31].

Encontrar atributos, tendências e anomalias nesses bancos é um dos grandes desafios da tecnologia da informação, isto é, transformar dados em informação e transformar informação em conhecimento. Nesse sentido, vêm ocorrendo contribuições importantes nas técnicas de análise de dados, tais como mineração de dados e aprendizagem de máquinas, tornando-as cada vez mais robustas e com base matemática consistente [32], [33], [34]. Dentre essas técnicas, a mineração de dados objetiva a extração de informações implícitas, previamente desconhecidas e potencialmente úteis. Isso é obtido através do desenvolvimento de algoritmos computacionais capazes de filtrar bancos de dados automaticamente em busca de regularidades ou padrões, extraindo informações para serem expressas em uma forma compreensível e que possam ser úteis para diversos propósitos, como por exemplo, a classificação de padrões [35].

A atividade de classificação de padrões é um processo inerente ao ser humano. Para distinguir diferentes objetos, usualmente, o ser humano armazena parâmetros marcantes, aqui chamados atributos característicos, de um determinado conjunto [31]. Assim, os métodos de seleção buscam identificar e reter apenas os atributos que mais contribuem para a discriminação em um determinado conjunto [31], [36].

Na literatura, diversas estratégias de seleção de atributos têm sido propostas. Em [37], a seleção de atributos provenientes da predição linear perceptual (*perceptual linear predictive* - PLP) e coeficientes cepstrais em escala mel (*mel-frequency cepstrum coefficients* - MFCC) tem como base o algoritmo *AdaBoost* [35]. Tais atributos são aplicado a sistemas de reco-

nhecimento automático de fala baseados em classificadores de máquina de vetores de suporte (*support vector machine* - SVM). Em [38], classificadores SVM são utilizados para a classificação de subconjuntos fonéticos, recebendo como entrada os atributos de *spectral peak locations*, *spectral rolloff*, *spectral centroid*, dentre outros. Em [39], classificadores baseados em *k-nearest neighbor* (*k*-NN) são utilizados em uma classificação fonética hierárquica, operando com atributos nos domínios do tempo e da frequência, tais como, a distância DTW (*Dynamic Time Warping*) e coeficientes MFCC, respectivamente.

Nos supramencionados trabalhos, os sinais considerados como referência são sinais de fala de banda larga, os quais possuem frequência de amostragem igual ou superior à 16000 Hz e contêm componentes de frequências superiores à 3400 Hz. Por essa razão, tais estratégias apresentam desempenhos severamente degradados quando operam com sinais de banda limitada. Em contraste, os algoritmos discutidos em [40], [41] e [42] utilizam, como referência, sinais de fala de banda limitada oriundos da PSTN. Nesse contexto, a busca pelo aprimoramento dos algoritmos de classificação fonética que apresentem desempenhos cada vez mais satisfatórios para sinais de banda limitada, utilizando uma quantidade reduzida de atributos, continua sendo destaque como um tópico ativo de pesquisa na área.

Para contornar as dificuldades de classificação em sinais de fala de banda limitada, em [43], [44] e [45], são apresentadas análises de atributos do sinal de fala para diferentes larguras de banda. Nesse contexto, atributos e grupos fonéticos são investigados objetivando o aprimoramento da classificação. Em [43] e [44], o uso de diferentes atributos durante o processo de classificação é proposto. Já em [45], assim como em [38], são propostas classificações com ênfase no grupos fonéticos fricativos, os quais são caracterizados por conterem energia espectral concentrada em altas frequências.

A investigação de grupos fonéticos também é discutida em [46], em que um classificador fonético, construído a partir de redes neurais de múltiplas camadas (*multilayer perceptron* - MLP), é proposto com base na especialização de classificadores em grupos fonéticos abrangentes (*broad group phonetic* - BGP) e na seleção de atributos de tempo e frequência. Tais classificadores enfatizam a capacidade de discriminação de quatro diferentes BGP (vogais, nasais, fricativas e oclusivas) para, posteriormente, discriminar os fonemas pertencentes a cada BGP estimado. Dessa maneira, são obtidas melhores taxas de desempenho quando comparado com a classificação isolada de fonemas. Para a redução de dimensionalidade, a medida de informação mútua (*mutual information - MI*) é usada como função objetivo para seleção dos atributos ótimos usados como entrada dos classificadores. Todavia, mesmo com a seleção de atributos ótimos, tal estratégia ainda necessita de

uma vasta quantidade de atributos para obter um desempenho satisfatório.

Portanto, objetivando a seleção de atributos para a classificação fonética de sinais de fala de banda limitada, neste capítulo, são investigadas métricas de seleção de componentes e são apresentadas análises resultantes das classificações efetuadas por classificadores de alto desempenho. A estratégia proposta tem como base algoritmos de árvores de decisão (J.48 e LMT [31]), MLP, SVM e *k*-NN, seguido da análise de MI, e seleção de parâmetros baseada em correlação (*correlation-based feature selection* - CFS) para o descarte dos atributos de menor relevância [32], [31]. Assim, a estratégia proposta consiste em identificar os atributos que facilitem a distinção entre grupos fonéticos diferentes. Os resultados de simulação permitem inferir acerca da seleção dos melhores atributos a serem considerados na composição de diferentes classificadores fonéticos, visando comprovar a eficácia da estratégia proposta.

4.1 Classificação Fonética

A tarefa de classificação é parte do processo de reconhecimento de padrões. Independente do sistema que está sendo implementado, a tarefa de classificação é considerada um ponto chave para o reconhecimento de padrões [33].

A classificação fonética é utilizada em uma ampla gama de aplicações nos mais variados tipos de sistemas de processamento de fala [39], [47], [38], [29] e [45]. O objetivo da classificação fonética é determinar a classe fonética produzida por um locutor a partir de uma amostra do sinal de fala. Para tal, é necessário o agrupamento de segmentos de sinais da fala em classes discriminativas.

O conjunto de sons componentes dos sinais de fala podem ser agrupados, de acordo com suas similaridades acústicas, em classes fonéticas. Tais classes representam um conjunto de características temporais e espectrais singulares. Essas singularidades são específicas de cada conjunto de sinais e proporcionam uma discriminação mais eficaz entre as diferentes classes fonéticas.

Neste trabalho de pesquisa, com base em [46], [38] e [45], discutimos a classificação do sinal de fala em BGP com ênfase no subgrupo fonético fricativo. Então, usando o conceito de classificação hierárquica discutido em [39] e de classes BGP, em [46], são propostas classes agrupadas hierarquicamente como ilustrado na Figura 15, a qual será utilizada como referência para o processo de classificação fonética dos sinais de fala. A Figura 15 mos-

tra diferentes níveis de classificação e os detalhes da classificação proposta são apresentados na Tabela 3.

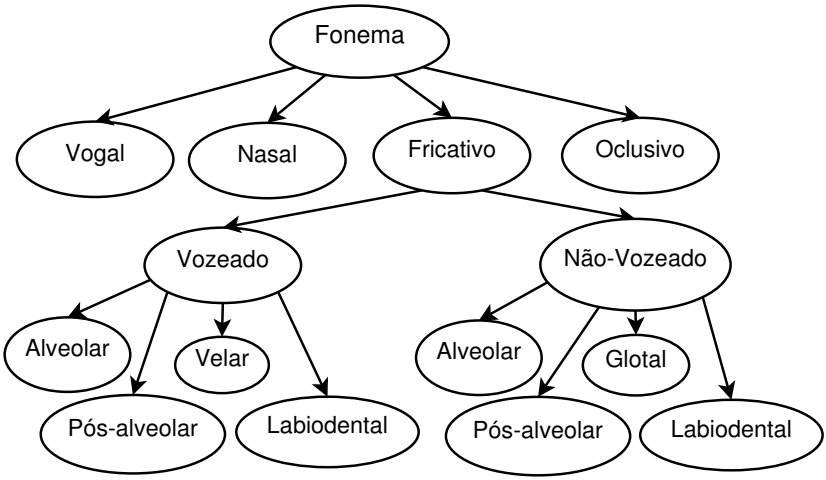


Figura 15 – Árvore fonética.

Tabela 3 – Distribuição de classes fonéticas para sinais de fala

Classificação hierárquica			
Classes	BGP	Vozeamento	Ponto de articulação
C_{l1}	Vogal	–	–
C_{l2}	Nasal		
C_{l3}	Oclusiva		
C_{l411}	Fricativa	Vozeado	Alveolar
C_{l412}			Velar
C_{l413}			Pós-alveolar
C_{l414}			Labiodental
C_{l421}		Não-vozeado	Alveolar
C_{l422}			Glotal
C_{l423}			Pós-alveolar
C_{l424}			Labiodental

4.2 Extração e Seleção de Atributos

Para o aprimoramento da classificação fonética, é necessária a realização de procedimentos de mineração de dados a fim de evidenciar as características relevantes das amostras do sinal de fala, isto é, selecionar os atributos mais característicos [37], [36].

A técnica de seleção de atributos tem como objetivo evidenciar os atributos mais relevantes em uma tarefa de classificação, possibilitando assim a exclusão de atributos redundantes e de atributos que apresentem pouca, ou quase nenhuma, contribuição à tarefa de classificação.

A redução da dimensionalidade dos dados permite que os algoritmos de classificação operem com maior eficiência [36], limitando o espaço de dados R^L e, consequentemente, também limitando as hipóteses de combinações dos atributos nos subespaços R^{s_n}, em que L denota o número total de atributos e $s_n < L$ representa a dimensão dos possíveis subespaços de R^L.

Nesse contexto, a proposta deste trabalho é aplicar os algoritmos, J.48, LMT, MLP, SVM e k-NN a classificações fonéticas e investigar os atributos característicos na classificação, aplicando análises MI e CFS para seleção de atributos extraídos a partir de sinais de fala de banda limitada (amostrados à taxa de 8000 Hz), contendo componentes de frequências entre 0 e 3400 Hz [32], [31].

4.2.1 Extração de Atributos

Assim, como em [38], [44], [48] e [49], os seguintes atributos vetoriais e escalares são elencados:

- Taxa de cruzamento por zero:

$$x_{zcr} = \frac{\sum_{k=2}^{N} \frac{1}{2} |\operatorname{sgn}\{s(k)\} - \operatorname{sgn}\{s(k-1)\}|}{(N-1)} \tag{4.1}$$

  onde N representa o número de amostras por quadro; $\operatorname{sgn}\{s(k)\} = 1$, quando $s(k) \geq 0$; e $\operatorname{sgn}\{s(k)\} = -1$, quando $s(k) < 0$.

- Energia normalizada do quadro:

$$x_{nrp}(m) = \frac{logE(m) - logE_{min}(m)}{log\bar{E}(m) - logE_{min}(m)} \tag{4.2}$$

com

$$\begin{aligned} E(m) &= \sum_{k=0}^{N-1} s^2(k) \\ E_{min}(m) &= \min_{\mu=0}^{N_{min}} E(m-\mu) \\ \bar{E}(m) &= \alpha\bar{E}(m-1)+(1-\alpha)E(m) \end{aligned} \tag{4.3}$$

onde m representa o quadro atual. O fator de suavização é ajustado para $\alpha = 0,96$ e a janela de busca mínima é de $N_{min} = 200$.

- Índice de gradiente:

$$x_{gi} = \sum_{k=2}^{N} \frac{\Psi(k)|s(k)-s(k-1)|}{\sqrt{\frac{1}{N}E(m)}} \tag{4.4}$$

onde

$$\Psi(k) = \frac{1}{2}|\psi(k)-\psi(k-1)|. \tag{4.5}$$

A variável $\psi(k)$ representa o sinal do gradiente $\psi(k) = \text{sgn}\{s(k) - s(k-1)\}$ [44].

- Estimativa do *kurtosis* local:

$$x_k = log\frac{1}{N}\sum_{k=0}^{N-1} s^4(k) - 2log\frac{1}{N}E(m). \tag{4.6}$$

- Centroide:

$$x_c = \frac{1}{N}\frac{\sum_{f=1}^{N} e_f log_2 f}{\sum_{f=1}^{N} e_f}. \tag{4.7}$$

- Centroide espectral:

$$x_{sc} = \frac{\sum_{i=0}^{M/2} i.|S(e^{j\Omega_i})|}{(\frac{M}{2}+1)\sum_{i=0}^{M/2}|S(e^{j\Omega_i})|}. \tag{4.8}$$

- *Flatness* espectral:

$$x_{sf} = log_2\left[1+\left(\frac{1}{N}\sum_{k=1}^{N}|s(k)|\right)\right]. \tag{4.9}$$

- Uniformidade:

$$u = -\sum_{f=1}^{N} \left(\frac{e_f}{\sum_{f=1}^{N} e_f} \right) log_N \left(\frac{e_f}{\sum_{f=1}^{N} e_f} \right). \tag{4.10}$$

- Estimativa de intensidade (*loundness*):

$$l = log_2 \left(1 + \frac{1}{N} \sum_{k=1}^{N} |s(k)| \right). \tag{4.11}$$

- Estimativa da largura de banda:

$$B = \sqrt{\frac{\sum_{f=1}^{N} (x_c - log_2 f)^2 e_f}{\sum_{f=1}^{N} e_f}}. \tag{4.12}$$

- Os doze primeiros coeficientes LPC.

- Os dez primeiros coeficientes de reflexão.

- Os treze primeiros coeficientes MFCC, seguidos de suas primeiras e segundas derivadas (velocidade e aceleração) sobre uma janela temporal de três sucessivos quadros.

Desta forma, a etapa de extração de atributos resulta em vetores $\mathbf{x}^{\mathrm{T}} = [x_1, \ldots, x_L]$, obtidos a uma taxa r, onde $1/r$ representa a duração do quadro com $L = 71$ atributos e $r = 100$ Hz.

4.2.2 Seleção de Atributos

Os métodos de seleção de atributos podem ser divididos em dois grupos: *wrappers* e *filters* [31], [36]. Métodos *wrappers* consistem na seleção de atributos com base no algoritmo de aprendizagem que é aplicado ao treinamento dos classificadores. Na maioria dos casos, é inviável que esse método explore todas as possibilidades do conjunto de treinamento. Sendo assim, métodos *wrappers* geralmente adotam uma busca heurística sub-ótima [37]. Métodos *filters* avaliam o valor do atributo x_i usando heurísticas como, por exemplo, a correlação de x_i com a classe y. Dessa forma, métodos *filters* demandam menor custo computacional do que os métodos *wrappers* e são independentes do algoritmo de aprendizagem.

Os métodos *filters* também podem ser divididos em métodos que avaliam um atributo individual, por exemplo, métodos MI e métodos que avaliam subconjuntos de atributos, como métodos CFS [36], [46]. Neste trabalho, MI e CFS são investigados e usados como base para a seleção de atributos.

4.2.2.1 Informação Mútua

A informação mútua, também conhecida como ganho de informação (*information gain* - IG), é descrita como segue. A função densidade de probabilidade $p(y)$ das classes fonéticas é estimada através do conjunto de treinamento. A variável aleatória Y associada às classes fonéticas tem entropia igual a

$$H(Y) = -\sum_{y \in Y} p(y) log_2[p(y)]. \tag{4.13}$$

Assumimos que o conjunto de treinamento seja representado por $\tau = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^T, y^T)\}$, onde T é o número total de quadros usados para o treinamento com $\mathbf{x} \supset x_i$ e $i \in \{1, \ldots, L\}$. Dessa forma, o conjunto de treinamento pode ser particionado de acordo com as funções densidades de probabilidade $p(x_i)$ e $p(y|x_i)$, estimadas através da frequência de ocorrência em τ. A entropia de Y condicionada à observação da variável aleatória X_i é expressa por

$$H(Y|X_i) = -\sum_{x_i \in X} p(x_i) \sum_{y \in Y} p(y|x_i) log_2[p(y|x_i)] \tag{4.14}$$

e a informação mútua $I(X_i;Y)$ é computada como

$$I(X_i;Y) = H(Y) - H(Y|X_i). \tag{4.15}$$

O método consiste em calcular $I(X_i;Y)\ \forall\ i = 1, \ldots, L$, e então selecionar os A atributos com maiores valores de $I(X_i;Y)$.

4.2.2.2 Seleção de Parâmetros Baseada em Correlação

A CFS é um método que avalia um subconjunto de atributos. O ponto chave desse método é a avaliação heurística de subconjuntos, examinando a utilidade de um atributo individual durante a predição da classe, juntamente com o nível de intercorrelação entre os demais atributos. Em (4.16), são atribuídos valores mais altos aos subconjuntos que contêm atributos fortemente correlacionados com a classe e apresentem baixa intercorrelação com os de-

mais atributos.

$$Merit_s = \frac{k_{\bar{r_{cf}}}}{\sqrt{k + k(k-1)\bar{r_{ff}}}} \tag{4.16}$$

onde $Merit_s$ é o 'mérito' heurístico do subconjunto de atributo A conter k atributos, $\bar{r_{cf}}$, a correlação média atributo-classe, e $\bar{r_{ff}}$, a intercorrelação média atributo-atributo. Para a aplicação de (4.16) é necessário o cálculo da correlação entre atributos. Para tal, em CFS, é utilizada a medida de incerteza simétrica para estimar o grau de associação entre os atributos X_i e X_j [31]. Assim,

$$SU = \left[\frac{H(X_i) + H(X_j) - H(X_i, X_j)}{H(X_i) + H(X_j)} \right]. \tag{4.17}$$

4.3 Algoritmos de Classificação

Em uma tarefa de classificação fonética, o objetivo central é observar uma amostra do sinal de fala e decidir em qual classe fonética ela se encontra. Essa tarefa pode ser estruturada em três blocos principais: pré-processamento, extração de atributos e classificação [35]. Costuma-se chamar os dois primeiros blocos de *front-end*, que tem como função reduzir o conjunto de dados a um conjunto específico de atributos que minimizem o erro no processo de classificação. O bloco seguinte - conhecido como classificador - tem a função de, a partir dos atributos apresentados, estimar a classe fonética a qual ele pertence.

Na fase de treinamento, o processo de "aprendizagem" do classificador é realizado através de conjuntos de amostras de treino, que apresentam atributos característicos para cada classe fonética. Independente se o sistema está em fase de treinamento ou de testes, o *front-end* está sempre presente, visando obter os atributos que servirão de entrada para o classificador.

Existem duas formas de se treinar um classificador: o aprendizado supervisionado e o aprendizado não-supervisionado. Na primeira, é fornecido um rótulo para cada classe durante a fase de treinamento, auxiliando o classificador a distinguir as propriedades acústicas de cada classe. Na segunda, apenas os exemplos de treinamento são disponibilizados ao classificador. Assim, ele forma agrupamentos naturais dos padrões de entrada [35].

Neste trabalho, utilizamos o método de aprendizagem supervisionada e investigamos os seguintes algoritmos de classificação: árvores de decisão J.48 e LMT, MLP, SVM e k-NN. A descrição detalhada desses algoritmos é apresentada em [31] e uma breve descrição é dada a seguir.

4.3.1 Árvore de Decisão J.48 e LMT

As árvores de decisões são os algoritmos de aprendizagem de máquinas mais populares em sistemas de classificação. Elas realizam uma análise dos atributos e consistem na implementação de um conjunto de limiares que avaliam (como verdadeiro ou falso) os atributos em cada um dos nós da árvore. Após a avaliação dos limiares, cada nó é ramificado em duas ou mais sub-árvores. Nos nós, são calculados resultados baseados nos valores individuais dos atributos. Cada resultado possível está relacionado com uma ramificação da árvore, apresentando as classes como resultado final dessa ramificação. O algoritmo J.48 é um método eficiente de implementação de árvores de decisão para a estimação e a classificação de dados incompletos e imprecisos. O algoritmo LMT é estruturado como uma árvore de decisão padrão contendo funções de regressão lógicas em cada nó ao invés da avaliação específica de um único atributo.

4.3.2 Rede Neural de Múltiplas Camadas (MLP)

Em termos práticos, as redes neurais são ferramentas não-lineares de modelagem estatística de dados. Elas podem ser utilizadas para modelar relações complexas entre entradas e saídas. Neste trabalho, utilizamos o algoritmo *backpropagation* MLP.

4.3.3 Máquinas de Vetores de Suporte (SVM)

O algoritmo SVM consiste em um método de aprendizado supervisionado que realiza a transformação de dados de entrada em um espaço de atributos de alta dimensionalidade e efetua a separação de classes através de superfícies de decisões (hiperplanos) nesse espaço. Dessa forma, os algoritmos SVM buscam a obtenção de hiperplanos ótimos capazes de separar diferentes classes de dados com as maiores margens de separação possíveis.

4.3.4 *k-Nearest Neighbor* (*k*-NN)

O algoritmo *k*-NN realiza a estimação da densidade local dos atributos de treinamento. Para a estimação da classe de um novo atributo x_a, o algoritmo *k*-NN calcula os *k*-vizinhos mais próximos a x_a e classifica-o como pertencente à classe que ocorre com maior frequência dentre os seus *k*-vizinhos.

4.4 Treinamento e Arquitetura dos Classificadores

Para implementar uma tarefa de classificação fonética e possibilitar o treinamento dos classificadores, torna-se necessária uma base de dados composta por arquivos de fala em que cada sinal de referência é caracterizado por uma série de atributos que descrevam características fonéticas. Neste trabalho de pesquisa, utilizou-se o *corpora* de fala disponibilizado em [50]. Visando a simulação de limitação em banda, os sinais de fala desse *corpora* foram aplicados à filtragem passa-baixas com banda passante de 0 a 3400 Hz. Dessa forma, 2 horas de áudio, de 13 locutores masculinos e 10 femininos, são utilizadas na etapa de treinamento. Para o treinamento dos classificadores, é usada a plataforma de código-livre WEKA (*Waikato Environment for Knowledge Analysis*) [31], que consiste em uma ferramenta de aprendizagem de máquina e mineração de dados do estado da arte. Os algoritmos de classificação citados na Seção 4.3 são utilizados no WEKA com suas configurações padrão, conforme descritas na Tabela 4.

Tabela 4 – Configuração dos classificadores

Classificador	Configuração
J.48	C=0,25; M=2
LMT	I=−1; M=15; W=0,0
MLP	L=0,3; M=0,2; N=500; V=0,0; S=0,0; E=20,0; H=a
SVM	S=G=R=0; K=2; D=3; N=0,5; M=40; C=1; E=10^{-3}
k-NN	K=1; W=0

4.4.1 Arquitetura de Classificação

O sistema proposto de classificação hierárquica primeiro analisa a qual BGP cada quadro pertence. Caso o quadro analisado pertença ao grupo fonético fricativo, um novo classificador é utilizado para a detecção de classes vozeadas ou não-vozeadas. Finalmente, de acordo com o nível de vozeamento estimado do quadro, um terceiro classificador é aplicado para identificação do tipo de fonema fricativo. A Figura 16 ilustra o diagrama de blocos do processo de classificação hierárquica proposto.

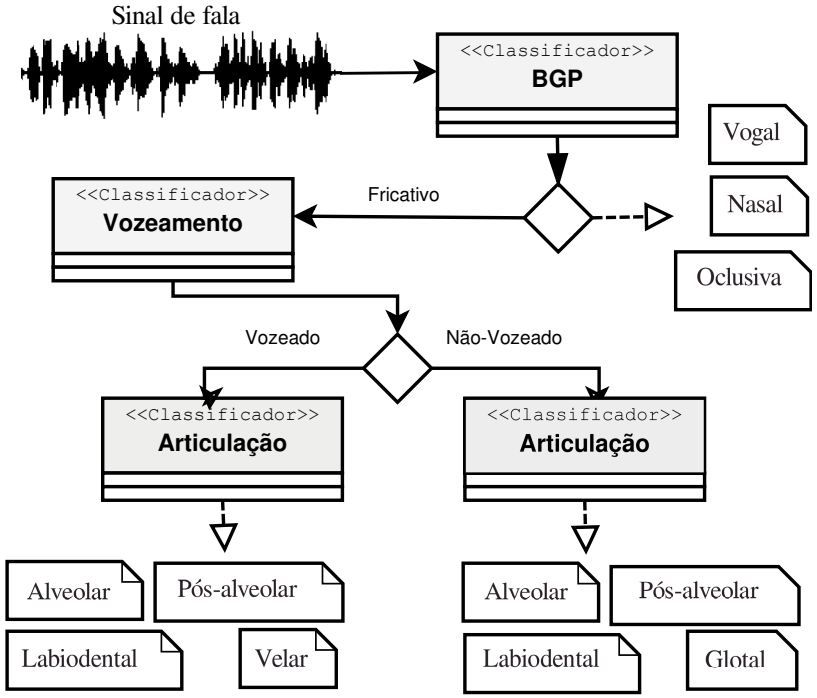


Figura 16 – Diagrama de blocos do processo de classificação hierárquica.

Para cada nível hierárquico de classificação, um subconjunto dos atributos de entrada $\mathbf{x}^{\mathrm{T}} = [x_1, \ldots, x_L]$ é utilizado. Esses subconjuntos são previamente analisados e definidos conforme as métricas de seleção de atributos descritas na Seção 4.2.2, constituindo a etapa de *front-end* dos classificadores. Para efeitos de comparação, um sistema de classificação direta dos grupos apresentados na Tabela 3 é obtido usando o algoritmo k-NN.

4.5 Resultados e Análise de Desempenho

Com o intuito de analisar a eficácia da seleção de atributos para os diferentes níveis hierárquicos, os conjuntos de atributos selecionados pelas técnicas MI e CSF são testados nos classificadores, citados na Seção 4.3, para cada grupo fonético. Os testes realizados avaliaram os desempenhos dos classificadores para 47 minutos de sinais de fala de 6 locutores masculinos e 3 femininos.

A Figura 17 apresenta a evolução do erro de classificação dos algoritmos *k*-NNs em relação ao número de atributos selecionados, bem como a matriz de confusão para cada nível de classificação. Assim, os diferentes níveis hierárquicos podem ser avaliados isoladamente. Dessa forma, os atributos mais relevantes são selecionados para comporem o *front-end* de cada classificador.

A Tabela 5 apresenta o desempenho dos classificadores e destaca os mais eficientes, os quais compõem a estratégia proposta de classificação. Além dos classificadores de melhor desempenho, também são apresentadas (na Tabela 5) as quantidades de atributos de cada nível hierárquico.

Na Tabela 6, observa-se o desempenho geral da estratégia proposta para classificação em contraste com o sistema padrão de classificação direta. Nesse contexto, as taxas de erro de classificação dos subgrupos fricativos $E_{rro}(C_{l4vf})$ são calculados a partir de (4.18).

$$\begin{array}{lcl} E_{rro}(C_{l4vf}) & = & 1 - P(C_{lBGP})P(C_{lvoze})P(C_{lfri}) \\ P(C_{lBGP}) & = & P(C_{lBGP} == Fricativo) \\ P(C_{lvoze}) & = & P(C_{lvoze} == v) \\ P(C_{lfri}) & = & P(C_{lfri} == p|v) \\ \forall v & \in & \{1,2\} \\ \forall p & \in & \{1,2,3,4\} \end{array} \tag{4.18}$$

A estratégia proposta de classificação, composta pelos atributos e classificadores em destaque na Tabela 5, é examinada através da comparação de suas matrizes de confusão fonética, apresentadas na Figura 17, com a matriz obtida pelo sistema padrão de classificação direta, mostrada na Figura 18. Nesse contexto, nota-se uma menor confusão na matriz resultante da estratégia aqui proposta, confirmando assim sua eficácia. A Figura 19 apresenta os desempenhos individuais do sistema padrão e o da estratégia proposta para cada classe fonética.

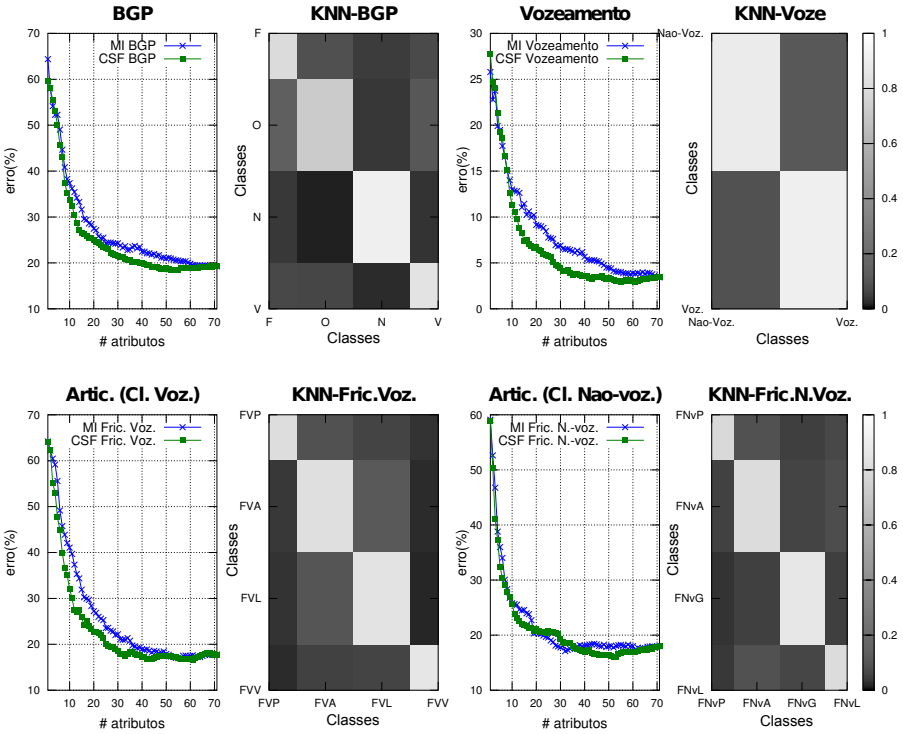


Figura 17 – Seleção de atributos e classificação fonética.

Tabela 5 – Seleção de atributos e desempenho específico dos classificadores

Classes	**CFS**	Erro de classificação(%)				
	Nº de atributos	J.48	LMT	MLP	**SVM**	***k*-NN**
BGP	**53**	30,61	24,40	23,50	22,91	**18,36**
Vozeamenento	**55**	10,10	05,72	05,60	**03,54**	03,78
Fricativa vozeada	**56**	32,96	24,96	22,51	20,81	**16,86**
Fricativa não-vozeada	**52**	24,43	17,51	15,56	**13,13**	16,08

Tabela 6 – Desempenho geral dos classificadores

Classificador	Nº de atributos	Erro de classificação(%)
Estratégia padrão	71	23,02
Estratégia proposta	**67**	**21,72**

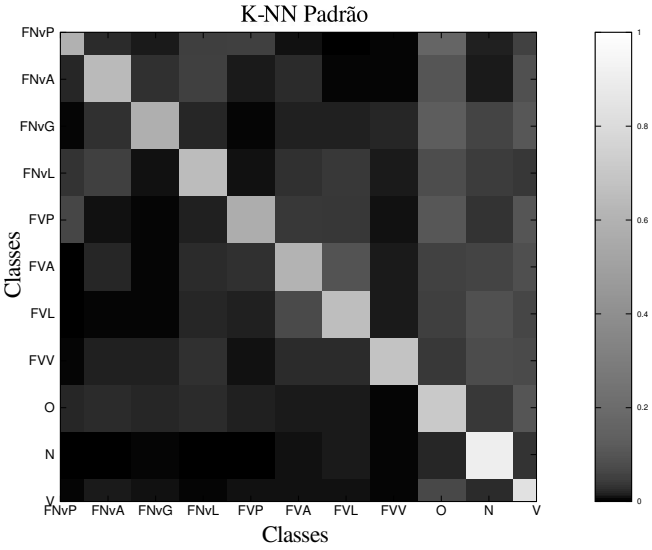


Figura 18 – Matriz de confusão do classificador *k*-NN padrão.

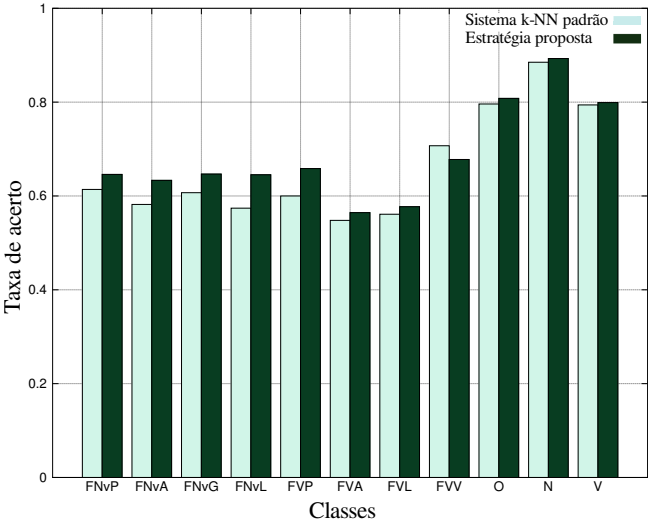


Figura 19 – Desempenho individual dos classificadores fonéticos.

4.6 Conclusões

Neste trabalho, uma nova estratégia para classificação fonética de sinais de fala de banda limitada foi apresentada. Essa estratégia é baseada na seleção de atributos aplicados a uma classificação fonética hierárquica. A partir da técnica de seleção de atributos, foram apresentados os atributos mais relevantes para a tarefa de classificação. A estratégia de classificação hierárquica permitiu a utilização de diferentes algoritmos em cada nível hierárquico. Assim, diversos algoritmos de classificação foram investigados e integrados à estratégia proposta. A implementação de algoritmos especialistas em cada nível proporcionou a redução do número de atributos, assim como a redução no erro de classificação. A estratégia proposta exibe um desempenho superior ao do sistema padrão de classificação direta. Os resultados de simulação apresentados permitiram uma análise acerca da seleção de atributos discriminativos para cada nível hierárquico, confirmando a eficácia da estratégia de classificação proposta.

Capítulo 5

Estratégia Proposta para Implementação de Algoritmos de Extensão Artificial de Largura de Banda

Conforme discutido no Capítulo 3, o objetivo de um sistema de ABWE é realçar o sinal de fala de NB, tornando-o mais agradável aos ouvintes e fazendo com que sua qualidade subjetiva se “assemelhe" a um sinal de WB. Isso é possível através da estimação dos componentes de frequências acima de 3400 Hz. Assim, a ideia básica de um sistema de ABWE é sintetizar artificialmente componentes de alta frequência, convertendo um sinal de NB em um sinal de WB, isto é, estimando as propriedades acústicas pertinentes a WB [45], [51].

Neste trabalho, visando compatibilidade com as redes de telefonia existentes e com os *codecs* de NB e de WB amplamente adotados no mercado de telecomunicações, tais como as Recomendações ITU-T G.729 [1] e G.729.1 [2], respectivamente, a estratégia para ABWE aqui desenvolvida é baseada nos procedimentos descritos em [2] e [9]. Para propiciar independência de *side information*, bem como redução de ocorrências e atenuação dos efeitos indesejáveis no sinal de fala reconstruído, são propostas alterações nas etapas de representação e estimação do trato vocal do sistemas propostos em [2] e [9]. Adicionalmente, sugerimos acrescentar a etapa de classificação fonética apresentada no Capítulo 4. Essa etapa é responsável por um tratamento específico em diferentes classes de sinais de fala, garantindo, dessa forma, uma melhor representação e, consequentemente, uma melhoria na clusterização dos parâmetros discriminativos do trato vocal, resultando em uma síntese mais “limpa” e agradável dos sinais de WB.

5.1 Proposta de Nova Estratégia para Extensão de Banda

O procedimento de ABWE adotado aqui é baseado no processo de codificação e decodificação de sinais de WB proposto em [9]. Entretanto, a estratégia implementada em [9] lança mão de *side information*, tais como fase e energia dos segmentos do sinal de fala para estimar os sinais de excitação e os envelopes temporais e espectrais correspondentes à banda alta do sinal. Para tornar a codificação independente de *side information* e para melhorar o

desempenho do decodificador em cenários de transmissões em NB, este trabalho propõe uma nova estratégia de processamento, incluindo uma etapa de classificação fonética e adotando, para a representação do envelope espectral, coeficientes LSF [52], ao invés de coeficientes de fase e energia.

A Figura 20 ilustra o diagrama de blocos do sistema de ABWE proposto. Assim como em sua versão original [9], esse diagrama tem como base uma estrutura de três estágios de processamento, descritos como segue:

1. **Estágio I**. Estimação do sinal de excitação através de predição linear (*code-excited linear prediction* - CELP) [1].

2. **Estágio II**. Filtragem do sinal de excitação resultante do primeiro estágio através dos envelopes temporal e espectral do trato vocal estimados a partir de uma consulta a um conjunto de *codebooks* baseados em classificação fonética.

3. **Estágio III**. Cálculo de ganho e pós-processamento para a estimação do sinal de WB.

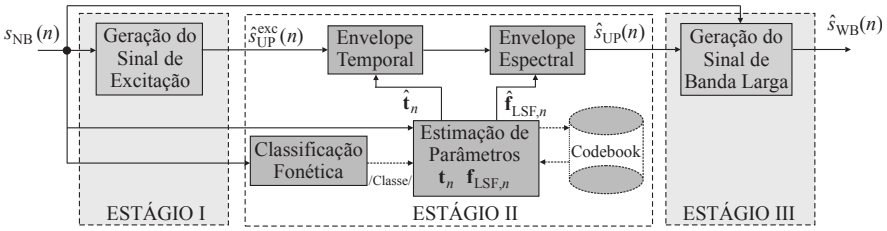


Figura 20 – Diagrama de blocos do sistema de ABWE proposto.

Os estágios que compõem o sistema de ABWE serão descritos com detalhes nas seções seguintes.

5.2 Estágio I - Estimação do Sinal de Excitação

A cada quadro (de aproximadamente 20 ms) do sinal de fala de NB $s_{\mathrm{NB}}(n)$, recebido como entrada no sistema de ABWE da Figura 20, são estimados sinais de excitação de UP $\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n)$. A estimação de $\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n)$ é obtida através de um modelo ideal de produção de fala quase-estacionário, no qual devem ser satisfeitos os seguintes critérios [9]:

- O sinal de excitação deve apresentar espectro plano.

- Para sinais de fala vozeados, a excitação deve conter harmônicos da frequência fundamental do sinal, F_0.
- Para sinais de fala não-vozeados, a excitação deve ser um ruído branco.
- Sinais de fala mistos (vozeados e não-vozeados) devem apresentar uma razão sinal-ruído (SNR) variável.
- A contribuição de parâmetros vozeados não deve ser dominante na banda de alta frequência.

Em [2] e [9], os critérios descritos anteriormente são satisfeitos e a mesma estratégia de estimação do sinal de excitação é também aqui adotada. A Figura 21 ilustra o diagrama de blocos do processo de estimação do sinal de excitação, no qual parâmetros de energia, ganho e *pitch* são obtidos através de procedimentos utilizados no CELP e aqui aplicados para estimar o sinal de excitação $\hat{s}_{\mathrm{UP}}^{\mathrm{exc}}(n)$, cuja composição é uma mistura de sinais não-vozeado $\hat{s}_{\mathrm{UP}}^{\mathrm{exc,nv}}(n)$, gerados a partir de um ruído branco, com sinais vozeados $\hat{s}_{\mathrm{UP}}^{\mathrm{exc,v}}(n)$, gerados a partir de um trem de pulsos periódicos com frequência fundamental F_0 [9].

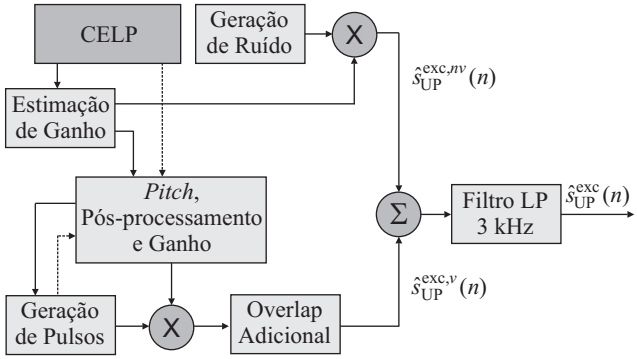


Figura 21 – Diagrama de blocos para geração do sinal de excitação.

5.3 Estágio II - Geração de Envelopes Temporais e Espectrais

Neste estágio do sistema de ABWE, são estimados os envelopes temporais e espectrais de UP que caracterizam o trato vocal no processo de geração da fala. Esses envelopes são representados pelos seguintes vetores de parâmetros

$$\mathbf{t}_n = [t_n(1), \ldots, t_n(16)]^{\mathrm{T}} \tag{5.1}$$

e

$$\mathbf{f}_{\mathrm{LSF},n} = [f_{\mathrm{LSF},n}(1), \ldots, f_{\mathrm{LSF},n}(19)]^{\mathrm{T}} \tag{5.2}$$

onde o vetor $\mathbf{t}_n$ representa o envelope temporal contendo energias logarítmicas de 16 subquadros (1,25 ms cada) [9] e o vetor $\mathbf{f}_{\mathrm{LSF},n}$, componentes LSF que caracterizam o envelope espectral. Como ilustrado na Figura 20, tais parâmetros são estimados através de consulta a um conjunto de *codebooks* selecionados de acordo com suas classes fonéticas. Esse processo de classificação reagrupa os vetores de parâmetros $\mathbf{t}_n$ e $\mathbf{f}_{\mathrm{LSF},n}$ em classes fonéticas, permitindo, dessa forma, uma estimação específica para cada classe.

5.3.1 Clusterização e Classificação Fonética

Conforme discutido no Capitulo 4, o conjunto de sons componentes dos sinais de fala podem ser agrupados, de acordo com suas similaridades acústicas, em classes fonéticas. Tais classes representam um conjunto de características temporais e espectrais singulares. Essas singularidades são específicas de cada conjunto de sinais e garantem maior discriminação entre as diferentes classes fonéticas.

Em [14] e [53], são descritas as propriedades acústicas dos sinais de fala. De acordo com a avaliação do comportamento de cada classe fonética quanto à distribuição de energia no domínio espectral, nota-se que, para o conjunto de classes fricativas, os componentes mais discriminativos encontram-se nas altas frequências, isto é, acima de 3400 Hz e, consequentemente, além da banda de frequência originalmente considerada pela PSTN. Dessa forma, dentre as demais classes fonéticas (veja [23]), a discriminação entre fonemas das classes fricativas torna sua percepção um tanto difícil para os usuários da PSTN. Assim, uma atenção especial é dedicada a tais classes de fonemas.

Com base no processo de classificação fonética apresentado no Capitulo 4, a Tabela 7 é aqui proposta e será utilizada como referência para o processo de clusterização dos sinais de fala $s_{\mathrm{NB}}(n)$ e $s_{\mathrm{WB}}(n)$.

A Tabela 7 apresenta quatro diferentes tipos de clusterização denominadas A, B, C e D, contendo duas, três, cinco e nove classes. Após o processo de clusterização e classificação fonética, a seleção de *codebooks* de NB e WB torna-se mais discriminativa e, consequentemente, mais apropriada para a etapa de treinamento.

Tabela 7 – Distribuição das classes fonéticas consideradas para os sinais de fala de NB e de WB

Classes				ex.	Descrição
2	3	5	9		
A1	B1	C1	D1	/z/	Fricativas vozeadas alveolares
			D2	/v/	Fricativas vozeadas labiodentais
			D3	/j/	Fricativas vozeadas pós-alveolares
		C2	D4	/V/	Demais fonemas vozeados
	B2	B3	D5	/f/	Fricativas não-vozeadas labiodentais
			D6	/s/	Fricativas não-vozeadas alveolares
			D7	/ξ/	Fricativas não-vozeadas pós-alveolares
		C4	D8	/U/	Demais fonemas não-vozeados
A2	B3	C5	D9	/Sil/	Silêncio

5.3.2 Extensão Usando Mapeamento via *Codebooks*

A Figura 22 mostra o diagrama de blocos da etapa de treinamento de *codebooks*, em que os filtros LP e HP representam filtros passa-baixas e passa-altas, respectivamente. Nessa etapa, o sinal de WB $s_{\mathrm{WB}}(n)$ é filtrado pelos correspondentes filtros, resultando nos sinais de NB $s_{\mathrm{NB}}(n)$ e de UP $s_{\mathrm{UP}}(n)$. A partir desses sinais, são extraídos parâmetros que representam os espaços de dados x^{NB}_{Classe} e y^{UP}_{Classe}, de acordo com suas correspondentes classes fonéticas.

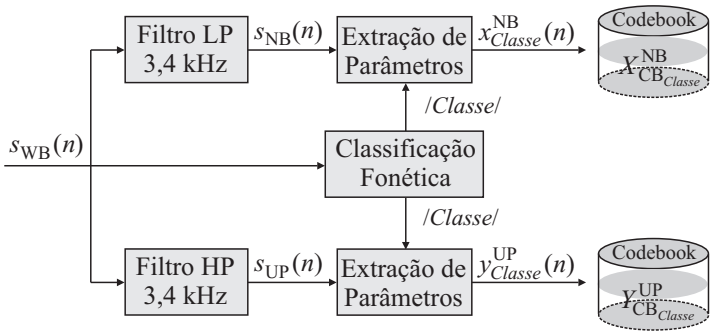


Figura 22 – Etapa do processo de treinamento de *codebooks*.

O processo de treinamento dos melhores *codebooks* para representação dos envelopes do trato vocal inclui o bloco de classificação fonética proposto, visando gerar os diferentes *codebooks* para as classes fonéticas definidas na Tabela 7. Em um primeiro momento, esse procedimento é governado por um processo de aprendizagem supervisionada [12], em que a classe fonética seja consultada e os respectivos espaços de dados x^{NB}_{Classe} e y^{UP}_{Classe} sejam gerados de acordo com suas classes correspondentes. Após a discriminação em classes, o algoritmo LGB [52] é utilizado para agrupar os vetores de parâmetros em torno de diferentes centroides e gerar os *codebooks*, $X^{\mathrm{NB}}_{\mathrm{CB}_{Classe}}$ e $Y^{\mathrm{UP}}_{\mathrm{CB}_{Classe}}$, de acordo com as suas correspondentes distâncias euclidianas, $d(S,\hat{S})$. Assim,

$$d(S,\hat{S}) = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|S(e^{j\Omega}) - \hat{S}(e^{j\Omega})\right|^2 d\Omega\right]^{1/2} \tag{5.3}$$

onde $S(e^{j\Omega})$ caracteriza o espaço de dados das centroides e $\hat{S}(e^{j\Omega})$, o espaço de dados do sinal recebido e analisado. Dessa forma, os *codebooks* gerados possuem maior discriminação entre suas diferentes centroides.

O processo de treinamento aqui utilizado considera as técnicas discutidas em [12] e [52] apresentadas na Seção 3.2.2.1, na qual *codebooks* duais de NB $X^{\mathrm{NB}}_{\mathrm{CB}}$ e de UP $Y^{\mathrm{UP}}_{\mathrm{CB}}$ são construídos. A construção desses *codebooks* duais se dá através do mapeamento um-para-um de índices equivalentes a *codewords* de UP para *codewords* de NB. Assim, índices de mapeamento das *codewords* de UP são usados para o mapeamento de seus correspondentes coeficientes em NB [12]. A utilização de índices de UP é adotada para evitar perda de informações de singularidades que resultassem em imprecisão dos envelopes do trato vocal, principalmente, nos casos em que os componentes espectrais sejam predominantemente de altas frequências. Entretanto, na etapa de teste indicada na Figura 20, durante o processo de estimação de parâmetros, uma imprecisão de singularidades acústicas pode ocorrer, por exemplo, devido a eventuais erros na classificação fonética e escolha do *codebook* mais apropriado para um determinado quadro do sinal de fala. Todavia, tal problema é atenuado devido à robustez da estratégia proposta com respeito a erros de classificação fonética, em que, mesmo ocorrendo erros de classificação, as versões estimadas dos parâmetros de UP não estariam tão distantes da versão correta, dado o agrupamento de classes similares.

5.3.3 Estimação de Parâmetros e Pós-processamento

Para a estimação dos parâmetros do vetor $\tilde{\mathbf{y}}$ que caracteriza o modelo do trato vocal, uma vez determinada a classe, denotada aqui como *Classe*, do quadro do segmento de fala de NB em análise, as K *codewords* mais similares aos vetores $\mathbf{f}_{\mathrm{LSF},n}$ e $\mathbf{t}_n$ do *codebook* $Y^{\mathrm{UP}}_{\mathrm{CB}_{Classe}}$ são selecionadas, isto é, $\mathbf{Q}(n|Y^{\mathrm{UP}}_{\mathrm{CB}_{Classe}}) = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_K\}$, e os correspondentes vetores $\hat{\mathbf{f}}_{\mathrm{LSF},n}$ e $\hat{\mathbf{t}}_n$ de UP são combinados linearmente com um peso w para cada *codeword*. Assim,

$$\hat{\mathbf{y}}_{S_{\mathrm{UP}}}(n) = \sum_{m=1}^{K} w_m^{S_{\mathrm{UP}}} \cdot \mathbf{q}_m(n|Y^{\mathrm{UP}}_{\mathrm{CB}}) \;\forall\; S_{\mathrm{UP}} \supset [\mathbf{f}_{\mathrm{LSF}}, \mathbf{t}] \tag{5.4}$$

onde $\mathbf{q}_m(n|Y^{\mathrm{UP}}_{\mathrm{CB}})$ representa as *codewords* do espaço de dados de UP S_{UP}.

A qualidade da estimação dos parâmetros como também a qualidade do sinal de fala reconstruído são aprimoradas através de filtragem de média móvel nos parâmetros estimados $\hat{\mathbf{f}}_{\mathrm{LSF},n}$ e $\hat{\mathbf{t}}_n$,

$$\tilde{\mathbf{y}}(n) = \frac{1}{2} \cdot \left[\hat{\mathbf{y}}_{S_{\mathrm{UP}}}(n) + \tilde{\mathbf{y}}(n-1)\right] \tag{5.5}$$

onde $\tilde{\mathbf{y}}(n)$ representa o vetor de parâmetros estimados do quadro corrente. Esse procedimento proporciona uma transição mais suave entre os parâmetros de cada quadro, atenuando artefatos presentes no processo de estimação variante no tempo [15].

5.4 Estágio III - Obtenção do Sinal de Banda Larga

Neste estágio, como ilustrado no diagrama da Figura 20, o sinal de UP $\hat{s}_{\mathrm{UP}}(n)$, resultante da convolução do sinal estimado de excitação $\hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n)$ com os envelopes temporal e espectral do trato vocal $\hat{\mathbf{t}}_n$ e $\hat{\mathbf{f}}_{\mathrm{LSF},n}$, respectivamente, é combinado com o sinal de NB $s_{\mathrm{NB}}(n)$ e, assim, o sinal estimado de WB $\hat{s}_{\mathrm{WB}}(n)$ é obtido. Então,

$$\hat{s}_{\mathrm{WB}}(n) = \hat{s}_{\mathrm{UP}}(n) + s_{\mathrm{NB}}(n) \;\forall\; \hat{s}_{\mathrm{UP}}(n) = \hat{s}^{\mathrm{exc}}_{\mathrm{UP}}(n) * \tilde{\mathbf{y}}(n)\,. \tag{5.6}$$

5.5 Resultados e Análise de Desempenho

Nesta seção, resultados de simulação são apresentados visando avaliar e comparar o desempenho da estratégia de implementação do sistema de ABWE proposta neste trabalho. Para efeito de comparação, o sistema com codificação em WB e os sistemas utilizando a PSTN são avaliados através de seus correspondentes sinais de fala, obtidos na saída dos terminais *far-end*. Especificamente, são considerados: um sistema usando codificação em WB, um sistema utilizando PSTN sem o uso de ABWE, um sistema utilizando PSTN com o uso de ABWE mas sem classificação fonética[5] [8], [17], e, finalmente, um sistema usando PSTN com a ABWE implementada a partir da estratégia aqui proposta.

A Figura 23 mostra o espectrograma de um sinal de WB original (enviado pelo terminal *near-end*), o espectrograma de um sinal de NB (recebido pelo terminal *far-end*) sem qualquer tratamento de ABWE e os espectrogramas dos sinais de WB sintetizados através de ABWE sem e com classificação fonética, para este último, considerando nove classes distintas, e *codebooks* contendo 1024 *codewords*.

A estratégia de ABWE aqui proposta também pode ser avaliada através da análise da densidade espectral de potência dos envelopes do trato vocal dos sinais de WB estimados e dos sinais em suas versões originais de NB e de WB. A Figura 24 mostra um exemplo de envelope espectral obtido a partir do sinal de WB original, do sinal de NB convencional sem o uso de ABWE e dos sinais de WB sintetizados através de ABWE sem e com o auxílio de classificação fonética.

De acordo com as representações dos envelopes estimados (comparados às suas versões originais de NB e de WB), pode ser constatado que o sinal de WB sintetizado através da utilização de classificação fonética com nove classes distintas apresenta uma representação mais fiel à versão de WB original.

[5] Sistema baseado no procedimento de quantização vetorial no qual é realizado um mapeamento de *codebook* (com 1024 *codewords*) sem que haja qualquer tipo de classificação fonética do sinal de fala.

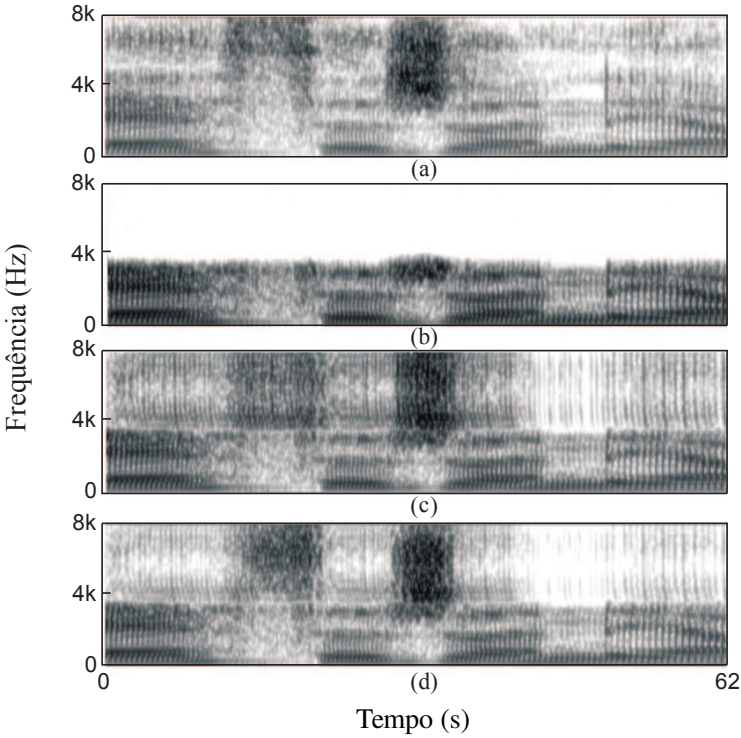


Figura 23 – Espectrogramas: (a) sinal original de WB; (b) sinal recebido de NB sem o uso de ABWE; (c) sinal de WB sintetizado através de ABWE sem classificação fonética; (d) sinal de WB sintetizado através de ABWE com classificação fonética.

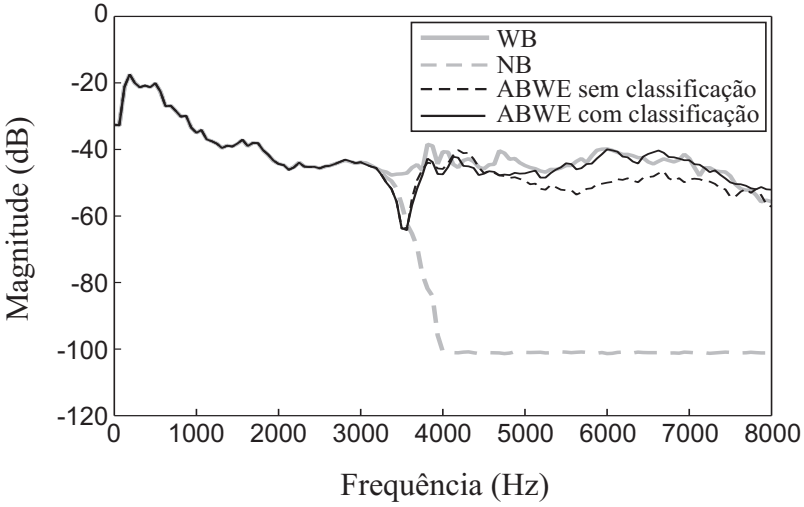


Figura 24 – Envelopes espectrais dos sinais de WB, de NB e sinais de WB sintetizados através de ABWE sem e com classificação fonética considerando 9 classes.

5.5.1 Análise Subjetiva do Sinal de Fala

Nesta etapa, foram considerados testes de avaliação (descritos no Apêndice) de acordo com as recomendações dadas em [54]. Para tal, foram coletados resultados de testes de 11 ouvintes. Em resumo, é verificada que a estratégia aqui proposta para o algoritmo ABWE é capaz de reduzir a ocorrência de artefatos nos componentes estimados de altas frequências quando comparado com procedimentos convencionais de mapeamento de *codebooks* não-supervisionados. O tratamento específico dos quadros do sinal de fala de NB em suas classes fonéticas correspondentes, em especial para as classes fricativas, possibilita um processo de estimação mais fiel aos componentes de WB originais. É evidente que, quando comparados os sinais sintetizados com suas versões originais de WB, ainda se nota a ausência de maior "brilho" no sinal de fala. Entretanto, quando comparado com sua versão convencional de NB sem tratamento de ABWE, pode ser verificada uma melhoria significativa da qualidade subjetiva do sinal de fala reconstruído.

5.5.2 Análise de Qualidade Usando Medidas Objetivas

Para a avaliação da qualidade do sinal de fala de WB sintetizado através da estratégia de ABWE utilizada neste trabalho de pesquisa, são adotadas medidas objetivas que simulam a análise perceptual logarítmica do ouvido humano: a medida W-PESQ (*perceptual evaluation of speech quality*) padronizada pela ITU-T [20], a razão sinal-ruído segmental calculada no domínio da frequência (FwSegSNR) e a distância espectral logarítmica (LSD) [18].

A Tabela 8 apresenta os resultados das medidas objetivas para os diferentes sistemas descritos no início desta seção. As medidas objetivas do sistema operando com os sinais de WB originais são consideradas como medidas de referência. Assim, os sistemas com melhores desempenho são aqueles que apresentam as medidas objetivas mais próximas das medidas de referência.

Tabela 8 – Desempenho dos sistemas de ABWE considerando diferentes medidas de qualidade

Sistemas	LSD (dB)	FwSegSNR (dB)	WB-PESQ
WB original	**00,00**	**35,00**	**4,50**
NB sem ABWE	11,82	19,90	1,85
ABWE sem classificação	07,08	19,58	2,52
ABWE 2 classes fonéticas	07,16	19,73	2,53
ABWE 3 classes fonéticas	06,90	21,05	2,64
ABWE 5 classes fonéticas	06,46	23,89	3,24
ABWE 9 classes fonéticas	06,44	23,94	3,49

A Figura 25 ilustra as curvas de desempenho dos sistemas de ABWE em relação às medidas WB-PESQ, considerando diferentes números de classes fonéticas com *codebooks* variando entre 8 a 2048 *codewords*.

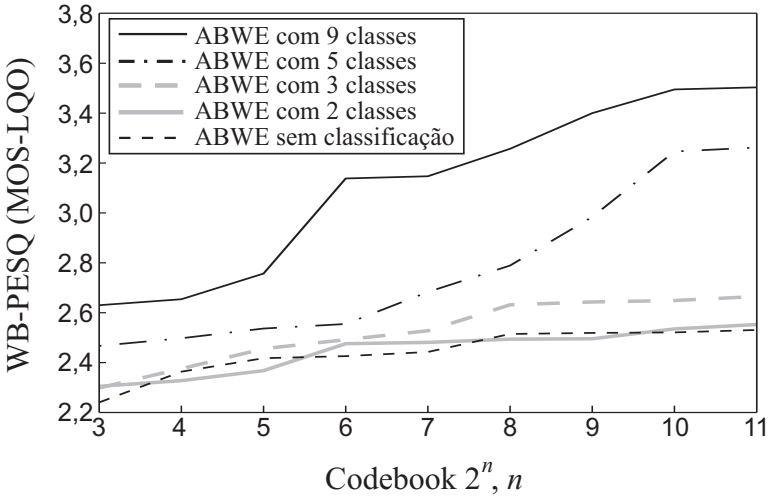


Figura 25 – Curvas de desempenho do algoritmo de ABWE utilizando *codebooks* para diferentes números de *codewords*.

5.6 Conclusões

Neste trabalho de pesquisa, uma nova estratégia para implementação de sistemas de ABWE foi apresentada. Essa estratégia utiliza coeficientes LSF para a representação do envelope espectral do trato vocal e inclui um procedimento de classificação fonética para os sinais de NB. A estratégia aqui proposta proporciona sinais sintetizados de WB significativamente melhores do que os sinais de NB e aqueles sintetizados a partir de ABWE convencionais. Resultados de avaliações subjetivas e objetivas também ratificam tais afirmações.

Capítulo 6

Extensão Artificial de Largura de Banda Aplicada a Sistemas de Reconhecimento Automático de Fala em Redes de Telefonia

Atualmente o mercado de telefonia dispõe de diversos serviços interativos, tais como os serviços presentes em sistemas automatizados de *help desk*, portais de voz, gerenciamento de diálogos em unidades de resposta audível (URA) e outros tipos de atendimento via *call centers* [55], como ilustrado na Figura 26. Tais serviços auxiliam o acesso de usuários da PSTN a informações via navegação em menus interativos. Esses menus podem ser acessados com a ajuda do teclado telefônico ou diretamente através de comandos de fala, em que o usuário é atendido por um assistente virtual controlado por reconhecimento automático de fala (*automatic speech recognition* - ASR) aplicado à PSTN [56]. Tendo em vista que os serviços comandados por fala são altamente dependentes do desempenho do ASR, as pesquisas sobre modelagem estatísticas desses sistemas continuam em franca evolução e se destacam como um tópico ativo em processamento digital de sinais [57], [58], [51].

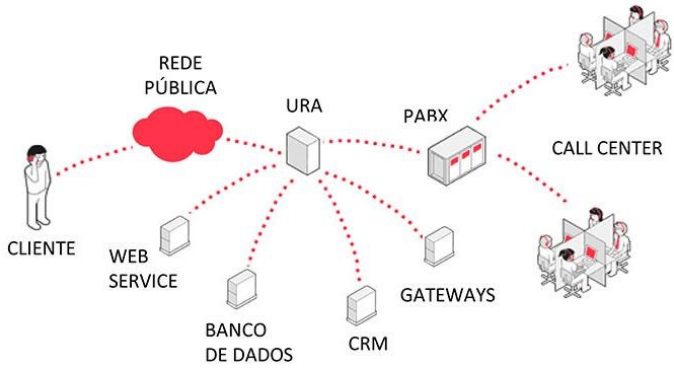


Figura 26 – Exemplo de serviços disponíveis na rede telefônica.

Resumidamente, um sistema típico de ASR é composto por duas etapas principais, o *front-end* e o *back-end*. A etapa de *front-end* recebe o sinal de entrada e efetua a extração de vetores de observação contendo informações codificadas da fala. A etapa de *back-end*, composta por um decodificador, dicionário fonético, modelo acústico e modelo de linguagem, é responsável pela decodificação dos vetores de observação e pelo processo de "busca" das informações linguísticas, por exemplo, fonemas e/ou palavras, contidas no sinal de fala [59], [22].

Os sistemas de ASR aplicados a redes de telefonia podem ser implementados de acordo com três diferentes metodologias: reconhecimento de fala embarcado (*embedded speech recognition* - ESR), que contêm as etapas de *front-end* e *back-end* integradas ao aparelho telefônico; reconhecimento de fala distribuído (*distributed speech recognition* - DSR), em que o *front-end* é integrado ao aparelho telefônico, tendo o *back-end* hospedado em um servidor externo; e o reconhecimento de fala em rede (*network speech recognition* - NSR), em que tanto o *front-end* quanto o *back-end* são hospedados em um servidor remoto e o aparelho telefônico apenas envia o sinal de fala para ser processado nesse servidor [60].

Visando a obtenção de melhores taxas de reconhecimento, sistemas de ASR desenvolvidos para *desktops* utilizam sinais de fala em WB amostrados geralmente a taxas maiores do que 16 kHz. Entretanto, sistemas aplicados à PSTN utilizam sinais de NB amostrados a 8 kHz e devido às características inerentes ao canal telefônico da PSTN (largura de banda de 300 a 3400 Hz) [45], além da perda de naturalidade e inteligibilidade, esses sinais quando comparados com sinais de fala de WB (com largura de banda entre 50 e 7000 Hz) também apresentam perdas de desempenho em sistemas de ASR [22].

Visando explorar os benefícios do sistema de ABWE apresentado no Capitulo 5, assim como os recursos computacionais em servidores remotos, neste trabalho, sugerimos a inclusão de um sistema de ABWE como uma etapa antecessora ao *front-end* de um ASR, utilizando a metodologia NSR. Nesse contexto, a metodologia NSR surge como uma alternativa promissora por apresentar como principal vantagem a utilização dos recursos computacionais disponíveis em um servidor [61], tornando possível um processamento independente dos componentes de *hardware* dos aparelhos telefônicos.

Portanto, neste capítulo, uma estratégia para sistemas de ASR com metodologia de NSR é proposta e o desempenho do ASR com o sistema de ABWE é discutido. A eficácia da estratégia proposta é verificada através de avaliações objetivas da taxa de erro de palavras (*word error rate* - WER).

6.1 Fundamentos de Reconhecimento Automático de Fala

O objetivo de um sistema de ASR é estimar satisfatoriamente as informações contidas em um sinal de fala. Geralmente o principal procedimento desse sistema é a conversão do sinal de fala em texto. Esta seção apresenta uma breve introdução aos fundamentos de ASR.

6.1.1 Arquitetura

Um sistema típico de ASR é composto por cinco blocos principais: *front-end*, dicionário fonético, modelo acústico, modelo de linguagem e decodificador, sendo que os quatro últimos blocos compõem uma estrutura usualmente chamada *back-end*, conforme ilustrado na Figura 27.

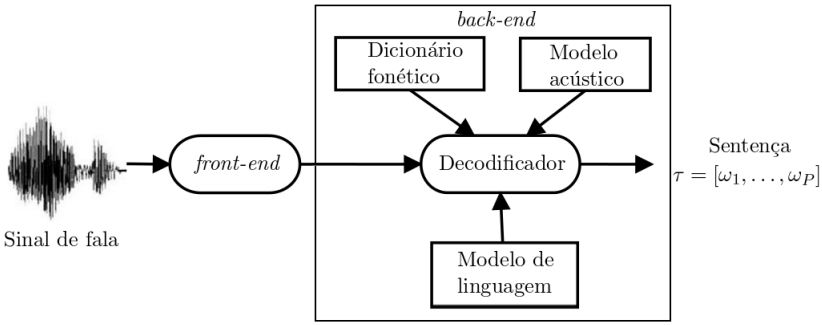


Figura 27 – Diagrama de blocos de um sistema típico de ASR.

O *front-end* extrai segmentos (quadros) a partir do sinal de fala e parametriza cada segmento em um vector $\mathbf{x}$ de dimensão L. Supõe-se aqui que T quadros são organizados em uma matriz $\mathbf{X}$, de dimensão $L \times T$, para representar uma sentença. O modelo de linguagem de um sistema de ASR fornece a probabilidade $p(\tau)$ de ocorrer uma sentença $\tau = [\omega_1, \ldots, \omega_P]$ de P palavras. Conceitualmente, o decodificador visa encontrar a sentença τ^* que maximize a probabilidade *a posteriori* dada por

$$\tau^* = \arg\max_{\tau} p(\tau|\mathbf{X}) = \arg\max_{\tau} \frac{p(\mathbf{X}|\tau)p(\tau)}{p(\mathbf{X})} \tag{6.1}$$

onde $p(\mathbf{X}|\tau)$ representa a verossimilhança acústica entre a matriz de observação $\mathbf{X}$ e as palavras da sentença τ. Essa verossimilhança é determinada por

um modelo acústico previamente treinado. Visto que $p(\mathbf{X})$ não depende de τ, (6.1) é equivalente a

$$\tau^* = \arg\max_{\tau} p(\mathbf{X}|\tau)p(\tau) \tag{6.2}$$

Devido ao grande número de possíveis sentenças, (6.2) não pode ser computada independentemente para cada sentença candidata. Portanto, os sistemas de ASR geralmente usam estruturas de dados, tais como árvores lexicais hierárquicas, quebrando sentenças em palavras e palavras em unidades básicas como fones ou trifones [22]. O mapeamento das palavras para as unidades básicas e vice-versa é realizado através de um dicionário fonético.

Em resumo, após o treinamento dos modelos, o ASR na fase de teste, usa o *front-end* para converter o sinal de entrada em atributos discriminativos e usa o *back-end* para estimar a sentença τ mais compatível ao contexto linguístico e ao sinal de entrada $\mathbf{X}$. Dessa forma, quanto melhor a qualidade do sinal de fala, maior o desempenho esperado do ASR. Nesse contexto, sistemas de ASR que utilizam sinais de fala de WB, com frequência de amostragem geralmente igual a 16 kHz apresentam em média desempenho 5% superior a sistemas que utilizam sinais amostrados em 8 kHz, como é o caso dos sinais provenientes da PSTN [22].

6.1.2 Extração de Atributos

Tendo em vista a obtenção de atributos ótimos discriminativos do sinal de fala, diversas alternativas para parametrizar as formas de onda desses sinais são utilizadas [22], [21]. Dentre elas, a parametrização utilizando coeficientes MFCC vem mostrando-se bastante eficaz e é comumente usada no bloco de *front-end* do ASR [22]. Essa técnica de análise utiliza a escala mel, expressa por

$$m = M(f) = 1125 \cdot \ln\left(1 + \frac{f}{700}\right) \tag{6.3}$$

onde f representa os componentes de frequência (em Hz).

O procedimento de extração de atributos consiste em dividir o espectro do sinal em B bandas com frequências centrais igualmente espaçadas na escala mel. Essas bandas de frequências são distribuídas através de bancos de filtros triangulares e, para cada banda, são computados os valores do logaritmo da energia e a transformada discreta do cosseno (*discrete cosine transform* - DCT). Os valores resultantes compõem os coeficientes MFCC, como ilustrado no diagrama da Figura 28.

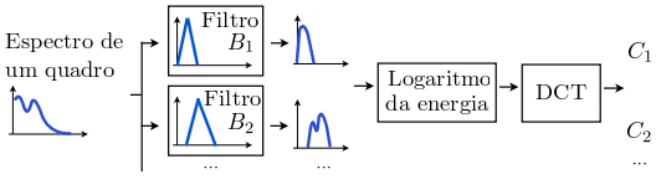


Figura 28 – Diagrama de bolcos ilustrativo da extração de atributos MFCC.

Assim, a distribuição (adequada) dos componentes de frequência nas B bandas é fundamental para o cálculo dos atributos ótimos discriminativos. Portanto, a largura de banda do espectro do sinal é diretamente proporcional à qualidade dos atributos MFCC. Então, por apresentarem uma maior faixa de frequência, sinais de WB possuem atributos de maior resolução na análise MFCC quando comparados a atributos obtidos a partir de sinais de NB [51].

Sinais de NB podem ser realçados através de algoritmos de ABWE que possibilitam a estimação de componentes de frequência adicionais capazes de ampliar a largura de banda do espectro. Assim, os componentes de frequência originais de NB e os estimados de WB podem ser redistribuídos eficientemente nos B bancos de filtros triangulares (da análise MFCC). A Figura 29 ilustra o processo de distribuição dos espectros de NB e de WB nos filtros triangulares correspondentes as B bandas de frequência.

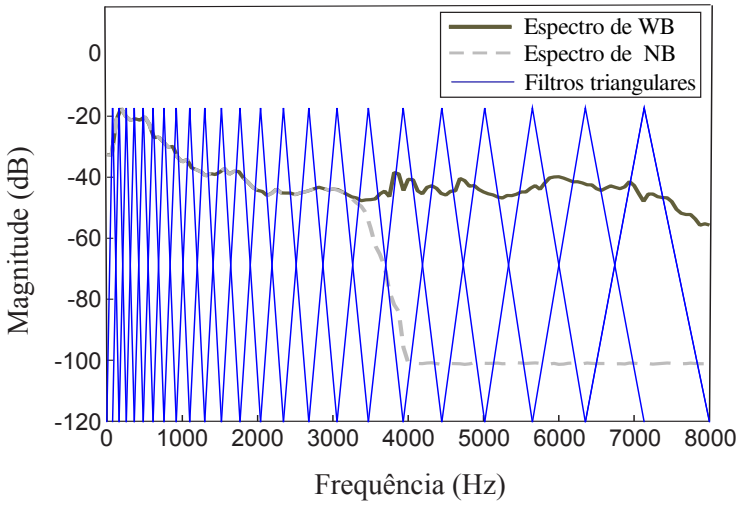


Figura 29 – Distribuição dos espectros de NB e de WB nos B bancos de filtros triangulares.

6.2 Estratégia NSR usando ASR com ABWE

No cenário de telefonia, sistemas NSR são adequados à utilização dos recursos computacionais de um servidor remoto. Assim, técnicas de processamento de sinais, tal como o realce da fala, podem ser melhor exploradas. Nesse contexto, este trabalho de pesquisa propõe uma estratégia em que o sinal de NB s_{NB}, fornecido pela PSTN, é realçado através de procedimentos de ABWE, gerando assim um sinal de WB sintético $\hat{s}_{\mathrm{WB}}$ na entrada do ASR, como ilustrado na Figura 30.

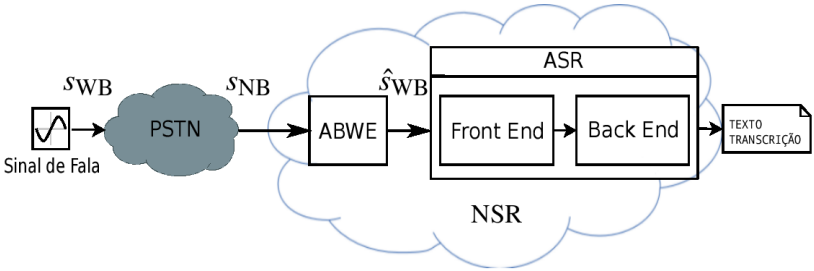


Figura 30 – Diagrama de blocos do ASR com ABWE em um sistema de NSR.

Portanto, o sistema NSR será composto pelos blocos de ABWE, seguido dos blocos do ASR: *front-end* e *back-end*.

6.2.1 Construção de Modelo Estatísticos para o ASR

Em (6.2), $p(\mathbf{X}|\tau)$ e $p(\tau)$ são determinados a partir dos modelos acústicos e de linguagem MA e ML, respectivamente. A estimação de um MA satisfatório é considerada a etapa mais desafiadora do projeto de um sistema de ASR. Nesse contexto, os procedimentos de ABWE discutidos nos capítulos anteriores são responsáveis pelo realce na matriz de observação $\mathbf{X}$, migrando de um espaço de dados de NB para WB. Assim, os atributos MFCC analisados pelo *front-end* contém informações espectrais mais discriminativas. Neste trabalho, utilizamos o software HTK [59] para construir o MA usando modelos escondidos de Markov (*hidden Markov models* - HMM) de acordo com as etapas descritas em [62]. Para a construção do ML, adotou-se a ferramenta MITLM [63].

Abaixo seguem alguns detalhes sobre as configurações utilizadas:

- Comprimento do quadro igual a 20 ms com sobreposição de 10 ms.

- Coeficientes por quadro contendo valores de energia, 12 coeficientes cepstrais em escala mel e, suas primeiras e segundas derivadas.

- MA baseado em HMM contínua de 5 estados e topologia esquerda-direita, constituídas por modelos trifônicos *cross-word* computados a partir de 38 monofones.

- ML baseado em tri-gramas treinados com 1534980 sentenças e utlizando a técnica de suavização de *Kneser-Ney* [63].

6.3 Resultados e Análise de Desempenho

Na maioria das aplicações de ASR, a figura de mérito usada para avaliar tais sistemas é a taxa de erro de palavra (*word error rate* - WER), definida como

$$WER = \frac{D+R}{W} \times 100\% \tag{6.4}$$

onde W é o número de palavras na sequência de entrada, e R e D são, respectivamente, o número de erros de substituição e deleção na sequência de palavra reconhecida quando comparado com a sequência correta.

Para que seja possível a obtenção de taxas aceitáveis de precisão, o estado da arte de ASR, usando MA baseados em HMM, necessita de um banco de dados de áudio (*corpora*) com grande variabilidade acústica, e os ML necessitam de milhões de sentenças para uma modelagem precisa do idioma em questão. Dessa forma, a disponibilidade de *corpora* é um fator primordial para o desenvolvimento de MA e ML precisos. Um *corpora* típico possui arquivos de fala com as suas transcrições associadas, grande quantidade de textos escritos e um dicionário fonético. Para atender tais requisitos, o desenvolvimento da etapa de ASR utiliza os *corpora* de fala e texto disponibilizados em [50] e [64].

Para a análise de resultados, o corpora *LapsBenchmark* [64] é adotado e os resultados das WER, definada em (6.4), são utilizados na avalização de três diferentes sistemas: sistemas de ASR com sinais de NB provenientes da PSTN; ASR com sinais de WB; e NSR com sistemas de ABWE. Os resultados obtidos com os dois primeiros sistemas, usando sinais de NB provenientes da PSTN e sinais de WB, são utilizados como referência para a avaliação do desempenho da estratégia de NSR com ABWE (NSR-ABWE) proposta neste trabalho.

A Tabela 9 apresenta o desempenho dos sistemas de referência (NB-PSTN e WB), assim como o desempenho do NSR-ABWE proposto, no qual os sinais de WB são sintetizados através do sistema de ABWE proposto na

Seção 5.1.

Tabela 9 – Descrição dos testes realizados

Duração do corpora	
Treinamento	Teste
4 horas	54 minutos
Desempenho do ASR	
Sistemas	**WER**
WB Original	27,77 %
NB-PSTN	30,69 %
NSR-ABWE	**29,37 %**

Como apresentado na Tabela 9, a estratégia de NSR-ABWE proposta neste trabalho apresenta um desempenho intermediário entre os sistemas de ASR usando NB-PSTN e WB-Original, resultando em ganho absoluto de $1,32\%$ na WER quando comparado à aplicação direta de sinais de NB provenientes da PSTN (NB-PSTN).

6.4 Conclusões

Neste trabalho de pesquisa, uma estratégia de NSR utilizando ABWE foi apresentada. Essa estratégia implementa o realce do sinal de NB (de entrada) e resulta em um sinal estimado de WB capaz de fornecer (artificialmente) atributos cepstrais mais discriminativos ao ASR. O sinal de WB estimado (de entrada) proporcionou maior riqueza espectral ao treinamento de modelos acústicos, quando comparado ao treinamento usando suas versões de NB provenientes da PSTN (resultando em MA mais precisos). Resultados de avaliações objetivas apresentados na Tabela 9 ratificam tais afirmações e confirmam a eficácia da estimação dos novos componentes de frequência do sinal sintetizado de WB $\hat{s}_{\mathrm{WB}}$, bem como a distribuição apropriada desses componentes nas B bandas de frequência da análise MFCC.

Capítulo 7

Conclusão e Comentários Finais

Neste trabalho de pesquisa, foi apresentada uma nova estratégia para implementação de sistemas de ABWE para sinais de fala aplicados à PSTN. Especificamente, a estratégia proposta foi baseada em um procedimento de classificação fonética de sinais de fala de NB com energia concentrada originalmente em altas frequências. Para tal, técnicas de seleção de atributos aplicadas a sinais de fala de banda limitada foram investigadas visando aprimorar a classificação em BGP com ênfase na discriminação de fonemas pertencentes ao grupo fonético fricativo. Adicionalmente, também foi discutida a integração do sistema de ABWE proposto em sistemas de ASR para o PB aplicados à PSTN. Neste capítulo, os principais pontos apresentados nesta dissertação são discutidos e algumas propostas de trabalhos futuros são sugeridas.

7.1 Sumário e Discussão dos Resultados

Conforme discutido nesta dissertação, a implementação de sistemas de ABWE se apresenta como uma alternativa interessante para contornar as limitações da comunicação em NB proveniente da PSTN. Dentre as diversas implementações de sistemas de ABWE encontradas na literatura e utilizadas em aplicações práticas, destacam-se aquelas que não necessitam de *side information*, por manterem a compatibilidade com as redes de telefonia existentes e com os principais *codecs* de NB e de WB. Tais implementações, entretanto, apresentam uma importante degradação de desempenho nos casos em que o sinal de fala contém energias concentradas em altas frequências. Para contornar essa deficiência, no Capítulo 4 foram discutidos os grupos fonéticos correspondentes ao PB e enfatizou-se a classificação em BGP com destaque para a discriminação de fonemas pertencentes ao grupo fonético fricativo, por apresentarem originalmente (em seus correspondentes sinais de WB) energia concentrada em altas frequências. Ainda neste capítulo, os efeitos da limitação de banda na classificação fonética de sinais de fala foram investigados e um conjunto de atributos característicos foi inicialmente avaliado. Posteriormente, uma estratégia para classificação fonética de sinais de fala de banda limitada foi proposta. Essa estratégia baseou-se na seleção de algoritmos de aprendizagem de máquina e de atributos característicos aplicados a uma

classificação fonética hierárquica. A partir dos resultados de simulação, utilizando as técnicas de seleção de atributos MI e CFS, foram investigados os algoritmos de aprendizagem de máquina mais adequados à tarefa de classificação fonética, bem como os atributos mais relevantes para a distinção entre os grupos fonéticos aqui considerados.

No Capítulo 5, com base nos resultados apresentados no Capítulo 4, uma nova estratégia para implementação de sistemas de ABWE utilizando classificação fonética para os sinais de NB foi proposta. Comparações de desempenho entre a estratégia de implementação proposta e a implementação de sistemas de ABWE sem classificação fonética foram realizadas, considerando simulações efetuadas no ambiente de desenvolvimento Matlab® e Octave®. Nessas simulações, diferentes cenários de operação foram analisados, a saber: um sistema usando codificação em WB, um sistema utilizando PSTN sem o uso de ABWE, um sistema utilizando PSTN com o uso de ABWE mas sem classificação fonética, e, finalmente, um sistema utilizando PSTN com a ABWE implementada a partir da estratégia proposta neste trabalho de pesquisa. A partir dos resultados de avaliações subjetivas e objetivas, verifica-se que, quando comparados os sinais sintetizados com suas versões originais de WB, ainda se nota a ausência de maior "brilho" no sinal de fala. Entretanto, quando comparado com sua versão convencional de NB sem tratamento de ABWE, pode ser verificada uma melhoria significativa da qualidade do sinal de fala reconstruído. Ainda com base nos resultados de simulações, a estratégia proposta proporcionou sinais sintetizados de WB significativamente melhores do que os sinais sintetizados a partir de implementações de ABWE convencionais sem classificação fonética.

Tendo em vista a versatilidade do sistema de ABWE aqui desenvolvido, no Capítulo 6 foi investigada a integração do sistema de ABWE em um sistema de ASR. Para tal, uma estratégia de NSR utilizando o sistema proposto de ABWE foi apresentada. Essa estratégia efetuou o realce dos sinais de NB (de entrada) e resultou em sinais estimados de WB capazes de fornecer (artificialmente) atributos mais discriminativos ao sistema de ASR. Especificamente, os sinais de WB estimados (de entrada) proporcionaram maior riqueza espectral ao treinamento de modelos acústicos, quando comparado ao treinamento usando suas versões de NB provenientes da PSTN (resultando em MA mais precisos). Assim, os resultados de simulações apresentados no Capítulo 6, também confirmam a eficácia da estimação dos componentes adicionais de frequência no sinal sintetizado de WB $\hat{s}_{\mathrm{WB}}$ obtido através da estratégia de implementação de sistemas de ABWE proposta neste trabalho de pesquisa.

Diante das contribuições apresentadas, acredita-se que os objetivos inicialmente estabelecidos foram alcançados com êxito durante a realização

do presente trabalho. As estratégias de implementações mostraram ser capazes de elevar tanto a qualidade subjetiva dos sinais de fala, quanto a qualidade objetiva medida através de diversas funções objetivo, incluindo o ganho de desempenho nos sistemas de ASR aplicados à PSTN. Portanto, acredita-se que tais características validam a aplicabilidade (nas atuais redes de telefonia) das estratégias formuladas neste trabalho.

7.2 Sugestões para Trabalhos Futuros

Como ideias de trabalhos futuros, sugere-se que a estratégia de implementação de sistemas de ABWE proposta nesta dissertação seja testada em aplicações práticas, visando assim uma análise de desempenho em condições reais de operação. Outra sugestão é a otimização e investigação de novos algoritmos de aprendizagem de máquinas para a tarefa de classificação fonética em níveis hierárquicos. Além disso, também pode ser considerada a utilização de uma função alternativa à função de mapeamento do trato vocal (mapeamento via *codebook*) usada na estratégia proposta, dada a restrição da dinâmica do segmento de fala limitada pela quantidade de *codewords*. Finalmente, considerando a integração de sistemas de ABWE em sistemas de ASR, a investigação de outros *front-ends* (além do MFCC) pode constituir um interessante tópico de pesquisa para a área de reconhecimento de fala.

APÊNDICE – Metodologia de Avaliação da Qualidade Subjetiva dos Sinais de Fala

Neste apêndice, de acordo com a Recomendação ITU-T P.800 [54], são descritos os métodos para a obtenção da qualidade subjetiva dos sinais de fala resultantes dos sistemas apresentados no Capítulo 5 deste trabalho de pesquisa.

A.1 Testes de Audição

Os testes de audição visam mensurar a capacidade de um sistema em transmitir adequadamente uma informação. Eles se baseiam em avaliações da qualidade subjetiva realizadas a partir da audição dos sinais de fala provenientes dos sistemas sob avaliação e dos sistemas de referência. Nesse contexto, os ouvintes devem receber apenas as instruções quanto à escala de avaliação a ser utilizada. Particularmente, a média aritmética dos valores obtidos nas avaliações de sinais de fala individuais é denominada "média de opiniões" (*mean opinion score* - MOS) e a média aritmética da comparação entre pares de sinais de fala é denominada "média de opiniões comparativas" (*comparative mean opinion score* - CMOS).

A.1.1 Classificação por Categoria Absoluta

Os testes de classificação por categoria absoluta (*absolute category rating* - ACR) se baseiam na avaliação direta da qualidade do sinal de fala, sem que o ouvinte (avaliador) disponha de material para comparação. Neste trabalho, são utilizadas duas estratégias de ACR:

- **Qualidade do sinal de fala**. Nesse teste, são analisados pequenos grupos de sinais de fala de curta duração. A Tabela 10 define a escala de avaliação considerada.

Tabela 10 – Escala de avaliação da qualidade de audição

Pontuação	Qualidade do sinal de fala
5	Excelente
4	Bom
3	Regular
2	Ruim
1	Péssimo

- **Esforço necessário para compreensão**. Nesse teste, há um maior interesse na avaliação subjetiva da inteligibilidade do sinal de fala do que de sua qualidade. A Tabela 11 apresenta a escala de avaliação utilizada.

Tabela 11 – Escala de avaliação do esforço de audição

Pontuação	Esforço necessário para compreensão
5	Sem esforço
4	Esforço mínimo
3	Esforço moderado
2	Esforço considerável
1	Esforço máximo (sem compreensão)

A.1.2 Classificação por Categoria de Comparação

Os testes de classificação por categoria de comparação (*comparison category rating* - CCR) avaliam pares de sinais de fala. No CCR, os ouvintes devem avaliar a qualidade do segundo sinal de fala com respeito ao primeiro. A Tabela 12 define a escala de avaliação utilizada.

Tabela 12 – Escala de comparação

Pontuação	**Qualidade do 2º em comparação ao 1º**
$+3$	Muito melhor
$+2$	Melhor
$+1$	Pouco melhor
0	Igual
-1	Pouco pior
-2	Pior
-3	Muito pior

A.2 Resultados Obtidos

A partir da análise das avaliações de 11 ouvintes, os resultados dos testes ACR e CCR são apresentados a seguir.

A.2.1 Avaliação MOS

Tabela 13 – Desempenho dos sistemas segundo a avaliação MOS

Sistemas	MOS	
	Qualidade de audição	Esforço de audição
WB original	$4,73$	$4,91$
NB sem ABWE	$2,82$	$4,09$
ABWE sem classificação	$3,36$	$4,36$
ABWE com 9 classes fonéticas	$3,82$	$4,45$

A.2.2 Avaliação CMOS

Tabela 14 – Desempenho dos sistemas segundo a avaliação CMOS

Qualidade do 2º em comparação ao 1º				
2º Sistema	1º Sistema			
	WB original	NB s/ ABWE	ABWE s/ class.	ABWE c/ class.
WB original	$0,00$	$2,00$	$1,64$	$1,36$
NB s/ ABWE	$-2,18$	$0,00$	$-1,73$	$-1,73$
ABWE s/ class.	$-1,18$	$1,27$	$0,00$	$-0,36$
ABWE c/ class.	$-0,91$	$1,45$	$0,73$	$0,00$

Referências Bibliográficas

[1] ITU-T, *Recomendation G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*. Geneve, Switzerland: Int. Telecomm. Union, 1996.

[2] ——, *Recomendation G.729.1: G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729*. Geneve, Switzerland: Int. Telecomm. Union, 2006.

[3] B. Iser, W. Minker, and G. Schmidt, *Bandwidth Extension of Speech Signals*, 1st ed. New York, NY, USA: Springer, 2008.

[4] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?" *IEEE Communications Magazine*, vol. 44, no. 5, pp. 106–111, May 2006.

[5] H. Pulakka, "Development and evaluation of artificial bandwidth extension methods for narrowband telephone speech," Ph.D. dissertation, Aalto University, Helsinki, Finland, 2013.

[6] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, South Brisbane, Australia, Apr. 2015, pp. 4395–4399.

[7] ITU, "Paired comparasion test of wide-band and narrow-band telephony," ITU-T, Tech. Rep. COM 12-9-E, Mar. 1993.

[8] T. K. Patel and P. Shrivastav, "Implementation of ITU-T G.729.ev wideband speech coder," *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, vol. 2, no. 3, pp. 27–32, Jun. 2012.

[9] B. Iser, P. Jax, P. Vary, H. Taddei, and S. Schandl, "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 15, no. 8, pp. 2496–2509, Nov. 2007.

[10] Y. Wang, S. Zhao, Y. Yu, and J. Kuang, "Speech bandwidth extension based on gmm and clustering method," in *Proc. International Conference on Communication Systems and Network Technologies (CSNT)*, vol. 1, Abr. 2015, pp. 437–441.

[11] M. Ghaderi and M. H. Savoji, "Wideband speech coding using adpcm and a new enhanced bandwidth extension method," in *Proc. IEEE International Symposium on Intelligent Signal Processing*, Floriana, Malta, Sept. 2011, pp. 1–4.

[12] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Philadelphia, PA, USA, Mar. 2005, pp. 805–808.

[13] D. M. Mohan, D. B. Karpur, M. Narayan, and J. Kishore, "Artificial bandwidth extension of narrowband speech using gaussian mixture model," in *Proc. Communication and Signal Processing (ICCSP)*, Calicut, India, Feb. 2011, pp. 410–412.

[14] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve, "The effect of highband harmonic structure in the artificial bandwidth expansion of telephone speech," in *Proc. International Conference on Speech Communication (INTERSPEECH)*, Antwerp, Belgium, Aug. 2007, pp. 2497–2500.

[15] T. Selvi and J. Pragatheeswaran, "Efficient speech enhancement technique by exploiting the harmonic structure of voiced segments," in *Proc. International Conference on Recent Trends Information Technology (ICRTIT)*, Chennai, Tamil Nadu, Jun. 2011, pp. 764–769.

[16] U. Koprnagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 86, no. 6, pp. 1296–1306, Jun. 2006.

[17] C. Yagli and E. Erzin, "Artificial bandwidth extension of spectral envelope with temporal clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Prague, Czech Republic, May 2011, pp. 5096–5099.

[18] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Dallas, TX, USA: Boca Raton: CRC Press, 2007.

[19] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[20] ITU-T, *Recomendation P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*. Geneve, Switzerland: Int. Telecomm. Union, 2007.

[21] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, 1st ed. Englewood Cliffs, NJ, USA: Pretice-Hall, 1978.

[22] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[23] T. C. Silva, *Fonética e Fonologia do Português - Roteiro de Estudos e Guia de Exercícios*. São Paulo, Brasil: Editora Contexto, 2010.

[24] M. O. Júnior, "Conversão do contorno de pitch por divisão de componentes para aplicação em sistemas de conversão de voz," Master's thesis, Pós-graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil, 2009.

[25] G. Miet, "Towards wideband speech by naeeowband speech bandwidth extension: magic effect or wideband recovery? application to mobile telephony." Ph.D. dissertation, Université du Maine, Le Mans, France, 2001.

[26] B. Iser and G. Schimidt, "Neural network versus codebooks in applications for bandwidth extension of speech signal." in *in Proc. European Conference on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 565–568.

[27] Y. Wang, S. Zhao, K. Mohammed, S. D. A. Bukhari, and J. Kuang, "Superwideband extension for amr-wb using conditional codebooks." in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 3695–3698.

[28] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies." in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, SaltLake City, UT, USA, May 2001, p. 665–668.

[29] S. Chen and H. Leung, "Speech bandwidth extension by data hiding and phonetic classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, Apr. 2007, pp. 15–20.

[30] Y. Qian and P. Kabal, "Wideband speech recovery from narrowband speech using classified codebook mapping." in *Proc. Australian International Conference on Speech Science and Technology*, Melbourne, Australia, Dec. 2002, p. 106–111.

[31] M. A. Hall, I. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[32] A. R. Webb and K. D. Copsey, *Statistical Pattern Recognition*, 3rd ed., I. John Wiley & Sons, Ed. San Francisco, CA, USA: Oxford University, Aug. 2011.

[33] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. New York, NY, USA: Springer Science, Nov. 2010.

[34] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proc. Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*, Tripolis, Greece, Jul. 2007, pp. 249–268.

[35] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2001.

[36] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions On Knowledge And Data Engineering*, vol. 15, pp. 1437–1447, Nov. 2003.

[37] A. Klautau, "Mining speech: automatic selection of heterogeneous features using boosting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, College Station, TX, USA, Apr. 2003, pp. 6–10.

[38] A. Frid and Y. Lavner, "Acoustic-phonetic analysis of fricatives for classification using svm based algorithm," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, Eliat, Israel, Nov. 2010, pp. 17–20.

[39] H. Hamooni and A. Mueen, "Dual-domain hierarchical classification of phonetic time series," in *Proc. IEEE International Conference on Data Mining (ICDM)*, Shenzhen, China, Dec. 2014, pp. 160–169.

[40] N. Morales, D. T. Toledano, J. H. L. Hansen, and J. Garrido, "Feature compensation techniques for asr on band-limited speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 758–774, May 2009.

[41] Y. He, J. Han, T. Zheng, and G. Zheng, "Compensation of partly reliable components for band-limited speech recognition with missing data techniques," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5476 – 479.

[42] J.-W. Lee, J.-Y. Choi, and H.-G. Kang, "Classification of stop place in consonant-vowel contexts using feature extrapolation of acoustic-phonetic features in telephone speech," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1536–1546, Dec. 2012.

[43] B. Chigier, "Phonetic classification on wide-band and telephone quality speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, May 1991, pp. 291–295.

[44] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada, May 2004, pp. 697–700.

[45] E. Silva and R. Seara, "Extensão artificial de largura de banda para sinais de fala usando classificação fonética," in *Anais do XXXI Simpósio Brasileiro de Telecomunicações (SBrT)*, Fortaleza, CE, Brasil, Set. 2013, pp. 1–5.

[46] P. Scanlon, D. P. E. Ellis, and R. B. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 15, no. 3, pp. 803–812, Mar. 2007.

[47] H. Huang, Y. Liu, L. ten Bosch, B. Cranen, and L. Boves, "Knowledge-based quadratic discriminant analysis for phonetic classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 25–30.

[48] S. Golub, "Classifying recorded music," Master's thesis, Division of Informatics, University of Edinburgh, Edinburgh ,UK, Nov. 2000.

[49] M. Algabri1, M. Alsulaiman, G. Muhammad, M. Zakariah, M. Bencherif, and Z. Ali, "Voiced and unvoiced classification using fuzzy logic," in *Proc.International Conference on Image Processing, Computer Vision and Pattern Recognition (IPCV)*, Las Vegas, NV, USA, Jul. 2015, pp. 416–420.

[50] C. A. Ynoguti and F. Violaro, "A brazilian portuguese speech database," in *Anais do XXXI Simpósio Brasileiro de Telecomunicações (SBrT)*, Rio de Janeiro, RJ, Brasil, Set. 2008, pp. 15–20.

[51] P. Bauer, J. Abel, V. Fisher, and T. Fingscheidt, "Automatic recognition of wideband telephone speech with limited amount of matched training data," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, Lisboa, Portugal, Sept. 2014, pp. 88–93.

[52] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, vol. 5, Yokohama, Japan, Sept. 94, pp. 1591–1594.

[53] L. M. T. de Jesus, "Acoustic phonetics of european portuguese fricative consonants," Ph.D. dissertation, Department of Electronics and Computer Science, University of Southampton, Southampton, England, 2001.

[54] ITU-T, *Recomendation P.800: Methods for subjective determination of transmission quality*. Geneve, Switzerland: Int. Telecomm. Union, 1996.

[55] A. Oliveira, E. Silva, H. Macedo, and L. Matos, "Brazilian portuguese speech-driven answering system," in *Proc. 6th Euro American Conference on Telematics and Information Systems (EATIS)*, Valencia, Spain, May 2012, pp. 1–8.

[56] N. Neto, C. Patrick, A. Klautau, and I. Trancoso, "Free tools and resources for brazilian portuguese speech recognition," *Journal of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, Mar. 2011.

[57] K. Livescu, E. Fosler-Lussier, and F. Metze, "Subword modeling for automatic speech recognition: Past, present and emerging approaches," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, Nov. 2012.

[58] D. O'Shaughnessy, "Acoustic analysis for automatic speech recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1038–1053, Abr. 2013.

[59] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book for HTK V3.4*, 1st ed. Cambridge, England: Cambridge University Press, 2006.

[60] D. Zaykovskiy, “Survey of the speech recognition techniques for mobile devices,” in *Proc. 11th International Conference on Speech and Computer (SPECOM)*, St. Petersburg, Russia, Jun. 2006, pp. 88–93.

[61] Y. Sunil and R. Sinha, “Exploration of class specific abwe for robust children’s asr under mismatched condition,” in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, Jul. 2012, pp. 1–5.

[62] E. Silva, L. Baptista, H. Fernandes, and A. Klautau, “Desenvolvimento de um sistema de reconhecimento automático de voz contínua com grande vocabulário para o português brasileiro,” in *Anais do XXV Congresso da Sociedade Brasileira de Computação*, São Leopolso, RS, Brasil, Jul. 2005, pp. 2258–2267.

[63] B.-J. Hsu and J. Glass, “Iterative language model estimation: Efficient data structure and algorithms,” in *Proc. International Speech Communication Association (INTERSPEECH)*, Brisbane, Australia, Sept. 2008, pp. 841–844.

[64] LaPS-UFPa. (2015) Grupo Fala Brasil. visited in nov., 2015. [Online]. Available: http://www.laps.ufpa.br/falabrasil