Pontifícia Universidade Católica
do Rio de Janeiro

Christian Dayan Arcos Gordillo

Realce e Reconhecimento de Voz Contínua em Ambientes Adversos

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós–graduação em Engenharia Elétrica da PUC-Rio.

Orientador : Prof. Marley Maria Bernardes Rebuzzi Vellasco
Co-orientador: Prof. Abraham Alcaim

Rio de Janeiro
Março de 2018

Pontifícia Universidade Católica
do Rio de Janeiro

Christian Dayan Arcos Gordillo

Realce e Reconhecimento de Voz Contínua em Ambientes Adversos

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós–graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Marley Maria Bernardes Rebuzzi Vellasco
Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Abraham Alcaim
Co-orientador
CETUC – PUC-Rio

Prof. Marco Antonio Grivet Mattoso Maia
CETUC – PUC-Rio

Prof.ª Mariane Rembold Petraglia
UFRJ

Prof. Fernando Gil Vianna Resende Junior
UFRJ

Prof. Paulo Roberto Rosa Lopes Nunes
IME

Prof. Márcio da Silveira Carvalho
Coordenador Setorial do Centro Técnico Científico – PUC-Rio

Rio de Janeiro, 23 de Março de 2018

Christian Dayan Arcos Gordillo

Graduou-se em Engenharia Eletrônica pela Universidade Francisco de Paula Santander (San José de Cúcuta, Colômbia) em 2010. Defendeu sua Dissertação de Mestrado em Sistemas de Telecomunicações em Março de 2013 pelo Departamento de Engenharia Elétrica da PUC-Rio. Trabalhou na empresa AGM Telecom no desenvolvimento de um sistema de reconhecimento de voz para deficientes auditivos.

Ficha Catalográfica

Arcos Gordillo, Christian Dayan

Realce e Reconhecimento de Voz Contínua em Ambientes Adversos / Christian Dayan Arcos Gordillo; orientador: Marley Maria Bernardes Rebuzzi Vellasco; co-orientador: Abraham Alcaim. – Rio de janeiro: PUC-Rio, Departamento de Engenharia Elétrica, 2018.

v., 179 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Engenharia Elétrica – Teses. 3. Reconhecimento de voz;. 4. realce de voz;. 5. robustez;. 6. máscara;. 7. histogramas;. 8. redes neurais profundas.. I. Rebuzzi Vellasco, Marley Maria Bernardes. II. Alcaim, Abraham. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

A Deus, pelos obstáculos que colocou neste longo caminho. Pois, quando as inclemências das dificuldades batiam na minha porta eu não compreendia, mas agora que os ventos cessaram e me encontro no topo da montanha, reconheço na paisagem a importância da lição.

A meus amados pais Guilmar Arcos e Luz Marina Gordillo fonte infinita de sabedoria e perseverança, por suas amorosas bênçãos que noite a noite recarregavam meu coração e por me demostrar que impossível é temporal, todos os meus triunfos são para vocês

A minha princesa Sarita, meu apoio, minha melhor amiga, minha alma gêmea, meu amor eterno.

Agradecimentos

Primeiramente e acima de tudo, dar infintas graças a **DEUS** por iluminar sempre o meu caminho, já que o resultado destes últimos quatro anos foram marcados por realizações diárias, que às vezes não dei o verdadeiro valor, mas sempre soube que a sua graça estava presente em todos os momentos deste árduo caminho de aprendizado.

Desejo expressar minha mais sincera gratidão ao Prof. Abraham Alcaim pelo tempo dedicado ao projeto, por suas valiosas intervenções, pela orientação, por sempre mostrar boa vontade e por confiar em mim desde o início e encorajar-me no trabalho que fiz durante esses anos. Foram 6 anos de aprendizado constante tanto no ambiente acadêmico quanto da vida em geral, compartilhando momentos prazerosos em que comemorávamos as nossas publicações ou simplesmente a nossa amizade. Para o senhor só tenho palavras de gratidão, salve professor.

Gostaria de agradecer ao Prof Bolsson. e à Prof.ª Marley por me dar o seu apoio e inestimável ajuda quando tudo parecia perdido.

A meus pais, por serem os principais promotores dos meus sonhos, seus conselhos seu amor e suas incontáveis historias de vida fazem com que cada dia acorde sendo uma melhor pessoa, nunca vou me cansar de escutar suas divertidas historias, que noite a noite alegravam meus intermináveis dias. O seu infinito amor tornou fácil os momentos que poderiam ter sido difíceis. Os AMO.

Ao meu amor eterno e princesa dos meus contos, é difícil expressar em estas poucas linhas o que você representa para mim: amizade, cumplicidade, paixão, alegria, fofura e tantos outros sentimentos que tomariam metade desta tese para descreve-los. Obrigado por teu infinito amor tua enorme paciência, teu apoio desinteressado e por acompanhar-me diariamente nesta longa aventura que estou preste a terminar, por tuas inesgotáveis palavras de encorajamento que me tiraram da tristeza nos momentos mais difíceis e por me mostrar dia após dia que o nosso amor vai além do tradicional. Tenho certeza que um a um dos nossos sonhos e projetos que temos para esta vida começara a se tornar realidade. Um futuro cheio de aventuras nos espera, te amo.

A Lorena Chamorro pelo imenso apoio em todos os aspectos da minha vida, pelas palavras que sempre tinha e que tanto me tocaram quando mais as precisava, por fazer-me acreditar que era possível e que não preciso de grandes coisas para fazer coisas grandes. Uma amizade dura uma média de 5 a 8 anos, se passar disso durara a vida inteira, é nóis.

Gostaria de agradecer também às organizações e agências de financiamento que possibilitaram minha pesquisa, incluindo o Centro de Estudos e

Telecomunicações CETUC e o departamento de engenharia elétrica da Pontifícia Universidade Católica de Rio de Janeiro (PUC–Rio), a CAPES, a FAPERJ e ao CNPq pelo apoio financeiro sem os quais este trabalho não poderia ter sido realizado.

Aos professores que participaram da banca examinadora.

Resumo

Arcos Gordillo, Christian Dayan; Rebuzzi Vellasco, Marley Maria Bernardes; Alcaim, Abraham. **Realce e Reconhecimento de Voz Contínua em Ambientes Adversos**. Rio de Janeiro, 2018. 179p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta tese apresenta e examina contribuições inovadoras no *front-end* dos sistemas de reconhecimento automático de voz (RAV) para o realce e reconhecimento de voz em ambientes adversos. A primeira proposta consiste em aplicar um filtro de mediana sobre a função de distribuição de probabilidade de cada coeficiente cepstral antes de utilizar uma transformação para um domínio invariante às distorções, com o objetivo de adaptar a voz ruidosa ao ambiente limpo de referência através da modificação de histogramas. Fundamentadas nos resultados de estudos psicofísicos do sistema auditivo humano, que utiliza como princípio o fato de que o som que atinge o ouvido é sujeito a um processo chamado Análise de Cena Auditiva (ASA), o qual examina como o sistema auditivo separa as fontes de som que compõem a entrada acústica, três novas abordagens aplicadas independentemente foram propostas para realce e reconhecimento de voz. A primeira aplica a estimativa de uma nova máscara no domínio espectral usando o conceito da transformada de Fourier de tempo curto (STFT). A máscara proposta aplica a técnica *Local Binary Pattern* (LBP) à relação sinal ruído (SNR) de cada unidade de tempo-frequência (T-F) para estimar uma máscara de vizinhança ideal (INM). Continuando com essa abordagem, propõe-se em seguida nesta tese o mascaramento usando as transformadas *wavelet* com base nos LBP para realçar os espectros temporais dos coeficientes *wavelet* nas altas frequências. Finalmente, é proposto um novo método de estimação da máscara INM, utilizando um algoritmo de aprendizagem supervisionado das *Deep Neural Networks* (DNN) com o objetivo de realizar a classificação de unidades T-F obtidas da saída dos bancos de filtros pertencentes a uma mesma fonte de som (ou predominantemente voz ou predominantemente ruído). O desempenho é comparado com as técnicas de máscara tradicionais IBM e IRM, tanto em termos de qualidade objetiva da voz, como através de taxas de erro de palavra. Os resultados das técnicas propostas evidenciam as melhoras obtidas em ambientes ruidosos, com diferenças significativamente superiores às abordagens convencionais.

Palavras-chave

Reconhecimento de voz; realce de voz; robustez; máscara; histogramas; redes neurais profundas.

Abstract

Arcos Gordillo, Christian Dayan; Rebuzzi Vellasco, Marley Maria Bernardes (Advisor); Alcaim, Abraham (Co-Advisor). **Enhancement and Continuous Speech Recognition in Adverse Environments**. Rio de Janeiro, 2018. 179p. Tese de doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This thesis presents and examines innovative contributions in front-end of the automatic speech recognition systems (ASR) for enhancement and speech recognition in adverse environments. The first proposal applies a median filter on the probability distribution function of each cepstral coefficient before using a transformation to a distortion-invariant domain, to adapt the corrupted voice to the clean reference environment by modifying histograms. Based on the results of psychophysical studies of the human auditory system, which uses as a principle the fact that sound reaching the ear is subjected to a process called Auditory Scene Analysis (ASA), which examines how the auditory system separates the sound sources that make up the acoustic input, three new approaches independently applied were proposed for enhancement and speech recognition. The first applies the estimation of a new mask in the spectral domain using the short-time Fourier Transform (STFT) concept. The proposed mask applies the Local Binary Pattern (LBP) technique to the Signal-to-Noise Ratio (SNR) of each time-frequency unit (T-F) to estimate an Ideal Neighborhood Mask (INM). Continuing with this approach, the masking using LBP-based wavelet transforms to highlight the temporal spectra of wavelet coefficients at high frequencies is proposed in this thesis. Finally, a new method of estimation of the INM mask is proposed, using a supervised learning algorithm of Deep Neural Network (DNN) to classify the T-F units obtained from the output of the filter banks belonging to a same source of sound (or predominantly voice or predominantly noise). The performance is compared with traditional IBM and IRM mask techniques, both regarding objective voice quality and through word error rates. The results of the proposed methods show the improvements obtained in noisy environments, with differences significantly superior to the conventional approaches.

Keywords

Speech recognition; speech enhancement; robustness; mask; histograms; deep neural networks.

Sumário

Lista de figuras

Lista de tabelas

1
Introdução

A partir das ideias e metodologias procedentes da teoria da informação, estatística, análise numérica, informática, processamento de sinais entre outras, surgiu desde os anos 60 uma área de pesquisa que envolve o estudo de sinais de voz e todos os métodos necessários para o seu processamento, chamada processamento de voz. O objetivo principal do processamento de voz é processar um conjunto finito de dados acústicos, obtidos através de sensores (microfones) e extrair as informações mais relevantes do sinal. Esse procedimento é levado a cabo mediante a implementação algorítmica de formulações matemáticas que modelam a capacidade humana para entender e processar o conteúdo da linguagem humana. Dentre as várias áreas de pesquisa abordadas no processamento de voz, algumas das mais representativas são:

- Reconhecimento automático de voz (RAV);
- Verificação e identificação de locutor;
- Processamento de voz para aprendizagem de línguas;
- Biometria por voz;
- Realce de voz;
- Codificação digital de voz;
- Transmissão de Voz sobre IP.

Um dos aspectos importantes na escolha das áreas acima expostas é definir quais informações do sinal capturado pelos microfones são relevantes. Por exemplo, a informação linguística será relevante se o objetivo é reconhecer a sequência de palavras produzidas pelo locutor. Nesse caso informações próprias do locutor (período fundamental) ou informações referentes ao ambiente acústico (ruído aditivo) serão irrelevantes para o sistema. Por outro lado, a informação do ambiente acústico será relevante se o objetivo é realçar o sinal de voz adquirida em ambientes adversos. Nesse caso será irrelevante a informação linguística.

Uma das áreas mais pesquisadas nas últimas décadas é o reconhecimento automático de voz (RAV), o qual situa-se dentro do marco mais geral do processamento de voz. O objetivo principal dos sistemas RAV é reproduzir de

forma automática o fluxo de informação da comunicação oral transformando-a em texto ou em comandos por voz para uma imensa variedade de aplicações [1], tais como:

- Acesso telefônico a sistemas de informação (*call centers*, *chatbots*);
- Ajuda para deficientes (pessoas tetraplégicas ou com dificuldade de movimento, deficientes auditivos, deficientes visuais);
- Interface amigável aos usuários de aplicações de realidade virtual;
- Transcrição automática de voz em texto (ditado);
- Assistentes virtuais inteligentes (*Alexa da amazon, Cloud Speech da google, S Voice e Bixby da samsung, Siri, fala abd tradução da Apple, Cortana da Midrosoft*) que possuem capacidade de interagir por voz, tocar músicas, fazer listas de tarefas, configurar alarmes, prover informações sobre tráfego, temperatura, entre outras informações.

Embora o reconhecimento de voz pareça ser uma tarefa muito simples para os seres humanos, tem-se visto ao longo das últimas décadas que é um processo muito complicado para as máquinas. Apesar dos grandes desenvolvimentos nos campos de informática, processamento digital de sinais e tecnologias da informação, o desempenho geral dos sistemas RAV, pode ser, em muitas situações, substancialmente menor que o dos seres humanos, já que segundo [2] existe uma série de problemas que incluem a alta variabilidade da voz [3], a ineficiência na modelagem da voz espontânea [4], a limitação da robustez dos sistemas RAV [5], entre outras.

Na atualidade o desempenho dos sistemas RAV melhorou consideravelmente devido a mais dados de treinamento, aumento do poder computacional e algoritmo de aprendizado profundo para modelagem acústica [6]. Por exemplo, os resultados que fornecem os melhores sistemas RAV desenvolvidos para o idioma inglês por parte do google atingem uma taxa de erro de palavras de $5,9\%$ em condições de laboratório, a mesma porcentagem que um transcritor humano profissional [7]. No entanto, esses sistemas ainda estão longe de atingir a mesma robustez dos seres humanos, já que em ambientes adversos as taxas de erro de palavra aumentam e o seu rendimento cai a níveis onde seu uso torna-se inaceitável, informou o cientista-chefe da Microsoft Research, Xuedong Huang. De acordo com esse cientista, *"The next frontier for voice recognition is to accurately transcribe speech even when it's coming over a lousy cell connection or an echoing McDonalds drive-thru speaker. Still has many challenges to address, such as achieving human levels of recognition in noisy environments with distant microphones, in recognizing accented speech,*

or speaking styles and languages for which only limited training data is available".

Ambientes adversos refere-se genericamente a todos aqueles que degradam o funcionamento do sistema de reconhecimento de voz. Basicamente, eles podem ser classificados em três grandes categorias: (i) fenômenos articulatórios devidos ao falante, como a velocidade de locução, estado de ânimo, entre outros; (ii) ambiente acústico, como ruído de fundo e efeitos de reverberação; (iii) características do canal de transmissão, como o microfone e a largura de banda disponível. Estes efeitos de distorção devido aos ambientes adversos serão analisados no Capítulo 2.

O processo de reconhecimento de voz baseia-se fundamentalmente em princípios de reconhecimento estatístico de padrões, onde os sinais acústicos são transformados em uma sequência de símbolos analisados e estruturados em unidades de sub-palavras (por exemplo, fones[1]), que os representem com a menor perda de informação possível. Os sistemas RAV são basicamente compostos por dois módulos: um $front-end$ (responsável pelo pré-processamento e extração de características ou atributos da voz) e um $back-end$ (responsável pelo reconhecimento a partir dos atributos extraídos através da comparação destes com uma série de padrões previamente estabelecidos)[8]. Mas para isso os sistemas exigem o desenvolvimento de dois estágios distintos: um de treinamento, em que os padrões são estabelecidos, e outro de testes para validar o sistema. O descasamento nesses dois estágios é principalmente devido ao fato de que a voz é apresentada de maneira diferente na fase de teste do que a prevista pela fase de treinamento. A fim de mitigar esse descasamento nos sistemas e atingir pequenas taxas de erro, nas últimas décadas tem-se proposto restrições nas tarefas de reconhecimento para assim limitar os problemas dos sistemas RAV. A Fig. 1.1 mostra as diferentes formas de restrição da voz impostas para cada tipo de reconhecimento. A medida que o sistema se afasta do centro do diagrama, encontram-se sistemas mais restritivos, enquanto os mais flexíveis são aqueles que cobrem uma superfície menor do diagrama.

De acordo com o diagrama de *Kiviat*, escolher o nível de reconhecimento segundo a necessidade do sistema é uma das dificuldades principais do reconhecimento automático de voz, devido que ele pode ser caracterizado por vários parâmetros, tais como palavras isoladas, palavras conectadas e de fala contínua. Esse último aumenta muito o nível de complexidade do reconhecedor. O reconhecedor deve ser capaz de lidar com limites temporais desconhecidos no sinal acústico, e de funcionar bem na presença de efeitos co-articulados e

[1]Unidade minima da palavra de características acústicas particulares e com duração típica.

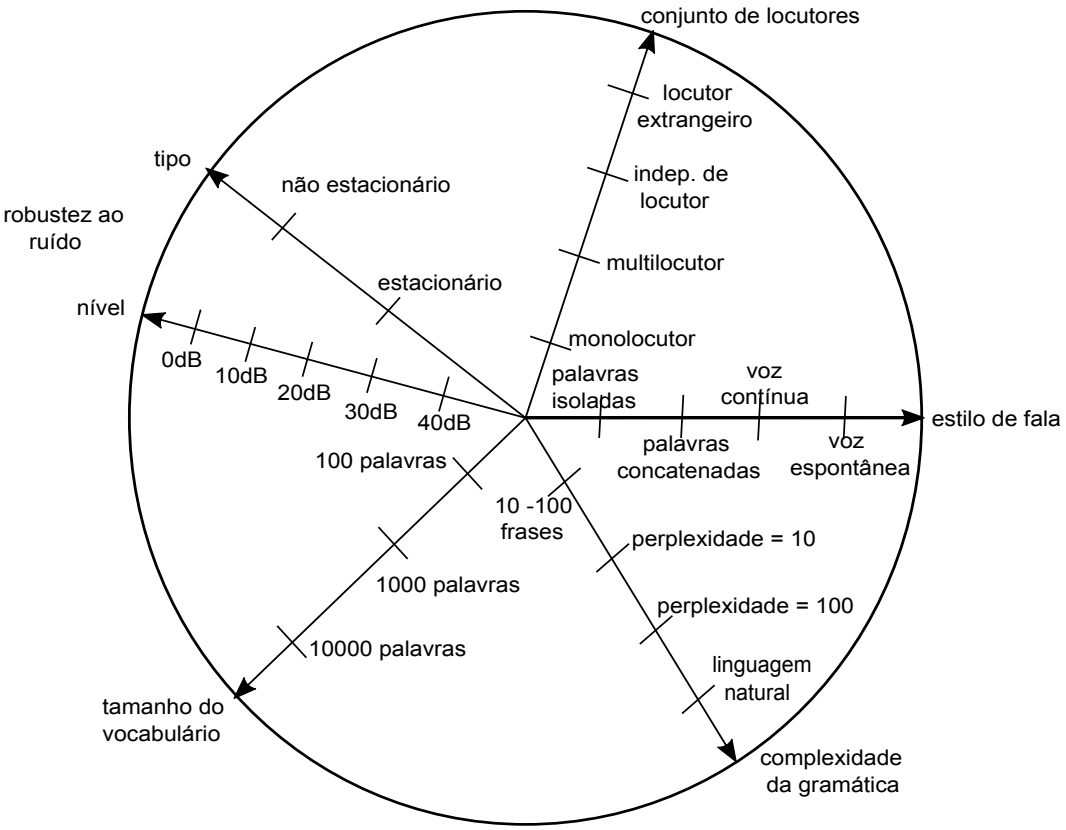

Figura 1.1: Diagrama de Kiviat da variabilidade de formas nas que pode-se representar o sinal de voz.

pronúncias descuidadas (inserções e omissões de fonemas) que são parte de um diálogo fluente e natural.

Na Tabela 1.1 apresenta-se uma visão global do diagrama de *Kiviat* e das variáveis que definem um sistema RAV e suas faixas de valores.

Parâmetro	**Variedade**
Forma de falar	Palavra isolada $\longleftrightarrow$ voz contínua
Estilo de fala	Texto lido $\longleftrightarrow$ Fala espontânea
Adaptação	Dependente de locutor $\longleftrightarrow$ Independente de locutor
Tamanho do vocabulário	Pequeno (< 20 palavras) $\longleftrightarrow$ Grande (>10.000 palavras)
Modelo da linguagem	Estados finitos $\longleftrightarrow$ Dependentes de contexto
Perplexidade	Pequena (<10) $\longleftrightarrow$ Grande (>100)
SNR	Alta (>40) $\longleftrightarrow$ Baixa(<0)
Transductor	Microfone de eliminação de eco $\longleftrightarrow$ Telefone

Tabela 1.1: Parâmetros típicos que caracterizam o sistema de reconhecimento de voz.

De acordo com esses problemas é difícil desenvolver um sistema que leve em conta todas as formas de se expressar de uma pessoa e de controlar a variabilidade dos diversos ambientes do mundo real. Por isso, nas últimas décadas tem-se desenvolvido técnicas e procedimentos com características específicas

que tentam cobrir, na medida do possível, o maior número desses problemas, visando dar uma solução mais geral aos sistemas RAV. Por exemplo, os sistemas de reconhecimento robusto, segundo o diagrama da Fig 1.1, focam sua atenção apenas em criar sistemas que estejam afetados pela variabilidade do ruído (nível, tipo de ruído), mas podem não considerar aspectos como modo de pronunciação, conjunto de locutores, tamanho do vocabulário e complexidade da gramática. Por outro lado, sistemas focados em reconhecimento de voz contínua englobam aspectos do diagrama *kiviat*, tais como tamanho do vocabulário, complexidade da gramática e modo de pronunciação. Embora essas duas abordagens tenham sido muito estudadas nas últimas décadas, a fusão entre elas para criar sistemas que trabalhem de maneira conjunta é relativamente nova [9][10].

Nas seguintes seções serão apresentadas, de maneira resumida a estrutura geral de um sistema RAV e as principais caraterísticas que o conformam, como o modelo de linguagem e a modelagem acústica.

1.1 Estrutura do sistema de reconhecimento de voz

Para entender as técnicas propostas mais adiante neste trabalho, é importante conhecer como funcionam os sistemas RAV. Esta seção descreve brevemente a estrutura empregada pelos sistemas RAV baseados em modelos ocultos de Markov (HMMs). Como foi acima mencionado, os sistemas RAV são divididos em dois blocos principais, como mostrado na Fig 1.2, e que serão brevemente analisados no restante deste capítulo. Uma revisão mas detalhada sobre as técnicas dos sistemas RAV e suas abordagens podem ser encontradas em [1][11][12][13][14].

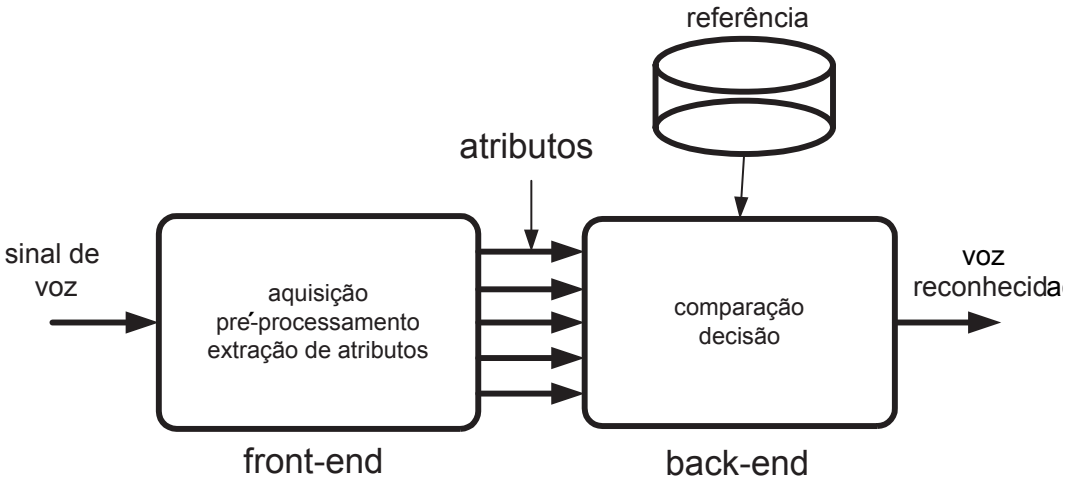

Figura 1.2: Diagrama de blocos geral de um sistema de reconhecimento.

1.1.1 Front-end

O objetivo principal do *front-end* é proporcionar uma representação paramétrica[2] do sinal de voz de entrada o suficientemente relevante para a identificação da mesma, removendo a redundância e a informação relacionada às fontes de variabilidade que ela tem.

Segundo [15], apesar da variedade de informações específicas do locutor e do ambiente acústico onde o sinal foi capturado, o conjunto de atributos deve ter características relevantes que permitam:

- Diferenciar os diversos fones da voz assim como também ter pouca variação dentro de um fone específico;
- Aportar dados sobre as variações da pronúncia, tais como sotaque, entonação, etc;
- Ser facilmente mensurável;
- Mostrar resistência ao disfarce;
- Não mudar ao longo do tempo.

Tendo em conta a informação acima exposta para poder caracterizar da melhor maneira possível os diferentes sons de cada língua, é preciso que os sistemas de reconhecimento façam uso da análise espectral de tempo curto. O objetivo é utilizar intervalos do sinal com propriedades estatisticamente constantes, quase estacionárias, que permitam analisar e processar o sinal de voz como um sinal estacionário. Uma forma mais detalhada de analisar o processo de parametrização é dividindo o *front-end* em três blocos principais, como apresentado na Fig. 1.3, onde cada bloco é composto por uma série de procedimentos que serão brevemente detalhados a seguir.

1.1.1.1 Pré-processamento

Para atingir a quase estacionariedade do sinal de voz, as seguintes etapas de pré-processamento são tipicamente aplicadas. Primeiro, é empregado um filtro digital passa-alta de primeira ordem sobre o sinal capturado, a fim de compensar os efeito dos pulsos glotais [5] e ressaltar as frequências dos formantes. Esse procedimento justifica-se por duas razões: (i) evitar a perda de dados durante o processo de segmentação, já que a maior parte da informação está contida nas frequências baixas e (ii) remover a componente

[2]Transformação do sinal de voz em uma forma compacta de atributos, que contém informações discriminativas da voz.

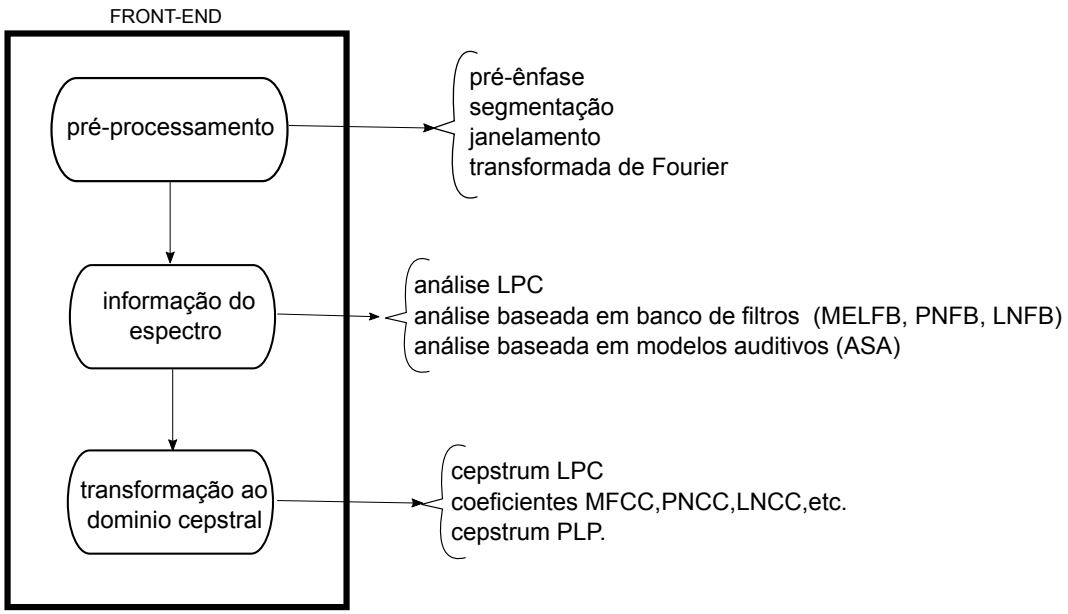

Figura 1.3: Processo de parametrização *Front-end*.

DC do sinal, aplainando-o espectralmente. O sinal resultante é segmentado em quadros superpostos com o objetivo de conseguir características de quase estacionariedade[16] e distinguir as partes do sinal que têm informação de voz daquelas que não têm, acompanhando assim as mudanças ao longo das frases.

Para obter uma estimativa espectral estável, o comprimento do quadro deverá ser ajustado convenientemente, de forma que seja o suficientemente curto para que apresente as características espectrais instantâneas da estimação e o suficientemente longo para minimizar a variância na estimação dos parâmetros causadas pelas variações do aparelho fonador[3]. Além disso, é altamente desejável minimizar a taxa de informação a enviar, com uma mudança (superposição) de quadro típica de 10ms [4] que permita capturar a maior parte da variabilidade temporal na voz.

Segmentar o sinal de voz traz o problema de descontinuidade ao início e ao final de cada quadro, devido ao fato de cada um começar e terminar bruscamente. Se simplesmente toma-se as amostras como estão em um segmento, quando aplica-se uma técnica de análise espectral como a transformada de Fourier, ela age como se estivesse operando em um sinal que é zero antes do início do segmento e, em seguida, salta bruscamente para o sinal durante o segmento e depois volta a zero quando o segmento termina. Isso introduz uma distorção significativa do sinal, fazendo com que pareça haver ruído de alta

[3]Em geral, blocos de 20 a 32 ms são adequados para a maior parte das aplicações, durante o qual a fala pode ser assumida quase que estacionária.

[4]Equivalente a uma taxa de quadros de 100 quadros por segundo a uma taxa de amostragem de 8000 amostras/segundo.

frequência no início e nos pontos finais de cada segmento. É necessário, então, diminuir esse efeito, multiplicando cada segmento por uma janela que seja adequada, visando suavizar as bordas do quadro até chegar a zero, e realçando a parte central para acentuar as propriedades características do segmento. No reconhecimento de voz, existem diferentes tipos de janelas. No entanto, a mais utilizada é a janela de Hamming[5] [17].

Finalmente é calculado o espectro de cada segmento da voz usando a transformada discreta de Fourier, com o objetivo de identificar características fonéticas e determinar sequências de fones no sinal original. As parametrizações utilizadas nos sistemas RAV são derivadas na sua totalidade a partir da análise espectral de potência dos segmentos de voz, as quais serão usadas para o treinamento dos HMMs de cada sub-unidade de palavra.

Por exemplo, a Fig 1.4 (a) e (c) mostra as formas de onda das vogais fechadas 'o' e 'u', respectivamente, no domínio do tempo. Pode-se ver que as formas de onda são quase similares, o que geraria um treinamento de HMMs parecido em ambos os casos, dificultando o processo de classificação no momento do reconhecimento. Visando solucionar este problema, a análise do sinal é passada para o domínio da frequência[6].

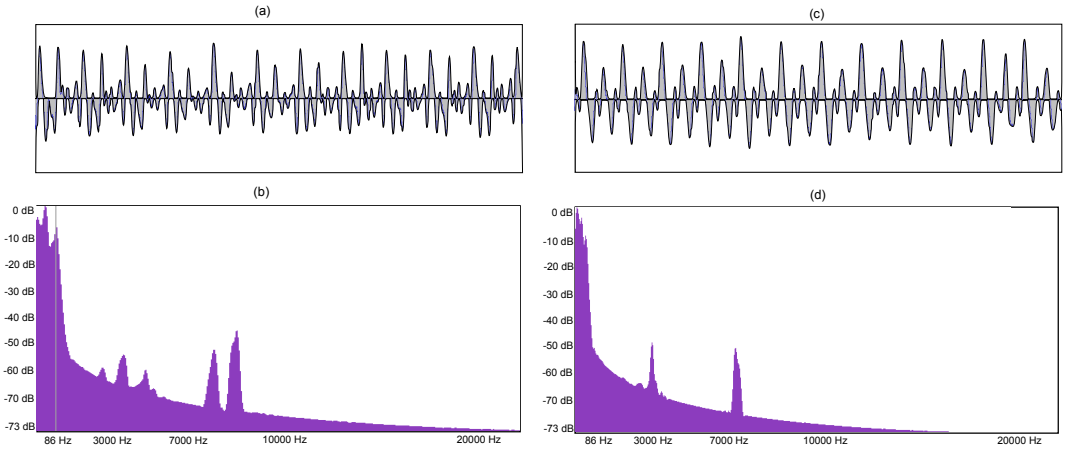

Figura 1.4: (a) e (c) formas de onda no domínio do tempo vogal 'o' e 'u', (b) e (d) espectro das vogais 'o' e 'u' respectivamente.

Em um modelo simples de produção de voz ela é caracterizada, no caso de sons sonoros, como as vogais, por excitação periódica de um sistema linear. Ou seja o espectro da voz é o produto do espectro de excitação e da resposta

[5]Janela que se adapta bem às características espectrais do sinal de voz, obtendo um bom compromisso entre a largura do lóbulo principal (resolução espectral) e as amplitudes dos lóbulos secundários.

[6]Domínio que permite a visão mais completa do sinal, trazendo facilidade na hora de analisar o seu comportamento e facilitando a aplicação de grande variedade de técnicas de filtragem digital.

em frequência do trato vocal [18]. Esse modelo gera vibrações que se repetem ao longo do tempo, e podem ser extraídas por meio da transforma discreta de Fourier (DFT) [19]. A Fig 1.4 (*b*) e (*d*) mostra como a diferença entre os dois sons fica mais clara agora: o 'o' é mais agudo e, portanto, apresenta um conteúdo de vibrações em frequências mais altas fazendo-o mais discriminativo na hora do treinamento do seu respectivo HMM.

1.1.1.2 Informação do espectro

Anteriormente, foi apresentado o bloco de pré-processamento de voz, onde o sinal de entrada é decomposto num conjunto de vetores no domínio da frequência, os quais fornecem informações significativas do espectro que caracterizara a configuração do trato vocal, permitindo uma boa diferenciação entre segmentos. Segundo [1], os seres humanos percebem o som em uma escala de frequência não linear. Por esse motivo para cada um dos vetores obtidos na etapa de pré-processamento estima-se a magnitude da densidade espectral de potência e passa-se por um banco de filtros com M canais distribuídos logaritmicamente e superpostos em uma escala perceptual auditiva (por exemplo, MEL, Bark, ERB, etc). O objetivo é obter uma representação do sinal mais suave e significativa sob o ponto de vista perceptual. Os vetores obtidos nas M saídas dos bancos de filtros são comprimidos usando um operador não linear, tipicamente o logaritmo neperiano, para modelar a sensibilidade perceptiva do ouvido humano.

A partir deste conjunto de vetores, obtém-se um conjunto de informações relevantes para os sistemas RAV, suprimindo as informações redundantes e as informações ligadas às fontes de variabilidade. Na literatura, existem várias técnicas de análise espectral. Dentre as mais comuns encontram-se as técnicas baseadas em modelos auditivos [20], análise baseada em banco de filtros [21][22][23][24] e análise LPC [25].

O objetivo das técnicas baseadas em *modelos auditivos* é fazer uma análise no domínio da frequência que seja consistente com as propriedades de seletividade do sistema auditivo humano. Segundo [26] os sons que atingem o ouvido humano estão sujeitos a um processo chamado análise de cena auditiva (ASA), que capta informações úteis como os contrastes temporais, as características não lineares do processo de audição, e a supressão lateral dos canais adjacentes. Um dos modelos de análise baseado em modelos auditivos mais conhecido é o *Perceptually-based Linear Prediction (PLP)*[27], onde se modela o espectro auditivo perceptual por meio de três conceitos da psicoacústica: (i) faz-se uma integração nas bandas críticas do espectro da potência da

voz, conseguindo assim o espectro auditivo; (ii) faz-se um pré-enfase sobre o espectro auditivo; (iii) aplica-se a raiz cúbica ou lei de intensidade potência, com o objetivo de transformar essa intensidade em uma medida perceptualmente mais significativa. Uma vez aplicadas essas três operações, o espectro resultante é analisado usando uma função *all-pole* [28]. Outro método baseado em modelos auditivos é o *Ensamble Interval Histogram (EIH)*[29]. O EIH oferece uma representação da voz com alta resolução espectral, através de um banco de filtros superior a 150 filtros que modelam a membrana basilar, onde a largura dos filtros cresce de forma não linear com a frequência central de cada filtro[7]. Finalmente tem-se os modelos auditivos síncronos [30]. Estes modelos usam um banco de filtros que aproxima as medidas fisiológicas da resposta da membrana basilar aos estímulos acústicos. No entanto, uma das grandes desvantagens dos modelos de análise baseados em modelos auditivos é ter um alto custo computacional.

Por outro lado, têm-se as técnicas baseadas na análise de banco de filtros. Elas consistem em um conjunto de filtros passa-baixas, cujas larguras de banda e espaçamentos são aproximadamente iguais às das bandas críticas[8] e cujo alcance das frequências centrais cobre as frequências mais importantes para a percepção da voz. Esses filtros imitam a filtragem feita pela cóclea no ouvido humano, é por isso que não devem ser lineares porque nas bandas de frequência mais baixas é onde ocorre a maior resolução espectral, que contém a informação mais relevante para a detecção de voz. Cada filtro ou canal modela a resposta da frequência associada a um ponto particular da membrana basilar [32]. Uma das técnicas referentes a esta categoria é a *Mel Frequency Spectral Coefficients (MFSC)*. Os coeficientes MFSC são o logaritmo da energia de saída dos filtros passa-faixa distribuídos na escala Mel. Isto é, a largura de banda dos filtros é incrementada de forma logarítmica a medida que aumenta sua frequência central. Assim, cada filtro fará uma média dos componentes espectrais presentes em sua banda, evitando a aparição de harmônicos a frequências múltiplas da frequência fundamental de cada segmento.

As diferentes escalas auditivas e diferentes distribuições dos filtros auditivos propostas na literatura [33][34] dão origem a novas técnicas de análise de banco de filtros com os mesmos conceitos dos MFSC. Por exemplo em [23] foi proposta uma nova técnica chamada *Locally Normalized Filter Banks* baseada nos modelos *Seneff* os quais realizam uma normalização local no domínio da frequência em cada canal auditivo [35]. Essa normalização local é feita di-

[7]Exemplos de escalas perceptuais para os bancos de filtros são: Escala MEL, escala BARK, escala ERB, etc.

[8] As bandas críticas referem-se às faixas de frequência correspondentes às regiões da membrana basilar que são estimuladas em frequências específicas [31]

vidindo a saída de cada filtro triangular (semelhante aos filtros triangulares dos coeficiente de MFCC convencionais) pela saída de um segundo filtro. Assim, são removidas as fortes variações na forma espectral que podem ser consideradas constantes em ambos os filtros.

Por último tem-se as representações baseadas em *Linear Predictive Coding* LPC [36]. A predição linear apresenta importantes propriedades, já que é uma técnica que modela o sistema do trato vocal para a produção da voz humana através de um filtro *all-pole*. Uma característica importante desta técnica é que permite encontrar os coeficientes do filtro sem fazer o cálculo do espectro explicitamente.

Desta forma, cada segmento do sinal de voz é representado por um vetor de coeficientes espectrais. Deve-se ter em conta que, embora a ordem da predição linear seja finita, podem ser calculados coeficientes espectrais infinitos. Por este motivo, é necessário limitar o número de coeficientes espectrais usando uma janela *liftering* [37].

1.1.1.3 Transformação ao domínio cepstral

Basicamente o objetivo deste bloco é representar o sinal de voz de forma adequada para o reconhecedor. Isso é feito através de conjuntos de vetores de N componentes que representam o espectro de cada segmento de voz. Como foi exposto anteriormente, as técnicas baseadas em análise espectral trabalham no domínio da potência espectral. Estas técnicas limitam sua análise devido ao fato de que os espectros dos filtros em bandas adjacentes estão bastante correlatados. É desejável eliminar essa correlação mantendo apenas a informação útil para o reconhecimento. Para isso a análise é geralmente completada com o cálculo do logaritmo da energia de cada faixa de frequência, levando os coeficientes espectrais ao domínio da *quefrencia*, para finalmente aplicar a Transformada Direta do Coseno (DCT), obtendo assim os coeficientes cepstrais. Analisando o Cepstrum sob o ponto de vista matemático, pode-se dizer que é um operador que transforma uma convolução no tempo em uma soma no domínio espectral. Consegue-se assim separar os dois componentes de informação do sinal de voz: a excitação e o filtro que modela o trato vocal.

A ideia principal dos coeficientes cepstrais é captar as mudanças temporais bruscas presentes no espectro. Devido a isto, utilizam-se além dos coeficientes extraídos até agora, chamados coeficientes "estáticos", os coeficientes delta e de aceleração, chamados coeficientes "dinâmicos", que capturam essas mudanças e incorporam informação relativa à transição dos coeficientes estáticos entre quadros vizinhos.

O cálculo dos coeficientes dinâmicos é feito através de regressão linear sobre uma janela, utilizando dois vetores antes e dois após o vetor calculado [38].

Ao longo dos anos, foram exploradas várias representações alternativas das características que satisfazem, em maior ou menor grau, as propriedades discriminativas e de robustez exigidas nos sistemas RAV. Algumas das técnicas mas conhecidas são *Mel frequency cepstral coefficients* (MFCC) [39] que são um conjunto de atributos baseados na escala auditiva Mel, amplamente utilizada no reconhecimento de voz introduzidos por *Davis e Mermelstein* na década de 1980, convertendo-se em estado da arte desde então.

Outro método baseado no domínio cepstral são os *Power-Normalized Cepstral Coefficients* (PNCC) [40]. Estes coeficientes são considerados como uma evolução dos MFCC. Sua eficiência é devida à adição de uma nova etapa de remoção de ruído, a qual, através da média das energias de uma banda ao longo de alguns quadros consecutivos, consegue remover a adição do ruído do sinal. Ao invés de usar filtros triangulares baseados na escala Mel, são aplicados filtros gammatone baseados na escala de Bandas Retangulares Equivalentes (ERB) [41], que representam bem a resposta impulsional da membrana basilar.

Os coeficientes *Locally Normalized Cepstral Coefficients* (LNCC)[24] são inspirados no detector de sincronização generalizada do seneff (GSD) [35]. Esses coeficientes executam uma normalização local no domínio da frequência em cada canal auditivo, sendo relativamente invariantes às mudanças na resposta de frequência do canal de transmissão. Essa normalização é conseguida dividindo a energia de dois tipos de filtros. O filtro do numerador que é definido como um filtro triangular, semelhante ao usado no banco de filtros Mel, enquanto o filtro que é denominador captura a energia de ambos os lados deste filtro.

Informação detalhada dos MFCC e PNCC, utilizados no *front-end* desta tese, é apresentada no apêndice A.

1.1.2 Back-end

O sistema de reconhecimento em seu bloco final compõe-se de três sub-estruturas fundamentais que visam comparar os vetores de características com os padrões de referência. Dando continuidade à seção anterior, o reconhecimento dos vetores de características extraídos no *front-end* ocorre no módulo de *back-end* ou etapa de comparação. Geralmente, uma máxima decodificação a posteriori (MAP) é realizada para encontrar a sequência de palavras mais provável e que melhor representa o sinal de voz observado. Isto significa buscar

a sequência de palavras $W = (w_1, w_2, ..., w_n)$, que melhor represente a sequência de vetores acústicos observados $O = (o_1, o_2, ..., o_T)$, onde o_t representa o vetor de atributos da voz obtido no instante de tempo t pelo *front-end*, e T é o numero total de segmentos da voz. Aplicando a regra de decisão baseada em MAP, a sequência de palavras $\widehat{W}$ reconhecidas a partir das observações O e dada por

$$\widehat{W} = arg \max_{W} P(W \mid O) \tag{1-1}$$

Esta maximização requer o calculo da probabilidade condicional $P(W \mid O)$, que através do teorema de Bayes, pode ser representada da seguinte forma

$$P(W \mid O) = \frac{P(W)P(O \mid W)}{P(O)} \tag{1-2}$$

Dessa forma a equação (1-1) pode-se reescrever como

$$\widehat{W} = arg \max_{W} \frac{P(W)P(O \mid W)}{P(O)} \tag{1-3}$$

onde $P(O)$ é a probabilidade de ocorrer uma determinada observação. Seu valor é constante e independente de W, de modo que é removido do processo de maximização.

$$\widehat{W} \simeq arg \max_{w} P(W)P(O \mid W) \tag{1-4}$$

Portanto, os elementos a avaliar são

- A probabilidade a *priori* $P(W)$ de que ocorra a sequência de palavras W, chamada de modelo de linguagem;
- A probabilidade da evidência acústica de cada frase $P(O \mid W)$, isto é, a probabilidade de que a transcrição W tenha a representação acústica O, chamada de modelo acústico.

A Fig. 1.5 apresenta o diagrama geral do *back-end* mostrando como misturam-se as informações das três fontes de conhecimento, que são o dicionário, o modelo de linguagem e o modelo acústico.

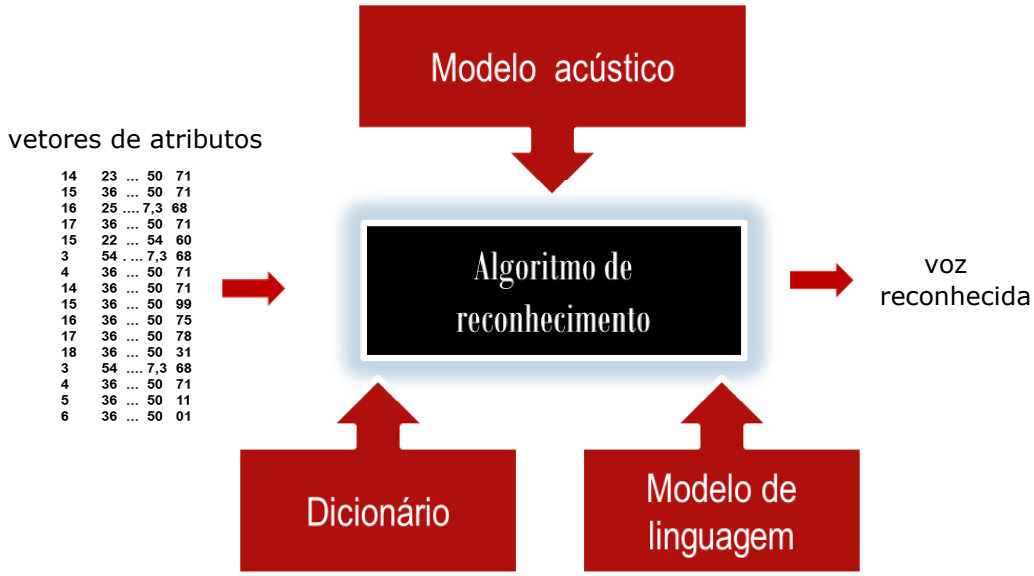

Figura 1.5: Diagrama de blocos do Back-end

1.1.2.1 Dicionário

O Dicionário, também conhecido como léxico, contém uma lista de palavras, juntamente com suas transcrições fonéticas. A representação fonética determina a sequência de fones que podem ocorrer durante o reconhecimento de cada palavra. Por exemplo, pode haver várias maneiras de pronunciar a mesma palavra devido à variabilidade de sotaques que existem no mesmo idioma, gerando assim várias transcrições fonéticas associadas. Para pequenas tarefas, como reconhecimento de comandos, o dicionário geralmente é muito simples (vocabulário pequeno) e o reconhecimento se faz com um mapeamento de um a um, do modelo acústico para a palavra. Por outro lado, em tarefas de reconhecimento maiores, por exemplo, transcrições de texto, onde os modelos acústicos podem representar unidades de sub-palavras (tipicamente fones, bifones[9], trifones[10]) requer-se um vocabulário com maior quantidade de palavras.

1.1.2.2 Modelo de linguagem

O modelo de linguagem determina a probabilidade a *priori* $P(W)$ da hipótese de palavras consideradas pelo reconhecedor. Essa hipótese é indepen-

[9]O bifone é um par adjacente de fones. Geralmente é usado para referir-se a uma gravação da transição entre dois fones.

[10]Os trifones são uma sequência de três fones. É chamado de trifone porque leva um contexto de três fones em conta (anterior, atual e próximo).

dente das observações e pode ser categorizada como determinística ou estatística. Os modelos determinísticos são definidos por gramática formal (regras) que restringem o idioma que o sistema pode reconhecer. Já os modelos estatísticos, usados neste trabalho, utilizam o contexto das palavras e a informação da frequência com que elas são pronunciadas, com o fim de encontrar opções prováveis que indiquem quais palavras têm mais chances de vir antes ou depois de uma outra.

Por exemplo, considere duas palavras com sons quase iguais "norte" e "morte". Se por exemplo, antes da palavra encontra-se a frase "no pólo..." o modelo da linguagem determina que "norte" é a palavra certa. Desta forma pode-se dizer que as restrições impostas pelo modelo de linguagem podem melhorar consideravelmente o rendimento do reconhecedor, reduzindo significativamente o espaço de busca da frase correta.

Em geral, o modelo de linguagem tem a tarefa de estimar a probabilidade de uma palavra w_i em uma sentença, dadas todas as palavras que a procedem $w_1, w_2...., w_n$.

Usando as regras elementares da teoria da probabilidade, pode-se expressar a probabilidade de cada hipótese da seguinte forma:

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1, w_2, ..., w_{i-1}) \tag{1-5}$$

onde $P(w_i|w_1, w_2, ..., w_{i-1})$ é a probabilidade de que a palavra w_i seja escolhida depois da sequência de palavras $(w_1, w_2, ..., w_{i-1})$.

A forma mais usada de se definir estas probabilidades é com a utilização de *n-gramas*, na qual a probabilidade de cada palavra em uma sentença depende apenas das $n-1$ palavras anteriores a ela, limitando assim o número de parâmetros do modelo de linguagem. Por exemplo, com n=2 tem-se o bigrama

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_2)...P(w_n|w_{n-1})$$

Com n=3 tem-se o trigrama

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_2w_1)...P(w_n|w_{n-2}...w_{n-1})$$

1.1.2.3 Modelo acústico

O modelo acústico é um conjunto de representações estatísticas dos diferentes sons do espaço acústico com o qual se está trabalhando. Ele é realizado a partir de um volume de dados de treinamento, pertencentes a dados de voz com seu correspondente rotulado (transcrições). Isso possibilita uma atribuição de cada som com sua representação escrita.

O objetivo do modelo acústico é calcular com precisão $P(O \mid W)$, ou seja, a probabilidade de que o sinal emitido seja O, dado que o locutor pronunciou as palavras W. Para modelar estas probabilidades o modelo acústico usa o fato de que as palavras faladas são compostas por sons, assim como as palavras escritas são compostas por letras. Usando esse conhecimento, podemos dividir as palavras em sons (representados pela pronúncia, fones, trifones,...,etc.) a fim de representar cada unidade acústica mediante um modelo estatístico independente que englobe a variabilidade acústica própria da pronúncia, e reuni-los novamente para o reconhecimento. Segundo [42] os modelos acústicos do sistema de reconhecimento podem ser gerados a partir de diferentes focos: (i) acústico-fonético, (ii) enfoque estatístico de padrões e (iii) enfoque baseado em inteligência artificial.

Nos sistemas RAV atuais, a técnica mais popular para a realização da modelagem acústica é baseada no enfoque estatístico de padrões através do uso de Modelos Ocultos de Markov (HMMs)[1][43] com funções de densidade contínua. Portanto, o modelo acústico do reconhecedor consiste em uma concatenação desses HMMs básicos (onde cada fone é representado por um HMM) para representar palavras. A introdução dos HMMs no campo da voz é usualmente creditada aos trabalhos independentes da *Carnegie Mellon University* [44] e da IBM [45]. Nesses trabalhos, foi percebida a necessidade de utilizar técnicas de modelamento estatístico que abordaram o problema de variabilidade da voz, a qual aumenta significativamente quando a complexidade e o tamanho do vocabulário.

É por isso que o sucesso destas estruturas deve-se, principalmente, à sua capacidade de modelar tanto as variabilidades acústicas como temporais do sinal de fala, e também por permitir a construção hierárquica dos modelos acústicos das sentenças.

Em geral, os HMMs podem ser considerados como um conjunto de estados ligados por transições com probabilidades associadas a cada transição, como ilustra a Fig. 1.6. O modelo começa com o estado inicial e, em cada passo de tempo discreto, ocorre uma transição a um novo estado, e um símbolo de saída é gerado. A transição e o símbolo de saída são aleatórios, sendo regidos

por um modelo probabilístico.

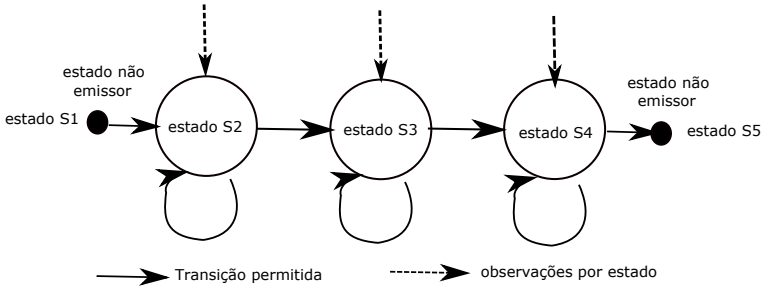

Figura 1.6: Representação de esquerda a direita do HMM

Para modelar efetivamente as características do sinal de voz que variam lentamente, o HMM de esquerda para direita é usado para modelar cada unidade fundamental no modelo acústico. Isto é, existem ligações entre estados que indicam as transições permitidas. Por exemplo, uma transição para o mesmo estado significa que o próximo segmento do sinal de voz permanece no estado atual. Uma transição para a direita indica que o sinal de voz do próximo segmento segue para o próximo estado do fone. Devido ao fato de que cada fone possui um padrão específico de evolução das distribuições dos atributos, apenas as transições de esquerda para direita e auto-transições são permitidas.

Os modelos acústicos clássicos são os GMM-HMM. Mais recentemente em [6] foi proposto o DNN-HMM, que representa o estado da arte em RAV.

1.1.2.4 GMM-HMM

Os HMMs são frequentemente usados como modelos acústicos para os sistemas RAV devido à sua capacidade de modelar estatisticamente a geração da voz. Um HMM geralmente representa uma unidade de sub-palavras, como um fone, e tem 3 estados onde cada um caracteriza um segmento quase estacionário da sub-palavra. No caso dos sistemas GMM-HMM, a cada estado é associada uma mistura de funções de probabilidade (pdf) gaussianas. Cada uma dessas misturas é usada para modelar a distribuição de probabilidade dos vetores de atributos da voz que pertencem ao estado[14].

Matematicamente, um HMM é caracterizado pelos seguintes elementos [46] [47]:

- $\boldsymbol{S} = \{ s_i \}, \quad i = 1, 2..., N$: um conjunto de todos os estados possíveis interligados entre si.
- $\boldsymbol{A} = \{ a_{ij} \}$: matriz de transição de probabilidade de estados, que representam a ordem dos estados em um HMM e a duração dos segmentos

do sinal em cada estado, onde a_{ij} é a probabilidade de ocorrer a transição do estado i ao estado j, ou seja,

$$a_{ij} = P\{\ q_t = s_j \mid q_{t-1} = s_i\} \qquad i, j = 1, 2..., N. \tag{1-6}$$

onde q_t representa o estado no instante t. Isso implica que a probabilidade de estar no estado do modelo em um instante de tempo t depende apenas do estado que foi visitado no instante de tempo anterior *t-1*.

– Um conjunto V com M símbolos de observação $V = (v_1, v_2, ..., v_M)$, onde em um instante t o modelo gera um símbolo $o_t \in V$

– $\boldsymbol{B} = \{\ b_j(v_k)\}$: matriz de probabilidades de observação, onde cada elemento representa a probabilidade de gerar um determinado símbolo em um certo estado j. Esta probabilidade é representada por

$$b_j(v_k) = P\{\ v_k = o_t | q_t = s_j\} \qquad j = 1, 2, ..., N; k = 1, ..., M. \tag{1-7}$$

Isso é a probabilidade de observação, que descreve as relações entre os vetores de atributos do sinal de voz o_t e os estados do modelo gerados pelo estado j no tempo t.

Caso a variável observada $v_k = O$ pertença a um espaço contínuo, as probabilidades de emissão $b_j(v_k)$, devem ser substituídas por funções densidade de probabilidade (pdf) contínuas. Neste caso a distribuição de vetores de atributos emitidos por cada estado é aproximado através do método de mistura de Gaussianas geralmente conhecidos como GMM, com M densidades Gaussianas caracterizadas pelos seus vetores média e a suas matrizes covariância. Isto é

$$b_j(O) = \sum_{k=1}^{M} P(k|s)\Lambda(O; \mu_s^{(k)}, \sigma_s^{(k)}) \tag{1-8}$$

onde M é o número de distribuições Gaussianas, $P(k|s)$ é o peso da *k-ésima* gaussiana e Λ representa a função densidade de probabilidade de um vetor gaussiano de média $\mu_s^{(k)}$ e matriz de covariância $\sigma_s^{(k)}$.

Assim, qualquer GMM-HMM é caracterizado pelo seguinte conjunto de parâmetros que representa o modelo acústico [46]:

$$\lambda = (a_{ij}, b_j, \pi_i) \tag{1-9}$$

onde π_i é o vetor com as probabilidades de estados iniciais. Nos modelos acústicos normalmente é assumido $\pi_1 = 1$ e $\pi_i = 0$ para todo $i \neq 1$, a sequência de símbolos que gera o modelo, $O = (o_1, o_2, .., o_T)$, é denominada observação e a sequência de estados q_t é denominada percurso.

Os parâmetros de cada modelo GMM-HMM, ou seja, as probabilidades de transição entre os estados e os parâmetros dos GMMs que modelam cada estado, são estimados a partir de dados de treinamento seguindo um determinado critério de otimização. Geralmente esse procedimento é feito através do método *Baum-Welch* [48]. Ele utiliza as probabilidades para frente e para atrás do algoritmo *forward-backward* [49], permitindo determinar de forma recorrente os parâmetros que definem o modelo acústico.

Finalmente, depois do treinamento do modelo acústico se faz o processo de decodificação. Dado o modelo acústico λ e uma nova sequência de observações $O = (o_1, o_2, .., o_T)$, aplica-se o algoritmo de Viterbi que é uma técnica que obtém a sequência mais provável de estados, para uma dada sequência emitida pelo HMM [50][51]. Informação detalhada dos algoritmos $Baum - Welch$ e *Viterbi* pode ser encontrada em [46].

Mais detalhes sobre a implementação das estruturas GMM-HMM para criar os modelos acústicos no reconhecimento de voz podem ser encontrados em [52][53][54][55][56][57].

1.1.2.5 DNN-HMM

Durante décadas os modelos acústicos baseados em GMM-HMM foram tão bem sucedidos para a modelagem acústica, que era difícil para qualquer novo método superá-los. Estes foram considerados o estado da arte em sistemas RAV durante varias décadas, devido ao fato de ter uma série de vantagens que os tornaram adequados para modelar as distribuições de probabilidade sobre vetores de atributos de entrada que estão associados a cada estado de um HMM, conseguindo descrever qualquer processo sequencial, como a voz. No entanto, há algumas décadas [58] foram desenvolvidas técnicas de modelagem de voz que empregam o mesmo conceito de modelo acústico baseado em HMM mas desta vez usando redes neuronais, conseguindo algum sucesso, já que através de uma única camada de unidades ocultas não-lineares podia-se prever estados HMM a partir dos coeficientes acústicos. No entanto, nem o *hardware* nem os algoritmos de aprendizagem foram adequados para o treinamento

de redes neurais com muitas camadas ocultas em grandes quantidades de dados e os benefícios de usar redes neurais com uma única camada oculta não eram suficientemente grandes para superar os tradicionais GMM-HMM. Ao longo dos últimos anos, os avanços nos algoritmos de aprendizagem em máquina e no *hardware* do computador, especialmente a aparição das unidades gráficas de processamento (GPUs), levaram a métodos mais eficientes para o treinamento de redes neurais profundas. Recentemente, a modelagem acústica usando redes neurais profundas (DNNs), denominadas sistemas híbridos DNN-HMM, ganhou popularidade em relação aos GMM-HMM tradicionais devido à sua robustez em condições realistas, considerado-se estado da arte dos sistemas RAV [6]. Os sistemas baseados em DNN de última geração contêm várias camadas ocultas, que permitem que a configuração aprenda informações de nível superior nos dados acústicos, juntamente com uma camada de saída treinada para fornecer probabilidades a *posteriori* para os estados de um HMM.

As redes neuronais profundas (DNN) são um conjunto de ferramentas que representam uma forma alternativa de aprendizagem e de processamento automático, baseado no funcionamento do sistema nervoso. Sua construção consta de várias camadas ocultas, entre a camada de entrada e a de saída. As DNN podem fazer a modelagem de complexas relações não lineares e suas arquiteturas geram modelos de composição onde as camadas extras permitem a composição das características das camadas anteriores. Isso possibilita uma grande capacidade de aprendizagem, o que gera um importante potencial na modelagem de complexos padrões de voz.

A ideia geral dos modelos híbridos baseados em DNN-HMM é aproximar a distribuição $P(o_t|q_t)$ que é a probabilidade de observar um curto intervalo de características acústicas (vetor de observações o_t), condicionadas a uma etiqueta de estado HMM L_i (cada estado $q_1, ..., q_t$ é representado por uma etiqueta L_i). Os atributos de entrada acústica representam cerca de 25ms a 32ms de áudio na maioria dos sistemas de reconhecimento de voz contínua de amplo vocabulário (LVCSR) que é considerado a duração de um fone. As etiquetas de estado HMM para LVCSR são senones[11]. Um sistema híbrido DNN-HMM usa um rede neural para aproximar $P(o_t|q_t)$ em vez de um GMM.

Uma rede neural não modela explicitamente a distribuição $P(o_t|q_t)$ requerida pelo HMM. Em vez disso, são treinadas redes neurais para estimar $P(q_t|o_t)$, o que permite visualizar a rede neural como um classificador de senones com entrada acústica. Usando a regra de Bayes pode-se obter $P(o_t|q_t)$ dada a distribuição de saída da rede neural $P(q_t|o_t)$ por

[11]estados sub-fonéticos agrupados dependentes de contexto. Originalmente, o termo "senone" significava um subtriphone generalizado

$$P(o_t \mid q_t) = \frac{P(o_t)P(q_t \mid o_t)}{P(q_t)} \tag{1-10}$$

onde $P(q_t)$ é a probabilidade *a priori* sobre os senones, que é calculada a partir das ocorrências do conjunto de treinamento, e $p(o_t)$ pode ser atribuída uma constante, já que os vetores de características de observação são considerados independentes um do outro. Desta forma pode-se executar uma decodificação HMM para maximizar uma combinação de modelos acústicos e de linguagem.

Tendo definido o sistema HMM híbrido o procedimento para construir redes neurais para modelar a distribuição de sênones $P(q_t|o_t)$ pode ser resumida da seguinte maneira:

- Geração de etiquetas (labels): Os bancos de dados com grandes vocabulários não tem alinhamento temporal, simplesmente transcrições; devido a este inconveniente é preciso usar um sistema GMM-HMM a fim de fazer-se um alinhamento forçado. O que resulta em um etiqueta para cada quadro[12];
- Configuração da arquitetura da rede neuronal: O tamanho e a estrutura das redes neurais utilizadas para a modelagem acústica são, de longe, a maior diferença entre os modernos sistemas HMM-DNN e os utilizados antes de 2010. Os DNNs modernos usam mais de uma camada oculta, tornando-as profundas; em geral 5 camadas ou mais, e 1000 neurônios ou mais por camada;
- Funções de ativação e algoritmo de otimização: dado um conjunto de treinamento acompanhado por suas respectivas etiquetas, escolhem-se funções tradicionais como *sigmoide* ou variante entre as camadas ocultas e *softmax* na camada de saída. Como função de perda, a *cross entropy* é usada a fim de maximizar a probabilidade da etiqueta observada dada a entrada. Finalmente se otimiza a DNN através do gradiente descendente estocástico (SGD).

Informações mais detalhadas sobre a implementação dos DNN-HMM para criar os modelos acústicos no reconhecimento de voz podem ser encontradas em [59][60][61][62].

Este tipo de rede neuronal, conhecida como arquitetura de aprendizagem profunda, tem sido usada na implementação de modelos acústicos para o reconhecimento de voz desde 2012 [63][64]. Recentemente em [6] mostrou-se que a probabilidade de emissão de cada estado pode ser modelada por

[12] trifones físicos (sênones), em geral milhares de sênones

uma rede neuronal profunda. A DNN demonstrou melhorar drasticamente o desempenho dos sistemas RAV em condições limpas, especialmente no cenário de treinamento multi-condição.

Nos últimos anos, as DNNs têm sido muito bem sucedidas nos trabalhos em reconhecimento de voz, onde pode-se ver que o número de trabalhos relacionadas a este tópico tem crescido exponencialmente desde 2012. Além disso, muitos sistemas de reconhecimento de voz comerciais como Microsoft, XBOX, Skype Translator, Google Now, Apple Siri, Deep Speech da mozilla, entre outros, são baseados atualmente neste tipo de técnica. Algumas das publicações mais recentes apresentam o sucesso destas ferramenta no elaboração de modelos acústicos. Por exemplo, em 2017 [62] os autores apresentam uma investigação empírica sobre quais aspectos são mais importantes no desenvolvimento de modelos acústicos com DNN, com o objetivo de melhorar o desempenho dos sistemas de reconhecimento de voz. Também fazem comparações entre os diferentes modelos híbridos (GMM-HMM, DNN-HMM) existentes que são utilizados no criação de modelos acústicos, através de diferentes métricas para quantificar fatores que influenciam diferenças no desempenho do sistema.

Em 2015 [65] as DNN foram utilizadas para modelagem das distribuições de emissões de HMMs, com o objetivo de aprender automaticamente as características adequadas para a classificação a partir dos dados fornecidos e em seguida estimar as probabilidades posteriores dos estados dos HMMs.

1.2 Motivação

O reconhecimento de voz é uma parte da inteligência artificial que tem como objetivo permitir a comunicação falada entre seres humanos e máquinas. A naturalidade com que os seres humanos se comunicam faz pensar que o reconhecimento de voz é uma tarefa simples. Porém, ele requer um processo complexo devido ao número de considerações a serem levados em conta para adequar o sinal e extrair suas características de forma eficiente.

Os avanços que ocorrem no campo do reconhecimento automático de voz (RAV) estão se tornando mais significativos dia a dia. Os reconhecedores atuais manipulam vocabulários cada vez maiores, operam em fala contínua e conseguem menores taxas de erro graças ao uso de algoritmos mais eficientes, equipamentos mais poderosos e mais baratos e a maior complexidade desses sistemas, permitindo com isso usar modelos mais sofisticados e refinados.

No entanto, apesar dos grandes avanços realizados, ainda está longe de se obter um sistema de reconhecimento de voz automático universal que funcione bem em qualquer aplicativo para qualquer locutor e em qualquer ambiente. Em

geral, o projeto e as características dos sistemas de reconhecimento automático de voz atuais dependem fortemente da aplicação a que se destinam e das condições de operação. Em geral, os RAV operam excepcionalmente bem em ambientes limpos, porém seu desempenho pode se tornar inaceitável em ambientes adversos.

Na atualidade têm se desenvolvido técnicas que melhoram a robustez dos sistemas de reconhecimento. Porém as taxas de reconhecimento dessas técnicas estão longe das atribuídas aos seres humanos, fazendo com que nos últimos anos a maior parte da pesquisa em sistemas de reconhecimento esteja centrada no reconhecimento de voz robusta.

Um sistema de reconhecimento precisa ser robusto a diferentes tipos e níveis de ruído para que seu desempenho não seja extremamente degradado.

Conforme as tecnologias da voz apresentam-se cada vez mais como parte integral de aplicações práticas em cenários reais (acesso a banco de dados por linha telefônica, discagem automática de números telefônicos, máquinas de ditado, etc.), tem-se observado a necessidade de desenvolver sistemas de reconhecimento de voz robusto. Ou seja, que mantenha seu rendimento dentro de uma ampla margem de condições ambientais, mesmo no caso em que essas condições variem de forma rápida. De fato, foi no ano 1997 [66] que através de levantamentos informais entre pesquisadores e empresas relacionadas com tecnologias de reconhecimento de voz, chegou-se à conclusão de que o limitante mais significativo dos sistemas atuais seria justamente a falta de robustez frente a condições adversas.

Em geral os sistemas de reconhecimento de voz atuais têm diversas limitações em sua construção e são muito sensíveis aos descasamentos entre condições de treinamento e teste. Grande parte das técnicas de processamento do sinal de voz tem sido desenvolvida em ambientes de laboratório, onde nenhum tipo de sinal indesejável se mistura com o sinal de voz. No entanto a nível prático, essas técnicas que apresentam ótimo desempenho em ambientes silenciosos, experimentam uma grande degradação quando agem em condições reais (escritório, rua, carro, avião, etc.) já que são afetadas pelo ambiente acústico. Por esse motivo, se o sistema for treinado com voz limpa não modelará corretamente o sinal de voz adquirido em condições ruidosas, sofrendo, portanto, uma alta degradação de desempenho.

Este problema de robustez tem sido a razão de que nos últimos anos não se tenha produzido uma transferência massiva da tecnologia de reconhecimento automático de voz no mundo empresarial. É por isso que na atualidade têm se desenvolvido técnicas que melhoram a robustez dos sistemas de reconhecimento, fazendo com que nos últimos anos a maior parte da pesquisa em

sistemas de reconhecimento esteja focada no reconhecimento de voz robusta. Porém, apesar dos esforços científicos e tecnológicos desenvolvidos, as taxas de reconhecimento dessas técnicas ainda estão longe das atribuídas aos seres humanos. É, portanto, fundamental resolver de forma eficiente essa falta de robustez.

Visando prover diversas soluções às necessidades de robustez para que o reconhecimento de voz possa se generalizar em qualquer aplicação real, esta tese foca em dois pontos de vista importantes quando busca desenvolver um sistema completamente eficiente, os quais são:

- *Funcional:* procura-se desenvolver algoritmos que sejam capazes de emular a comunicação oral, conseguindo que mediante a devida sequência de funções cumpra o objetivo para o qual foi projetado, de modo que melhore seu rendimento, facilitando assim o trabalho do usuário;
- *Econômico:* procura-se uma forma de implementar algoritmos ainda mais eficientes que contribuam às necessidades do mundo moderno, acelerando as operações e reduzindo custos, garantindo a maior confiabilidade e melhorando assim o grau de precisão e desempenho que possa-se alcançar.

Devido à análise anterior, nasce a motivação do presente trabalho, abordando os problemas correspondentes às distorções do sinal de voz devido às condições de ruído, desenvolvendo algoritmos que sejam capazes de fornecer maior robustez e que proporcionem o reconhecimento de voz com a melhor qualidade possível.

1.3 Objetivos

Esta tese tem por finalidade aprofundar no estudo de técnicas de reconhecimento robusto em condições adversas. Para isso, será utilizado o estado da arte em reconhecimento, e propostas novas estruturas com base nos algoritmos de robustez apresentados na literatura nos últimos anos, a partir da exploração e mistura de diferentes técnicas.

O principal objetivo consiste em propor e analisar técnicas e atributos que venham obter um melhor desempenho do sistema de reconhecimento de voz robusto. Em particular o nosso foco consiste em forma mais robusta à etapa de extração de características dos sistemas de reconhecimento para que esta seja menos sensível aos efeitos do ruído.

Como primeira ideia, será proposta a mistura de duas técnicas conhecidas na literatura (equalização de histogramas e filtro de mediana) com o objetivo

de diminuir as distorções do espaço de representação, o qual modificará as médias e variâncias dos vetores que as representam. O principal foco é melhorar os vetores de atributos, através de uma suavização das funções densidade de probabilidade (pdf) utilizando uma filtragem não linear (filtragem por mediana MED), composta de uma janela deslizante de um número de elementos ímpares de cada pdf. Essa técnica será seguida de um procedimento de mapeamento de histogramas com o objetivo de levar a pdf ruidosa a uma pdf comum. Os resultados numéricos mostraram que as taxas de acerto com o método proposto MED-HEQ melhoraram as tarefas de reconhecimento robusto em comparação com outras misturas apresentadas na literatura.

Esta tese introduz também um novo tipo de mascaramento auditivo, baseado nos sistemas *Computational Auditory Scene Analysis* (CASA), propondo-se um novo mascaramento baseado no uso dos *Local Binary Pattern* (LBP), que tem como objetivo principal separar as distintas fontes sonoras que compõem a entrada acústica. Através deste tipo de mascaramento busca-se reter as unidades T-F (tempo-frequência) onde a voz predomina sobre o ruído. Em particular, o objetivo através deste tipo de mascaramento é fortalecer o sinal de voz antes e depois da extração de atributos dos sistemas RAV, tanto no domínio espectral quanto no domínio *wavelet*, a fim de melhorar a inteligibilidade e a qualidade do sinal de voz que será entregue ao reconhecedor.

Conforme mencionado na caracterização do problema, diversos aspectos e estratégias relativos aos sistemas de reconhecimento de voz contínua foram considerados. Para tal, propõe-se alcançar com esta tese os seguintes objetivos particulares:

- Pesquisa bibliográfica e estudo teórico sobre sistemas de reconhecimento de voz e métodos de robustez desenvolvidos até a atualidade;
- Revisar as estratégias utilizadas para o reconhecimento de voz baseadas em mistura de gaussianas GMM e em *deep neural networks* DNN;
- Revisar as técnicas e procedimentos de extração de atributos do sinal de voz;
- Analisar o efeito do ruído sobre o sinal de voz e no reconhecimento. Este é um ponto importante, já que o desempenho dos reconhecedores de voz diminui quando o ambiente acústico distorce o espaço de representação do sinal de voz;
- Proposta e simulação de uma variedade de algoritmos de robustez de reconhecimento de voz, cobrindo uma vasta gama de níveis de ruído, com o propósito de avaliar a eficiência e a capacidade de cada um deles para

eliminar os efeitos do ruído sobre o sinal de voz, e superar os principais esquemas relatados na literatura;

– Apresentação e discussão dos resultados experimentais.

1.4 Estrutura da tese

Esta tese foi organizada em sete capítulos, sendo o primeiro esta introdução.

No Capítulo 2 faz-se uma apresentação do estado da arte em reconhecimento de voz contínua robusto.

O Capítulo 3 inclui uma breve revisão do método tradicional de mapeamento de histogramas e a filtragem de média temporal sobre as PDFs (funções distribuição de probabilidade). Se apresenta o algoritmo de filtragem não-linear proposto nesta tese (que nada mais é do que uma filtragem por mediana das PDFs), o procedimento experimental, a discussão dos resultados e conclusões.

No Capítulo 4 serão descritas as técnicas de mascaramento espectrais clássicas existentes, *Ideal Binary Mask* (IBM), *Ideal Ratio Mask* (IRM) e será exposta em detalhe a técnica *Ideal Neighbourhood Mask* (INM) proposta baseada em mascaramentos através de *Local Binary Patterns* (LBP), procedimento experimental e a discussão dos resultados. Finalmente, serão apresentadas algumas conclusões.

O Capítulos 5 fornece, inicialmente, uma breve visão geral dos trabalhos anteriores relacionados ao algoritmo de realce de voz baseado em *wavelets* e à técnica LBP. Será apresentado o método de realce de voz proposto usando máscaras baseadas nos padrões binários locais (WLBP) aplicadas ao esquema *wavelet-denoising*. Finalmente, os resultados da simulação e as conclusões são fornecidas.

O Capítulo 6 foca-se na proposta de mascaramento baseada em LBPs e DNN (Deep Neural Networks), serão discutidos os testes realizados e apresentados os resultados obtidos e conclusões.

Finalmente, o Capítulo 7 provê uma série de conclusões e contribuições do trabalho e também são fornecidas possíveis linhas para trabalhos futuros.

2
Estado da Arte em Reconhecimento de Voz Contínua Robusto

2.1
Introdução

Os avanços na área da inteligência artificial são dia a dia mais significativos. No campo do reconhecimento automático de voz (RAV), os sistemas atuais têm conseguido passar do reconhecimento de palavras isoladas próprias de um vocabulário limitado a situações de reconhecimento de voz contínua onde o vocabulário é de grande dimensão, atingindo taxas de precisão de palavra até 95% para a língua inglesa em condições de laboratório (caso do sistema RAV do google) de acordo com o relatório anual de tendências da internet da Mary Meeker de março de 2017[67]. No entanto, esses sistemas são sensíveis à alteração das condições acústicas, o que pode causar degradação significativa do desempenho, já que o sinal sofre distorções que não têm sido contempladas na etapa de treinamento.

Uma das principais causas é a existência de vários tipos de ruído de fundo, que fazem com que o desempenho dos sistemas em aplicações do mundo real, tais como, serviços de comunicações de voz sem fio, dispositivos de aparelhos auditivos digitais, telefonia móvel com mãos-livres, transmissão de voz, entre outros, esteja longe de ser satisfatório, já que o ruído degrada os sistemas até níveis em que seu uso se torna definitivamente inaceitável. Estas alterações acústicas limitam significativamente o funcionamento dos sistemas RAV em ambientes reais, já que se superpõem ao sinal de voz, mascarando e alterando as suas características. É por isso que na atualidade tem se acrescentado métodos de robustez nos sistemas RAV que não exigem uma carga computacional excessiva e que melhoram o desempenho dos sistemas evitando o descasamento entre as condições de treinamento e as de reconhecimento. Estes métodos constituem uma área de pesquisa fundamental no processamento de voz conhecida como robustez, e estão divididas em 3 categorias principais, segundo seu enfoque principal [68].

Neste capítulo é feita uma breve análise das fontes de variabilidade que degradam os sistemas RAV, focando no efeito do ruído aditivo sobre o sinal de

voz, apresentando um modelo matemático que permitirá analisar quantitativamente o grau de degradação do sinal de voz quando capturada em ambientes reais. Finalmente, é apresentado o estado da arte dos principais métodos RAV robustos ao ruído, que foram propostos e publicados nos últimos anos, e que criaram um impacto significativo na pesquisa, focando especialmente nos métodos que estão diretamente relacionados com as propostas apresentadas nesta teses.

2.2 Modelo geral do ambiente acústico

A existência de ruído é inevitável em aplicações do mundo real. Segundo a teoria da comunicação, o ruído é um som inarticulado ou distúrbio anômalo que causa uma sensação de audição desagradável no sistema auditivo humano e que gera uma interferência não desejada no processo comunicativo, distorcendo a informação transmitida pela onda acústica portadora de informação, dificultando sua correta percepção e evitando que a informação chegue de forma clara [69].

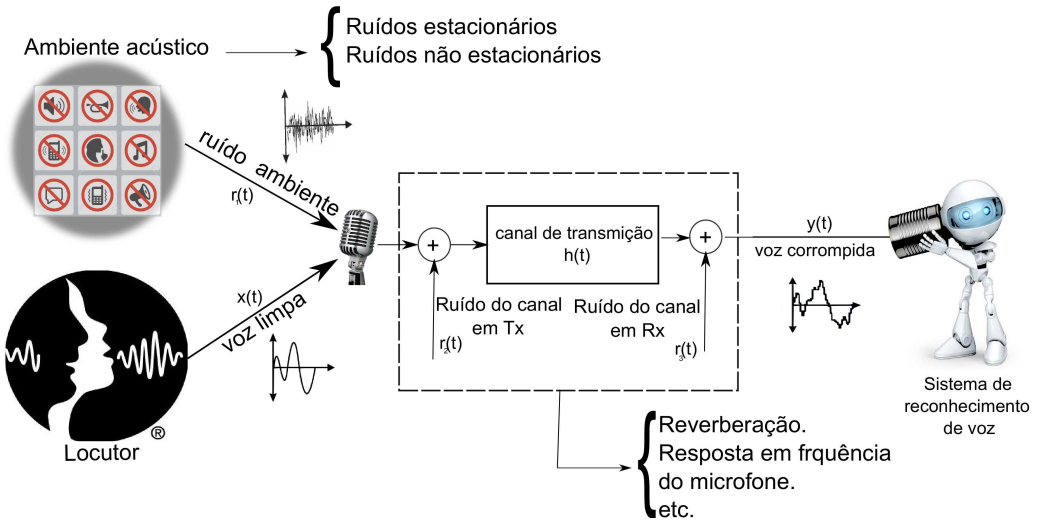

Figura 2.1: Modelo do ambiente acústico para ruído aditivo e convolutivo. x(t) representa o sinal de voz limpo, r(t) representa o ruído aditivo, h(t) representa a distorção do canal (o ruído convolutivo) e y(t) representa o sinal corrompido resultante.

Na Fig. 2.1 se ilustra uma visão simplificada de um sistema de comunicação comumente usado para o ambiente acústico [70], onde é representado o problema geral dos sistemas RAV em condições reais. Nesse sistema as fontes de distorção ou variabilidade podem se classificar em três categorias:

– *Variações devido ao locutor:* nessa categoria a fonte de informação $x(t)$ que gera variações nos sistemas tem a ver com as características do trato vocal e o gênero do mesmo, já que fatores intrínsecos ao locutor (estado emocional, estado físico, etc.) afetam diretamente as características da voz, por exemplo, a amplitude e a distribuição da frequência fundamental. Essas variações podem ser atenuadas por meio de parametrizações que isolam as informações acústicas com treinamentos dependentes do contexto, métodos de adaptação de locutor ou normalização do trato vocal do locutor;

– *Características do canal:* os efeitos do canal variam lentamente ao longo do tempo. Geralmente estas distorções são determinadas pelas características da resposta em frequência do microfone que captura o sinal e do canal de transmissão. Esse tipo de distorção também chamado ruído convolutivo $h(t)$ que mistura-se de forma convolucional com o sinal de voz no tempo. Normalmente, ele é modelado como um filtro linear e invariante ao longo do tempo, que corresponde a um distúrbio convolucional com o sinal no domínio temporal ou como a adição de um deslocamento no domínio log-spectral, fazendo com que o sinal sofra uma série de distorções espectrais antes de ser processado pelos sistemas RAV. Alguns exemplos deste tipo de distorção são: (i) as características do canal de transmissão, tais como o tipo e localização do microfone de gravação que geram um impacto significativo no espectro de voz, já que o microfone usado para coletar dados de teste é diferente do usado para coletar dados de treinamento, produzindo uma incompatibilidade no espectro médio [70]; (ii) os efeitos devido à codificação da voz para sua transmissão eficiente sobre o canal de transmissão (VoIP) e o comprimento de banda disponível; (iii) assim como as paredes e outros obstáculos na sala de gravação que produzem reflexões múltiplas criando um grau de reverberação ou eco que influenciam o espectro do sinal;

– *Ambiente acústico:* Finalmente a terceira fonte de distorção, que é o foco principal de estudo desta tese e que tem se tornando o motor da investigação no campo de reconhecimento automático robusto de voz nos últimos anos, é a distorção devido ao ambiente acústico tendo como principal representante o ruído aditivo $r(t)$ (ronco do motor, ventiladores, ar condicionado, falatório, etc.). Esse tipo de ruído mistura todos os sinais ambientais emitidos pelas fontes de som presentes enquanto o locutor está falando, perdendo qualidade e inteligibilidade. O ruído aditivo é adicionado ao sinal de voz no domínio do tempo, e pode ser categorizado por possuir grande quantidade de energia não harmonicamente

distribuída nas diferentes frequências. Segundo [71] os tipo de ruído podem ser:

- **Estacionário:** Dentro deste grupo encontram-se todos aqueles ruídos cujas propriedades permanecessem constantes ao longo do tempo ou, pelo menos, em um período de tempo suficientemente longo, possuindo uma densidade espectral que não varia com o tempo, mantendo suas propriedades constantes como o caso do ruído aditivo branco, o qual tem um espetro de potência plana.
- **Não estacionário:** são ruídos imprevisíveis, que estão continuamente mudando ou que aparecem intermitentemente sem qualquer periodicidade, suas densidades espectrais mudam com o tempo dificultando sua caracterização. Por exemplo, as vozes espontâneas que no caso do ruído de restaurante fazem com que as características espectrais (e temporais) estejam constantemente mudando.

O paradigma mais utilizado e simplificado para modelar esses efeitos ou distorções apresentados na Fig. 2.1 e que modificam o sinal capturado pelo reconhecedor, é o definido como uma mistura do ruído aditivo e filtragem não linear [1], isto é, agrupando todas as distorções aditivas numa única fonte de ruído $r(t)$ e todas as distorções convolutivas representadas por meio da resposta impulsiva em uma única resposta $h(t)$. Cabe salientar que mesmo não inclusa de forma explícita, a distorção devida às características próprias do locutor podem ser modeladas como uma distorção multiplicativa do espectro do sinal de voz normalizado.

Como pode se ver na Fig. 2.1 um sinal de voz ao se propagar por meio de um canal acústico e capturado por um microfone é corrompido por ruídos indesejados, o que pode resultar em degradação apreciável ou perda na qualidade e inteligibilidade da voz. Em outras palavras, para cada instante t o sinal de voz degradado $y(t)$, pode ser expresso em termos de uma versão filtrada do sinal de voz $x(t)$ emitida pelo locutor, misturada de forma aditiva com o ruído ambiental $r(t)$ [5], por meio da equação

$$y(t) = x(t) \otimes h(t) + r(t) \tag{2-1}$$

onde a distorção convolucional devido ao microfone, ao canal e às características próprias do locutor apresentam-se por meio da resposta ao impulso de um filtro linear e invariante $h(t)$, e $\otimes$ representa o operador de convolução. Tradicionalmente, são feitas suposições que fazem com que o modelo seja compreensível sob o ponto de vista matemático. Por exemplo, segundo [72] é comum

supor que o ruído $r(t)$ é um processo aleatório de média zero descorrelatado do sinal de voz $x(t)$. Essa suposição permite que os termos de correlação cruzada sejam ignorados nas equações que descrevem a operação desses algoritmos, tornando a matemática mais atrativa. Por outro lado, [73][74] estabelecem que a distorção do canal $h(t)$ torna-se aditiva no domínio log-spectral, considerado-a como um deslocamento nos parâmetros da voz limpa, sendo muitas vezes removido por meio de algum processo de filtragem linear, como por exemplo filtragem RASTA [75] ou algum outro filtro passa-banda.

Com as suposições acima expostas e baseando-se nos trabalhos [76][77], será apresentada na seguinte seção a formulação matemática do modelo de distorção. Cabe salientar que ao longo desta tese não será considerado o efeito do canal, somente o de ruídos ambientes.

2.3 Modelagem matemática do ruído e seus efeitos nos sistemas RAV

Partindo da equação (2-1), que formula o problema descrito na Fig. 2.1 e que relaciona o sinal de voz $x(t)$ com as distorções $r(t)$ e $h(t)$, o sinal contaminado $y(t)$ pode ser reformulado no domínio espectral aplicando a transformada discreta de Fourier (DFT, Discrete Fourier Transform) de N amostras, no ponto de frequência $\omega-$ésimo do quadro t-ésimo da voz corrompida, respectivamente, da seguinte maneira

$$Y(t,\omega) = X(t,\omega)H(\omega) + R(t,\omega) \tag{2-2}$$

onde ω é o índice no domínio da DFT e $H(\omega)$ é a função de transferência do canal, que como mencionado em [73] é invariante no tempo. A partir deste ponto e para maior clareza o subíndice t será omitido, levando em consideração que as seguintes expressões aplicam-se a cada quadro. Partindo de (2-2), o efeito do ruído sobre o sinal de voz limpo no domínio espectral de potência pode ser representado como

$$\begin{aligned} |Y(\omega)|^2 &= |X(\omega)H(\omega) + R(\omega)|^2 \\ &= |X(\omega)|^2|H(\omega)|^2 + |R(\omega)|^2 + 2|X(\omega)H(\omega)||R(\omega)|cos_(\theta_\omega) \end{aligned} \tag{2-3}$$

onde θ_ω representa o ângulo no plano complexo entre $|Y(\omega)||H(\omega)|$ e $|R(\omega)|$ e coincide com a diferença entre as fases no ponto de frequência $\omega-$ésimo existente entre o espectro do sinal limpo e o ruído. Embora existam autores que consideram esta informação relevante [76] [78] , nesta tese assume-se a

informação de fase nula, devido ao fato de que no trato vocal não existe um mecanismo que sincronize a fase e o ruído se desconsidera-se o efeito Lombard[1], ficando a equação (2-3) reescrita como

$$S_Y(\omega) = S_X(\omega)S_H(\omega) + S_R(\omega) \tag{2-4}$$

onde $S_Y(\omega)$ representa o espectro de potência do sinal corrompido $y(t)$, $S_X(\omega)$ representa o espectro de potência do sinal limpo $x(t)$, $S_H(\omega)$ representa o espectro de potência do canal $h(t)$ e $S_R(\omega)$ representa o espectro de potência do ruído aditivo $r(t)$.

O seguinte passo em um algoritmo de extração de atributos, envolve a análise do sinal através de um banco de filtros em uma escala perceptiva que imita a resolução em frequência do ouvido humano. A filtragem mais conhecida é aquela que utiliza D filtros com resposta triangular e espaçada linearmente no domínio da frequência logarítmica seguindo a escala Mel [79]. De acordo com isso, para calcular as energias em cada banda, um banco de filtros é aplicado à densidade espectral segundo [1]. Portanto, se $B_i(\omega)$ denota a resposta em frequência da banda i-ésima da voz corrompida, $Y_i(\omega)$, a saída dessa filtragem para essa banda i do banco de filtros pode ser calculada como

$$\begin{aligned}\widehat{Y_i} &= \sum_{\omega} B_i(\omega)S_Y(\omega) \\ &= \sum_{\omega} B_i(\omega)(S_X(\omega)S_H(\omega) + S_R(\omega)) \\ &= \sum_{\omega} B_i(\omega)(S_X(\omega)S_H(\omega)) + \sum_{\omega} B_i(\omega)S_R(\omega)\end{aligned} \tag{2-5}$$

onde i = 1,....,D

Com o objetivo de passar a equação (2-5) ao domínio cepstral é necessário definir as seguintes equações, que são a representação em função da saída do banco de filtros para cada um dos termos da equação acima exposta:

$$\begin{aligned}\widehat{X_i} &= \sum_{\omega} B_i(\omega)S_X(\omega) \\ \widehat{R_i} &= \sum_{\omega} B_i(\omega)S_R(\omega)) \\ \widehat{H_i} &= \frac{\sum_{\omega} B_i(\omega)S_X(\omega)S_H(\omega)}{\widehat{X_i}}\end{aligned} \tag{2-6}$$

[1]O efeito Lombard é a tendência involuntária de aumentar o esforço vocal ao falar em um local ruidoso para melhorar a audibilidade da voz. Essas mudanças não só afetam a intensidade da voz, se não também o tom a frequência e a duração do som dos fones

onde $\widehat{X_i}$, $\widehat{H_i}$ e $\widehat{R}_i$ representam os vetores correspondentes à saída do banco de filtros do sinal de voz limpa, ruído convolutivo e ruído aditivo, respectivamente. Assim, a partir das equações (2-5) e (2-6), e aplicando logaritmos em ambos os lados da equação, podemos chegar em uma expressão do sinal corrompido $\widehat{Y_i}$ que busca imitar a resolução em amplitude do ser humano dada por

$$\begin{aligned} log\widehat{Y_i} &= log(\widehat{X_i}\widehat{H_i} + \widehat{R}_i) \\ log\widehat{Y_i} &= log\left(\widehat{X_i}\widehat{H_i}\left(1 + \frac{\widehat{R}_i}{\widehat{X_i}\widehat{H_i}}\right)\right) \\ log\widehat{Y_i} &= log\left(\widehat{X_i}\widehat{H_i}\right) + log\left(1 + \frac{\widehat{R}_i}{\widehat{X_i}\widehat{H_i}}\right) \end{aligned} \tag{2-7}$$

Aplicando os logaritmos da expressão 2-7 e definindo os vetores correspondentes ao sinal corrompido, o ruído aditivo, o ruído do canal e o sinal limpo como

$$\begin{aligned} \widehat{y}_i &= log(\widehat{Y_i}) \\ \widehat{x}_i &= log(\widehat{X_i}) \\ \widehat{r}_i &= log(\widehat{R}_i) \\ \widehat{h}_i &= log(\widehat{H_i}) \end{aligned} \tag{2-8}$$

resulta na equação

$$\begin{aligned} log\widehat{Y_i} &= log\left(\widehat{X_i}\right) + log\left(\widehat{H_i}\right) + log\left(1 + exp\left(log\frac{\widehat{R}_i}{\widehat{X_i}\widehat{H_i}}\right)\right) \\ \widehat{y}_i &= \widehat{x}_i + \widehat{h}_i + log\left(1 + exp^{(\widehat{r}_i - \widehat{x}_i - \widehat{h}_i)}\right) \end{aligned} \tag{2-9}$$

Finalmente a equação (2-9) pode se expressar da seguinte forma

$$\widehat{y}_i = \widehat{x}_i + \widehat{h}_i + log(1 + exp^{\widehat{r}_i - \widehat{x}_i - \widehat{h}_i}) = \widehat{x}_i + \widehat{h}_i + g(\widehat{r}_i - \widehat{x}_i - \widehat{h}_i) \tag{2-10}$$

onde 1 é o vetor de uns de comprimento D e as funções *exp(.)*, *log(.)* e *g(.)* se aplicam em cada componente do vetor. A função *g(.)* é referida na literatura como descasamento (*mistmach function*) e é definida como:

$$g(z) = log(1 + exp(z)) \tag{2-11}$$

Devido ao fato do ruído convolucional ser um deslocamento nos parâmetros da voz limpa: no domínio da log-energia (ver equação (2-10)), este deslocamento é constante para todas as frequências, e pode ser removido por meio de algum processo de filtragem linear. A fim de simplificar a análise e como foi referido anteriormente neste trabalho consideraremos o ruído convolutivo nulo $H(\omega) = 0$.

Finalmente, seguindo os passos da extração de atributos pode-se obter os M coeficientes cepstrais aplicando à equação 2-10 a transformada discreta do coseno (DCT) e sua inversa representadas por ζ e ζ^{-1} respectivamente. A DCT é aplicada à sequência de logaritmos do item anterior, a fim de descorrelatá-los[2], resultando em

$$\widehat{y}^{\zeta} = \widehat{x}^{\zeta} + \zeta log(1 + exp^{\zeta^{-1}(\widehat{r}^{\zeta} - \widehat{x}^{\zeta})}) \tag{2-12}$$

onde os vetores desta expressão são os projetados dos vetores com as energias na escala logarítmica das seguintes equações

$$\begin{aligned} \widehat{y}^{\zeta} &= \zeta(log\widehat{Y}_1 \quad log\widehat{Y}_2 \quad \cdots \quad log\widehat{Y}_M) \\ \widehat{x}^{\zeta} &= \zeta(log\widehat{X}_1 \quad log\widehat{X}_2 \quad \cdots \quad log\widehat{X}_M) \\ \widehat{r}^{\zeta} &= \zeta(log\widehat{R}_1 \quad log\widehat{R}_2 \quad \cdots \quad log\widehat{R}_M) \end{aligned} \tag{2-13}$$

A partir dessas equações, pode-se formular uma relação entre as características cepstrais que representam a voz limpa e a voz com ruído, observando como o sinal limpo é modificado no domínio log-Mel (Fig. 2.2) pela transformação não linear das energias logarítmicas e cepstrais através da equação (2-12), as quais são frequentemente usadas nos sistemas RAV.

Pode-se ver como a relação entre o sinal limpo e o sinal corrompido com ruído aditivo é não linear para valores de energia da voz corrompida próximos ou menores que o ruído. Isso faz com que o sinal corrompido decresça assintoticamente ao nível da energia do ruído, enquanto para energia alta (superior à energia do ruído) a voz corrompida se aproxima da voz limpa, ficando praticamente inalterada. Ou seja, a energia ruído não tem efeito sobre o sinal de voz.

[2]A razão pela qual o DCT é usada na extração de atributos, deve-se a que ela gera uma saída aproximadamente descorrelatada. E os atributos descorrelatados podem ser modelados eficientemente como uma distribuição gaussiana com uma matriz de covariância diagonal.

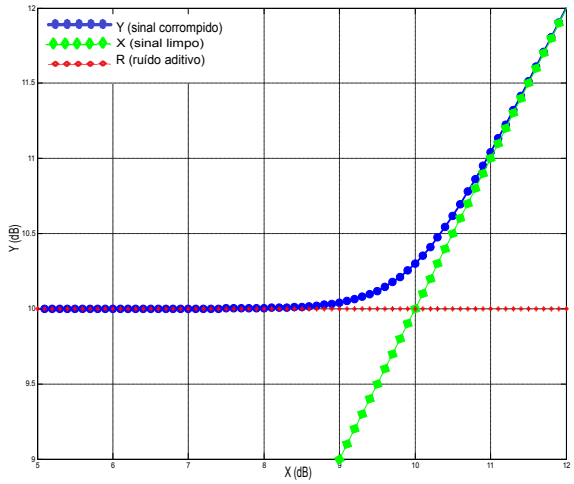

Figura 2.2: Modificação da energia logarítmica pela adição de um ruído de 10 dB.

2.3.1 Efeitos do ruído aditivo sobre as distribuições estatísticas dos parâmetros da voz

Uma maneira simples de entender como o ruído aditivo afeta o discurso é observando como os vetores de atributos da voz limpa são transformados pelo ambiente acústico. Em [5] e [80] foi realizado um estudo qualitativo do efeito do ruído aditivo sobre esses parâmetros onde foram observados os seguintes fenômenos sobre as distribuições do sinal corrompido:

- Em um ambiente acústico corrompido com ruído as distribuições estatísticas dos parâmetros não são normalmente distribuídas, já que as distribuições de probabilidade sofrem deformações, devido à não-linearidade da transformação produzida pelo nível da SNR e a variância do ruído;
- A média das distribuições sofre um aumento e a variância uma diminuição (supondo as variâncias do ruído menores que as do sinal de voz original).

Para visualizar os fenômenos acima expostos foi simulada uma distribuição de probabilidade dos atributos da voz limpa $p(x)$ por meio de uma distribuição normal de média $\mu = 3$ e desvio padrão $\sigma = 1$ e corrompida com ruído artificial de 5 e 10 dB (Fig. 2.3) Os dados artificiais do sinal de voz e do ruído foram misturados de acordo com a equação 2-10. Pode-se ver como o ruído faz com que a distribuição de probabilidade dos sinais ruidosos não sejam normalmente distribuídas [81]. Por outro lado, se o ruído for considerável pode-se observar a compressão da distribuição de sinal ruidoso resultante. Em geral, o efeito do ruído no sinal de voz gera nos atributos uma mudança na média e uma diminuição na variância das distribuições de probabilidade.

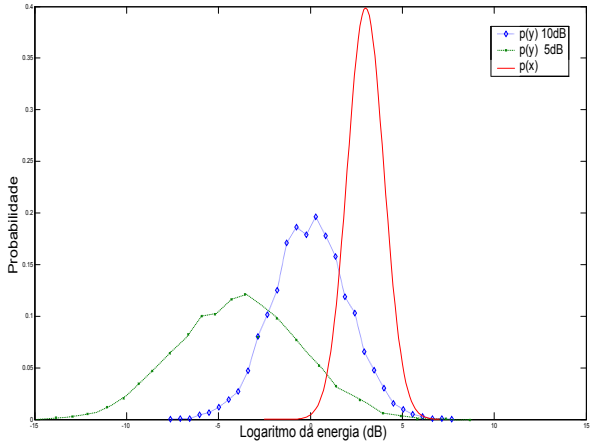

Figura 2.3: Modificação da distribuição de probabilidade de um sinal limpo devido ao efeito do ruído. A distribuição de voz limpa $p(x)$ é considerada Gaussiana de média $\mu = 3$ e desvio padrão $\sigma = 1$, corrompida com ruído de 5 e 10dB.

Uma característica distintiva dos diversos tipos de ruído é a forma do seu espectro, particularmente no que diz respeito à distribuição de energia de ruído no domínio da frequência. A Fig. 2.4 mostra como o ruído no caso *babble*[3] mascara completamente o sinal limpo, gerando um descasamento entre as etapas de treinamento e teste e prejudicando o desempenho do sistema RAV. Pode-se ver como quanto menor é a SNR maior é o mascaramento do sinal original, degradando a inteligibilidade e a qualidade da voz o que resulta numa perda de informação no sistema.

Analogamente a Fig. 2.6 mostra o efeito do ruído aditivo sobre a distribuição de probabilidade correspondente ao coeficiente cepstral C_0 (ligado com o valor da energia na escala logarítmica) de uma frase do banco de dados AURORA-4 (440c020b). Neste exemplo, o sinal de voz foi artificialmente contaminado com ruído *babble* para SNRs 0dB e 10dB. Os sinais de voz limpo e corrompido foram parametrizados e as distribuições de probabilidade foram aproximadas pelos histogramas correspondentes a cada componente do vetor de características. Pode-se observar que o ruído causa um deslocamento das médias das distribuições, bem como uma compressão ou modificação das variâncias. Além das alterações desses momentos das distribuições, pode-se apreciar o efeito não-linear que origina o ruído, que se manifesta em uma modificação da forma das distribuições de probabilidade (ou o que é equivalente, uma alteração dos momentos de ordens superiores). Esses efeitos não-lineares são mais importantes para o coeficiente de baixa ordem, e eles se tornam menos pronunciados quando a ordem do coeficiente cepstral aumenta (vide Fig.2.5). Portanto, as distribuições de probabilidade dos atributos do

[3]ruído de murmúrio de fala

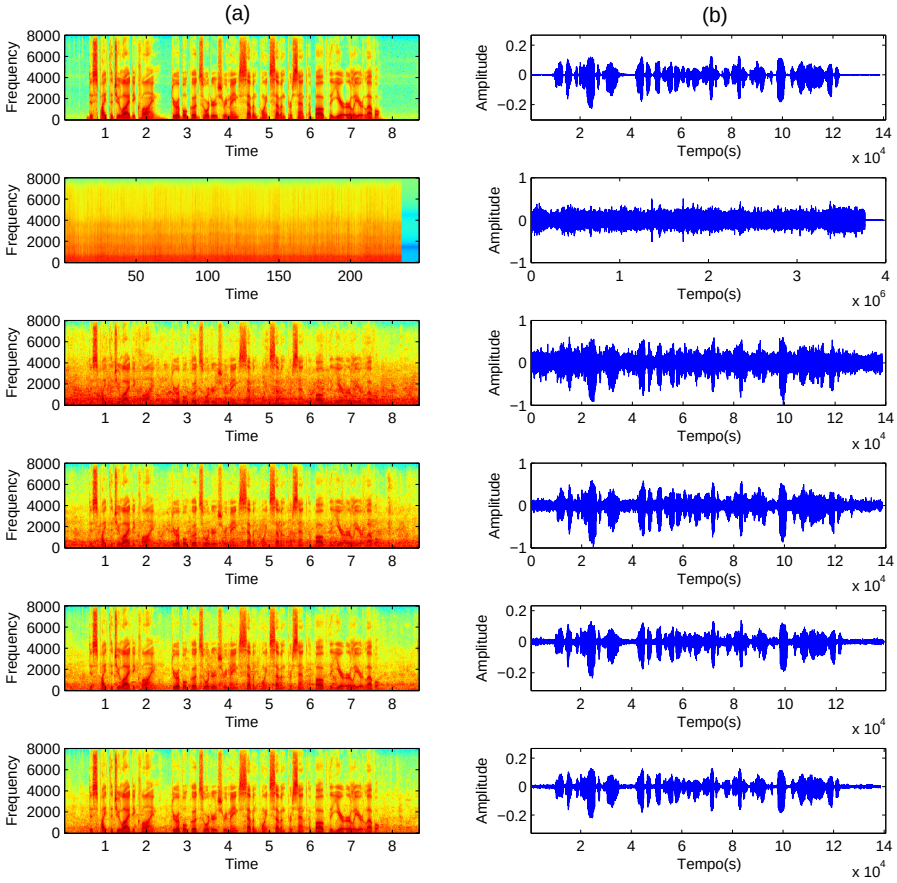

Figura 2.4: Frase do banco de dados AURORA-4 (440c020b) corrompida com ruído *babble*. De cima para baixo: sinal limpo, ruído *babble*, sinal corrompido com SNR = 0dB, sinal corrompido com SNR = 5dB, sinal corrompido com SNR = 10dB, sinal corrompido com SNR = 15dB.

sinal corrompido serão muito mais complexas do que a dos atributos da voz original.

Observa-se que a não linearidade causada pela função de descasamento dos parâmetros estáticos, equação (2-12) é altamente não linear e como resultado, é muito difícil prever a distribuição das características limpas dada a distribuição de características ruidosas. Isso gera um descasamento dos sinais que diminui as taxas de reconhecimento nos ambientes reais. Já que os testes recebem um sinal $y(t)$ bem diferente do que era esperado com $x(t)$. Esse descasamento entre as etapas de treinamento e teste nos sistemas RAV provoca sérios erros no reconhecimento, prejudicando o desempenho do sistema. A Fig. 2.7 mostra como o efeito do ruído aditivo degrada o sinal de voz reduzindo as taxas de acerto dos sistemas RAV para os diferentes ruídos das tarefas do

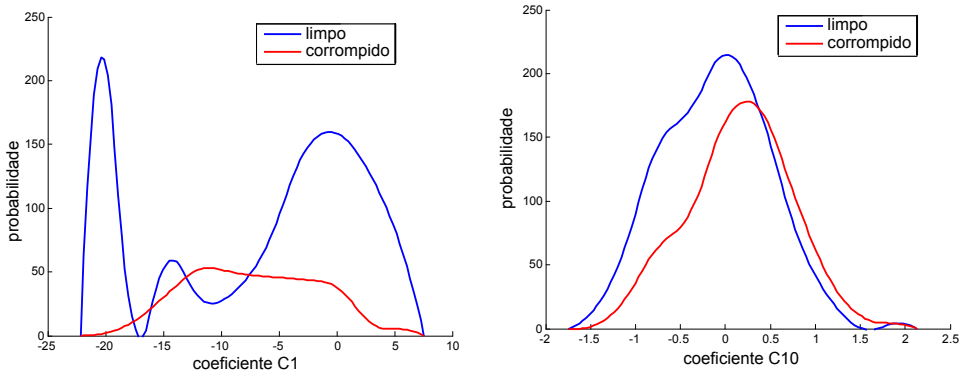

Figura 2.5: Distribuição de probabilidade dos coeficientes cepstrais C1 e C10 da frase 440c020b do banco de dados Aurora-4, limpo (azul) corrompido com ruído babble de 5dB (vermelho).

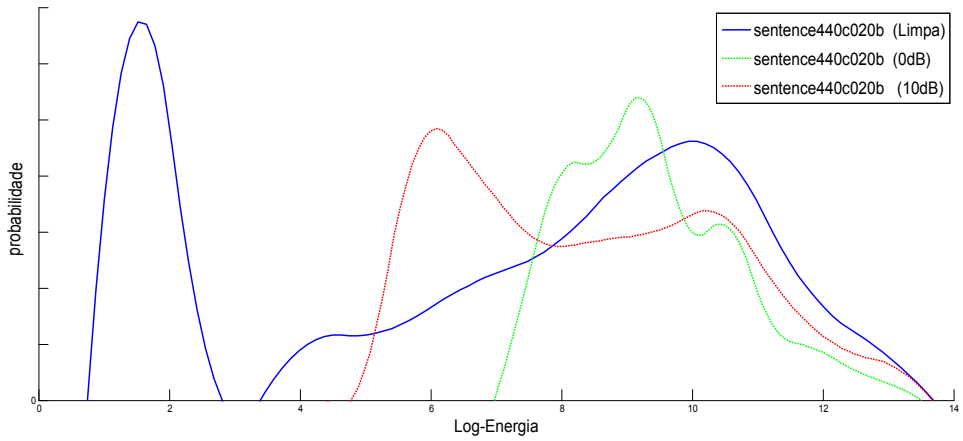

Figura 2.6: Distribuição de probabilidade do coeficiente cepstral C0 da frase 440c020b do banco de dados Aurora-4 com relações sinal ruído de 0 e 10dB

banco de dados AURORA-4 em um sistema hibrido DNN-HMM. Pode-se ver que para ruídos como *babble* e *restaurant* o sistema apresenta o pior rendimento. Nessa experiência, os sinais de voz ruidosos foram criados adicionando ruídos artificialmente sem filtragem linear envolvida, ou seja supondo desprezível a presença do ruído convolutivo $h(t)$ e outras distorções próprias do locutor (ex. efeitos dos pulsos glotais e a impedância de radiação dos lábios). Mesmo sendo de importância nos sistemas robustos de reconhecimento de voz, estes ruídos e distorções não serão abordados nesta tese.

Segundo [82], uma solução a este problema é ter um banco de dados de treinamento igual ao número de condições adversas que possam ocorrer. Porém, é uma tarefa difícil reunir dados de todos os ambientes possíveis. Devido a essas limitações vários trabalhos abordam o problema desenvolvendo técnicas robustas, as quais serão apresentadas a seguir.

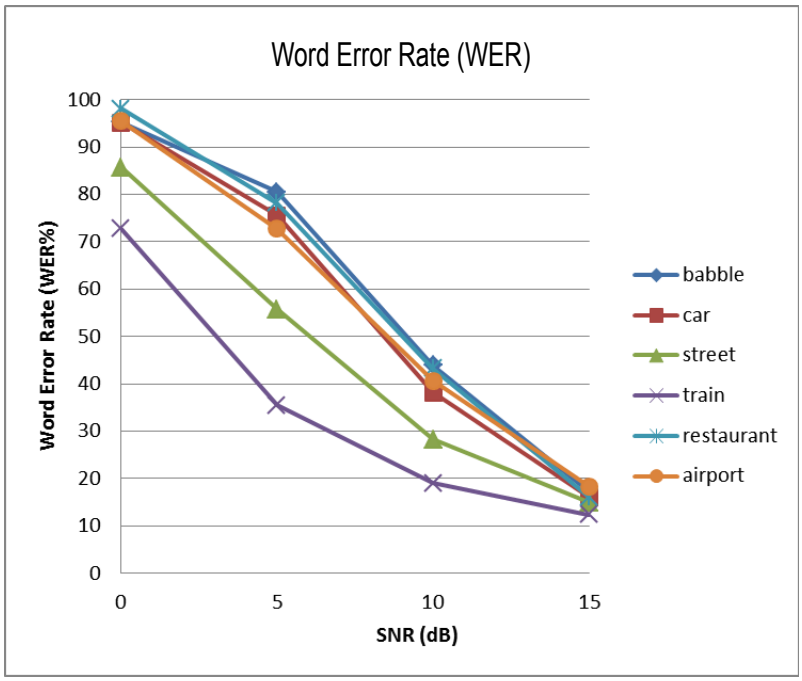

Figura 2.7: Word error rate (WER) em diferentes tipos de ruído

2.4 Revisão das Técnicas de Reconhecimento Robusto

Na seção anterior foi analisado o efeito do ambiente acústico sobre os parâmetros estatísticos da voz, e mostrou-se como este modifica as distribuições de probabilidade da voz, gerando um descasamento entre as condições de treinamento e teste dos sistemas RAV. Visando melhorar o desempenho dos sistemas RAV e minimizar os efeitos do ambiente acústico, têm sido realizados nas ultimas décadas importantes esforços para tornar mais robustos os sistemas frente a essas situações. Nesse sentido têm sido desenvolvidos diversos métodos que procuram reduzir o descasamento entre as condições de projeto e operação do sistema. Estes métodos podem ser agrupados em três abordagens principais:

- Técnicas de realce do sinal de voz;
- Técnicas de compensação de atributos;
- Técnicas de adaptação de modelos.

Nas seguintes seções, serão feitas breves descrições dos aportes mais relevantes de cada uma destas abordagens, salientando-se que esta tese esta focada nas técnicas de realce de voz e na compensação de atributos. Para realce de voz, foram desenvolvidas técnicas baseadas em mascaramentos através de *Local Binary Pattern* (LBP) sobre diferentes domínios (espectrais, Capítulos 4 e 5, e cepstrais, Capítulo 6). E para compensação de atributos foi proposta uma filtragem de mediana sobre as funções de distribuição dos atributos

corrompidos antes de equalizá-los para uma função de referência gaussiana, cuja abordagem será apresentada no Capítulo 3.

2.4.1 Técnicas de realce do sinal de voz

Muitas das técnicas de realce de voz foram desenvolvidas inicialmente mais para melhorar (realçar) a qualidade da voz do que para aplicações em robustez de reconhecimento de voz. As técnicas pertencentes a esta categoria foram as primeiras a serem aplicadas nos métodos de robustez entre os anos 60 e 70 [83], e desde então, e ao longo das últimas décadas, muitas pesquisas focaram-se nessa área. Grande parte dessas técnicas são amplamente utilizadas nos sistemas RAV atuais por sua simplicidade e por permitirem solucionar situações adversas que não sejam muito desfavoráveis, por meio de métodos simples de filtragem passa baixa, atenuando o ruído que está por fora da banda em análise. O objetivo principal destas técnicas, denominadas técnicas de realce de fala, é eliminar o ruído do sinal antes da extração de atributos e seu posterior reconhecimento, como se mostra na Fig.2.8.

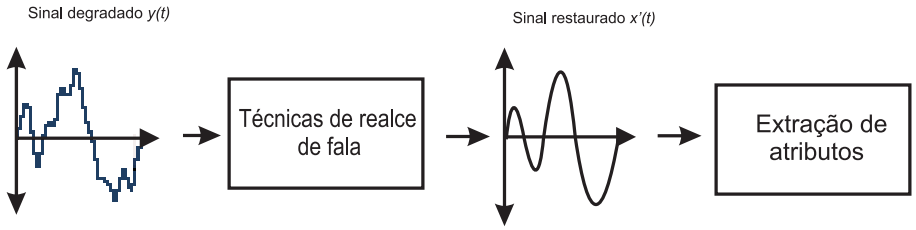

Figura 2.8: Restauração do sinal por meio de técnicas de realce de voz

Uma característica importante desse tipo de abordagem é a sua forma de tratar os efeitos do ruído tanto no domínio do tempo quanto no domínio da frequência, através de métodos de realce de voz e restauração de voz.

Nos métodos de restauração de voz estão incluídas todas as técnicas que processam o sinal a fim de separar as distorções causadas pelo ruído. Segundo [84] frequentemente ouve-se a voz em um ruído de fundo onde os fonemas individuais podem estar mascarados, mas a compreensão é possível. Desse modo as técnicas de restauração procuram que o sinal de voz processado seja o mais parecido com o original (ideal), onde a voz "ideal" dependerá do ambiente de aplicação. Por outro lado segundo [85] as técnicas de realce da voz apresentam como objetivo básico o fato de que a voz processada tenha um melhor som do que a não processada. Um outro ponto importante segundo [71] é o fato da voz poder ser classificada de acordo com o número de microfones disponíveis. Ela pode usar um único ou multicanal. O realce de voz multicanal proporciona um melhor desempenho do que a de um canal único, mas, devido

às suas implementações e facilidade computacional, a técnica de realce de fala de canal único ainda é uma área de pesquisa significativa [86],[87].

O acima exposto mostra que o objetivo das técnicas de realce de voz varia de acordo com a aplicação em questão, já que sempre será possível diminuir o ruído, mas sempre às custas de introduzir distorções, como o conhecido ruído musical que por sua vez prejudica a inteligibilidade da voz. Com base nesse contexto e procurando manter a inteligibilidade e a qualidade do sinal de voz contaminado por ruído aditivo, foram propostas ao longo das ultimas décadas inúmeras técnicas de realce de voz, algumas das quais são mencionadas a seguir.

Um trabalho clássico e bastante relevante nessa área é o presentado por Boll [88] chamado subtração espectral (SS). A subtração espectral é uma das técnicas mais simples e efetivas que permitem melhorar a relação sinal-ruído do sinal de voz degradado com ruído aditivo [89][90]. Esse método está baseado na premissa de que o sinal de voz e o ruído são decorrelatados e aditivos no domínio do tempo, de modo que o espectro em potência do sinal degradado é a soma dos espectros da potência da voz e do ruído. A técnica clássica de SS realiza a melhoria do sinal de voz estimando o ruído nos momentos em que não há presença de voz. O princípio básico desta técnica é resumido no diagrama em blocos da Fig. 2.9.

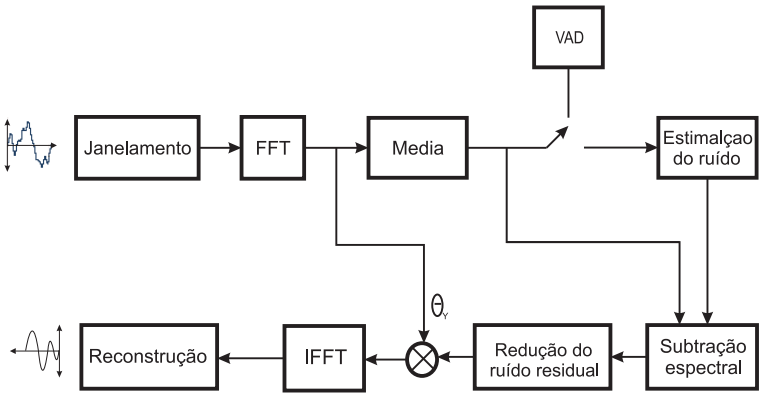

Figura 2.9: Diagrama de blocos do processo de subtração espectral.

Segundo o modelo de ambiente acústico apresentado na equação (2-1), e supondo desprezível a presença do ruído convolutivo e outras distorções próprias do locutor, a expressão analítica da SS que proporciona uma estimação do espectro do sinal restaurado é dada por

$$| \widehat{X}_{(\omega)} |^{\alpha_s} = \left(| Y_{(\omega)} |^{\alpha_s} - \beta_s (E \mid \widehat{R}_{(\omega)} |)^{\alpha_s} \right) \theta_y \qquad (2\text{-}14)$$

onde $\widehat{X}_{(\omega)}$ é a estimativa do espectro do sinal limpo menos a estimativa do espectro ruído ($E \mid \widehat{R}_{(\omega)} |$) que é calculado nos instantes onde a atividade vocal

é nula, por meio de um detector de atividade de voz / não voz (VAD)[91], ω o índice da frequência e α_s indica o domínio onde sera efetuada a subtração (domínio da magnitude $\alpha_s = 1$ ou domínio da potência $\alpha_s = 2$), β_s é um fator de sobre subtração o qual compensa parcialmente as deficiências da estimativa do ruído configurado geralmente com valores entre $1 \leq \beta_s \leq 2$, e θ_y é a fase do sinal, que devido ao fato de que o ouvido humano não é sensível à fase do sinal, pode-se utilizar a fase do espetro do sinal ruidoso como estimativa da fase para reconstruir o sinal limpo.

Nas últimas décadas surgiram vários estudos que melhoraram a ideia original correspondente à subtração espectral. Um estudo comparativo publicado por *Chaudhari e Dhonde* em 2015 [92] mostra as contribuições mais bem sucedidas ao longo do tempo sobre a subtração espectral e chega à conclusão que os sistemas baseados nessa técnica são amplamente utilizados como estado da arte por serem os mais efetivos quando agem sobre ruídos aditivos decorrelatados. No entanto, este método sofre duas importantes limitações. Por um lado as técnicas baseadas em subtração espectral são altamente dependentes da estimativa do ruído, já que um erro na estimativa dos períodos de silêncio dá lugar a uma estimativa "errônea" do espectro de ruído. Por outro lado as técnicas baseadas em SS dão lugar a valores negativos da potência espectral que ocasionam uma operação não linear que produz o conhecido ruído musical que é composto de vários sinais de banda estreita (tons) cujas amplitudes e frequências variam com o tempo, produzindo uma sensação auditiva desagradável.

Para contornar essa limitação, outras técnicas de realce foram propostas na literatura, como a filtragem de Wiener [93] [94] [95]. O filtro de Wiener possui a característica de lidar com estatísticas de primeira e segunda ordem. Ele busca minimizar a variância da diferença entre o sinal realçado e o sinal limpo considerando atividade de voz e ruído. O filtro Wiener assume que o sinal de voz e o ruído aditivo são processos estocásticos estacionários com características espectrais conhecidas, sendo um filtro ideal para recuperar a voz limpa, já que tem a particularidade de ser ótimo de acordo com o mínimo erro médio quadrático (MMSE) entre o sinal limpo e o sinal obtido pelo processo de filtragem [96] [97], minimizando o erro entre o sinal de interesse e a sua estimação. Em [98] foi usado o ganho de filtro de Wiener para aumentar os pesos da amplitude do espectro e melhorar a atenuação dos sinais de interesse.

A ideia fundamental do filtro de Wiener é estimar um filtro ideal da voz corrompida. Este filtro age como um peso de frequência no módulo de espectro do sinal ruidoso. Considerando um sinal de voz estacionário corrompido $y(t)$, o sinal realçado $\widehat{y}(t)$ através de um filtro linear invariante no tempo com resposta

ao impulso é dado pela equação

$$\widehat{y}(t) = \sum_{k=0}^{M-1} h_k y(t-k) \quad t = 0, 1, 2... \tag{2-15}$$

onde h_k são os coeficientes do filtro e M é o numero de coeficientes. O erro $e(t)$ entre o sinal estimado $\widehat{y}(t)$ e o sinal desejado $x(t)$ é dado por:

$$e(t) = x(t) - \widehat{y}(t) \tag{2-16}$$

O objetivo do filtro de Wiener é minimizar esse erro. Para isso o valor médio quadrático do erro estimado é comumente usado como critério de minimização. Informações detalhadas da estimativa ideal do filtro, assim como do mínimo erro quadrático, podem ser encontradas em [71][99][100][101].

Outra técnica que surgiu como uma poderosa ferramenta para remover o ruído do sinal corrompido foi proposta por *Donoho e Johnston* em [102]. Esta técnica chamada *Wavelet-denoising* baseia-se na aplicação de um limiar às componentes de ruído aditivo do sinal corrompido em cada subbanda de alta frequência a fim de minimizar o ruído ou eliminar as componentes indesejadas (dependendo do limiar utilizado), mantendo as informações importantes do sinal de voz. Através desta técnica é possível realizar uma filtragem do sinal degradado $y(t)$ para eliminação do ruído e posteriormente, restaurar o sinal original $x(t)$ ou pelo menos gerar um similar, aproveitando as características das transformadas *wavelet* que podem dividir um determinado sinal em diferentes componentes de escala, conseguindo encontrar informações de frequência sem perder informações temporais, de modo que os coeficientes *wavelet* de alta frequência podem ser ajustados de acordo com um certo limiar. Esse processo pode ser representados por meio das equações

$$\begin{aligned} \widehat{W}_{(a,b)} &= W(y(t)) \\ Z_{(a,b)} &= D(\widehat{W}_{(a,b)}, \lambda) \\ \widehat{x}(t) &= W^{-1}(Z_{(a,b)}) \end{aligned} \tag{2-17}$$

onde $W(.)$ e $W^{-1}(.)$ representam os operadores da transformada wavelet e sua inversa respectivamente, $D(., \lambda)$ representa o operador de *denoising* com limiar λ, e os parâmetros $a, b \in R$, sendo $a \neq 0$, representam os parâmetros de escala e deslocamento, respectivamente. Informação detalhada da implementação da técnica *wavelet-denoising* assim como algumas modificações podem ser

encontradas em [103][104].

Segundo [105] e [106], um ponto chave dessa abordagem é que por meio da transformada *wavelet* é possível modelar o comportamento do canal auditivo humano, representando a informação no domínio tempo-frequência, permitindo melhorar algumas das limitações da transformada de Fourier. Por exemplo, a largura da janela de tempo que uma vez escolhida é mantida fixa durante toda a análise, ao contrário da transformada *wavelet* que varia mostrando exatamente a localização de pequenas descontinuidades do sinal no tempo, trazendo assim mais poder e flexibilidade para analisar o sinal de voz. Outra caraterística interessante que oferece esta transformada, além da representação no plano tempo-frequência, é que faz a descorrelação dos parâmetros espectrais. Devido a essas características, a transformada *wavelet* tem sido usada no lugar da transformada de Fourier janelada clássica, para extração de características espectrais do sinal de voz [74]. Na literatura diversas variações de *wavelet denoising* tem sido propostas. Em [107] o método de realce baseado em *wavelet-denosing* é usado para atenuar mais o sinal após aplicar uma função de ganho para minimizar o erro médio quadrático, a fim de separar quadros de voz e não voz com base na probabilidade calculada pela razão de verossimilhança de dois modelos de mistura de Gaussianas (GMMs). Em [108] uma seleção de canal de modulação é usada como função de limiar para *denoising*. Nessa proposta se faz uma decomposição de três níveis obtendo 8 sub-bandas usando a função *wavelet* mãe *symlet4* onde as sub-bandas de alta frequência são passadas pela função de limiar para supressão de ruído.

Uma outra técnica pertencente aos métodos de supressão de ruído baseados no domínio da frequência é a bem conhecida decomposição modal empírica (EMD) [109], que foi proposta para análise de processos não-lineares e não estacionários. Esse método, quando aplicado sobre ruídos, resulta em uma decomposição semelhante àquela obtida por um banco de filtros [110]. Um ponto importante desse método em comparação com a decomposição em sub-bandas *wavelet* é que as funções base são uma decomposição do próprio sinal. A ideia principal por trás desse método é que o sinal ruidoso é decomposto de forma adaptativa em componentes oscilatórios chamados funções de modo intrínseco (IMFs), usando uma decomposição temporal chamada processo de peneiração. Em um estudo mais recente [111] os autores apresentam uma modificação da técnica de filtragem baseada em EMD (EMDF) dos componentes de ruído de baixa frequência inspirada pela aproximação de baixa classificação tipicamente usada em algoritmos de reconhecimento de voz subespacial. Esse método separa a voz do ruído ao analisar as estatísticas de segunda ordem das funções do modo intrínseco (IMF) formadas a partir da

decomposição EMD dos sinais de voz. Este método é particularmente eficaz em ambientes de ruído de baixa frequência. Em [112] os autores substituem o filtro da proposta EMDF por um novo esquema de realce multivariado usando a decomposição do modo empírico (MEMD). Esse esquema levou à proposta de uma técnica que pode alinhar os modos de frequência comuns em vários canais de dados multivariados, facilitando assim a decomposição direta de dados multicanal.

Outro conjunto de técnicas que têm mostrado bons resultados nos métodos de realce de voz em ambientes ruidosos nas últimas décadas são derivados a partir de estudos sobre a percepção auditiva humana. Esses métodos baseados na estimativa de máscara de segregação [113][114] ou técnicas de dados perdidos [115][116], utilizam as propriedades dos sistemas auditivos humanos para calcular o limiar de mascaramento do ruído [117] a fim de selecionar da matriz de mascaramento quais unidades contêm energia predominante de voz considerando-as como úteis, e o resto como energia de voz dominada pelo ruído e por tanto não contendo informação útil [118]. Em [119] foi apresentado um estudo que comprovou que as taxas de reconhecimento não melhoravam quando o número de unidades úteis a serem reconhecidas era mínimo. Esse estudo mostrou que as técnicas de mascaramento dependem altamente do detector de componentes, ou seja de uma boa estimação da SNR.

Um dos pontos-chave a ter em conta na concepção dessas técnicas de mascaramento é a escolha do tipo de máscara, já que existem máscaras contínuas e máscaras discretas. As primeiras são de tipo binário onde cada unidade do espectro é representada com 1 quando a voz domina sobre o ruído e 0 caso contrário. Por exemplo, considerando o nosso ambiente acústico, equação (2-2), desprezando o ruído convolutivo, a máscara discreta $D\aleph_{(t,\omega)}$ pode ser representada da seguinte maneira

$$D\aleph_{(t,\omega)} = \begin{cases} 1, & se \quad SNR_{(t,\omega)} > Th \\ 0, & caso \quad contrario \end{cases} \tag{2-18}$$

onde

$$SNR_{(t,\omega)} = log\frac{S_x(t,\omega)}{S_r(t,\omega)} \tag{2-19}$$

e $S_x(t,\omega)$ e $S_r(t,\omega)$ representam a energia do sinal e do ruído, respectivamente no tempo t e frequência ω, e Th é um limiar que decidirá se a unidade correspondente a esse tempo t e a essa frequência ω é considerada útil ou

não.

Por outro lado, as unidades de uma máscara contínua, diferentemente da máscara discreta, estão na faixa 0 e 1, indicando a probabilidade de que uma unidade esteja dominada pela energia da voz. A forma de representar esse tipo de mascaramento é por meio da função sigmoide definida da seguinte maneira

$$C\aleph_{(t,f)} = \frac{1}{1 + exp^{-\alpha(SNR_{(t,\omega)} - \beta_{\aleph})}} \tag{2-20}$$

onde α controla a inclinação da função sigmoide e $\beta_{\aleph}$ representa o limiar.

Uma outra distinção importante na escolha de tipo de máscara é estabelecer se a máscara será ideal ou estimada. No primeiro caso as máscara ideais, ou também chamadas oráculo, são usadas para máscaras binárias que usam um conhecimento prévio que não está disponível na maioria das aplicações da vida real. Uma grande objeção ao conceito de máscaras oráculo é que elas não são usadas em aplicações reais devido ao conhecimento a priori requerido. No entanto, as máscaras oráculo estabelecem um limite superior de desempenho, o que as torna úteis como referências para algoritmos de mascaramento desenvolvidos em aplicações reais. O cálculo deste tipo de máscara só pode ser realizado nas situações em que é possível estimar, precisamente, a densidade espectral da potência de ruído. Por outro lado, as máscaras estimadas permitem estimar a confiabilidade de cada unidade da matriz de mascaramento das unidades analisadas. Como esperado, o desempenho da máscara estimada é menor, procurando-se aproximar o máximo possível à máscara ideal.

Wang *et al.* em [120] introduziram uma abordagem de segregação de voz, que é a tarefa de separar a voz do ruído de fundo. Ela considera que os sons que atingem o ouvido estão sujeitos a um processo chamado Análise da Cena Auditiva (ASA para o acrônimo em inglês) [26]. Com base nesse processo foi proposta a *Ideal Binary Mask* (IBM) [121] [122] que foi sugerida como objetivo computacional dos algoritmos *Computational Auditory Scene Analysis* (CASA) [32], cujo principal objetivo é fazer a separação das distintas fontes de som que compõem a entrada acústica. Continuando com a mesma linha de mascaramento discreto, Srinivasan *et al.* propuseram uma alternativa de mascaramento chamada *Ideal Ratio Mask* (IRM) [123]. Os autores mostraram que em uma banda de frequência estreita, existe uma relação sistemática entre SNR *a priori* e valores de *Interaural Time Differences*(ITD) e *Interaural Intensity Differences*(IID), o que os motivou a estimar uma razão ideal de mascaramento usando estatísticas coletadas para ITD e IID em cada faixa de frequência individual. Matematicamente a IRM é estreitamente ligada ao

ganho do filtro de Wiener [124].

Mais recentemente, um novo tipo de mascaramento baseado em *deep neural networks* (DNNs) foi apresentado [125][126] com o objetivo de extrair características acústicas do sinal corrompido, a fim de treinar um algoritmo de aprendizado supervisionado. Nessa configuração o alvo de treinamento ou o sinal de aprendizado está configurado para aprender uma função que mapeie as características de uma máscara binária construída a partir do sinal corrompido e do sinal de ruído. Os resultados apresentados mostraram que o rendimento quando o treinamento com DNN é usado supera os métodos tradicionais. Em [127] uma máscara IRM baseada em *deep neural network* foi proposta para melhorar a robustez dos sistemas de reconhecimento. Essa técnica faz uma estimativa suavizada da tradicional IRM no domínio de frequência MEL filtrando o ruído de um espectrograma do sinal ruidoso de cada saída do banco de filtros antes de realizar o cepstrum. Em [128] o mascaramento IRM é substituido por um mascaramento baseado em ativações ocultas IHM. Nessa proposta, *Bo Li. et al.* propõem que ao invés de mascarar as unidades dominantes de ruído, descartem-se as unidades ocultas que geram ativação inconsistente para voz de diferentes condições.

Na literatura podem ser encontradas inúmeras variações referentes a essas técnicas de estimação de máscaras aplicadas tanto aos coeficientes de potência espectral em escala logarítmica quanto aos coeficientes cepstrais, proporcionando em ambos os casos melhorias similares nos resultados de reconhecimento [71] [129] . Uma análise mais detalhada das técnicas de mascaramento mais relevantes dos últimos tempos, que compõem o estado da arte dos métodos de mascaramento, são reunidos nos artigos [130] e [131].

2.4.2 Técnicas de compensação de atributos

Em contraste com as técnicas de realce de voz vistas na seção anterior, as técnicas de compensação de atributos atuam sobre as características parametrizadas com o objetivo de recuperar o melhor possível os vetores de atributos limpos, preservando o poder discriminativo destes. A Fig. 2.10 mostra um diagrama em blocos genérico dessa categoria de técnicas. Através deste conjunto de técnicas procura-se conseguir uma representação do sinal de voz que seja mais robusta que os parâmetros extraídos tradicionalmente (LPC, MFCC, etc.). Para realizar essa tarefa as técnicas são divididas em duas abordagens: redução de ruído por meio de filtragem e normalização dos atributos. A seguir descrevem-se brevemente algumas das técnicas usadas nessa abordagem.

Quando o sinal de voz é corrompido pelo ruído, tanto a distribuição

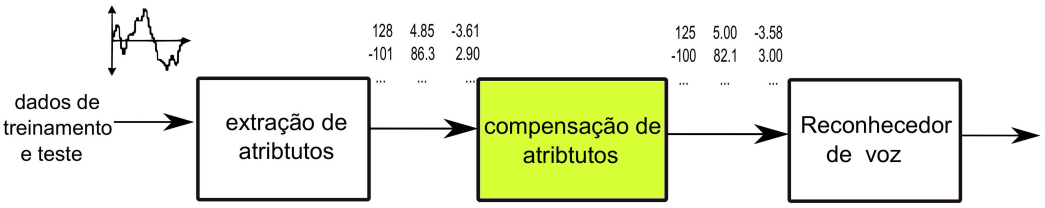

Figura 2.10: Restauração de atributos por meio de técnicas de compensação.

estatística quanto a estrutura temporal são distorcidas. Portanto é desejável filtrar e normalizar também a estrutura temporal dos atributos. Um dos principais inconvenientes na hora da parametrização do sinal de voz são as mudanças no ambiente acústico devido aos diferentes tipos de ruído que variam rapidamente em comparação com as variações temporais nas características da voz[4]. Para lidar com isso, no primeiro caso, foi proposto em [132] uma técnica conhecida como RASTA, que executa uma filtragem passa-banda para eliminar as distorções devido à resposta em frequência do canal, suavizando os atributos da voz ao longo do eixo temporal, removendo as frequências menores que 1Hz e atenuando as frequências maiores que 16Hz, já que segundo [133] a maior parte da informação linguística encontra-se compreendida nas frequências de modulação na faixa de 1 a 16 Hz, suavizando assim a correlação entre as componentes de características por meio de uma filtragem das sequências temporais dos parâmetros espectrais. Essa filtragem pode ser realizada em cada componente no domínio log espectral [134] ou no domínio cepstral [135]. Eliminando essas variações das componentes espectrais em comparação com a faixa típica de variação do sinal de voz, melhora-se significativamente a precisão dos sistemas RAV em condições adversas, já que se conseguem suavizar as trajetórias resultantes dos cepstrum em comparação com as originais. Matematicamente o filtro RASTA [75] é definido como

$$H(z) = 0.1 \cdot z^4 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - Kz^{-1}} \tag{2-21}$$

Na equação 2-21 pode-se observar dois termos importantes: (i) no denominador existe uma constante K que está associada à menor frequência de corte, cujo valor indica o grau de integração das informações dos quadros anteriores e é definida empiricamente; (ii) no numerador, observa-se uma regressão de segunda ordem semelhante à dos coeficientes delta-cepstrum.

[4]A taxa de mudança dos componentes não linguísticos presentes no sinal de voz é muitas vezes diferente da do trato vocal que codifica a mensagem linguística [73].

A aplicação do RASTA geralmente pode reduzir a incompatibilidade entre condições de treinamento e teste [136]. No entanto o RASTA pode não ser útil em situações em que não há incompatibilidade entre essas condições, ou seja, em condições acústicas onde o treinamento e o teste são semelhantes, pois aplicar uma filtragem aos vetores de características cepstrais de um sinal limpo elimina informações necessárias para o reconhecimento [137].

Recentemente em [138] foi proposta uma nova abordagem de filtros temporais chamada *Temporal Structure Normalization* (TSN). Nesse método os autores tentam normalizar a estrutura temporal dos atributos ruidosos, representados pela função densidade espectral de potência (PSD), para uma função de referência (atributos limpos) por meio de filtragem linear.

Por outro lado, os métodos que usam normalização são mais eficientes e simples na hora de remover as variações lentas da voz. Essas técnicas visam eliminar o *offset* ou deslocamento da média, transformando as características extraídas da voz para um domínio em que a variabilidade introduzida pelo ruído seja minimizada, ou os momentos estatísticos sejam normalizados para uma referência comum, geralmente as distribuições estatísticas dos atributos limpos, reduzindo assim o descasamento entre as distribuições dos dados utilizados para treinamento e teste.

Uma das técnicas mais conhecidas na literatura é a apresentada por F. Liu *et al.* em [139] chamada normalização de média cepstral (CMN), a qual subtrai o valor médio dos coeficientes cepstrais sobre toda a expressão (combinação de vetores cepstrais) e subtrai essa média de cada segmento (vetor cepstral único). Ou seja, dada uma sequência de vetores de atributos no domínio cepstrais $(C_y(1), ..., C_y(N))$ onde N é o número de segmentos, a sequência normalizada pela técnica CMN $(\widehat{C}_y(1), ..., \widehat{C}_y(N))$ é obtida subtraindo a média amostral $\overline{C_y}$ de cada vetor $C_y(n)$ da seguinte forma

$$\begin{aligned} \overline{C}_y(n) &= \frac{1}{N}\sum_{n=1}^{N} C_y(n) \\ \widehat{C}_y(n) &= C_y(n) - \overline{C}_y \end{aligned} \tag{2-22}$$

A diferença de desempenho entre CMN e RASTA pode ser devido a uma possível perda de informação, pois a operação de filtragem no RASTA remove não apenas o componente constante, mas também algumas outras mudanças lentas em cada componente cepstral. Como consequência dessa operação, resulta que o primeiro momento da distribuição dos dados observados, ou seja, a média cepstral, seja zero. No entanto, apenas a normalização da média de características para zero é muitas vezes insuficiente para melhorar a robustez

do RAV em casos de ruído aditivo. Por esse motivo foi apresentada em [140] uma técnica de normalização como extensão do método CMN, chamada normalização da variância cepstral (CVN), proporcionando maior robustez ao ruído, já que a variância dos atributos de voz é dimensionada de forma diferente no domínio cepstral causada pelo ruído aditivo devido ao operador logaritmo no processo de extração de atributos [5]. Da mesma forma que a CMN, a normalização da variância cepstral é uma técnica que utiliza implicitamente a dinâmica (ou seja, as variações temporais) do espectro da voz, suprimindo as partes de menor variação da representação espectral, com base na hipótese de que as variações mais lentas são geralmente devidas aos efeitos da canal de transmissão e carregam pouca informação.

Anos depois, uma tendência emergente no domínio do aprimoramento da voz surgiu com a mistura das duas técnicas de normalização cepstral CMN e CVN a fim de eliminar o descasamento produzido pelo canal, os efeitos de filtragem linear, a variabilidade do locutor, e o ruído aditivo [11]. Essa nova técnica chamada normalização de média e variância cepstral (CMVN) [141] [142] normaliza o primeiro e segundo momento da distribuição dos dados observados, fazendo com que a média da sequência de vetores de atributos seja zero e a variância um. Dessa forma, ao calcular e subtrair o valor médio de cada um dos vetores e normalizar pela variância, assegura-se que o efeito do ruído (aditivo) sobre os coeficientes cepstrais seja reduzido. Isso ajuda a minimizar a variabilidade dos dados e aumenta a eficácia dos sistemas RAV. Em [143] apresenta-se uma análise mais detalhada das diferentes técnicas de normalização cepstral.

De forma análoga à forma como a técnica CVN veio complementar a técnica CMN, incluindo uma normalização da variância, que reduz apenas os efeitos lineares do ruído, foi apresentada por *Segura et al.* em [144] uma técnica mais robusta de normalização de atributos, que transforma não linearmente a função de distribuição de probabilidade dos coeficientes cepstrais, modificando-a para que coincida com a distribuição de dados de treinamento. Esta técnica, chamada equalização de histogramas (HEQ), foi apresentada como uma extensão da técnica CMVN, na qual normaliza-se não só os primeiros momentos dos vetores de características, mas todos os momentos da função de distribuição de cada atributo. A eficácia do método depende de uma estimativa adequada dos histogramas da voz a ser compensada, e a suposição de que o processo de ruído distorce o espaço da característica aplicando-se à transformação monotônica. Assim, as transformações estimadas dessa maneira podem corrigir a distorção do espaço de representação, já que igualam a distribuição de probabilidade da voz ruidosa para a distribuição de probabilidade da voz em condições de

referência. Em geral o que o HEQ faz é normalizar as funções de densidade de probabilidade dos dados de treinamento e teste, transformando-os em um terceiro pdf comum que se torna uma referência. Ao longo dos últimos anos várias têm sido as modificações da técnica original, visando uma melhor representação dos vetores de atributos que levem a melhorar as taxas de acerto dos sistemas RAV. Em primeiro lugar é escolhida a distribuição de referência com as quais os vetores de atributos serão igualados. Em [145] foi proposta uma distribuição estatística gaussiana de média zero e variância um. Segundo [146] a função densidade de probabilidade do sinal de voz tem uma forma de onda similar a uma Gaussiana, além de que grande parte dos sistemas de reconhecimento as distribuições de saída dos HMMs são modeladas como mistura de Gaussianas. Já em [147] foi proposta uma função de distribuição empírica estimada a partir dos histogramas acumulativos dos dados de treinamento. Dessa forma a função de distribuição de referência terá uma forma muito semelhante a uma mistura gaussiana, cujos valores de média e variância têm a ver com características próprias do banco de dados e do processo de treinamento. Por outro lado deve decidir-se em que domínio se aplica a técnica HEQ. Num trabalho publicado em 2006 [148], foi proposto fazer a equalização na saída do banco de filtros Mel conseguindo uma melhora nas taxas de erro de palavras, reduzindo de $45,7\%$ para $25,5\%$ (treinamento limpo) e de $19,5\%$ para $17,0\%$ (treinamento multi condição). Já em [149] os autores propuseram a equalização sobre os coeficientes cepstrais, já que possuem propriedades estatísticas diferentes e capacidades discriminativas decrescentes ao aumentar a ordem do coeficiente, ajudando na eficiência computacional assim como nas taxas de acerto dos sistemas RAV. Os autores propõem que a equalização dos coeficientes C_0 e C_1 seja responsável pelo maior incremento nas taxas de acerto, aumentando o número de coeficientes quando existe ruído de canal adicionado ao ruído aditivo. Uma melhoria adicional do HEQ foi proposta em [150], onde os parâmetros são estimados para normalizar os histogramas dos atributos e maximizar a probabilidade dos atributos normalizados avaliados no modelo acústico. Recentemente em [151] foi proposta uma mistura de uma filtragem temporal de média com a normalização baseada em HEQ chamada *filtered-based histogram equalization* (FHEQ) com o objetivo suavizar a sequência de funções de distribuição de probabilidades antes do mapeamento.

Uma série de esquemas populares [152][153] são baseados na compensação vectorial da série Taylor (VTS), que usa uma aproximação linear para representar a influência do ruído no sinal de voz limpo. Segundo [154] o objetivo de VTS é estimar a função de densidade de probabilidade da voz degradada $p(y)$, dada a função densidade de probabilidade da voz limpa $p(x)$ um segmento do

sinal corrompido e o desenvolvimento da série de Taylor que relacionara a voz limpa com a voz corrompida. Assim que a função densidade de probabilidade da voz degradada for calculada, se faz uma estimativa baseada no minimo erro médio quadrático a fim de prever a sequência da voz limpa não observada. Uma das grandes desvantagens deste método é seu alto costo computacional [155].

Uma das mais recentes contribuições baseadas nas técnicas de normalização foi apresentada em 2017 por *Fredes et al.* [23]. Essa técnica conhecida como *Locally Normalized Filter Banks* (LNFB) é o procedimento anterior ao cálculo da DCT dos atributos *Locally Normalized Cepstral Coefficients* (LNCC), na qual aplica-se uma normalização local no espaço do banco de filtros, dividindo a saída de um filtro triangular de ponderação de frequência pela saída de um segundo filtro de ponderação de frequência. Essa normalização remove grandes variações na forma espectral que podem ser consideradas constantes em ambos os filtros, como as oscilações, que assumem-se surgir principalmente da variabilidade do canal. Nessa contribuição os autores mostram que misturando essa técnica no domínio log-espectral com as técnicas acima expostas como CMN e CMVN, consegue-se uma redução significativa nas taxas de erro de palavra.

2.4.3 Técnicas de compensação de modelos

Quando o sinal de voz é corrompido por ruído, a distribuição dos atributos de teste é totalmente diferente dos atributos de treinamento, isto é, há descasamento entre condições de treinamento e teste. Como resultado, o modelo acústico não representa mais exatamente os atributos de teste e, portanto, o desempenho do reconhecimento é degradado. Para melhorar a robustez, diversos métodos foram propostos baseados em compensação de modelos, a fim de melhorar as representações dos atributos de teste. Estes métodos têm o mesmo objetivo das técnicas anteriores, que é minimizar ou compensar o descasamento entre as condições dos ambientes de treinamento e teste. No entanto, a diferença dos esquemas de compensação de atributos é que eles agem diretamente no *back-end* dos sistemas RAV modificando os modelos acústicos geralmente criados com base em mistura de Gaussianas dos atributos na fase de treinamento. Segundo [156], os métodos de compensação de modelos geralmente requerem dados de compensação e sua transcrição para adaptar o modelo acústico. De um modo geral, quanto mais dados de compensação, mais eficaz é o procedimento.

A Figura 2.11 representa um diagrama de blocos do processo de com-

pensação realizado no módulo de classificação que geralmente é usado para modificar as distribuições treinadas com voz limpa de acordo com as condições acústicas do ambiente.

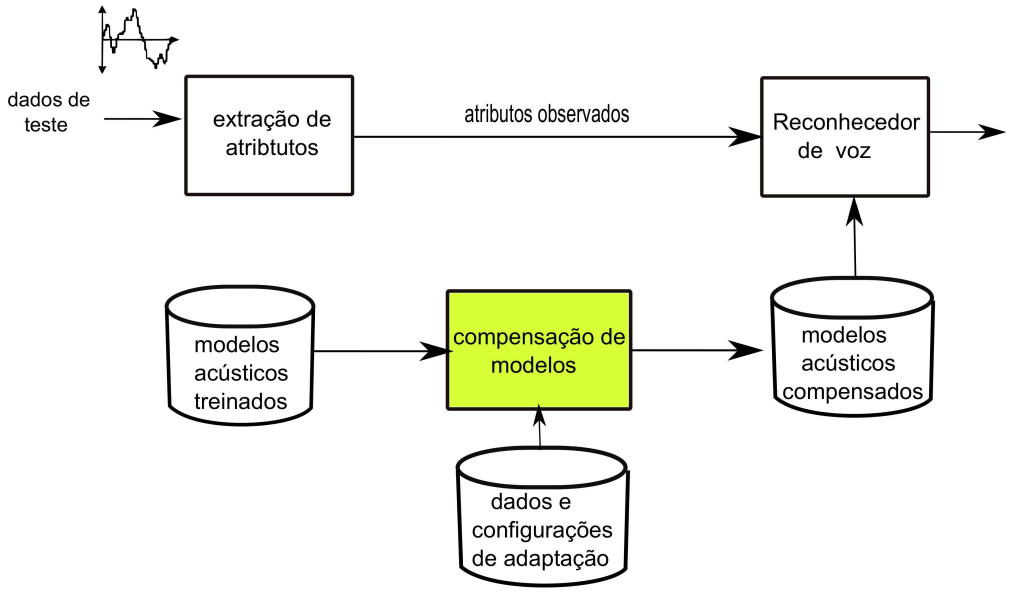

Figura 2.11: Restauração de atributos por meio de técnicas de compensação.

O principal objetivo da maioria dessas técnicas é obter um novo conjunto de modelos que seja o mais parecido possível com o que seria obtido se treinado nas mesmas condições. Note-se que quando o sinal de voz é corrompido por ruído de canal ou ruído aditivo, a distribuição dos atributos de teste é diferente dos atributos de treinamento, gerando incompatibilidade entre condições de treinamento e teste. Como resultado, o modelo acústico não representa mais exatamente os atributos de teste que degradam o desempenho dos sistemas. A principal vantagem desta abordagem é, portanto, que não é necessário fazer hipótese sobre o sinal de voz, pois o que é feito é usar a voz corrompida para adaptar os parâmetros dos modelos disponíveis para reconhecimento. Além disso, os dados de teste não sofrem nenhuma alteração, ou seja, este tipo de abordagem é aplicado simplesmente na etapa de treinamento. O uso de modelos analíticos permite que essas técnicas sejam muito eficientes no tempo e eficazes em memória computacional. Entre os métodos que se enquadram nesse foco ao longo dos últimos anos, destacam-se: adaptação estatística de modelos por meio de *regressão linear de máxima probabilidade* (MLLR) e *adaptação de máximos a posteriori* (MAP) que podem ser supervisionados ou não supervisionados, correspondendo as situações em que se tem as transcrições dos dados ou não, respectivamente [5][152]. Destaca-se também a decomposição de modelos por meio da *Parallel Model Combination* (PMC) pudendo ser de modo estático

(*offline*), quando se usa os dados de adaptação antes de fazer o reconhecimento, ou de modo incremental quando se adapta os modelos conforme avança o reconhecimento (*online*)[47][157] .

Uma das alternativas mais simples de adaptação de modelos baseia-se em treinar os modelos acústicos com dados corrompidos. Este tipo de treinamento, chamado multicondição [158], oferece altas taxas de reconhecimento devido ao fato de treinar os HMMs com voz adquirida em ambientes reais, usando diferentes tipos de ruídos e diferentes relações sinal ruído. No entanto, um dos maiores problemas deste tipo de técnica é que deveria se conhecer todas as condições adversas do mundo real, o que exigiria um grande número de dados adquiridos em múltiplos ambientes, já que em presença de algum tipo de ruído desconhecido as taxas de acerto do sistemas começam a diminuir.

Um pouco mais recente, em [159] foi proposta uma nova forma de treinamento chamada *noise adaptive training (NAT)* que pode ser aplicada a todos os dados de treinamento, normalizando a distorção ambiental como parte do modelo de treinamento. Os parâmetros dos modelo pseudo-limpos aprendidos com o NAT são usados para adaptá-los por meio da técnica de compensação de modelos da *série de Taylor Vectorial (VTS)* [160] para decodificar os sinais corrompidos no momento do teste

Procurando dar uma solução ao treinamento multicondição em sistemas reais, onde a quantidade de dados de treinamento é pequena, foi proposta a técnica MAP [161], que considera os parâmetros a serem adaptados como variáveis aleatórias, permitindo incorporar informações *a priori* que orientam o processo de estimativa. Ou seja, encontrar aqueles atributos que maximizam a probabilidade dos mesmos. Sob este pressuposto, pode-se fazer uma estimativa, em um quadro estatístico, dos parâmetros para obter a voz limpa. Porém, o MAP mesmo sendo melhor que o treinamento multicondição ainda continua com a grande desvantagem de conjuntos de adaptação pequenos. A razão disso é que ele simplesmente modificará os atributos das unidades acústicas observadas no conjunto de adaptação, o que significa que as unidades acústicas não observadas permanecerão inalteradas. Por outro lado, é uma técnica que só pode ser usada de modo *offline*, já que precisa fazer integrações numéricas que envolvem a distribuição de voz limpa, ruído e voz contaminada, o que leva a problemas práticos para sua aplicação em tempo real.

Uma outra técnica de adaptação de modelos acústicos baseada em transformações lineares foi apresentada por *Leggetter et al.* em [162] chamada MLLR, onde assume-se que a matriz média e a matriz de covariância do modelo adaptado são transformações lineares do modelo inicial. Ela encontra um conjunto de matrizes de transformação que maximizam a probabilidade de ob-

servar os vetores cepstrum ruidosos. Essa técnica foi proposta principalmente para adaptar sistemas independentes de locutor a um determinado locutor a fim de adaptar as médias da mistura de gaussianas do modelo previamente treinado para vários locutores, mas também provou ser eficaz para compensação de ruído ambiental [163]. Uma desvantagem desse tipo de técnica é que as matrizes de transformação para adaptar a média são diferentes aquelas usadas para adaptar as matrizes de covariância, o que eleva o custo computacional dos sistemas. Para dar solução a este problema foi proposta em [164][165] uma adaptação de MLLR chamada fMLLR, onde a transformação é usada para adaptar tanto a matriz de médias como as de covariâncias. Uma vantagem dessas técnicas é que pode ser usada tanto para adaptar os modelos acústicos quanto para modificar o espaço de atributos, reduzindo o custo computacional, já que não precisa modificar os HMMs.

Foi desenvolvida Também uma técnica de adaptação de modelos onde a voz corrompida é modelada através de um HMM de MXN onde M é o número de estados para modelar a voz limpa e N para modelar o ruído. Essa técnica, conhecida como PMC [47][157] cria um modelo para o ruído e um para a voz corrompida a fim de misturá-los para obter um modelo adaptado do ambiente acústico com o qual será avaliado o sistema. O novo modelo HMM resultante é muito semelhante ao que seria obtido por meio do treinamento com os vetores contaminados do ambiente acústico avaliado. Como foi exposto na seção 2.3, a interação entre voz e ruído é expressa mais naturalmente no domínio do banco de filtros, assim o modelo HMM final é transformado ao domínio cepstral mediante a aplicação da transformada discreta do coseno DCT. Dessa técnica inúmeras variações têm sido apresentadas nas últimas décadas [166][167][168][169].

Em [170] foi proposta uma outra alternativa para compensar modelos baseada na mistura de Gaussianas. Essa técnica, chamada *subspace gaussian mixture models* (SGMM), utiliza um modelo de subespaços globalmente compartilhado entre os estados, de forma a capturar as maiores virações do modelo, provendo uma representação compacta dos modelos acústicos resultando em uma estimativa robusta de atributos, o que permite obter melhoras no desempenho dos sistemas de reconhecimento.

Recentemente, aproveitando a capacidade que as redes neurais têm de aprender comportamentos não-lineares a partir de um conjunto de dados foram apresentadas em [6][171][172] novas formas de modelar o ambiente acústico baseadas em redes neuronais profundas DNN. Essas redes representam uma maneira alternativa de avaliar quão bem cada estado de cada HMM se encaixa em um quadro ou uma janela curta de quadros de coeficientes que representam

a entrada acústica.

Aproveitando esta nova forma de modelagem acústica, que tem mostrado maior eficácia no estado da arte em comparação com os modelos acústicos baseado em GMM-HMM [6], foram desenvolvidas técnicas utilizadas para adaptar modelos limpos a ambientes com determinados dados de adaptação usando DNN. Algumas das mais relevantes são apresentadas a seguir.

Em [173] foi desenvolvido um sistema de adaptação de modelos chamado *adaptation methods for context dependent deep neural network hidden Markov models* (CD-DNN-HMM) que faz uma transformação dos parâmetros na camada oculta superior das DNN conseguindo diminuir as taxas de erro de palavra de 17% para 14%. A motivação dessa adaptação é baseada em uma visão do DNN como um processo em duas etapas: extração de recursos não-lineares em camadas inferiores seguido de uma camada de classificação log-linear no topo.

Seltzer *et al.* em [172] fizeram um estudo sobre o desempenho da robustez do ruído por meio das técnicas de modelagem acústica baseadas em DNN e apresentaram três métodos do estado da arte na modelagem acústica, que melhoram as taxas de reconhecimento dos sistemas RAV, baseados no treinamento adaptativo do ruído. Os dois primeiros são focados no espaço de características e no espaço dos modelos. Esses métodos usam informações sobre a distorção acústica, seja através de aprimoramento de atributos antes do treinamento da rede ou durante o treinamento da mesma. O terceiro refere-se ao recentemente proposto método de treinamento chamado *dropout* o qual é usado nos conjuntos de dados onde o sobre ajuste (*over-fitting*) é uma preocupação.

A ideia principal do trabalho apresentado por Seltzer *et al.* [172] é incorporar métodos de robustez no momento do treinamento das DNNs. Para isso, começa com um treinamento multicondição tomado como referência, o qual permite que a rede aprenda atributos de nível superior que sejam mais invariantes para os efeitos do ruído em relação à precisão da classificação. Com esse treinamento, é alcançado um melhor rendimento em ambientes adversos de teste com ruídos de 5 e 15dB em comparação com os modelos acústicos tradicionais baseados em GMMs que fornecem 8% a mais de erro para o mesmo conjunto de teste. Fica comprovado assim, que os modelos acústicos baseados em DNN são mais robustos que os GMM.

Visando melhorar esse rendimento os autores propuseram reduzir o efeito do ruído sobre as DNNs usando algoritmos de realce de voz antes do treinamento desses atributos com DNN, já que é uma maneira óbvia de reduzir a variabilidade nos atributos causada pela distorção do ambiente, ou seja tentar

removê-la das observações. Por exemplo, *Bayesian feature enhancement* (BFE) [174] que é um método que melhora de forma eficiente os atributos de voz corrompidos por ruído aditivo ou reverberação nos *logarithmic mel-frequency power spectral* LMPS, sendo considerado um pré-processamento eficiente para os sistemas RAV robusto com menos nós na rede.

Ainda em [172] outra alternativa foi proposta, mas dessa vez o objetivo foi incorporar informação referente ao ambiente acústico no momento do treinamento das DNNs. Essa nova abordagem chamada $DNN-NAT$ faz uma estimativa do modelo do ruído a fim de adaptar os parâmetros gaussianos do reconhecedor com base em um modelo físico que define como o ruído corrompe a voz limpa. O desempenho dessa nova abordagem gerou melhor desempenho do sistema, reduzindo os erros de reconhecimento em 3% em comparação com o sistema de referência.

Como último método de robustez proposto por Seltzer *et al.* [172] está o treinamento dos DNNs com *dropout*. Essa nova abordagem busca reduzir o problema de *over-fitting* presente nos treinamentos de algoritmos baseados em redes neuronais e que usam um conjunto de treinamento relativamente pequeno. O *dropout* é um algoritmo relativamente novo para treinamento de redes neurais, que se fundamenta na eliminação aleatória de neurônios durante o processo de aprendizagem, para evitar a sobre adaptação aos dados (*over-fitting*), conseguindo assim uma considerável melhoria nos resultados dos sistemas RAV. Essa abordagem melhorou em quase 6% as taxas de erro de palavra em comparação com os melhores resultados publicados, baseados simplesmente em decodificação DNN, ou seja, sem nenhuma técnica de robustez adicional.

Mais recentemente, em [175] os autores propuseram uma adaptação de modelo acústico através de duas configurações: a adaptação tradicional não supervisionada e uma adaptação supervisionada, onde alguns minutos de discurso transcritos estão disponíveis. Por meio desta última abordagem e fazendo uso de uma modificação da técnica MLLR para agir no espaço dos atributos, foi possível aumentar a robustez dos modelos acústicos baseados na mistura das transformações fMLLR com a modelagem DNN.

Em [176] [177] [178] são apresentadas em detalhe outras técnicas de compensação de modelos acústicos baseadas em DNN.

2.5 Conclusões

Neste capítulo, foi apresentado o modelo geral do ambiente acústico dos sistemas RAV, descrevendo uma expressão analítica que modela o efeito das duas principais fontes de ruído, aditivo e convolutivo, no sinal de voz.

Com base nessa descrição se fez uma análise de como os efeitos do ambiente acústico no sinal de voz afeta severamente as características principais do sinal no domínio espectral e cepstral, degradando drasticamente o rendimento dos sistemas, já que se gera uma modificação das distribuições de probabilidade deslocando a sua média e reduzindo a variância em relação àquelas da voz limpa, produzindo assim um descasamento entre as condições de treinamento e as de teste. Apresenta-se também uma breve visão geral das estratégias clássicas de robustez mais relevantes nos últimos tempos, com o objetivo de oferecer um melhor desempenho frente ao ruído aditivo nos sistemas RAV. Essas estrategias foram agrupadas em três categorias (i) técnicas de realce de voz, as quais buscam melhorar a inteligibilidade e qualidade da voz; (ii) técnicas de compensação de atributos, que têm como objetivo principal fornecer aos sistemas RAV atributos mais discriminativos; e (iii) técnicas de compensação de modelos, que buscam criar modelos acústicos que se adaptem a qualquer ambiente. Como foi expressado no início do capítulo, as contribuições originais dessa tese se concentrarão no desenvolvimento de algoritmos de robustez enquadrados nas técnicas de realce de voz e compensação de atributos.

3
Reconhecimento de Voz Robusto Baseado em Filtragem por Mediana da Função Distribuição de Probabilidade

Foi visto no capítulo anterior que dentre o grande número de técnicas de reconhecimento de voz robusta propostas na literatura, as várias abordagens adotadas são divididas em três categorias. Uma delas é a compensação de atributos, onde cada vetor de características é modificado para ficar mais parecido com o de um sinal limpo, como se mostra na Fig. 3.1. Este conjunto de técnicas tem como objetivo reduzir o descasamento entre os atributos de treinamento e os de teste, em particular modificando suas distribuições de probabilidade, sem perder seu poder discriminativo.

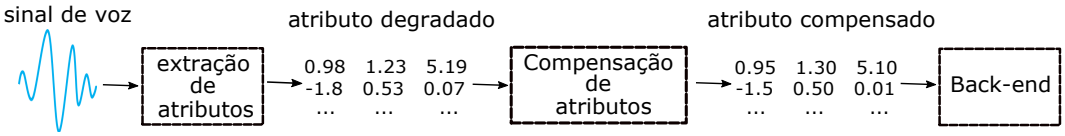

Figura 3.1: Diagrama de blocos do sistema de reconhecimento robusto baseado em compensação de atributos.

Neste capítulo segue-se essa filosofia e propõe-se uma nova abordagem de compensação de atributos para aumentar a robustez dos coeficientes cepstrais PNCC usados nos sistemas RAV. A razão de usar PNCC (e não MFCC) neste capítulo será vista posteriormente. A proposta apresentada neste capítulo é realizada através de uma filtragem não linear sobre as funções de distribuição de probabilidade (*PDF*s) dos atributos corrompidos no *front-end* do reconhecedor. Em seguida é aplicada uma transformação não linear para um domínio invariante às distorções que o ruído provoca nas *PDF*s de cada atributo, a fim de eliminar as elevadas intensidades dos ruídos locais. Os resultados das simulações e a eficácia do método proposto são comparados com outras técnicas recentes da literatura.

3.1
Equalização de histogramas

A corrupção do sinal de voz causada pelo ruído aditivo degrada o rendimento dos sistemas de reconhecimento, devido à distorção do espaço

de representação, que modifica os diferentes momentos estatísticos de cada atributo utilizados para representar o sinal de voz, gerando um descasamento entre as condições de treinamento e de teste, como foi apresentado no capítulo 2. Essas modificações aleatórias causadas pelo ruído podem ser descritas em termos dos valores médios do sinal e modeladas através de funções de distribuição de probabilidade (*PDF*s).

Como foi descrito no Capítulo 2, algumas das técnicas que compensam os atributos através da normalização dos momentos das distribuições dos dados observados são a normalização da média cepstral (CMN) e a normalização da média e da variância (MVN), que normalizam o primeiro e os dois primeiros momentos da distribuição dos dados observados, respectivamente [139][140]. Porém, por se tratar de transformações lineares, a CMN não é suficiente para compensar os efeitos não-lineares do ruído aditivo. Isso torna esse método efetivo apenas para níveis moderados de ruído aditivo. Por outro lado, a MVN [141] tem demostrado melhores resultados que a técnica CMN, já que como foi apresentado no Capítulo 2, o ruído aditivo além de gerar um deslocamento da média produz uma compressão da variância e o MVN normaliza esses dois momentos da distribuição dos dados observados trazendo uma redução no descasamento e aumentando as taxas de reconhecimento. No entanto, a distorção não linear causada pelo ambiente acústico (ruído aditivo) não afeta apenas a média e a variância das distribuições de probabilidade, mas também os momentos de ordem superior.

Um outro método que transforma os atributos extraídos do sinal de voz, para um domínio no qual a variabilidade introduzida pelo ruído é mais reduzida, é a técnica conhecida como Equalização de Histogramas (HEQ). A HEQ é capaz de lidar com o efeito não-linear causado pelo ruído aditivo através da normalização de todos os momentos da distribuição de probabilidade de cada vetor de atributos, isto é, ela normaliza os histogramas dos atributos da voz limpa e corrompida para uma distribuição uniforme. Existem várias motivações por trás do uso de técnicas que normalizam os momentos de ordem superior no processamento de sinal. Estas incluem (i) eliminar o ruído aditivo de um espectro de potência desconhecida; (ii) detectar e caracterizar as propriedades não-lineares nos sinais, bem como identificar sistemas não-lineares.

A HEQ [14] [179]é uma técnica frequentemente usada em processamento de imagens. Embora continuemos denominá-la HEQ, ela tem por objetivo aplicar uma transformação não linear que modifique o histograma da imagem original para um histograma de referência, melhorando o brilho e o contraste, otimizando a faixa dinâmica da escala de cinzas, mas não está limitado a este

tipo de aplicação. Em [180] *Balchandran e Mammone* introduziram a HEQ no reconhecimento de locutor para compensar as características da voz e aumentar a robustez do reconhecimento, removendo as distorções não lineares no cepstrum LPC de um sistema de identificação de locutor, ajustando os parâmetros dentro de uma faixa comum e usando como distribuição de referência os dados de treinamento limpos. Já em [144][145][181] a implementação de HEQ no *front-end* dos sistemas RAV, fez-se para compensar os efeitos do ruído no domínio cepstral dos coeficientes MFCC, onde foi mostrado que este método fornece melhorias substanciais na taxa de reconhecimento de voz em condições de ruído, quando usado como uma técnica de compensação isolada ou quando aplicado em combinação com outros métodos como VTS.

Ao contrário das técnicas baseadas em realce de voz, a HEQ não faz nenhuma suposição a *priori* sobre as características do ruído, nem a forma como este distorce a voz. Isso permite que eles abordem com sucesso diferentes tipos de distorções que podem degradar o sinal de voz.

Quando a técnica HEQ é aplicada sobre os atributos da voz, cada componente do vetor de atributos tanto da fase de treinamento como do reconhecimento é mapeada independentemente, aplicando uma transformação não linear que modifica as distribuições dos dados de teste fazendo com que estes assemelhem-se às distribuições dos dados de treinamento.

O fundamento teórico da técnica HEQ está relacionado com as propriedades das variáveis aleatórias, no qual para uma variável aleatória x com função de densidade de probabilidade $p_x(X)$ e função de distribuição de probabilidade *PDF* $C_x(X)$, ela possa ser transformada para uma variável y através da função $Y_{HEQ} = F(X)$ com uma *PDF* de referência $C_y(Y)$, mantendo idêntica *PDF* ($C_x(X) = C_y(Y)$), contanto que a transformação aplicada $F(X)$ seja inversível. Para isso, define-se uma variável aleatória uniforme ν. Da teoria de probabilidade, sabe-se que ν satisfaz às seguintes equações:

$$\nu = \int_0^x p_x(X)dX = C_x(X) \tag{3-1}$$

$$\nu = \int_0^y p_y(Y)dY = C_y(Y) \tag{3-2}$$

Logo, substituindo Y por $F(X)$ e igualando as equações (3-1) e (3-2), resulta que

$$C_x(X) = C_y(F(X)) \tag{3-3}$$

O que significa que a função F é dada por

$$Y_{HEQ} = F(X) = C_y^{-1}(C_x(X)) \tag{3-4}$$

onde C_y^{-1} representa a *PDF* de referência inversa. Ou seja, $z = C_y^{-1}(\rho) \Longleftrightarrow C_y(z) = \rho, \rho \in [0, 1]$. A transformação $F(X)$ definida em (3-4) é uma função monotônica não decrescente ao longo do tempo e como pode-se ver em (3-4) sua expressão é definida em função da *PDF* da variável que se transforma. O processo da transformação é detalhado em [180]. O objetivo é aplicar uma transformação não linear, que transforme a função densidade de probabilidade da voz contaminada em uma função densidade de probabilidade de referência, como mostrado na Fig.3.2.

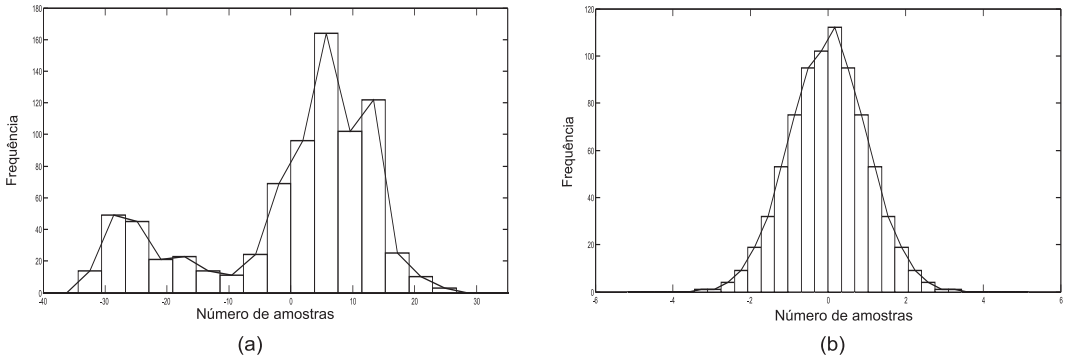

Figura 3.2: Mapeamento de histogramas do coeficiente C_0 dos atributos MFCC *(a) pdf* do coeficiente cepstral original *(b) pdf* do coeficiente cepstral mapeado.

Em outras palavras, o reconhecimento foi levado para um domínio onde os dados são idealmente invulneráveis às transformações lineares e não lineares que o ruído aditivo possa provocar.

São diversas as configurações que deve-se levar em conta na hora de equalizar os vetores de atributos. As mais importantes são:

- Escolha do domínio de equalização;
- Escolha da distribuição de referência;
- Estimação dos histogramas dos dados observados.

3.1.0.1
Escolha do domínio de equalização

Como foi apresentado em capítulos anteriores, a representação das características principais do sinal de voz passam por uma análise espectral onde cada segmento passa por um banco de filtros com caraterísticas próprias do sistema auditivo humano. Em [148][182] os autores estudaram os efeitos da equalização nesse domínio, tomando como vetor de equalização a saída logarítmica dos bancos de filtros. O fundamento principal de equalizar no domínio do banco de filtros é que pode-se compensar distorções específicas de algumas frequências que têm efeitos independentes em determinados componentes dos bancos de filtros. Por outro lado, trabalhos como [145] e [147] fazem a equalização no domínio da quefrencia, atuando sobre os coeficientes Mel e suas derivadas, mostrando melhores resultados em domínios menos correlatados devido ao fato de HEQ aplicar-se por separado a cada componente do vetor de atributos.

Segundo [183], quando a equalização do histograma é implementada no domínio cepstral pode-se considerar como uma extensão natural da técnica MVN. O HEQ fornece uma equalização da média e variância (como MVN) e iguala o resto dos momentos de ordem superior, que afetam a forma das distribuições de probabilidade dos atributos do sinal de voz.

Em [184] foi feita uma análise comparativa da aplicação da técnica HEQ no domínio cepstral tradicional MFCC e nos atributos PNCC, mostrando maior robustez à equalização sobre os atributos PNCC. Por esse motivo os atributos PNCC serão usados para os testes referentes a este capítulo.

3.1.0.2
Escolha da distribuição de referência

As duas formas de usar a referência são através de uma distribuição estatística ou uma distribuição empírica. No primeiro caso a distribuição estatística mais usada é a distribuição gaussiana de média zero e variância unitária. Segundo [146], essa função de referência tem uma vantagem fundamental frente a outras funções, devido ao fato de que na maior parte dos sistemas de reconhecimento as distribuições de saída dos HMM serão modeladas com mistura de gaussianas. Por outro lado, as distribuições empíricas são estimadas a partir dos dados de treinamento construídas empiricamente mediante a utilização de histogramas cumulativos. No entanto, esse procedimento requer uma carga computacional maior, já que essa técnica exige que o número de dados para equalizar seja suficiente para representar de maneira ótima as estatísticas principais do sinal de voz.

Com essas considerações, neste trabalho decidiu-se realizar o mapeamento através de funções de referência gaussianas com função densidade de probabilidade dada por

$$p_y(Y) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{-Y^2}{2}\right) \tag{3-5}$$

3.1.0.3 Estimação dos histogramas dos dados observados

A fim de fazer uma implementação eficiente, utiliza-se um número finito de observações e, portanto, os histogramas cumulativos são utilizados no lugar de distribuições de probabilidades. Por esse motivo, o procedimento é chamado de equalização de histograma em vez de equalização de distribuição de probabilidade [145]. O procedimento é realizado da seguinte maneira:

- Calculam-se os histogramas cumulativos da função de referência $p_y(Y)$;
- Para cada componente do vetor de atributos, tanto no treinamento quanto no teste, calculam-se os histogramas cumulativos, os quais serão usados para estimar aproximadamente sua *PDF* $C_x(X)$. Para obter a transformação de cada componente, o histograma cumulativo foi estimado segundo as especificações de [145]. Dado um conjunto de N observações correspondente aos valores de um coeficiente cepstral, a *pdf* pode ser aproximada através de seu histograma por meio da seguinte equação:

$$p_x(X \in I_i) = \frac{n_i}{N} \tag{3-6}$$

  onde n_i é o número de observações do intervalo I_i. A função de distribuição pode ser aproximada por

$$C_x(x_i) = C_x(X \in I_i) = \sum_{j=1}^{i} \frac{n_j}{N} \tag{3-7}$$

  Para construir o histograma foram considerados 100 intervalos uniformes na faixa de $[\mu - 4\sigma, \mu + 4\sigma]$, onde μ e σ são a média e o desvio padrão do componente equalizado;
- Aplica-se a equação (3-4) a cada ponto do histograma cumulativo.

É importante ressaltar que essa estimativa da transformação a partir dos histogramas cumulativos pode ser tediosa e computacionalmente menos eficiente. Uma forma mais simples e de menos gasto computacional para obter a transformação é através da estatística ordenada dos dados, proposta por Segura *et al.* em [147].

A Fig. 3.3 mostra os efeitos da equalização de histogramas sobre a representação da voz no segundo coeficiente cepstral (PNCC) para distintas condições de SNR.

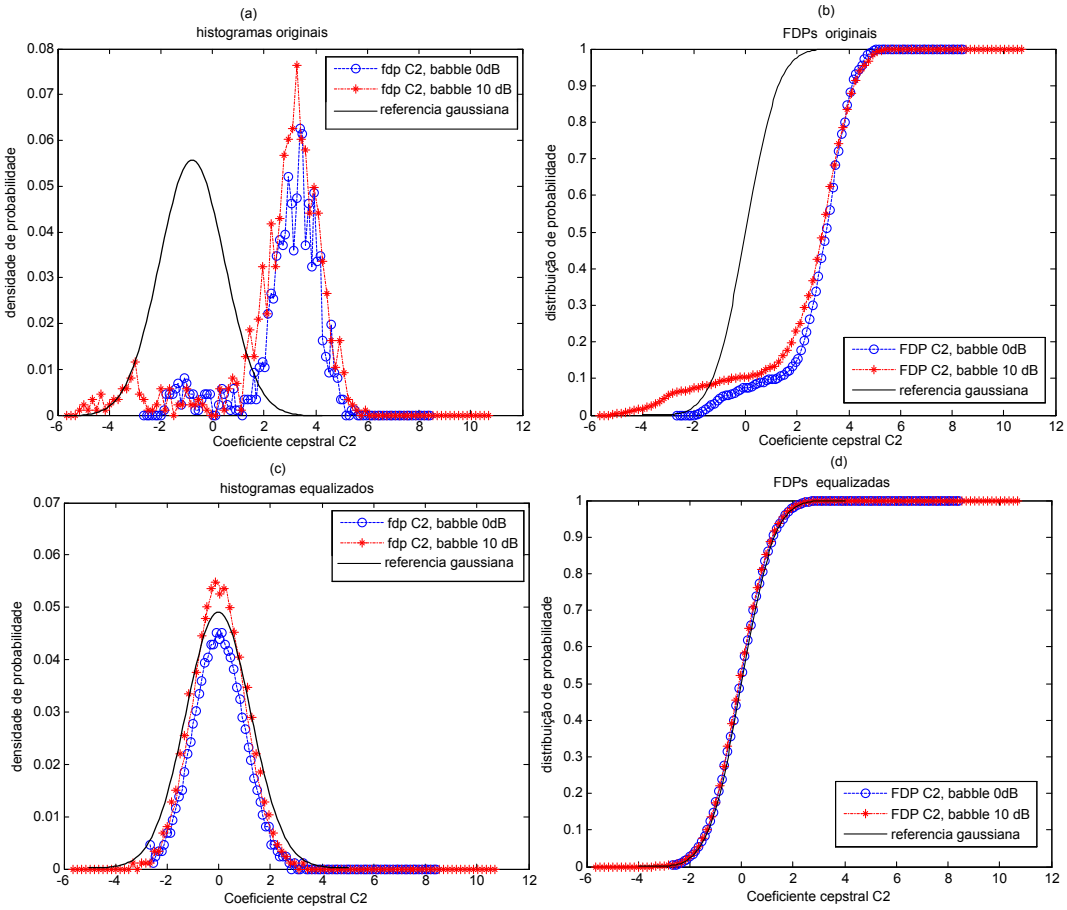

Figura 3.3: Efeitos da HEQ sobre a frase "440c020a" do banco de dados Aurora-4 com SNR de 0 e 10 dB (a) e (c) funções densidade de probabilidade dos coeficientes corrompidos e equalizados, respectivamente, e (b) e (d) funções de distribuição dos coeficientes originais e os transformados com HEQ respectivamente .

Pode-se ver como o ruído gera um deslocamento das médias assim como uma modificação das variâncias. Além das alterações desses momentos das distribuições, pode-se apreciar o efeito não-linear originado pelo ruído sobre o sinal de voz original que se manifesta em uma modificação da forma das distribuições de probabilidade. Isto é, uma alteração dos momentos de ordem superior. Na Fig.3.3 (c) pode-se ver como através da transformação realizada

em cada componente do vetor atributos consegue-se reduzir o descasamento entre as diferentes relações SNR, convertendo os histogramas originais do coeficiente *C2* no seu respectivo histograma de referência, conservando a mesma *pdf* de acordo com o acima exposto.

3.2 Filtro de suavização através da média temporal das funções de distribuição de probabilidade

Como foi visto na seção anterior, a HEQ compensa todos os momentos estatísticos das funções de probabilidade, minimizando o descasamento entre as condições de treinamento e as condições de teste. No entanto, HEQ produz uma perda de informação na hora de equalizar os histogramas dos dados observados para os histogramas de referência. Segundo [151] a sequência dos atributos corrompidos apresenta uma característica mais oscilante do que a sequência de atributos limpos, introduzindo distorções de frequência de alta modulação na hora da equalização, produzindo sequências de distribuições de probabilidade mais oscilantes. Buscando dar solução a esses desajustes causados pelas oscilações de alta frequência, recentemente, em [151] os autores propuseram um método que integra técnicas de filtragem linear temporal com HEQ. O método é chamado *filter-based histogram equalization* (FHEQ) e é motivado pelos conceitos de técnicas de filtragem temporais como RASTA [75] e ARMA [185]. FHEQ usa um filtro passa-baixa de primeira ordem com coeficientes 0,25 e 0,75, que suaviza a sequência original de *PDFs* de cada coeficiente MFCC antes da equalização, reduzindo erros de reconhecimento, já que cada ponto da nova *PDF* é uma soma ponderada dos dois pontos vizinhos do *PDF* original. Esta proposta modifica a equação (3-4) para

$$Y_{FHEQ_i} = C_y^{-1}(\sum_k h_k C_x(X_{i-k})) \tag{3-8}$$

onde h_k representa os coeficientes do filtro temporal, $C_x(X_i), i = 1, ..., N$ são as *PDF*s de uma sequência de atributos arbitrária $X_1, ..., X_N$, sendo N o número total de quadros (que é visto como o conjunto de amostras de uma variável aleatória x) e $Y_{FHEQ_1}, ..., Y_{FHEQ_N}$ é a nova sequência de atributos com *PDF*s aproximadas às *PDF*s de referência.

Assim a técnica FHEQ funciona como um filtro temporal passa-baixa que suaviza as funções de distribuição de probabilidade dos atributos antes de ser equalizados.

3.3 Filtro de mediana das funções de distribuição de probabilidade

Na seção anterior, apresentou-se um filtro de média temporal através de um filtro passa-baixa de dois pontos que foi utilizado a fim de obter uma nova sequência suavizada das *PDF*s originais. No entanto, nesses filtros cada ponto da função de probabilidade que foi degradado pode variar significativamente. Consequentemente, a média também pode ser muito diferente dos valores das *PDF*. Isto significa que este tipo de filtro é muito sensível a alterações locais.

Visando aumentar a robustez dos sistemas e melhorar os problemas apresentados pelos filtros lineares, é proposta uma nova abordagem de filtragem não linear das *PDF*s antes da equalização. Nessa nova abordagem em vez de filtrar as informações referentes às *PDF*s através de filtros lineares temporais, as respectivas *PDF*s dos dados de teste são suavizadas por meio de um filtro de mediana. Cada ponto da *PDF* de referência é gerado através do cálculo da mediana dos valores das *PDF*s da vizinhança ao redor do ponto correspondente da *PDF* original. Através da filtragem por mediana é esperada uma melhor suavização do que a filtragem por média em ambientes altamente ruidosos. Isto é porque a suavização por filtragem de mediana é menos sensível a elevadas intensidades de ruído local.

A filtragem por mediana sobre a *PDF* de cada atributo é obtida através de uma janela deslizante com um número ímpar de amostras através das *PDF*s. Substituindo o valor do meio pela mediana das *PDF*s na janela. Usando este conceito e a equação (3-4) para levar em consideração a equalização de histograma, obtém-se

$$Y_{MED-HEQ_i} = C_y^{-1}(Med[C_x(X_{i\pm k})], k \in W) \tag{3-9}$$

onde W é a janela em torno da *PDF* de cada coeficiente de característica sendo que o conjunto de valores de W devem estar ordenados em forma ascendente e *Med* é a função médiana que tomará o valor intermediário como o novo valor da *PDF*. Neste trabalho, foi selecionado $k = 3$.

A Fig. 3.4 mostra o procedimento do algoritmo MED-HEQ proposto nesta tese. A Fig. 3.4 (a), (b), (c), (d) e (e), respectivamente, apresenta a sequência original do coeficiente *C2* dos atributos PNCC correspondentes à frase "440c020a"do banco de dados Aurora-4, a distribuição de probabilidade original, a distribuição de probabilidade filtrada através do filtro de mediana, o coeficiente equalizado para a função de referência, e a sequência equalizada do coeficiente *C2*.

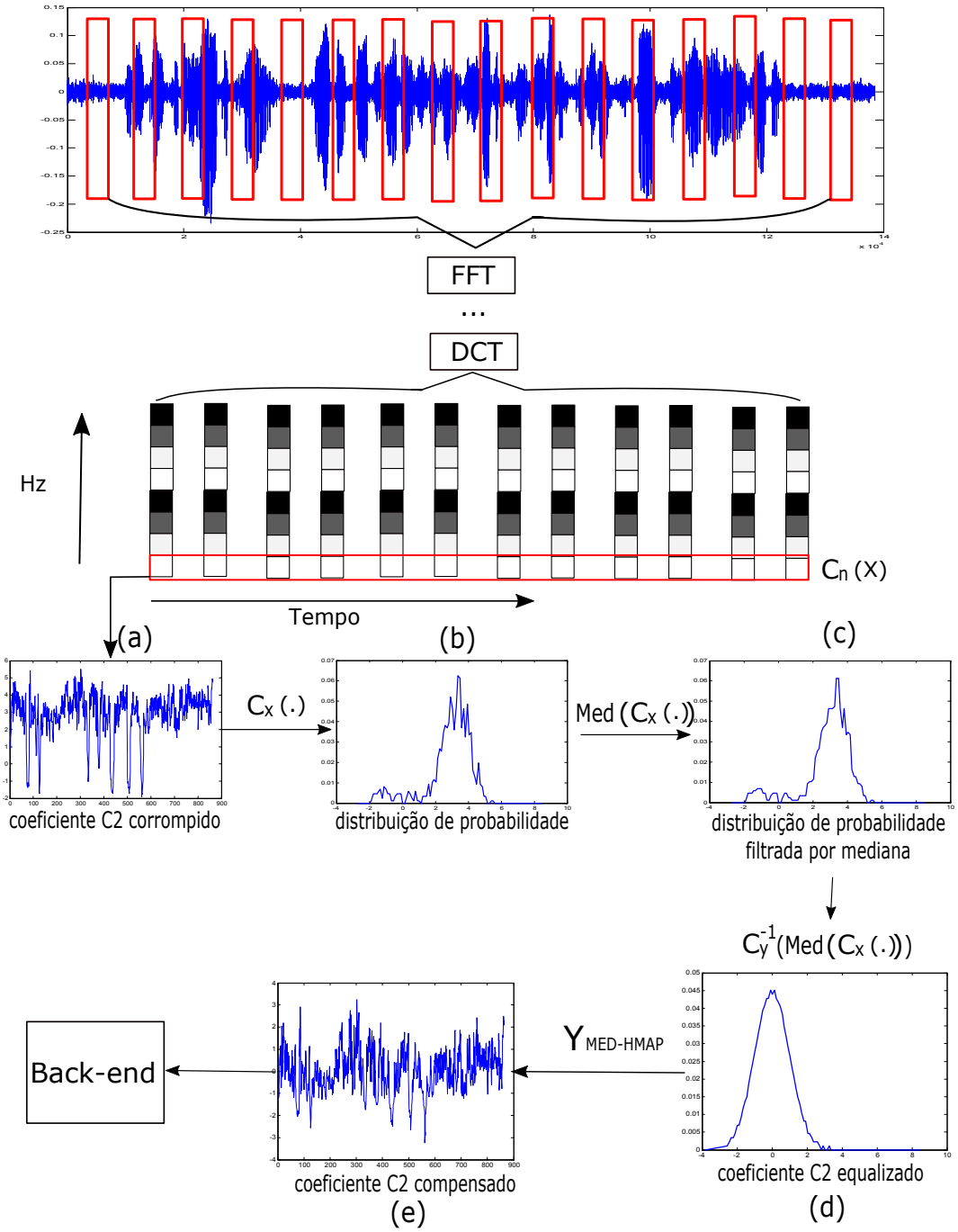

Figura 3.4: Procedimento do algoritmo MED-HEQ.

3.4 Configurações experimentais

A análise dos benefícios do método de filtragem de mediana sobre as *PDF*s proposto nesta tese é realizada usando 2 bancos de dados usados tradicionalmente em reconhecimento de voz contínua, o TIMIT e o AURORA-4. Nesta seção serão aprestadas de maneira geral as características mais importantes dos bancos de dados, as configurações do *front-end* e as condições

do sistema proposto usadas ao longo deste trabalho.

3.4.0.4 Bancos de dados de voz e de ruído

O banco de dados TIMIT [186] foi projetado para utilizar seus dados na aquisição de conhecimento fonético e acústico da voz, e para desenvolver e avaliar sistemas de reconhecimento de voz contínua independente de locutor. Ele foi criado com ajuda de diferentes grupos de pesquisa, respaldados pelo *Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO)*

Sua sigla vem de *Texas Instrument*(TI) e *Massachusetts Institute of Technology*(MIT). Possui um total de 6300 frases pronunciadas por 630 pessoas, das quais 70% são homens (438) e 30% são mulheres (192), onde cada um pronuncia 10 frases, abrangendo os diversos sotaques do inglês americano para ambos os sexos. Para este trabalho foram utilizadas em total 4620 sentenças para treinar e para criar o modelo de linguagem e 1000 sentenças para teste.

Por outro lado, o banco de dados Aurora-4 contém frases do *Wall Street Journal* (WSJ)[187]. O banco de dados de voz contínua que foi inicialmente lançado pelo grupo de trabalho STQ AURORA do Instituto Europeu de Normas de Telecomunicações (ETSI), a fim de avaliar os padrões DSR de reconhecimento de fala distribuídos e contém um vocabulário fechado de 5.000 palavras.

O conjunto de dados de treinamento consiste de 7318 frases gravadas com microfones de proximidade, sem adição de ruído. Os testes que formam a tarefa de avaliação foram construídos corrompendo as frases limpas com diversos tipos de ruído. Este trabalho selecionou sinais de ruído a partir de uma segunda base de dados: A NOISEX-92, que contém arquivos de som de diversas naturezas com níveis de SNR variando de 0dB a 15dB.

3.4.0.5 Configurações do sistema

A configuração utilizada para avaliar a robustez do sistema proposto é a empregada na maioria dos trabalhos vistos na área, citados ao longo desta tese, e é normalmente referida como configuração padrão (standard). Abaixo está uma lista que detalha a configuração utilizada:

- **Parâmetros do reconhecimento de voz**
  - Número de estados do HMM: 3 estados com emissão de saída
  - Unidades acústicas: trifones

  - Quantidade de componentes GMM utilizados na modelagem da voz: inicialmente 1 gaussiana, acrescentando uma a uma até alcançar o total de 8.

- **Pré-processamento do sinal de voz**

  - Taxa de amostragem: 8KHz
  - Filtro pré-enfase: $\alpha = 0.97$
  - Tamanho do quadro de fala considerado: 25ms
  - Atualização de segmentos: 10ms
  - Janelamento dos segmentos: janela de Hamming

- **Parâmetros de extração de atributos**

  - PNCC: B= 40 bandas com filtros de ordem n= 1, expoente a0= 1/15, somente os 20 primeiros valores da DCT foram considerados e apenas coeficientes delta foram incluídos (os de aceleração não afetavam o desempenho).

A função distribuição de probabilidade de referência escolhida para o HEQ é a distribuição de Gauss com média zero e variância unitária, já que segundo [146] esta oferece uma maior vantagem devido ao fato de que em diversos sistemas de reconhecimento a saída da distribuição dos HMMs são modeladas como mistura de gaussianas.

Em avaliações de sistemas de reconhecimento de voz contínuas, o reconhecimento é executado sentença por sentença, onde erros de reconhecimento de três tipos podem aparecer: inserções de palavras na frase reconhecida; palavras substituídas; palavras excluídas que não aparecem na frase reconhecida. Desta forma, a taxa a avaliação do reconhecedor é dada pela taxa de acerto das palavras (WAR) nas frases de teste. Esse valor é o número total de palavras subtraindo os erros e depois normalizado pelo número total de palavras, matematicamente representado por

$$WAR(\%) = 100\frac{N - (S + D + I)}{N} \tag{3-10}$$

onde N é o número de palavras esperadas no teste, S é o número de palavras substituídas, D é o número de palavras deletadas e I é o número de palavras inseridas.

Alternativamente, pode-se expressar o desempenho do reconhecedor em termos de taxa de error de palavra *Word Error Rate* (WER), da seguinte forma

$$WER(\%) = 100 - WAR(\%) \tag{3-11}$$

O desempenho dos sistemas RAV é medido através de uma determinada probabilidade de erro de palavra $\widehat{p} = WER$. No entanto, o valor médio indicado por $\widehat{p}$ tem uma margem de erro. E essa margem é medida pelo intervalo de confiança dado por

$$\Delta = \left[\widehat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right] \tag{3-12}$$

onde α é o nível de confiança e z é o valor da distribuição normal para o nível de confiança selecionado (por exemplo, para $\alpha = 95\%$, $z = 1,96$).

A Fig 3.5 mostra a amplitude Δ do intervalo de confiança de 95% ($P(\widehat{p} \in \Delta)$) para as bases de dados AURORA-4 e TIMIT. Neste caso para taxas de erro em torno de 20%, 30%, 40%, e 50%, serão consideradas melhorias estatisticamente significativas àquelas menores a 0,78, 0,89, 0,96 e 0,98 para AURORA-4 e menores a 0,34, 0,39, 0,42 e 0,43 para TIMIT respectivamente. Visando facilitar a leitura e não sobrecarregar de informação as tabelas de resultados, os intervalos de confiança não serão apresentados, recomenda-se a consulta da Fig. 3.5 para mais detalhes.

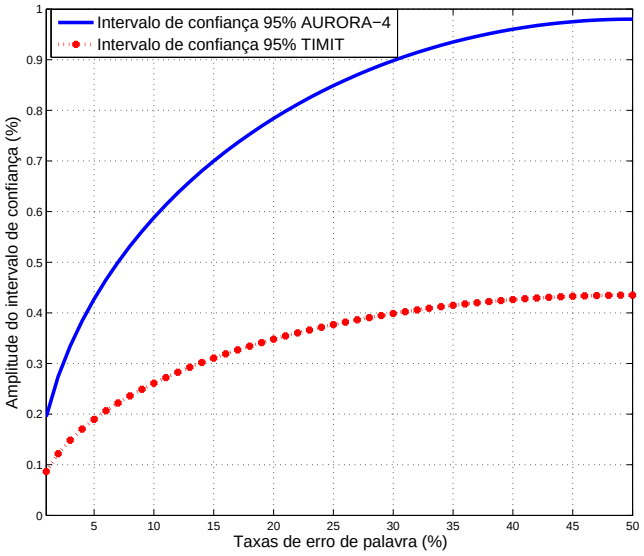

Figura 3.5: Amplitude do intervalo de confiança de 95% em função das taxas de erro de palavras WER(%) para os testes de reconhecimento sobre o conjunto de voz limpa das bases de dados AURORA-4 e TIMIT.

3.5 Resultados de simulações

A fim de avaliar os métodos de compensação de atributos descritos nesta tese, foram realizados quatro tipos de testes. No primeiro caso, um sistema de referência *baseline*, com base unicamente na características PNCC é usado em condições limpas e subsequentemente corrompido com diferentes tipos de ruído (*white, babble, f16* e *factory*) em quatro diferentes SNRs (0, 5, 10 e 15dB). Os resultados do sistema *baseline* baseados unicamente em atributos PNCC (sem técnicas de robustez) são apresentados na Tabela 3.1. Os outros testes foram realizados a fim de comparar o desempenho do método proposto com dois sistemas baseados em HEQ: i) o método tradicional HEQ usando a estatística ordenada, e ii) o método recentemente proposto FHEQ em [151]. Foram utilizadas as mesmas condições do sistema, mas aplicando a equalização de histogramas, a filtragem por média temporal e o método proposto MED-HEQ.

Os dados da Tabela 3.1 mostram as taxas de erro de palavra (WER%) dos métodos de compensação baseados em equalização de histograma, para cada tipo de ruído sobre o banco de dados TIMIT. Tomando-se a média sobre 0, 5, 10 e 15 dB de SNR

Tabela 3.1: Resultados de reconhecimento obtidos para o banco de dados TIMIT. Tomando-se a média sobre as diferentes condições de SNR.

	white	**babble**	**f16**	**factory**
Baseline	40,36	33,48	29,98	41,39
HEQ	37,12	30,78	26,37	37,92
FHEQ	34,02	30,18	25,34	39,20
MED-HEQ	**33,50**	**29,58**	**25,27**	**38,18**

Da Tabela 3.1 pode-se ver que quando são aplicados filtros sobre as *PDF*s antes da equalização o sistema oferece melhores resultados do que as técnicas de equalização tradicionais. Observa-se a melhoria introduzida pelo mapeamento dos coeficientes PNCC para um *PDF* de referência, proporcionando melhorias no desempenho do reconhecimento em relação ao sistema *baseline*. Por outro lado, os resultados mostram a importância de suavizar *PDF*s antes de aplicar a respectiva equalização. Para o método proposto (MED-HEQ), observa-se que há uma diminuição da média dos erros de palavra de $37,12\%$ para $33,50\%$ com ruído *white*, de $30,78\%$ para $29,58\%$ com ruido *babble*, e de $26,37\%$ para $25,27\%$ com ruído *f16* e para o ruído de *factory* não apresenta melhora em relação à HEQ tradicional.

A Tabela 3.2 mostra as taxas de erro de palavra (WER%) dos métodos de compensação baseadas em equalização de histograma, para cada tipo de ruído sobre o banco de dados AURORA-4, tomando-se a média sobre 0, 5, 10 e 15 dB de SNR.

Tabela 3.2: Resultados de reconhecimento obtidos para o banco de dados AURORA-4. Tomando-se a média sobre as diferentes condições de SNR

	white	**babble**	**f16**	**factory**
Baseline	49,88	40,48	37,54	51,88
HEQ	38,54	34,22	34,93	48,13
FHEQ	29,52	31,44	33,55	46,56
MED-HEQ	**26,97**	**25,34**	**31,11**	**45,88**

Na Tabela 3.2, vê-se que os resultados obtidos a partir da aplicação das técnicas baseadas em HEQ sobre o banco de dados AURORA-4 reduzem a degradação devida ao ambiente acústico sobre as características da voz, superando o método proposto em [151], diminuindo a média das taxas de erro de palavra de $29,52\%$ para $26,97,03\%$ com ruído *white*, de $31,44\%$ para $25,34\%$ com ruído *babble*, de $33,55\%$ para $31,11\%$ com ruído *f16* e de $46,56\%$ para $45,88\%$ com ruído *factory*. Também é importante destacar o desempenho excepcional atingido pelo sistemas RAV no ruído *babble*. A taxa de erro de palavra média para HEQ é de $34,22\%$ uma melhora relativa de $6,2\%$ com relação ao resultado *Baseline*. Quando o filtro temporal de média é aplicado antes da equalização, a taxa de erro é de $31,44\%$ o que representa uma melhora relativa de $9,04\%$ sobre o sistema *Baseline*. Finalmente, quando aplicado o filtro não linear de mediana sobre as *PDF*s de cada vetor de atributos, a taxa de erro do sistema fornece um rendimento de $25,34\%$ uma molharia representativa de $15,34\%$ com relação ao sistema *Baseline*, de $9,08\%$ sobre os resultados tradicionais de HEQ e de $6,30\%$ sobre os resultados obtidos com FHEQ.

3.6 Conclusões

Com o objetivo de conseguir sistemas robustos frente à degradação do sinal de voz devido ao ambiente acústico e tratar o problema de descasamento entre as etapas de treinamento e teste, neste capítulo foi apresentado um método de compensação de atributos baseado na suavização não linear das funções de distribuição de probabilidade sobre a conhecida técnica de equalização de histogramas. O objetivo é buscar remover as possíveis oscilações do vetor de atributos, causadas pela adição de ruído aditivo. Foram revisadas al-

gumas das principais propostas de normalização de atributos, mostrando que enquanto técnicas como CMN e CVN normalizam o primeiro e os dois primeiros momentos das distribuições do vetor de atributos, respectivamente; a técnica HEQ normaliza todos os momentos da função de distribuição de cada atributo, modificando as características da voz dos dados de teste, a fim de que sejam o mais parecidas possíveis às características usadas na fase de treinamento.

Oscilações de alta frequência devido aos efeitos do ruído sobre o sinal da voz geram novos desajustes no processo de equalização. Por isso, foi proposto na literatura um método de compensação de características baseado na suavização temporal da função distribuição de probabilidade. Porém, por serem filtros lineares não conseguem compensar as distorções não lineares causadas pelo ruído aditivo sobre o sinal original, já que os filtros lineares sobre as *PDF*s são muito sensíveis a alterações locais. Por isso, utilizando o novo sistema proposto (filtragem por mediana das *PDF* (MED-HEQ)) as funções de distribuição são passadas por um filtro não linear que é menos sensível a elevadas intensidades de ruído local. Com isso, o descasamento entre treinamento e teste é mais adequadamente reduzido, conservando as componentes principais da voz antes da respectiva equalização. Grandes melhorias foram registradas nas tarefas de reconhecimento robusto sobre os bancos de dados Aurora-4 e TIMIT utilizando o sistema proposto (filtragem por mediana das *PDF* (MED-HEQ)).

4
A Máscara INM (Ideal Neighbouhood Mask) Sobre o Sinal de Voz para Realce e Reconhecimento de Voz em Ambientes Adversos

No Capítulo 2 foi apresentada uma revisão das diferentes metodologias da literatura em sistemas RAV robustos. Dentre elas, as técnicas de realce de voz mostraram ser uma solução atraente para problemas dessa natureza. Estas técnicas têm como objetivo principal melhorar a qualidade e a inteligibilidade do sinal corrompido tanto no domínio do tempo quanto no domínio da frequência, melhorando o desempenho dos sistemas RAV em ambientes adversos.

Neste capítulo introduzimos uma abordagem inovadora, baseada no realce de voz, que melhora a inteligibilidade e a qualidade do sinal a partir de aplicações de uma estimativa de máscara espectral. O novo esquema emprega uma máscara de vizinhança ideal ou *Ideal Neighborhood Mask* (INM) que tem a capacidade de usar eficientemente os *Local Binary Pattern* (LBP) que indicarão quais unidades Tempo-Frequência (T-F) da voz corrompida são dominadas pelo ruído. Ao longo deste capítulo, será descrito o novo algoritmo INM e apresentada uma análise comparativa do esquema proposto com as máscaras tradicionais *Ideal Binary Mask* (IBM) e *Ideal Ratio Mask* (IRM).

4.1
Ideal Binary Mask (IBM)

Sob o ponto de vista da psicoacústica o mascaramento é definido como o efeito que é gerado no ouvido quando exposto a dois ou mais sons de intensidades diferentes simultaneamente. Um deles, de maior energia, pode mascarar os outros, afetando a sua percepção [188].

Quando o sinal de voz limpo é transformado para o domínio da frequência as energias do sinal tendem a concentrar-se em torno de frequências de interesse (por exemplo, altas frequências para as fricativas). Porém, quando o sinal de voz é corrompido com algum tipo de ruído, este gera uma distribuição das componentes do sinal que geralmente agrupam-se nas regiões onde a energia do sinal de voz é baixa, deixando muitas vezes inalteradas regiões onde a sua energia é alta. Na Fig. 4.1 pode-se ver como o ruído de alta intensidade, como o "babble 0dB" quando misturado com voz, distorce a parte baixa do

espectrograma.

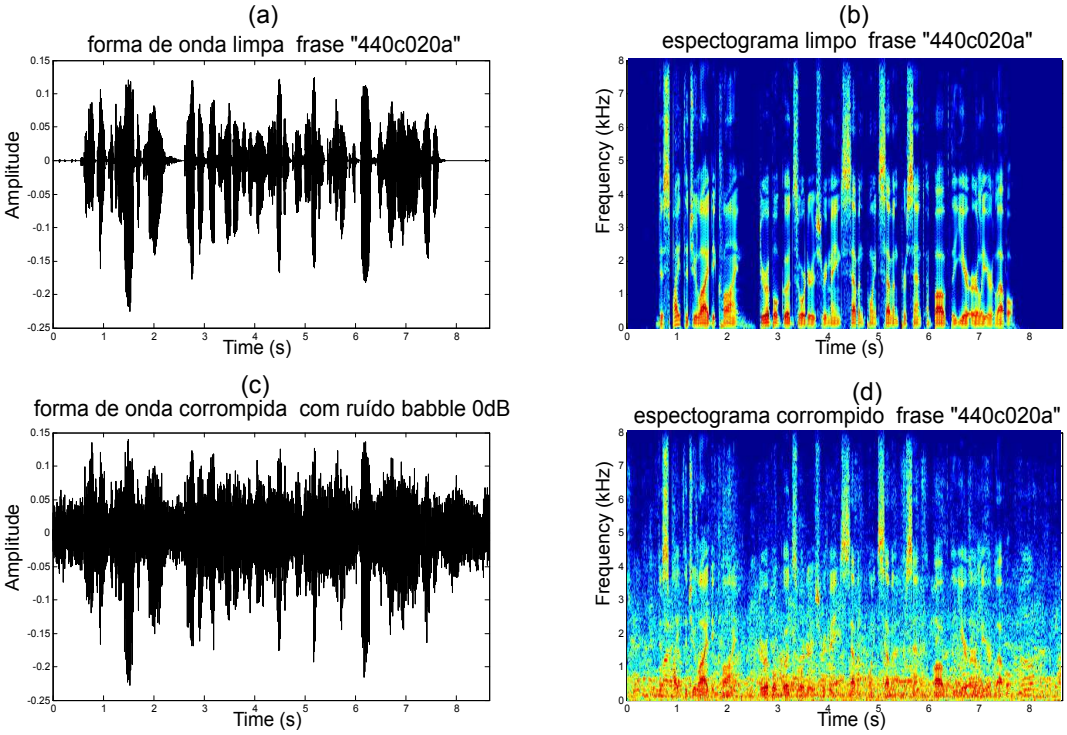

Figura 4.1: Exemplo de mascaramento de duas fontes de som. As figuras (a) e (c) representam as formas de onda da frase "440c020a"do banco de dados AURORA-4 limpa e corrompida com ruído babble de 0dB, respectivamente, e (b) e (d) são seus respectivos espectrogramas.

Este efeito de mascaramento faz com que a inteligibilidade da voz seja significativamente reduzida. No caso do sistema auditivo humano, este efeito é muito menos problemático, já que ele é bem mais robusto aos ruídos do meio ambiente.

O problema de mascaramento conhecido como *cocktail party*[189] surgiu nos anos 50 quando *Cherry* colocou duas perguntas: Como é possível reconhecer o que uma pessoa está falando quando outras estão falando ao mesmo tempo? Com que base lógica pode-se construir um filtro que possa resolver esta operação? Estas duas questões no processamento de sinais de voz podem ser traduzidas em como distinguir em um espectro da voz, regiões dominadas pela energia do ruído das energias dominadas pela voz e como solucionar o problema do reconhecimento do sinal de voz com esses espectros observados. Baseados nestas questões, os pesquisadores desenvolveram nos últimos anos métodos para lidar com esse problema do mascaramento auditivo através de abordagens de separação de voz, também conhecidas como *speech segregation*. Estas abordagens estruturam-se com base nos resultados dos estudos psicofísi-

cos do sistema auditivo humano, que propõe que os sons que atingem o ouvido são submetidos a um processo chamado análise de cena auditiva (ASA)[26].

Baseado nesse processo, Wang em [121] propôs a *Ideal Binary Mask* (IBM) como objetivo principal dos sistemas CASA [190] a fim de separar as diferentes fontes de som que compõem a entrada acústica. Tipicamente, nesses sistemas o sinal de entrada é transformado em uma série de segmentos chamados unidade tempo-frequência (T-F), onde cada unidade pertencente a um determinado tempo e a uma frequência específica é atribuído o valor 1 se a energia da voz excede a energia do ruído e 0 caso contrário. A fim de obter as unidades T-F são dois os métodos comumente usados [191]: (i) transformada em blocos, e (ii) transformada baseada em banco de filtros. Neste capítulo será usado o primeiro enfoque . Especificamente, será utilizado o modelo do ambiente acústico sem considerar o ruído convolutivo, ou seja, o sinal corrompido é expresso por

$$y(t) = x(t) + r(t) \tag{4-1}$$

onde t denota o domínio do tempo, $x(t)$ é o sinal limpo (*target*), e $r(t)$ o ruído do ambiente. Sua decomposição em série de unidades T-F se faz usando o mesmo conceito da transformada de Fourier de curto tempo (STFT). Isto é, primeiro divide-se o sinal em quadros sucessivos de 32 ms com superposição entre eles de 10 ms e, em seguida, transforma-se cada quadro para o domínio de frequência através da transformada rápida de Fourier FFT. Cria-se, então, um mapa de unidades T-F, onde para cada unidade T-F se sua relação sinal ruído (SNR) local é maior que um valor predefinido (LC), ou seja, a energia da voz é maior que a energia do ruído, pode-se considerar esse valor como voz dominante e à IBM é associado o valor 1. Caso contrário, à IBM assume o valor de 0, significando que é ruído dominante. Matematicamente a IBM e definida como

$$IBM_{(t,\omega)} = \begin{cases} 1 & se \quad SNR_{(t,\omega)} > LC \\ 0 & c.c \end{cases} \tag{4-2}$$

onde LC é o critério local ou limiar determinado empiricamente para cada técnica usando conjuntos de validação (geralmente toma faixas de valores de $LC \in [-6, 6]$ [192][193]) e $SNR_{(t,\omega)}$ é a relação sinal-ruído instantânea para cada unidade T-F dada por

$$SNR_{(t,\omega)} = 10log\frac{X_{(t,\omega)}}{R_{(t,\omega)}} \tag{4-3}$$

onde $X_{(t,\omega)}$ e $R_{(t,\omega)}$ são a energia instantânea do sinal limpo e do ruído respectivamente em um tempo t e uma frequência ω.

A Fig. 4.2 mostra o mascaramento binário para diferentes limiares LC

de um sinal de voz corrompido com ruído *babble* de 0dB. A Fig. 4.2 (a) mostra o espectrograma do sinal limpo, onde as partes claras representam as diferentes intensidades de energia da voz, (b) mostra o espectrograma do sinal corrompido com ruído *bable* de 0 dB onde pode-se ver como a energia do sinal fica mascarada pelo ruído de alta intensidade, (c) e (d) são os mascaramentos binários do sinal corrompido com diferentes limiares LC, onde a cor cinza indica 1 e a cor preta indica 0. Pode-se ver que a figura (c) fornece uma melhor recuperação da energia da voz a partir do sinal original corrompido.

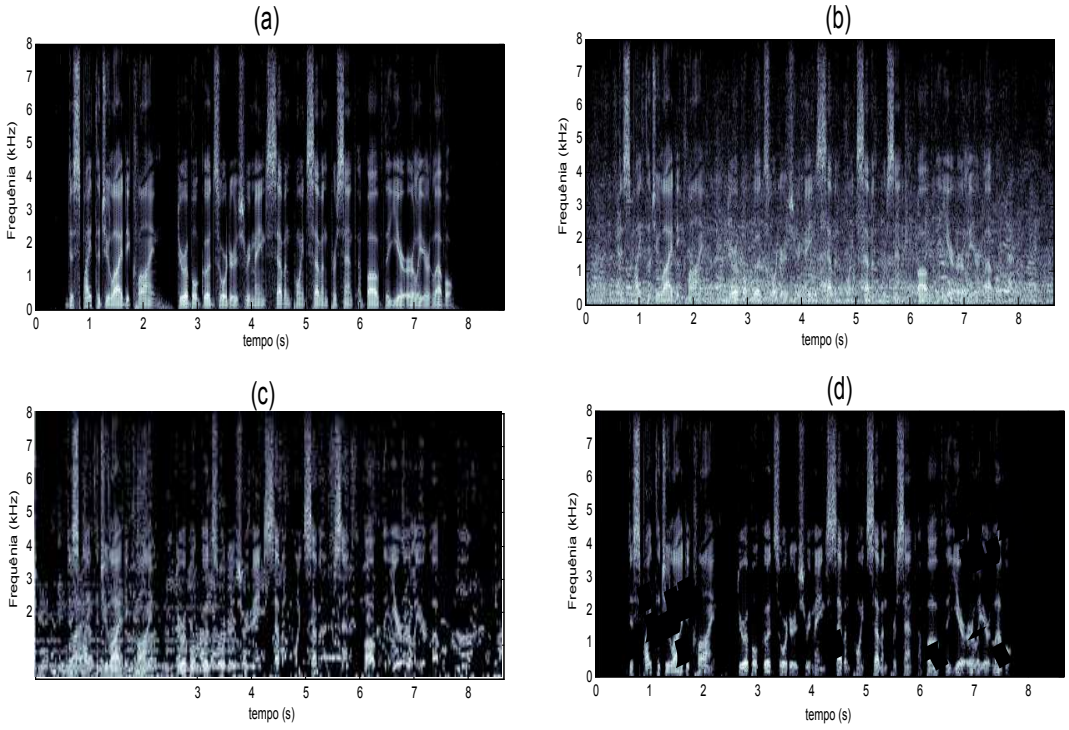

Figura 4.2: Exemplo de IBM, as figuras (a) e (b) representam os espectrogramas da frase "440c020a" do banco de dados AURORA-4 limpa e corrompida com ruido babble de 0dB respectivamente, (c) IBM com $LC = -6$, (d) IBM com $LC = 6$.

Como pode-se observar na Fig.4.2, processar sinais ruidosos usando IBM melhora a robustez dos sistemas [194]. Cabe salientar que a definição da equação 4-3 para o cálculo da SNR instantânea supõe um conhecimento ideal. Na prática, a IBM tem que ser estimada diretamente a partir do sinal degradado, isto é, usar algoritmos que façam uma estimativa *a priori* da SNR a qual possa ser usada para obter o valor de cada unidade T-F e fazer o realce.

A IBM tem sido amplamente utilizada na literatura [192] [195], mostrando que sob certas restrições, ela é uma ótima máscara binária em termos de relação sinal ruído (SNR)[196].

4.2 Ideal Ratio Mask (IRM)

Como apresentado na seção anterior, alguns anos atrás introduziu-se uma abordagem de supressão de ruído baseada nos resultados de estudos psicofísicos dos sistemas auditivos humanos. Com base nesse processo, Wang [121] propôs a IBM que é definida no domínio tempo frequência (T-F) com o objetivo de identificar unidades de voz dominante (etiquetada com 1) e ruído dominante (etiquetada com 0) de um sinal corrompido. No entanto, em uma grande parte dos métodos baseados em supressão de ruído, os dados perdidos ou distorção artificial, produzem o chamado *ruído musical* que degrada a qualidade do sinal original. Segundo [100] a aplicação de uma máscara binária aos espectros de voz corrompida pode afetar a qualidade da voz nesse processo de remoção de componentes espectrais, ou seja, quando às unidades T-F é atribuído 0, esse procedimento pode potencialmente produzir o *ruído musical*. A redução a zero das unidades T-F (ou remoção de componentes espectrais) pode criar picos pequenos e isolados no espectro que ocorrem em locais de frequência aleatórias em cada quadro. Convertidos ao domínio do tempo, esses picos são semelhantes a tons com frequências que mudam aleatoriamente de quadro para quadro e produzem o *ruído musical*.

Para resolver este problema, a *Ideal Ratio Mask* (IRM) foi proposta em [197] com o objetivo de suavizar as unidades T-F ao invés de removê-las. A IRM proporciona um melhor desempenho porque está intimamente relacionada com o filtro de Wiener[123], onde um valor de SNR alto indica baixa atenuação da energia das unidades T-F, enquanto um valor de SNR baixo indica alta atenuação, suavizando todas as unidades T-F em vez de removê-las como o caso da IBM. A IRM é definida por

$$IRM_{(t,\omega)} = \frac{10^{(SNR_{(t,\omega)}/10)}}{10^{(SNR_{(t,\omega)}/10)} + 1} \tag{4-4}$$

onde o $SNR_{(t,\omega)}$ é a relação sinal ruido instantânea para cada unidade T-F (equação 4-3).

A Fig. 4.3 (a) e (b) representam os espectrogramas do sinal de voz limpo e corrompido com ruído *babble* de 0 dB, enquanto as Fig. 4.3 (c) e (d), representam os espectrogramas das mascaras IBM com $LC = -6$ e IRM. Observa-se como o mascaramento baseado em IRM melhora o espectro do sinal.

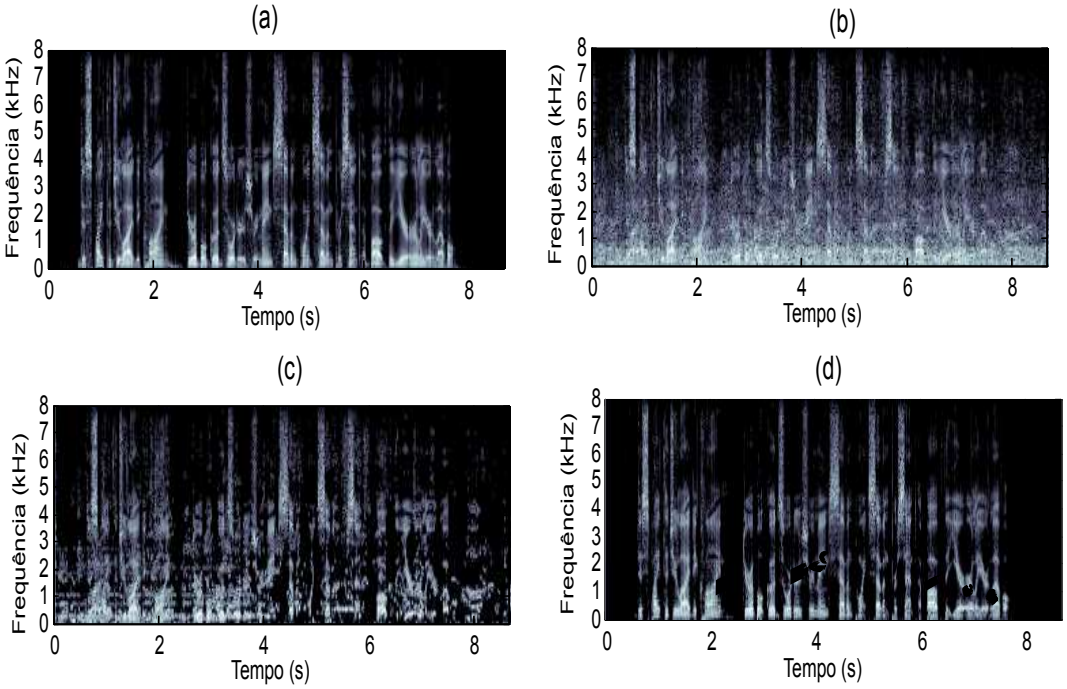

Figura 4.3: Exemplo de IRM, as figuras (a) e (b) representam os espectrogramas da frase "440c020a" do banco de dados AURORA-4 limpa e corrompida com ruido babble de 0dB respectivamente, (c) IBM com $LC = 6$ (d) IRM.

4.3 Ideal Neighborhood Mask (INM)

Nesta seção é proposta uma nova técnica de mascaramento para realçar o sinal de voz e melhorar sua qualidade e inteligibilidade, além de ser aplicada com sucesso em um reconhecedor de voz baseado em DNN-HMM em presença de ruído ambiente. Esta nova abordagem, chamada *Ideal Neighborhood Mask* (INM), usa a técnica *Local Binary pattern* (LBP)[198][199], que é frequentemente usada no processamento de imagens em 2-D para a descrição da textura. Os LBPs tornaram-se como uns dos melhores descritores de textura, em termos de desempenho e habilidades altamente discriminativas e computacionalmente simples e eficientes [200][201]. O objetivo deste esquema é resumir a estrutura local de uma imagem comparando cada pixel com seus p vizinhos.

Em processamento de imagens o operador LBP original normalmente funciona em um bloco de 3 x 3 pixels de uma imagem (ver Fig. 4.4 para ilustração). Cada pixel da imagem é considerado centro e limiar em relação aos seus vizinhos. Se a intensidade do pixel central for maior ou igual à intensidade do seu pixel vizinho a este será atribuído o valor 0. Caso contrario, será atribuído o valor 1.

Este procedimento resulta em um número binário para cada pixel que é transformado em um número decimal. Conforme mencionado em [202], com 8

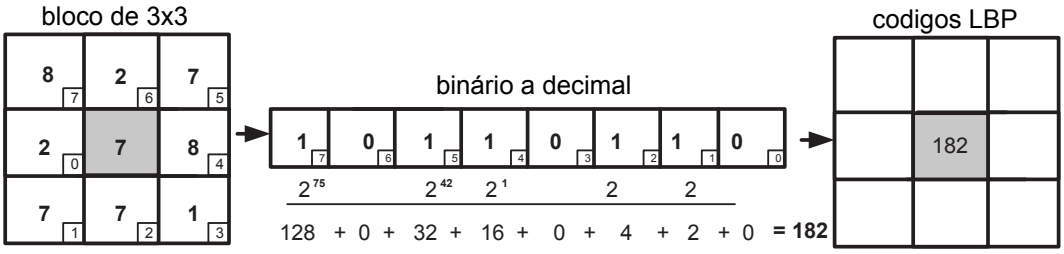

Figura 4.4: Calculo do código binário de oito pixeis vizinhos

pixeis vizinhos obtém-se um total de $2^8 = 256$ diferentes combinações possíveis, que representam os valores relativos da escala de cinza do pixel central e sua vizinhança. O código LBP para o pixel central é dado na forma decimal como

$$LBP_p = \sum_{i=0}^{p-1} sgn(f_i - f_c)2^p \tag{4-5}$$

onde f_c representa o valor de cinza do pixel central, f_i, $i = 0, ..., p-1$ os valores de cinza dos $p = 8$ pixeis vizinhos, e a função sgn é $+1$ se o argumento for positivo e -1 caso contrário.

Em [203], *Chatlani et al.* adaptaram o operador 2-D LBP usado em processamento de imagens para um operador 1-D LBP para processamento de voz, e apresentaram uma abordagem LBP 1-D, teoricamente muito simples, ainda eficiente, para detecção de atividade de voz (VAD). O conceito do método LBP unidimensional consiste em um código binário que descreve as mudanças locais abruptas do sinal 1-D, com o objetivo de estimar períodos de voz e não-voz. O código LBP 1-D é obtido a partir de uma janela deslizante com um número ímpar de amostras através do sinal, onde cada amostra vizinha é limiar em relação às amostras centrais da janela de processamento. Um exemplo do operador 1-D LBP e seus códigos binários são dados na Fig. 4.5, onde p é o número de vizinhos, definido como 8 (1x8 padrão de máscara).

Continuando com essa metodologia, o objetivo do método INM proposto nesta tese é adaptar o operador LBP 1-D a fim de trabalhar sobre cada segmento da matriz de unidades T-F obtidas na seção 4.1. Ou seja, dado o sinal de entrada corrompido $y(t)$ a nova máscara é implementada para segmentos que são extraídos de $y(t)$ aplicando uma janela de Hamming com um comprimento de 32 ms e superposição de 10 ms entre quadros, e computando a transformada de Fourier discreta (DFT) de cada quadro com 512 coeficientes DFT. Isso resulta em uma decomposição do sinal em uma matriz bidimensional $\Gamma_{(M,N)}$, onde M representa o número de quadros e N é o número de coeficientes

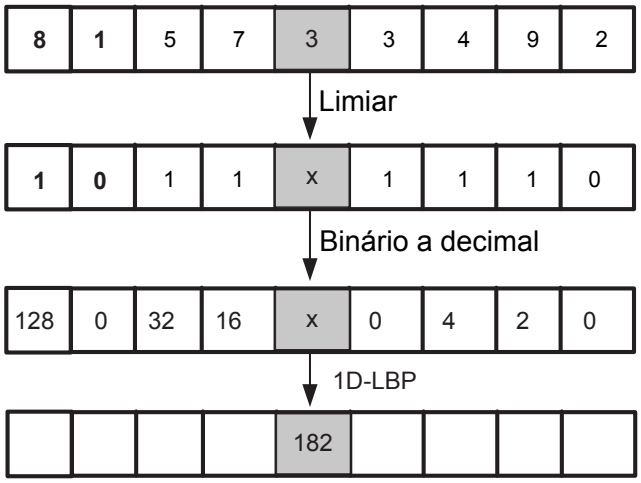

Figura 4.5: Cálculo do operador 1-D LBP de oito amostras vizinhas

da DFT. Isso produz uma matriz de unidades T-F, onde cada valor T-F da matriz $\Gamma_{(M,N)}$ representa a relação sinal ruído (SNR) instantânea do sinal corrompido. Após obter a matriz $\Gamma_{(M,N)}$, estima-se a máscara INM, adaptando o operador LBP para trabalhar em cada quadro da matriz $\Gamma_{(M,N)}$ (uma linha dessa matriz). Assim, para cada unidade T-F, definimos o seu código LBP correspondente que é obtido através de uma janela com um número par de unidades T-F vizinhas da matriz $\Gamma_{(M,N)}$, comparando cada unidade T-F desta janela em relação ao vizinho e considerando o resultado como um número binário[1]. Desta forma, a SNR de cada unidade T-F da matriz $\Gamma_{(M,N)}$ é codificada usando a informação desses valores vizinhos. O número binário resultante é transformado em um número decimal, que representara um código LBP exclusivo com um nível mais alto de informação, criando uma nova matriz $LBP_{(M,N)}$ com códigos LBP

$$LBP[k] = \sum_{i=1}^{p/2} \Big\{ sgn[\gamma[k-i] - \gamma[k]] 2^{p/2-i} + sgn[\gamma[k+i] - \gamma[k]] 2^{p/2+i-1} \Big\} \tag{4-6}$$

onde p é o número de unidade T-F vizinhas que envolvem cada unidade T-F na análise, $\gamma[k]$ representa a SNR instantânea estimada para cada unidade T-F em dB diretamente das unidades T-F corrompidas em um dado quadro. O cálculo de SNR é o mesmo de (equação 4-3), e k é o índice de cada unidade T-F, $k = 0$ a $N - 1$. Este intervalo de valores k está associado a cada linha da matriz $\Gamma_{(M,N)}$, a função $sgn[.]$ é definida como 1 se seu argumento for maior do que um limiar α_U e 0 caso contrário. Isto significa que

[1]Para o caso de duas unidades T-F vizinhas, ou seja, uma unidade T-F anterior e uma unidade T-F posterior à unidade de análise, o número binário será de 2 bits.

$$sgn[\gamma[k \pm i] - \gamma[k]] = \begin{cases} 1, & se \quad \gamma[k \pm i] - \gamma[k] > \alpha_U \\ 0, & caso \quad \text{contrário} \end{cases} \tag{4-7}$$

Note-se que o algoritmo acima foi desenvolvido levando em consideração o método de eliminação de ruído denominado *Visushrink* introduzido por *Donoho* em [105] onde α_U representa o limiar $softthreshold$, proposto como estimativa de limiar universal, dado por

$$\alpha_U = \delta\sqrt{2ln(N)} \tag{4-8}$$

onde N representa o tamanho dos segmento da matriz $\Gamma_{(M,N)}$ e δ é uma estimativa aproximada do nível de ruído. Esta estimativa é dada por

$$\delta = \frac{med(|\gamma[k]|)}{0.6745}, \quad k = 0, ..., N-1 \tag{4-9}$$

onde *med* é a mediana de cada segmento da matriz $\Gamma_{(M,N)}$. O uso do limiar α_U é importante porque evita a influência de ruídos muito pequenos.

Para o caso $p = 2$, obtém-se números binários de 2 bits, ou seja, códigos LBP decimais variando de 0 a 3. Para cada código LBP, estabelece-se o valor da máscara INM correspondente para 1 se o código LBP for 3. Isso representa a situação em que a energia da fala é significativamente maior do que a do ruído. Quando os códigos LBP são 1 ou 2, esses valores são suavizados usando a raiz quadrada do filtro de Wiener [2]. Finalmente, uma suavização temporal por meio de um filtro de média ponderado é realizado quando o código LBP é 0 para reduzir as flutuações entre a energia local da voz ruidosa e aquela quando a energia da voz é maior do que o ruído. Este procedimento realça a presença de voz em unidades T-F vizinhas, suavizando todas as unidades T-F com energia de ruído dominante, em vez de removê-las, como na IBM. Quantitativamente, para o caso em que $p = 2$ a máscara INM é definida como

$$INM[k] = \begin{cases} \dfrac{\gamma[k-1] + 2\gamma[k] + \gamma[k+1]}{4} & para LBP[k] = 0 \\ \sqrt{\dfrac{\gamma[k]}{1+\gamma[k]}} & para LBP[k] = 1 \; or \; 2 \\ 1 & para LBP[k] = 3 \end{cases} \tag{4-10}$$

Em contraste com as máscaras ideais com base no verdadeiro espectro de fala, referidas como máscaras Oracle [71], consideramos a situação em que a máscara proposta faz uso de um estimador SNR diretamente do sinal

[2]Que é escolhida por ser fácil de implementar, requer menos carga computacional e é muito eficaz para melhorar a qualidade e a inteligibilidade da voz [204].

corrompido. Observe que isso corresponde ao cenário de aplicativos reais, onde a única informação disponível é o discurso ruidoso. Vários algoritmos de estimativa são descritos em [71]. Neste trabalho, usamos o bem conhecido algoritmo de média recursiva controlada modificada (IMCRA), proposto por Cohen [205] para estimar o espectro de ruído de fundo e calcular a SNR. Informação detalhada do algoritmo IMCRA pode-se encontrar em [205][206].

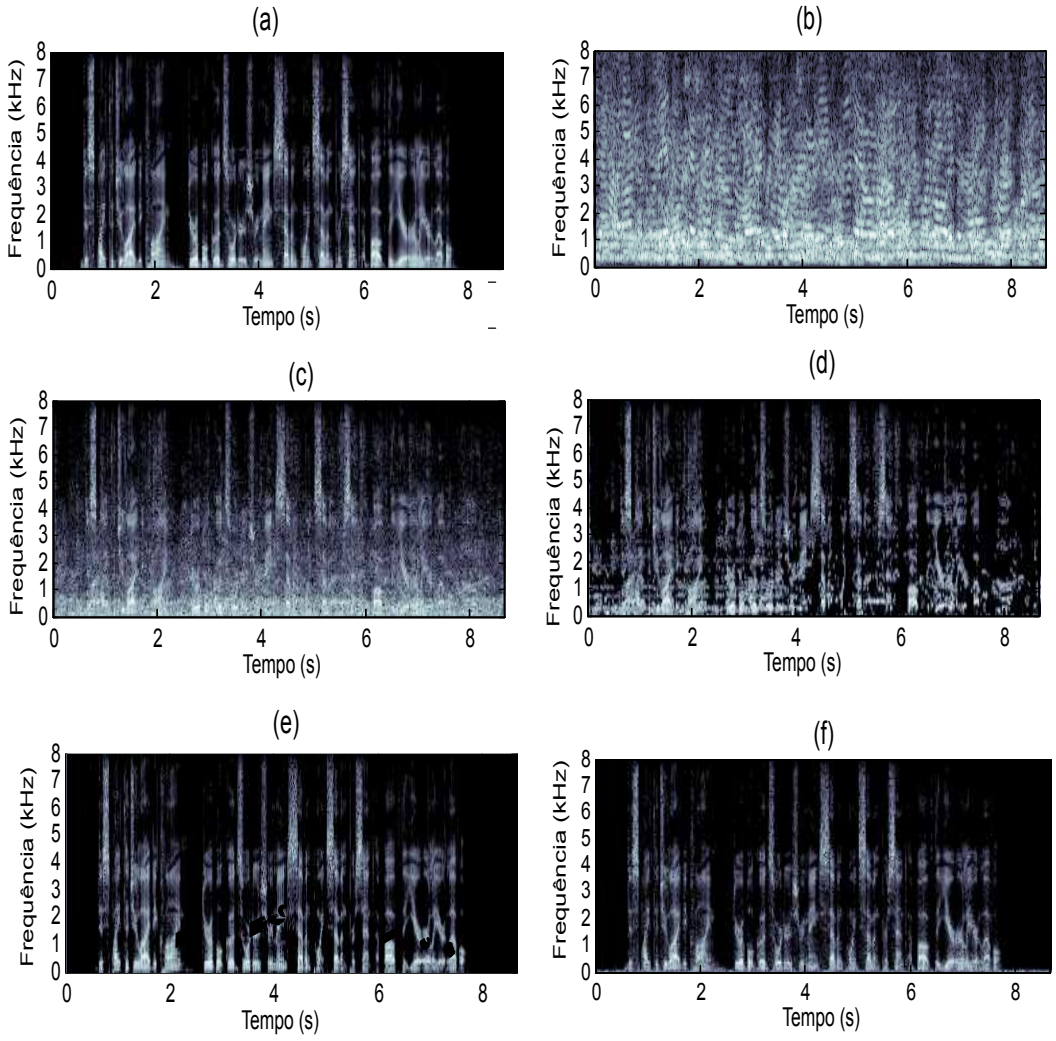

Figura 4.6: Exemplo de INM, a figuras (a) representa o espectrograma da frase "440c020a" do banco de dados AURORA-4 limpa (b) ruído bable (c) frase "440c020a" corrompida com ruído *babble* com SNR de 0dB, (d) IBM com $LC = 6$ (e) IRM e (f) INM

.

A Fig. 4.6 apresenta os espectrogramas das diferentes configurações de máscaras vistas até este ponto: (a) sinal de voz limpa "440c020a" do banco de dados AURORA-4 (b) ruído *babble*, (c) sinal de voz corrompido com ruído *babble* com SNR de 0dB, (d) espectrograma do sinal restaurado com a máscara IBM com $LC = 6$, (e) espectrograma do sinal restaurado com a máscara IRM,

e (f) espectrograma do sinal restaurado com a máscara INM. Pode-se ver como o ruído gera uma grande incompatibilidade entre a voz limpa e a corrompida, levando a graves problemas de inteligibilidade da voz. As figuras (d), (e) e (f) mostram a importância das técnicas de mascaramento IBM, IRM, e INM e como elas reduzem a degradação da voz causada pelo ruído melhorando a qualidade do sinal e realçando sua inteligibilidade. A partir de uma comparação visual, pode-se observar que a INM preserva informações mais detalhadas do que a IBM e o IRM.

A título de exemplo, no Algoritmo 1 é apresentado o procedimento para o caso de $p = 2$ vizinhos, detalhando os passos para obter a máscara INM.

Algorithm 1 Cálculo da máscara INM para p=2

Input: Sinal de entrada y(t).
Output: $INM[k]$.

1: Segmentar a voz em quadros de 32-ms (256 amostras em uma frequência de amostragem de 8kHz) com 10-ms de superposição entre quadros.
2: Aplicar janela Hamming a cada quadro, realizando a sua respectiva transformada de Fourier.
3: Calcular $SNR[k]$ *a priori*, ou seja, $\gamma[k]$ de cada segmento.
4: **while** $\gamma[k]$ True (para $k = [0 : N - 1]$) **do**
5: Calcular códigos LBP na janela de análise deslisante W de comprimento p.
6: **if** $\gamma[i \pm 1] \geq \gamma[i]$ **then**
7: p = 0.
8: **return** $W_p = 1$; incrementar p
9: **else**
10: **return** $W_p = 0$; incrementar p
11: **end if**
12: **if** $p = 2$ **then**
13: **return** $\gamma[k] = 2^{W_p} + 2^{W_{p+1}}$
14: **end if**
15: **end while**
16: Separar todos os segmentos com diferentes valores LBP
17: Calcular INM[k] de acordo com (4-10)

4.4 PESQ, ganho de SNR, e taxa de erro de palavra (WER) em reconhecimento de voz baseado em DNN

Nesta seção apresentam-se e discutem-se os resultados da simulação do algoritmo proposto tanto no caso ideal (oracle) INM quanto no caso real usando o algoritmo da média mínima avançada recursiva controlada (IMCRA) proposto por Cohen. Os resultados foram comparados com as máscaras tradicionais IBM e IRM nas mesmas condições. Todos os experimentos foram realizados

no subconjunto ruidoso das tarefas Aurora-4. O subconjunto escolhido consiste de 330 sentenças de fala limpos misturados com 6 ruídos ambientais *babble, airport, restaurant, street, car, train* variando de 0dB a 15dB. O sinal original foi amostrado a uma frequência de 8kHz. As avaliações de desempenho são realizadas em termos da medida objetiva da qualidade da fala p.862, conhecida como avaliação perceptiva do padrão de qualidade da voz (PESQ), o ganho da relação sinal ruído (SNR gain), e a taxa de erro de palavra (WER) em um sistema de reconhecimento de voz contínuo baseado em DNN.

4.4.1 ITU-T P.862 Perceptual Evaluation of Speech Quality (PESQ)

O PESQ [71],conforme definido na norma UIT-T P.862, é um método objetivo para testar a qualidade da voz. A objetividade baseia-se na comparação com o método tradicional MOS (Mean Opinion Score) [207] no qual um grupo de ouvintes é usado para classificar a qualidade da voz para um valor que varia de 1 (ruim) a 5 (excelente).

4.4.2 Ganho da relação sinal ruído SNR

O ganho da relação sinal ruído (SNR-gain) é definido como a média das diferenças entre a SNR em dB, antes e depois da aplicação da técnica de aprimoramento em todos os quadros do sinal de voz, ou seja, ela compara o nível de um sinal desejado com o nível de ruído de fundo.

4.4.3 Taxa de Erro de Palavra WER

A precisão de um sistema de reconhecimento de voz contínua baseado em *Deep Neural Network* (DNN) foi avaliado através de testes de reconhecimento, onde são contabilizados o número de erros de palavras cometidos. A taxa de erro de palavras (*Word Error Rate*) (WER) é uma métrica comum do desempenho de um sistema de reconhecimento de voz e é definida como

$$WER(\%) = \frac{n_i + n_b + n_s}{n_w} \tag{4-11}$$

onde n_i é o número de erros de inserção, n_b número de palavras deletadas, n_s número de palavras substituídas e n_w é o número de palavras totais na transcrição de referência associada a uma determinada frase de teste. Estes erros são contabilizados através de um alinhamento baseado num algoritmo de programação dinâmico entre a transcrição de referência e a obtida no momento do reconhecimento.

As taxas de erro de palavras são fornecidas por um sistema de reconhecimento de voz híbrido, usando uma rede neural profunda e os modelos ocultos de markov (DNN-HMM). O sistema é implementado usando o kit de ferramentas de reconhecimento de voz *Kaldi* [208]. Os parâmetros experimentais, incluindo o tipo de atributos, são os mesmos que para experimentos padrão usando a formula de Kaldi $s5$, ou seja, nenhum desses parâmetros foi otimizado.

O Modelo acústico foi treinado usando o conjunto de treinamento limpo si84 (7138 sentenças) da AURORA-4 que consiste em sinais gravados usando um microfone Sennheiser e processados usando um filtro P.314. Utiliza-se o conjunto de dados Nov92 (330 sentenças) como o conjunto de teste que consiste em 7 subconjuntos de teste: 1 subconjunto limpo e 6 subconjuntos corrompidos por seis diferentes tipos de ruído *babble, airport, restaurant, street, car, e train* variando de 0dB a 15dB. O sistema de reconhecimento empregou em todos os casos o MFCC como vetores de atributos com deltas e delta-delta, que são agrupados com um contexto temporal de ś5 quadros em torno do quadro em análise. Isso resulta em uma representação de $11 \times 39 = 429$ atributos cepstrais os quais são transformados com uma análise discriminante linear (LDA) e uma transformação linear de máxima verossimilhança (MLLT). O modelo de linguagem utiliza os trigramas fornecidos nas tarefas da WSJ. Finalmente, os alinhamentos forçados foram gerados a partir da fórmula tri4b de Kaldi.

O classificador é treinado com trifones dependentes do contexto como alvo, onde cada fone é modelado com três estados HMM. O DNN é implementado usando a configuração *nnet* padrão da Kaldi que possui sete camadas ocultas com 2048 unidades por camada. Todos os nós ocultos usam função de ativação *sigmoidal*, à exceção da camada de saída que usa a função *softmax*. Os pesos das camadas ocultas são inicializados usando pré-treinamento *Restricted Boltzmann Machine* (RBM) que é um aprendizado não supervisionado, onde é modelada a dependência entre um grupo de variáveis aleatórias usando uma arquitetura de duas camadas (uma camada visível ou entrada e uma camada oculta)[209]. A primeira camada oculta é pré-treinada por 100 épocas, as camadas subsequentes são pré-treinadas por 35 épocas cada. A taxa de aprendizagem para a primeira camada está definida em 0,004, e para as restantes camadas em 0,01. A rede está ajustada com base no critério de erro de entropia cruzada de 25 épocas sem *early stopping*. O tamanho do *minibatch* é configurado para 256 para as primeiras 5 épocas. Para as restantes 20 épocas, está configurado para 1024. A taxa de aprendizado é ajustada para 0,08 para as primeiras 5 épocas, 0,32 para as próximas 10 épocas e depois reduzida para 0,008 para as 10 épocas finais. O *momentum* é fixado em 0,9 nos estágios de pré-treinamento.

4.4.4 Resultados de simulações

A Tabela 4.1 mostra os resultados SNR-Gain do algoritmo de mascaramento proposto neste capítulo (INM) em comparação com os métodos tradicionais IBM e IRM em condições ideais (oraculo). Esta condição nos mostrara o limiar máximo até onde as técnicas podem realçar o sinal, já que para estes testes assume-se que o sinal de voz e o ruído são conhecidos. Nos outros experimentos é empregada uma estimativa de ruído através do algoritmo IMCRA [71] a fim de estimar a SNR do sinal corrompido, simulando condições reais (neste caso identificaremos os métodos como EBM, ERM e ENM).

Tabela 4.1: Média do $SNR - Gain$ em dB sobre os diferentes tipos de ruído.

SNR	**noisy**	**IBM**	**IRM**	**INM**	**EBM**	**ERM**	**ENM**
0	-4,41	5,25	7,06	8,57	-1,876	0,50	1,13
5	-1,60	7,11	9,01	9,83	0,641	2,37	3,39
10	1,50	9,29	11,60	12,34	3,504	6,21	7,50
15	4,82	11,77	13,71	14,83	6,634	8,32	9,04

O ponto de partida dos experimentos é mostrar como os resultados obtidos sem qualquer técnica de mascaramento (coluna noisy) apresentam o pior desempenho, mostrando como o ruído afeta severamente o sinal de voz. Eles serão tomados como referência para os testes posteriores, a fim de verificar como as técnicas de mascaramento melhoram a qualidade do sinal de voz. Pode-se ver claramente a tendência do SNR-Gain ser consideravelmente inferior quanto maior for a adição de ruído. Experimentos também foram realizados visando obter um sistema mais robusto ao efeito da adição do ruído. Para isso foram acrescentados ao sistema os métodos *IBM*, *IRM* e *INM*, testados nas mesmas condições anteriores. Como esperado, a melhora da SNR-Gain é considerável quando são usadas as técnicas de mascaramento em todos os cenários, apresentando resultados muito promissores especialmente no mascaramento baseado nos LBP. Pode-se ver que em condições reais a ENM fornece a maior SNR-Gain em todos os casos.

Num segundo experimento, consideramos a avaliação objetiva perceptiva da qualidade da voz (PESQ) como medida objetiva recomendada pela UIT-T (Recomendação P. 862) para a avaliação da qualidade da voz de telefonia celular e codecs de voz de banda estreita.

A Tabela 4.2 apresenta os resultados das medidas PESQ obtidas com os três algoritmos de mascaramento considerados tanto em condições ideais quanto em reais. Do mesmo modo que foi feito anteriormente, os resultados da tabela correspondem a uma média sobre os 6 tipos de ruído. Da mesma

Tabela 4.2: Média do *PESQ* sobre os diferentes tipos de ruído.

SNR	**noisy**	**IBM**	**IRM**	**INM**	**EBM**	**ERM**	**ENM**
0	1,068	2,088	2,135	2,815	1,110	1,149	1,243
5	1,199	2,595	2,660	3,250	1,226	1,300	1,466
10	1,459	3,124	3,192	3,624	1,475	1,591	1,817
15	1,870	3,562	3,619	3,906	1,893	2,039	2,278

forma que o SNR-Gain, o experimento foi realizado com 4 níveis da entrada SNR, ou seja, 0 dB, 5 dB , 10 dB e 15dB. A partir desses resultados, pode-se ver facilmente que o algoritmo de mascaramento INM proposto é superior aos algoritmos tradicionais.

Num terceiro experimento, o desempenho de um sistema de reconhecimento de fala contínuo baseado em DNN foi avaliado pela taxa de erro de palavra (WER) (com as condições experimentais que foram relatadas). Em condições limpas, o sistema produz uma WER de $4,71\%$. A Fig. 4.7 mostra os resultados obtidos para condições ideais de mascaramento em termos de média de WER sobre as SNR de 0, 5, 10 e 15 dB. A partir dessa figura pode-se ver como o sistema de mascaramento proposto INM fornece melhores resultados que os mascaramentos IBM e IRM em todos os cenários. Da Tabela 4.1 sabe-se que quanto menor a SNR pior a inteligibilidade da voz. A mesma ideia já era esperada nos sistemas RAV, sendo o desempenho reduzido quando se trabalha em ambientes com maior intensidade de ruído .

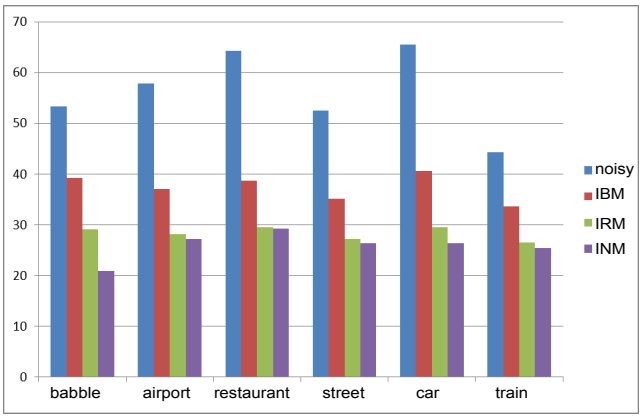

Figura 4.7: Resultados de reconhecimento das mascaras ideais (oraculo) obtidas para o banco de dados AURORA-4 tomando-se a média sobre as diferentes condições de SNR

O desempenho das máscaras estimadas é mostrado na Tabela 4.3, onde cada método tem seu desempenho calculado como a média sobre as SNR de 0, 5, 10 e 15 dB. Pode-se observar que a máscara proposta ENM oferece melhores resultados que EBM e ERM em todos os cenários. Tomando a média de todos os tipos de ruído e valores SNR, observamos que a máscara proposta diminui em

média os erros de reconhecimento de $52,04\%$, $49,92\%$, $43,60\%$ para $36,52\%$, em comparação com noisy, EBM e ERM, respectivamente. É importante observar que esses resultados experimentais mostram o desempenho superior da técnica proposta em condições reais, ou seja, sem conhecer todos os sinais a priori.

Tabela 4.3: Resultados de reconhecimento obtidos para o banco de dados AURORA-4, tomando-se a média sobre as diferentes condições de SNR. MR significa melhoria relativa em relação ao sistema ruidoso

system	**babble**	**airport**	**restaurant**	**street**	**car**	**train**	**avg.**	**MR**
Noisy	59,24	56,80	58,88	46,18	56,27	34,90	52,04	0%
EBM	56,37	54,28	57,34	44,67	54,02	32,86	49,92	4,07%
ERM	51,16	48,52	53,22	39,47	45,16	24,11	43,60	16,21%
ENM	**43,86**	**41,79**	**46,68**	**32,45**	**37,48**	**16,91**	**36,52**	29,82%

Em vista dos resultados da Tabela 4.3, deduz-se que depois de utilizar o mascaramento ENM o rendimento do sistema RAV melhora significativamente as taxas de erro de palavra para cada um dos cenários dados nos sistemas. Também é importante observar a relevância das melhorias das três máscaras em comparação com a situação em que nenhum aprimoramento é usado (primeira linha da Tabela). Também pode-se ver a melhora relativa (MR) do mascaramento ENM com uma diferença de $13,61\%$ e $25,75\%$ para ERM e EBM, respectivamente. Em resumo, o ENM atinge menores erros de palavra em comparação com a EBM e a ERM.

4.5 Conclusões

Neste capítulo, propusemos uma nova estrutura de mascaramento para melhorar o realce da voz. O algoritmo estima uma máscara de vizinhança ideal (INM), que baseia-se na técnica *Local Binary Patterns* (LBP), originalmente empregada em processamento de imagens. Com esta máscara, exploramos as características espectro-temporais da voz para realizar o aprimoramento do sinal. Comparamos nossa máscara com as técnicas de mascaramento espectral tradicionais da literatura em um ambiente real onde a máscara não depende da condição verdadeira ou ideal de conhecer todos os sinais *a priori* (condição oráculo). Mostramos que os resultados da máscara proposta em condições reais são significativamente melhores nos desempenhos objetivos de qualidade PESQ, mostrando ser uma boa técnica para o realce da voz. Os experimentos realizados com o reconhecedor de voz baseado em DNN revela que, em termos de taxa de erro de palavra, o ENM também é mais eficaz na redução de ruído. O ENM oferece vantagens significativas de desempenho de reconhecimento de voz em relação às outras técnicas de mascaramento.

Propusemos uma nova estrutura que estima um mascaramento e melhora os sistemas de reconhecimento. Ao contrário dos métodos do estado da arte que trabalham sobre os coeficientes ceptrais, nosso sistema executa o mascaramento LBP sobre o sinal de voz, antes da extração de atributos. Observou-se que a máscara estimada pelo método proposto é mais robusta que a conhecida ERM, atingindo em média uma taxa de erro abaixo de $7,08\%$ nos subconjuntos ruidosos do corpus Aurora-4.

Finalmente, a principal motivação do método de mascaramento baseado em LBPs foi o fato de que, com os códigos LBP, a energia da unidade central T-F original é codificada com a energia das unidades T-F vizinhas, de modo que a informação codificada será realizada em um nível mais alto. Além disso, a máscara INM (que no caso de condições reais é denominada ENM) não depende da parametrização utilizada para a representação da fala em reconhecimento de voz, o que torna possível usá-la em diferentes aplicações de realce.

5
Realce e Reconhecimento de Voz Robusto Usando Mascaramento INM Sobre a Técnica Wavelet Denoising

Neste capítulo, é apresentado um algoritmo de mascaramento dos coeficientes *wavelet* com base na abordagem *Local Binary Pattern* (LBP) para realçar os espectros temporais dos coeficientes *wavelet* das altas frequências e aprimorar o sinal de voz. A nova técnica consiste basicamente da aplicação do mascaramento INM, descrito no Capítulo 4 sobre o esquema de *wavelet-denoising*. A técnica *wavelet-denoising* explora o esquema de realce de ruído através da divisão da voz degradada em forma piramidal (sub-bandas), extraindo informações de frequência sem perder informações temporais e reduzindo a influência do ruído presente em sub-bandas de alta frequência. Na proposta descrita neste capítulo, o realce da voz em cada sub-banda de alta frequência é realizado através do mascaramento INM baseado nos LBPs, que codificam a relação entre o valor original de cada coeficiente *wavelet* e os valores dos coeficientes vizinhos. Esta abordagem reduz de forma eficiente o ruído contido nos espectros de alta frequência da transformada *wavelet* em vez de eliminá-los através de um limiar. A eficiência do LBP é devida essencialmente ao fato do coeficiente *wavelet* original ser codificado com os valores dos seus vizinhos. Portanto, a informação codificada leva em consideração um nível mais alto de informação. Os resultados das simulações e a eficácia do método proposto são comparados com outras técnicas apresentadas na literatura.

5.1
A Técnica Wavelet Denoising

5.1.1
Introdução

No Capítulo 2 foi brevemente descrita a técnica *wavelet-denoising* como uma ferramenta para remover o ruído do sinal corrompido. Nesta seção será apresentada uma visão mais detalhada deste tipo de transformada (*wavelet*) e do processo de realce (*denoising*) que é feito sobre ela.

Antes de descrever as características da análise dos sinais por meio de transformações *wavelet*, é preciso ressaltar a importância de sua utilização. A

análise espectral baseada em transformada de *Fourier* é a ferramenta analítica dominante para análise de domínio de frequência. No entanto, a transformada de *Fourier* não fornece informações sobre as mudanças do espectro em relação ao tempo. Por outro lado, a transformada de *Fourier* assume que o sinal é estacionário, mas na realidade inúmeros sinais de interesse do mundo real apresentam características não estacionárias, como por exemplo o sinal de voz. Para superar esta deficiência, em [210] foi apresentado um método modificado da transformada de *Fourier*, chamado análise de *Fourier* por intervalos ou transformada de *Fourier* de tempo curto (STFT) que permite representar o sinal em ambos os domínios de tempo e frequência através de uma função de uma janela no tempo. Porém, através desta abordagem simplesmente pode-se obter informação no tempo e frequência com uma precisão limitada, já que uma vez que se um determinado tamanho da janela for escolhido, essa janela será a mesma para todas as frequências. Buscando dar solução a esses problemas foi proposta uma alternativa matemática de análise, denominada transformada *wavelet* [211], como a evolução da transformada de *Fourier*. A análise *wavelet* permite representar um sinal em diferentes resoluções permitindo controlar que porções do sinal serão mais o menos afetadas pelo processo. Isso é, permite usar grandes intervalos de tempo nos segmentos onde é necessária maior precisão (baixa frequência) e menores intervalos de tempo em alta frequência. Ou seja, o domínio *wavelet* mistura as informações do domínio do tempo com as do domínio da frequência permitindo uma maior flexibilidade já que traz muito mais detalhes para analisar o sinal.

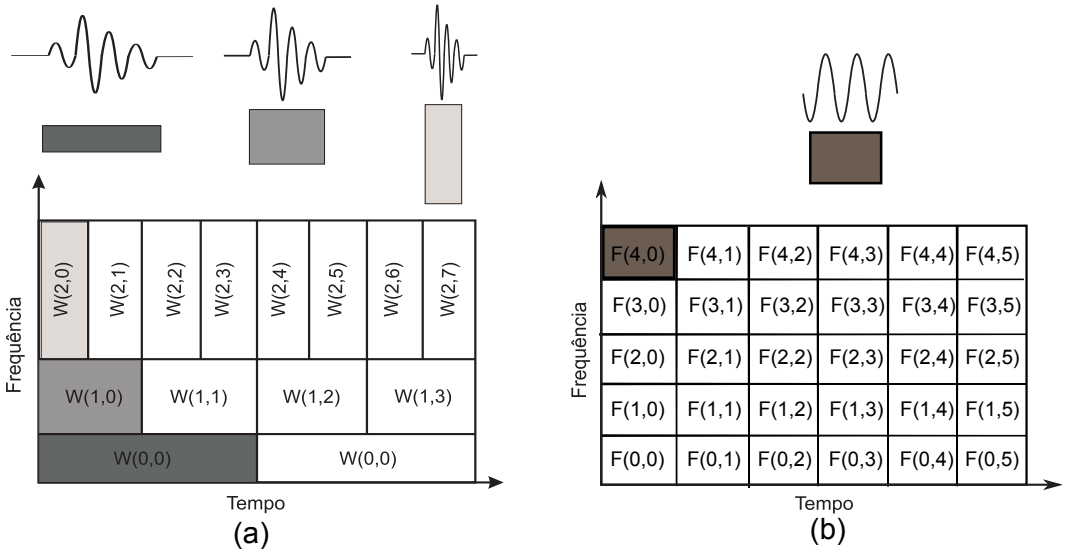

Figura 5.1: Esquemas de transformadas (a) Wavelet (b) Fourier de tempo curto (STFT).

A Fig. 5.1 apresenta uma comparação entre o esquema da transformada de *Fourier* por intervalos (STFT) e a transformada *wavelet*. Pode-se ver da Fig. 5.1 como a transformação *wavelet* permite analisar sinais em diferentes intervalos de tempo e frequência. O sinal é decomposto em versões deslocadas (no tempo) e escaladas da *wavelet* original ψ conhecida como *wavelet* mãe.

Uma das principais características da transformada *wavelet* é a sua irregularidade, diferentemente das funções senoidais que são bases da transformada de Fourier, como pode-se ver na Fig.5.2. Estas características dependem da escolha da função mãe, a qual sera definida de acordo com o tipo de sinal analisado.

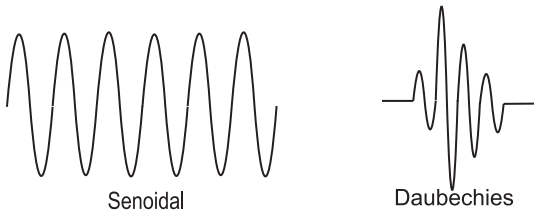

Figura 5.2: Comparação do sinal senoidal com o sinal *wavelet Daubrchies*.

Na literatura existem várias famílias de funções *wavelet* que têm mostrado ser úteis. Um estudo realizado por *Long et al.* em [212] apresenta uma seleção das melhores transformadas, entre elas pode-se citar *Haar, Daubechies, Biortogonal, Coiflets,* e a *Symlets*.

Neste capítulo a função mãe *Daubechies* e sua transformada de forma discreta será usada como base da análise.

5.1.2 Transformada Wavelet Discreta

A transformada *wavelet* discreta (TDW) é uma ótima ferramenta usada para realce de voz e compressão de imagem. Ele fornece informações satisfatórias tanto para análise quanto para reconstrução [213]. A TDW decompõe o sinal em diferentes sub-bandas com diferentes resoluções. A forma geral é dada por

$$f(t) = \frac{1}{\sqrt{T}} \sum_{j=0}^{J-1} \sum_{K} cD_j(k)\psi_j(t) + \sum_{K} cA_J(k)\phi_J(t) \qquad (5\text{-}1)$$

onde $\psi_j(t)$ e $\phi_J(t)$ representam a *wavelet* mãe e a função de escala para o nível J, e cA e cD representam os coeficientes de aproximação e detalhe do j-ésimo nível, respectivamente.

A TDW foi introduzida por *Mallat* [214], com o objetivo de analisar a decomposição em multi-resolução de imagens, desenvolvendo um algoritmo baseado em bancos de filtros passa-alta e passa-baixa que permitiu-lhe obter uma decomposição do sinal em diferentes faixas de frequência com diferentes resoluções. Deste modo, obteve coeficientes com uma aproximação do sinal original a partir dos filtros passa-baixa, produzindo os chamados coeficientes de aproximação (cA) enquanto que dos filtros passa-alta, obteve os coeficientes *wavelet* para cada detalhe do sinal (cD) (por exemplo, o ruído) originando assim os coeficientes de detalhes.

A Fig. 5.3 mostra a decomposição dos sinais através de bancos de filtros. Uma característica importante a se levar em conta é que a saída de cada filtro tem que passar por um processo de *down-sampling* (a cada duas saídas do filtro, descarta-se uma delas sem perder informação do sinal). Dessa forma a resolução é reduzida à metade do tempo. Além disso, cada saída tem metade da faixa de frequência da entrada, de modo que a resolução da frequência foi reduzida à metade.

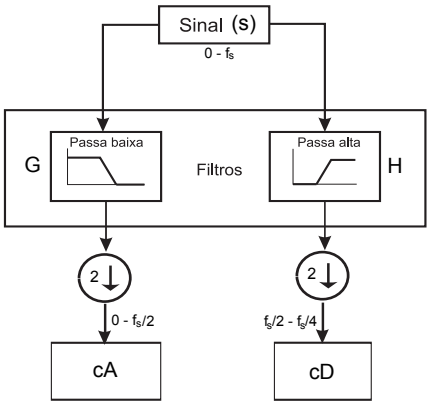

Figura 5.3: Diagrama de decomposição de sinais através de banco de filtros.

Este procedimento repete-se iterativamente até que o nível contenha apenas uma amostra do sinal ou até o j-ésimo nível de decomposição desejado pela análise. Ou seja, se produz uma análise piramidal ou de multi-resolução através de filtros multiníveis. Isto significa passar os coeficientes de aproximações (cA) por outro banco de filtros, de modo que o sinal é dividido em várias componentes de resolução mais baixa, como apresentado na Fig 5.4(a). Note-se que $cD1$ representa os coeficientes de detalhe da componente de mais alta frequência do sinal e $cA3$ a de menor frequência.

As saídas de cada nível são dadas pelas seguintes equações:

$$cA_{j+1}(t) = \sum_{k} H(2t-k)cA_j(k) \tag{5-2}$$

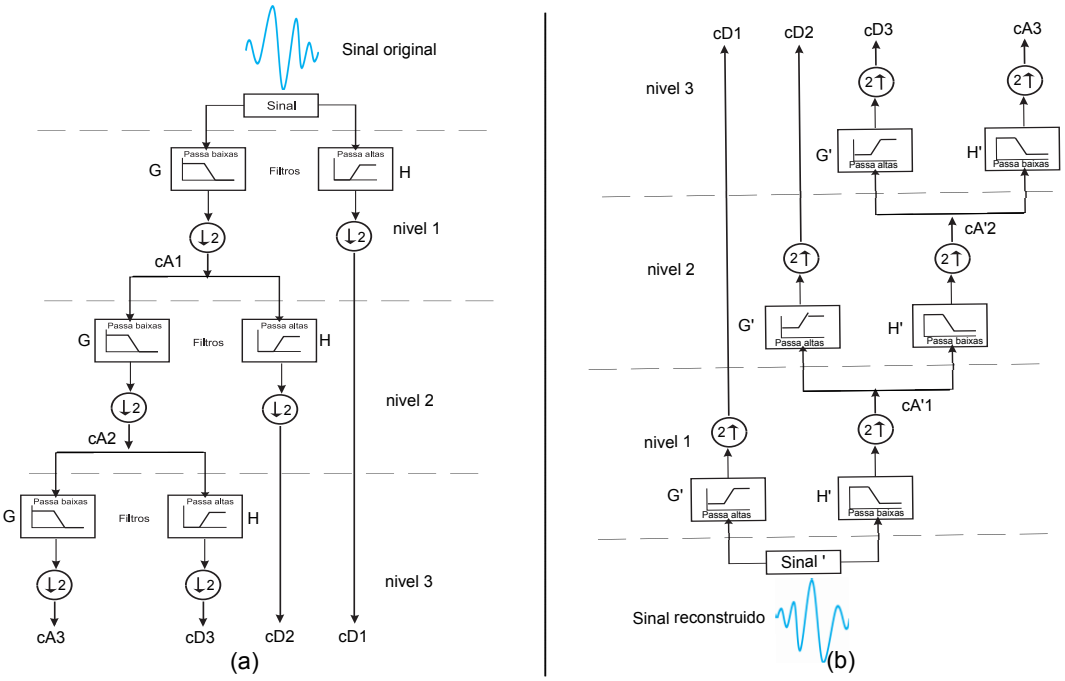

Figura 5.4: (a) decomposição multirresolução através de filtros multiníveis $Sinal = cA3 + cD3 + cD2 + cD1$ (b) reconstrução multi-resolução através de filtros inversos wavelet, cA1, cA2, cA3 representam os coeficientes de aproximação do sinal original nos níveis 1, 2, 3 respectivamente. cD1, cD2, cD3 representam os coeficientes de detalhe.

$$cD_{j+1}(t) = \sum_{k} G(2t-k)cA_j(k) \tag{5-3}$$

onde $H(2t-k)$ e $G(2t-k)$ são as respostas impulsionais dos filtros *wavelet* passa-altas, e passa-baixas, respectivamente, com uma dizimação da resolução de fator 2.

A reconstrução do sinal é feita a partir dos dados de cada nível da etapa de decomposição, usando a transformada *wavelet* inversa (Fig. 5.4(b)). Primeiro, nos coeficientes de cada nível tanto de aproximação quanto de detalhe é aplicado um processo de *up-sampling* que faz uma interpolação do sinal para compensar o *down-sampling* da decomposição. Em seguida se passa o sinal interpolado pelos filtros de sínteses H' e G' para finalmente reconstruir o sinal. De acordo com [214], o projeto dos filtros tanto para decomposição quanto para reconstrução devem satisfazer as condições estabelecidas pelos filtros espelhados em quadratura (*quadrature mirror filters,* QMF) [215].

5.1.3 Wavelet Denoising

Segundo [105], a partir da análise *wavelet* é possível realizar a filtragem dos sinais degradados para eliminação do ruído e posteriormente, restaurar o sinal original ou pelo menos, gerar um similar. Este processo de redução de ruído é conhecido como *denoising* e precisa de três etapas básicas, ilustradas no diagrama de blocos da Fig. 5.5.

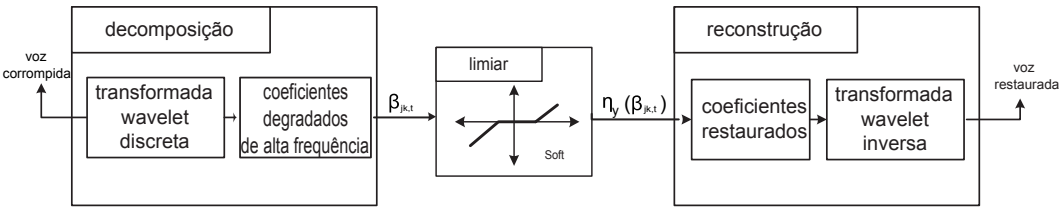

Figura 5.5: Diagrama de blocos da técnica wavelet-denoising.

Dado o sinal corrompido $y(t) = x(t) + r(t)$ a técnica *wavelet-denoising* proposta em [102][105], busca recuperar o sinal $x(t)$ dos dados ruidosos $y(t)$. O procedimento geral desta técnica representada na Fig. 5.5 é o seguinte:
Passo 1: Aplicar uma decomposição *wavelet* de J-níveis ao sinal ruidoso para produzir os coeficientes de *wavelet* corrompidos;
Passo 2: Aplicar o limiar apropriado aos coeficientes de detalhe (alta frequência) de forma a reduzir adequadamente os coeficientes *wavelet* do sinal ruidoso. A função do limiar pode ser *soft* ou *hard*. Neste trabalho, usa-se a função *soft* definida por

$$\eta_y(\beta_{jk,t}) = \begin{cases} sgn(\beta_{jk,t})(|\beta_{jk,t}| - \delta_j), & |\beta_{jk,t}| \geq \delta_j \\ 0, & caso \quad contrario \end{cases} \tag{5-4}$$

onde $\beta_{jk,t}$ é o k-ésimo coeficiente de detalhes (alta frequência) do sinal ruidoso no nível j, no t-ésimo quadro (tempo) em uma resolução específica, e a função *sgn(.)* é $+1$ se o argumento for positivo e -1 caso contrário. Note-se que $\eta(\beta_{jk,t})$ representa os coeficientes $\beta_{jk,t}$ restaurados;
Passo 3: Aplicar a transformada inversa *wavelet* dos coeficientes restaurados através do limiar *soft* para obter o sinal com redução de ruído.

O algoritmo de *wavelet-denoising* aqui utilizado foi desenvolvido levando em consideração o método de eliminação de ruído denominado *Visushrink* introduzido por Donoho [102], onde δ_j na equação (5-4) representa o *soft-theshold*, que é proposto como uma estimativa de limiar universal dado por

$$\delta_j = \sigma_j \sqrt{2ln(N_j)} \tag{5-5}$$

onde N_j representa o número dos coeficientes no nível j e σ_j é uma estimativa aproximada do nível de ruído. Esta estimativa é dada por

$$\sigma_j = \frac{med(\beta_{jk})}{0,6745}, \quad k = 0, ..., N_j - 1 \tag{5-6}$$

onde *med* é a mediana dos coeficientes de detalhe no nível de resolução mais alto $j = J$.

5.2 Mascaramento INM sobre wavelet denoising

O objetivo aqui é substituir os limiares usados em *wavelet-denoising* (ver equação (5-4)) pela máscara INM proposta no Capítulo 4. Os novos limiares serão definidos pela sigla $WLBP_{jk,t}$ a ser explicados posteriormente.

Como foi mencionado no Capítulo 4, o objetivo das técnicas de mascaramento é separar a voz das fontes de ruído. Para atingir esse objetivo foi proposta na literatura a máscara IBM [121], que consiste de uma matriz binária de unidades T-F construída a partir de voz corrompida, onde cada unidade T-F é definida como 1 se a SNR local for maior do que um limiar e 0 caso contrário (conforme detalhado na Seção 4.1). Esta máscara tem sido amplamente utilizada na literatura e demonstrou que, sob certas restrições, é a máscara binária ideal em termos de relação sinal-ruído (SNR). No entanto, esse método apresenta uma limitação significativa que afeta a qualidade da voz. Quando os componentes da frequência espectral são reduzidos a zero, eles produzem um ruído musical muito desagradável. Por outro lado, uma decisão crítica deste tipo de método é escolher o domínio T-F adequado para representar as variações do tempo no sinal. Tradicionalmente, esses métodos usam a transformada de *Fourier* de curto prazo (STFT) para produzir uma representação em unidades T-F do sinal corrompido.

Neste capítulo propõe-se uma nova abordagem para realce e reconhecimento de voz robusta. Ela é caracterizada pela aplicação da máscara INM, descrita no Capítulo 4, sobre o esquema de *wavelet-denoising*. No método proposto, aplica-se o mascaramento LBP (descrito na Seção 4.3) às sub-bandas de alta frequência da decomposição *wavelet*. O diagrama desta estimativa de máscara com base nos coeficientes *wavelet* é mostrado na Fig. 5.6.

Conforme mencionado na Seção 5.1.3, no passo 1, calcula-se a transfor-

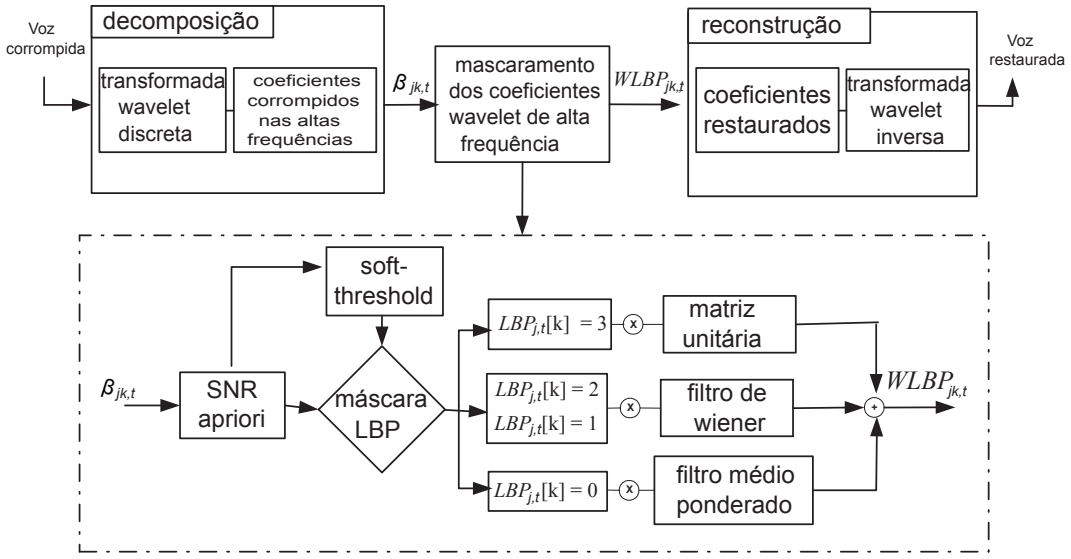

Figura 5.6: Diagrama de blocos do mascaramento proposto baseado nas técnicas local binary patterns (LBP) e transformadas wavelet.

mada *wavelet* do sinal de entrada usando a função *wavelet* mãe *Daubechies* através da seguinte equação

$$\beta_{jk,t} = \frac{1}{\sqrt{T_j}} \sum_{t=0}^{T-1} y(t)\psi_{j,k}(t) \tag{5-7}$$

onde $\beta_{jk,t}$ são os coeficientes da transformada *wavelet* para cada j-ésimo nível de decomposição (escala j) e k-ésima posição do coeficiente *wavelet* em análise, e $\psi_{j,k}(t)$ é a família de funções wavelet (*Daubechies)* com a escala j, e posição k [216] como apresentado na Seção 5.1.1. Na equação (5-7), t é o instante de tempo associado ao quadro em análise.

Como foi explicado na seção anterior, a decomposição do sinal em diferentes bandas de frequência é obtida por meio de filtragem sucessiva do sinal de entrada através de filtros passa-baixa e passa-alta no domínio do tempo. O sinal de voz corrompido de entrada $y(t)$ é primeiro filtrado por um filtro passa-baixa e um filtro passa-alta. O resultado será um sinal de aproximação $a1$ e um sinal de detalhe $d1$, cada um contendo metade das amostras do sinal de entrada $y(t)$. O filtro passa-alta produz os coeficientes *wavelet* de detalhe onde a máscara LBP será aplicada. O filtro passa-baixa produz a função de escala para o próximo nível da decomposição hierárquica. Quando as bandas de baixa frequência são introduzidas em outro sistema de banco de filtros, idêntico ao primeiro, é criada uma estrutura em árvore que divide o espectro do sinal original em oitavas. A decomposição produz J níveis

de coeficientes *wavelet* (ver Fig. 5.7) correspondentes a sinais individuais onde os de alta frequência serão usados para o mascaramento proposto. Na Fig. 5.7, o sinal de entrada é decomposto em 5 níveis, onde o sinal $y(t)$ (Fig. 5.7(b)) corresponde ao sinal a ser analisado. O sinal $a5$ é o componente de baixa frequência do sinal de entrada, que corresponde à saída do último filtro passa-baixa da árvore de decomposição. Os sinais dj $(j = 1...5)$ são os componentes de alta frequência.

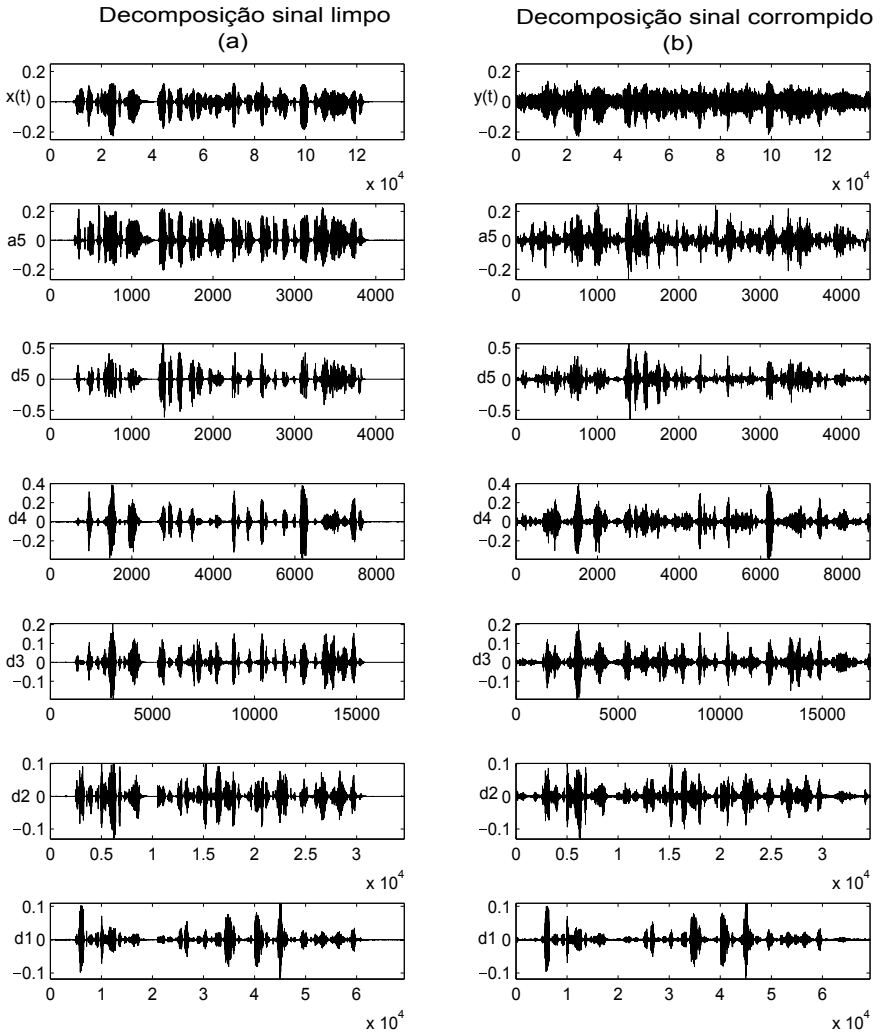

Figura 5.7: Decomposição wavelet do sinal de entrada em 5 níveis (a) sinal limpo (b) sinal corrompido com ruído babble de 0dB.

A saída de qualquer filtro de alta frequência é subdividida em quadros de tempo consecutivos de 32 ms com 10 ms de superposição. Esse processo gera uma matriz bidimensional $\Gamma_{(M,N)}$, onde M representa o número de quadros e N é o número de coeficientes *wavelet* em cada quadro. Isso produz uma

matriz de unidades T-F, para cada nível de decomposição. O operador LBP é adaptado para trabalhar em cada linha da matriz $\Gamma_{(M,N)}$ para cada j-ésimo nível de decomposição. A descrição matemática do operador LBP adaptado para cada j-ésimo nível, para o k-ésimo coeficiente em um particular tempo t (quadro) é dado da seguinte forma

$$LBP_{j,t}[k] = \sum_{i=1}^{p/2} \Big\{ sgn[\gamma[k-i] - \gamma[k]] 2^{p/2-i} + sgn[\gamma[k+i] - \gamma[k]] 2^{p/2+i-1} \Big\} \tag{5-8}$$

onde p é o número de coeficientes vizinhos em torno a cada unidade T-F de análise. A função $sgn[.]$ é definida como 1 se a diferença entre os coeficientes vizinhos e a unidade de análise T-F for maior do que o limiar δ_j dado em (5-5) e 0 caso contrário, $\gamma[k]$ representa o SNR *a priori* estimada para cada unidade T-F em dB diretamente dos coeficientes ruidosos (como explicado na Seção 4.3).

A estimação de $\gamma[k]$ é realizada através do algoritmo IMCRA relatado em [205]. Esse estimador é usado para substituir as condições ideais de pré-conhecimento dos sinais de voz e ruído usados nas máscaras ideias IBM e IRM.

Para o caso $p = 2$, obtém-se números binários de 2 bits, ou seja, códigos LBP decimais variando de 0 a 3. Para cada código LBP, estabelece-se o valor da máscara INM correspondente para 1 se o código LBP for 3. Isso representa a situação em que a energia da fala é significativamente maior do que a do ruído. Quando os códigos LBP são 1 ou 2, esses valores são suavizados usando a raiz quadrada do filtro de Wiener [204]. Finalmente, uma suavização temporal por meio de um filtro de média ponderado é realizado quando o código LBP é 0 para reduzir as flutuações entre a energia local da voz ruidosa e aquela quando a energia da voz é maior do que o ruído. Este procedimento realça a presença de voz em unidades T-F vizinhas, suavizando todas as unidades T-F com energia de ruído dominante, em vez de removê-las, como na IBM. Quantitativamente, para o caso em que $p = 2$ a máscara proposta WLBP para cada j-ésimo nível, do k-ésimo coeficiente e um tempo particular t é definida por

$$WLBP_{jk,t} = \begin{cases} \dfrac{\gamma[k-1] + 2\gamma[k] + \gamma[k+1]}{4} & para LBP_{j,t}[k] = 0 \\ \sqrt{\dfrac{\gamma[k]}{1+\gamma[k]}} & para LBP_{j,t}[k] = 1 \, ou \, 2 \\ 1 & para LBP_{j,t}[k] = 3 \end{cases} \tag{5-9}$$

Note-se, então, que os coeficientes dos sinais de detalhe corrompidos $\beta_{jk,t}$ são substituídos pelos coeficientes restaurados $WLBP_{jk,t}$. Finalmente, a transformação inversa é aplicada para obter a síntese do sinal. Nesta etapa, o sinal de voz processado pelo mascaramento LBP das *wavelets* de alta frequência é passado através dos filtros de síntese passa-alta e passa-baixa. Para os filtros passa-alta, o sinal reconstruído é derivado dos coeficientes retidos. A transformada *wavelet* inversa dos sinais de detalhe é dada por

$$\beta_{jk,t}^{-1} = \frac{1}{\sqrt{T_j}} \sum_j \sum_k WLBP_{jk,t} \psi_{j,k}(t) \tag{5-10}$$

No Algoritmo 2 resume-se a técnica proposta para o caso em que o número de vizinhos é $p = 2$.

Algorithm 2 Cálculo da máscara WLBP para p = 2

Entrada: Sinal de voz corrompido $y(t)$.
Saída: $WLBP_{jk,t}$.

1: Calcular a transformada wavelet para o sinal ruidoso usando a função mãe Daubechies 10 com $J = 5$ níveis de decomposição.
2: Segmentar cada j-ésimo nível de detalhe, $j = 1, ..., 5$, em quadros de 32 ms (256 amostras em uma frequência de amostragem de 8kHz) com intervalos de 10 ms.
3: Calcular a SNR *a priori* $\gamma[k]$ para cada sub-banda wavelet de alta frequência de acordo com [205].
4: **while** $\gamma[k]$ True (for $k = [0 : 255]$) **do**
5: Executar o código $LBP_{jk,t}[k]$ na janela de análise deslisante de comprimento $p = 2$.
6: **if** $\gamma[k \pm 1] \geq \gamma[k]$ **then**
7: p=0.
8: **return** $W_p = 1$; incrementar p
9: **else**
10: **return** $W_p = 0$; incrementar p
11: **end if**
12: **if** $p = 2$ **then**
13: **return** $LBP_{jk,t}[k] = 2^{W_p} + 2^{W_{p+1}}$
14: **end if**
15: **end while**
16: Separar todos os segmentos com diferentes valores de LBP.
17: Calcular $WLBP_{jk,t}$ de acordo com a equação (5-9)

5.3 PESQ e taxa de erro de palavra (WER) em reconhecimento de voz baseado em DNN

Nesta seção são apresentados e discutidos os resultados de simulação do algoritmo proposto (WLBP). Os resultados foram comparados com três

métodos tradicionais de realce de voz: subtração espectral clássica (SS), *wavelet-denoising* (WD) e a máscara binária estimada EBM. No caso dos algoritmos WLBP e EBM, estimou-se a SNR *a priori*, usando o algoritmo IMCRA proposto por Cohen [205]. Todos os experimentos foram realizados no subconjunto ruidoso da tarefa Aurora-4, conforme descrito em capítulos anteriores. As avaliações de desempenho foram realizadas com o PESQ para avaliar a qualidade da voz realçada e com a taxa de erro de palavra WER através do sistema RAV baseado em DNN. As métricas de avaliação, assim como a configuração utilizada para o sistema RAV, foram apresentadas no Capítulo 4.

Os resultados de desempenho em termos de qualidade PESQ, tomando-se a média sobre os diferentes tipos de ruído, são apresentados na Tabela 5.1. Nessa tabela, o sistema *noisy* significa que nenhum dos métodos de realce foi utilizado. A partir dessa tabela, observa-se que os resultados ruidosos obtidos sem qualquer técnica de realce, afetam significativamente o sinal de voz. A maior eficiência dos algoritmos é quando os LBPs são usados no processo de mascaramento *wavelet*. Como pode ser visto, o desempenho do esquema de mascaramento proposto melhora a medida PESQ em todos os cenários.

Tabela 5.1: Média do *PESQ* sobre os diferentes tipos de ruído

SNR	**noisy**	**SS**	**WD**	**WLBP**	**EBM**
0	1,06	1,19	1,20	**1,30**	1,11
5	1,19	1,40	1,42	**1,56**	1,22
10	1,45	1,75	1,78	**1,94**	1,47
15	1,87	2,20	2,25	**2,40**	1,89

É importante destacar como o mascaramento sobre as transformadas *wavelet* supera consideravelmente o mascaramento tradicional baseado em transformadas de *Fourier* (o método EBM) com taxas de qualidade superiores a $0,20$ para cenários de 0 e 5 dB e superiores a $0,5$ em cenários onde a SNR é menor.

Embora não tenha sido realizada uma avaliação subjetiva formal, observou-se a partir de testes de escuta informais que o método proposto não apresenta o ruído musical desconfortável, presente nas outras técnicas de realce.

Finalmente, o desempenho do sistema de reconhecimento de voz contínuo baseado em DNN foi avaliado pela média do desempenho de medida WER sobre as diferentes condições de SNR. Em condições limpas, o sistema produz uma WER de $4,71\%$. A Tabela 5.2 mostra o desempenho da máscara proposta em comparação com as demais técnicas para seis ruídos ambientais.

Para cada método a WER é calculada pela média sobre 0, 5, 10 e 15dB. Pode-se observar que o mascaramento de coeficientes *wavelet* com base em padrões binários locais (WLBP) supera a subtração espectral (SS), o *wavelet-denoising* (WD) e a máscara EBM em todos os cenários de ruído em média sobre as diferentes condições de SNR. Tomando a média da WER em todos os tipos de ruído na Tabela 5.2, resulta a coluna denominada *avg.* É importante ressaltar que para o esquema WLBP essa medida é $32,6\%$, enquanto que para os métodos de subtração espectral, *wavelet-denoising* e EBM as taxas são $37,11\%$, $35,08\%$ e $49,92\%$, respectivamente.

Tabela 5.2: Taxas de erro de palavras WER obtidas para o banco de dados AURORA-4, tomando-se a média sobre as diferentes condições de SNR

system	babble	airport	restaurant	street	car	train	avg.
Noisy	59,24	56,80	58,88	46,18	56,27	34,90	52,04
SS	43,95	38,48	46,57	31,94	39,15	22,62	37,11
WD	36,29	35,82	42,63	34,84	39,94	21,00	35,08
WLBP	**36,09**	**34,29**	**40,24**	**31,14**	**34,29**	**18,77**	**32,47**
EBM	56,37	54,28	57,34	44,67	54,02	32,86	49,92

A Tabela 5.2 mostra que as técnicas baseadas em *wavelet-denoising* são mais robustos em todos os cenários. Porém, se aplicado um mascaramento LBP sobre os coeficientes *wavelet* de alta frequência, a técnica oferece resultados superiores às técnicas de *wavelet-denoising* tradicionais. Os resultados mostram a importância de utilizar esse mascaramento sobre as altas frequências na decomposição *wavelet*. As taxas de erro médias do WLBP em relação ao WD são reduzidos de $36,29\%$ para $36,09\%$ no ruído *babble*, de $35,82\%$ para $34,29\%$ no ruído *airport*, de $42,63\%$ para $40,24\%$ no ruído *restaurant*, de $34,84\%$ para $31,14\%$ no ruído *street*, de $39,94\%$ para $34,29\%$ no ruído *car* e de $21,00\%$ para $18,77\%$ no ruído *train*.

5.4 Conclusões

Neste capítulo foi proposta uma nova abordagem para realçar a voz corrompida e melhorar desempenho de sistemas de reconhecimento de voz em ambientes adversos. O esquema proposto (WLBP) explora as características espectro-temporais da voz para realizar o realce do sinal, empregando a máscara de padrões binários locais (LBP) ou mascaramento INM em cada sinal de sub-banda de alta frequência da decomposição *wavelet*. O novo método WLBP foi comparado com técnicas de realce tradicionais da literatura (SS, WD, EBM) em seis ambientes ruidosos reais (*babble, airport, restaurant, street, car e train*). É empregado um algoritmo de estimativa de SNR, que é utilizada em todas

as técnicas de mascaramento que não depende das condições ideais dos sinais conhecidos *a priori*. Mostrou-se que os resultados fornecidos pelo esquema proposto são melhores em termos da medida de qualidade PESQ, mostrando ser uma boa técnica para o realce de voz. Os resultados experimentais obtidos com um sistema RAV baseado em DNN em ambientes ruidosos corroboram a superioridade do esquema proposto no cenário robusto de reconhecimento de voz. O algoritmo WLBP produz resultados de reconhecimento de voz superiores em comparação com os esquemas SS, WD e EBM. Isso revela que não só em relação à qualidade objetiva da voz, mas também em termos da taxa de erro da palavra de um reconhecedor de voz baseado em DNN, o WLBP é mais efetivo na redução de ruído. Observou-se que a máscara estimada pelo método proposto é mais robusta que a tradicional EBM, atingindo em média uma WER inferior a $17,45\%$ em todos subconjuntos ruidosos do corpus Aurora-4. Os testes de escuta informal também mostraram que o método proposto em um contexto acústico melhora a qualidade da fala, evitando o ruído musical, altamente desagradável, presente em outras técnicas de realce.

6
Segregação de Voz Usando a Máscara INM Baseada em Bancos de Filtros com Estimadores IMCRA e DNN

6.1
Introdução

A técnica de mascaramento INM proposta nesta tese, aplicada sobre o sinal de voz através da transformada de Fourier e da transformada *wavelet* tem mostrado ser uma ferramenta importante para melhorar a qualidade do sinal corrompido e consequentemente reduzir as taxas de erro de palavra nos sistemas de RAV. Esta máscara baseada na técnica de processamento de imagens Local binary Pattern (LBP) primeiro transforma o sinal de voz em representações T-F através da divisão em quadros consecutivos do sinal de voz. No Capítulo 4 foi usada a transformada de Fourier de tempo curto e no Capítulo 5 a máscara foi aplicada a cada sub-banda de alta frequência da transformada *wavelet* do sinal de voz com o objetivo de obter uma matriz de unidades T-F sobre a qual decide-se a predominância da voz ou a predominância do ruído.

Neste capítulo é apresentada uma nova abordagem da máscara INM, dessa vez usando os conceitos do modelo auditivo CASA, fazendo uma representação em tempo-frequência (T-F) do sinal de entrada, através do modelo computacional do sistema auditivo humano. Esse modelo consiste de um banco de filtros passa-banda que imitam a filtragem feita na cóclea do ouvido humano com o objetivo de realizar a classificação de unidades T-F a partir de uma estimação local da SNR. Inicialmente, é utilizado o estimador IMCRA [71] dentro do contexto geral do capítulo, que consiste de máscaras para realce de voz baseadas em bancos de filtros. A segunda parte do capítulo apresenta uma abordagem inovadora que utiliza um algoritmo de aprendizagem supervisionado das *Deep Neural Networks* (DNNs), que tem como entrada um conjunto de atributos extraídos de cada unidade T-F e tem como saída a SNR estimada. A abordagem proposta foi avaliada no banco de dados AURORA-4 sobre um sistema RAV baseado em DNN e os resultados demonstram que as taxas de erro de palavra do método proposto foram significativamente inferiores às das máscaras tradicionais IBM e IRM. O capítulo encerra com uma comparação

geral entre a nova proposta e aquelas apresentadas nos capítulos 4 e 5, onde a técnica de mascaramento INM (ou ENM, em sua versão que aplica estimação da SNR local) é aplicada sobre o sinal de voz. Novamente verifica-se a grande superioridade da proposta apresentada neste capítulo.

6.2 Mascaramento INM através de modelos auditivos

Conforme discutido nos Capítulos 2 e 4, para resolver o problema do *cocktail party*, foram propostos diversos métodos de segregação da voz ao longo dos anos, que usam os modelos auditivos do sistemas CASA (Computational Auditory Scene Analysis)[194] com o objetivo de separar as diferentes fontes de som que compõem a entrada acústica. Segundo *Bregman*[26], o processo de separação tem duas etapas principais: segmentação e agrupamento. Tipicamente estes sistemas de mascaramento na primeira etapa transformam o sinal de entrada em unidades T-F conhecidas como segmentos, onde espera-se que cada T-F seja gerado por uma mesma fonte de som. A segunda etapa é o agrupamento, onde os segmentos que provavelmente pertencem à mesma fonte são agrupados em um único vetor. A Fig. 6.1 apresenta um diagrama de blocos de um sistema CASA.

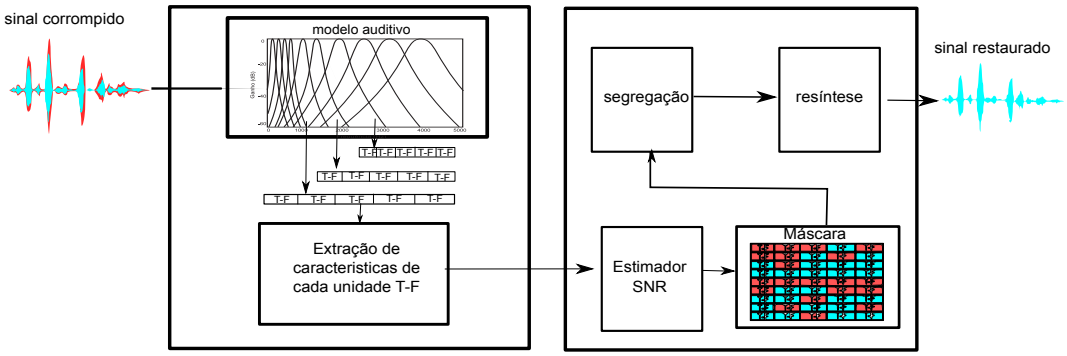

Figura 6.1: Diagrama de blocos sistema CASA.

6.2.1 Unidades tempo-frequência T-F

Como pode-se ver na Fig. 6.1, o primeiro procedimento dos sistemas CASA consiste em uma análise T-F do sinal de entrada. Na literatura existem dois tipos de transformações T-F que são comumente usados [191]. A primeira, chamada transformada em blocos usada no capítulo 4, que aplica a *short time Fourier transform* (STFT) para dividir o sinal de entrada em segmentos superpostos, transformando cada quadro para o domínio da frequência. A segunda

chamada transformada baseada em banco de filtros, que faz uma análise no domínio da frequência buscando que seja consistente com as propriedades de seletividade do sistema auditivo. Esta segunda transformada consiste na utilização de filtros Gammatome baseados na escala de Bandas Retangulares Equivalentes (ERB) [41]. Esses filtros possuem bandas de passo não uniformes e superpostas, onde cada filtro representa a resposta em frequência relacionada com um ponto particular da membrana basilar [217]. Tipicamente o banco de filtros é composto por 64 filtros gammatone com frequências centrais variando de 50 a 8000 Hz. A resposta ao impulso de cada filtro é dada por

$$g(t) = \begin{cases} t^{l-1}e^{-2\pi tb}\cos(2\pi f_c t), & t \geq 0 \\ 0, & caso \quad contrario \end{cases} \tag{6-1}$$

onde $l = 4$ é a ordem do filtro, b é a largura de banda retangular equivalente ERB e f_c é a frequência central associada a ela num instante t, e cuja resposta em frequência é dada por

$$G(\omega) = [1 + j(\omega - \omega_0)/b]^{-n} + [1 + j(\omega + \omega_0)/b]^{-n}, \quad (-\infty < \omega < \infty) \tag{6-2}$$

onde ω_0 representa a frequência central de cada filtro e n é a ordem do filtro. Para n fixa, b atua como um parâmetro de escala, de modo que a largura de banda de cada filtro aumenta com b, o parâmetro de ordem n controla a forma geral do filtro. Para b fixo, a largura de banda diminui à medida que n aumenta [218]. A largura de banda retangular de cada filtro e outras informações relativas aos filtros gammatone encontram-se no Apêndice A.

A resposta em frequência da Fig. 6.2 (b) mostra que os filtros são passa-banda, e que suas frequências centrais e larguras de banda aumentam logaritmicamente com a frequência.

Na saída de cada filtro é realizado um janelamento no tempo com janelas de 20 ms de duração e 10 ms de superposição, resultando em uma matriz de unidades T-F de 64 linhas e m colunas (m é o número de quadros considerando a superposição), onde cada unidade corresponde a um tempo e uma frequência determinada. Este processo de filtragem e janelamento é conhecido como *cochleagram* [217]. O número de linhas corresponde ao número de filtros utilizados.

Segundo *Roneel et al.* [219], o *cochleagram* consiste em representar o logaritmo da energia de cada unidade T-F. Uma característica importante que o difere do *spectrogram* é que o *cochleagram* possui mais componentes de frequência (resolução maior) na faixa de frequência mais baixa com menor largura de banda e menos componentes de frequência na faixa de frequência mais alta com maior largura de banda. Isso o torna uma ferramenta importante

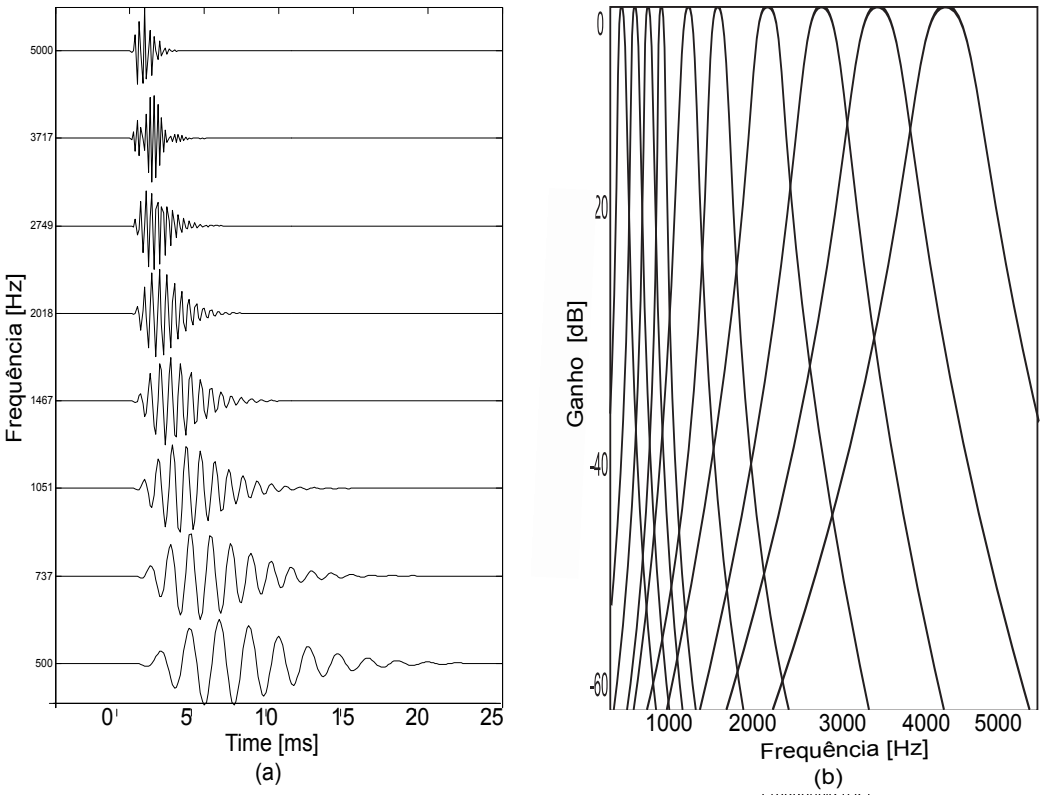

Figura 6.2: Filtros gammatone. (a) resposta impulsiva para 8 filtros gammatone, (b) resposta em frequência desses filtros.

para trabalhar com modelos auditivos. A Fig. 6.3 apresenta a diferença entre a representação da sentença "440c020a" da base AURORA-4 em unidades tempo-frequência através do *cochleogram* e o *spectrogram*.

Neste capítulo esta segunda abordagem *cochleagram* será usada com o objetivo de obter as unidades T-F para fazer os mascaramentos INM, IRM e IBM.

6.2.2 Máscaras baseadas na estimativa da SNR local

Existe uma mistura de unidades T-F confiáveis e não confiáveis quando a voz é contaminada por ruído aditivo. Se diz que uma unidade T-F é confiável quando sua energia é maior que um determinado limiar, que é quando a unidade T-F contém predominantemente energia de voz. Caso contrário, as unidades T-F são dominadas pela energia do ruído sendo consideradas como dados não confiáveis. Uma medida eficiente e amplamente utilizada de saber se uma determinada unidade T-F pertence à voz dominante ou ao ruído dominante, consiste em calcular a relação sinal-ruído (SNR) local dessa unidade. No

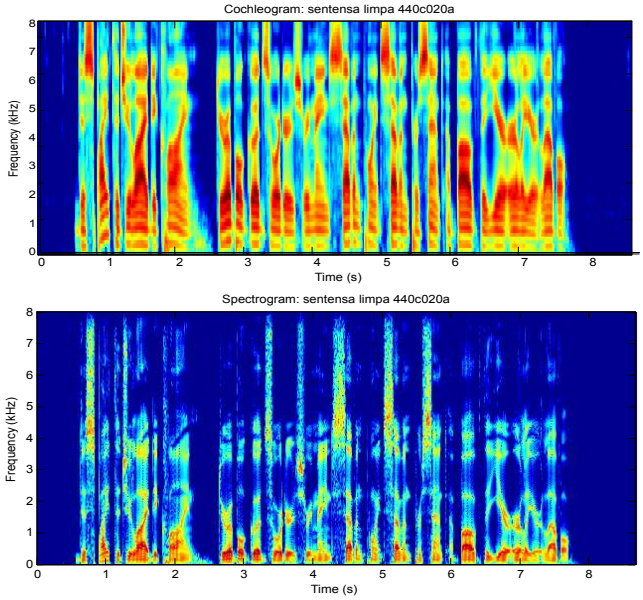

Figura 6.3: Comparação entre o cochleogram (acima) e spectrogram (abaixo) da sentença "440c020a".

Capítulo 4 a SNR foi definida como

$$SNR_{(t,\omega)} = 10log\frac{X_{(t,\omega)}}{R_{(t,\omega)}} \tag{6-3}$$

onde $X_{(t,\omega)}$ e $R_{(t,\omega)}$ são a energia instantânea do sinal limpo e do ruído respectivamente em um tempo t e uma frequência ω. Nesse caso a unidade T-F é só uma amostra. Como explicado no Capítulo 4, estas unidades T-F são calculadas através de transformação por bloco. Ou seja, apenas um único valor espectral é observado em cada unidade T-F. Quando usada a transformação de unidades T-F através de banco de filtros, cada unidade T-F passa a ser um quadro, ou seja, a equação (6-3) transforma-se em

$$SNR_{(t,\omega)} = 10log\frac{\sum_k X^2_{(t,\omega)}(k)}{\sum_k (Y_{(t,\omega)}(k) - X_{(t,\omega)}(k))^2} \tag{6-4}$$

onde $X_{(t,\omega)}$ e $Y_{(t,\omega)}$ são o sinal limpo e corrompido, respectivamente, e k representa o índice de cada amostra em um segmento.

Com base nessa SNR, podemos calcular a confiabilidade de cada elemento do espectro. Se o que é desejado é uma máscara binária, essa confiabilidade pode ser calculada pelo limite de valores SNR com base em um determinado limiar LC como apresentado no capítulo 4, onde valores de SNR com voz dominante serão representados com 1 e 0 caso contraio, ou seja,

$$IBM_{(t,\omega)} = \begin{cases} 1 & se \quad SNR_{(t,\omega)} > LC \\ 0 & caso \quad contrario \end{cases} \tag{6-5}$$

A equação 6-5 representa a conhecida *ideal binary mask* (IBM), considerada ideal, já que se dá uma separação perfeita entre voz e ruído, devido

ao fato de conhecer a *priori* os componentes do sinal. Este tipo de mascaramento é útil para conhecer o nível máximo de confiabilidade que pode chegar a ter a técnica, e por que são utilizadas como *targets* (alvos) em processos de separação supervisionados.

Por outro lado, o que é realmente importante é aplicar estes métodos em ambientes reais, onde consiga-se estimar a confiabilidade de cada um dos elementos do espectro do sinal capturado em ambientes adversos sem conhecer sinal e ruído a *priori*. Para isso, é preciso estimar a SNR local de cada unidade T-F. Esta SNR desempenha um papel importante na etapa de mascaramento, já que o desempenho das máscaras depende de sua estimação. Em nossa proposta, o método *improved minima controlled recursive averaging* (IMCRA) e um algoritmo de aprendizagem supervisionado DNN são usados para estimar a SNR e realizar o respectivo mascaramento. As novas abordagens são apresentados a seguir.

6.2.2.1 A. Máscara INM sobre o banco de filtros com estimador IMCRA

O mascaramento INM possui a capacidade de usar eficientemente os LBPs para estimar uma máscara ideal que identifica quais unidades T-F do sinal de voz corrompido são dominadas pelo ruído. Nesta seção se faz uma estimativa da máscara sobre o banco de filtros auditivos com o objetivo de diminuir as taxas de erro de palavra e comparar seu desempenho com as máscaras tradicionais IBM e IRM. Esta nova proposta é representada em diagrama de blocos na Fig. 6.4.

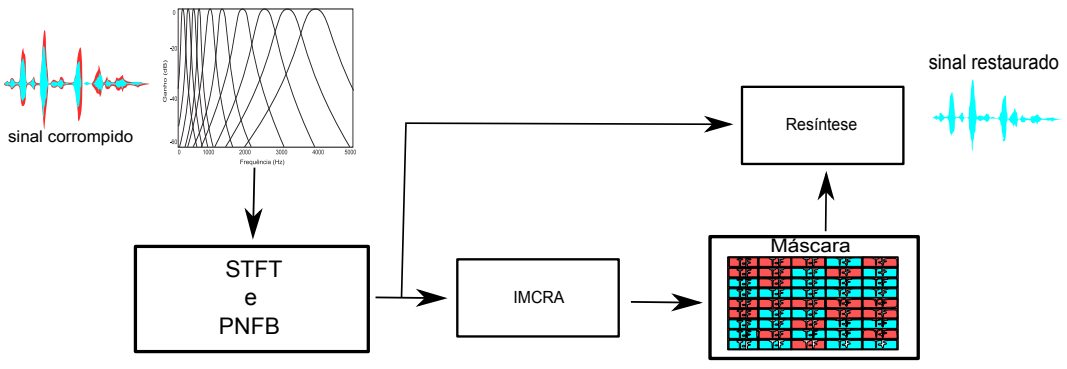

Figura 6.4: Diagrama em blocos de ENM sobre o banco de filtros com estimador IMCRA.

As unidade T-F são extraídas dos bancos de filtros através de um procedimento de filtragem e janelamento chamado *cochleogram*. Através deste procedimento é possível conhecer a SNR ideal das unidades T-F quando o sinal limpo e o ruído são conhecidos. No entanto, o que se busca nesta seção é utilizar

um algoritmo que estime a SNR de cada unidade T-F do sinal corrompido a fim de criar a máscara INM proposta nesta tese.

Nesta nova abordagem usamos o algoritmo IMCRA a fim de estimar a SNR *a priori* do sinal corrompido de cada unidade T-F. Isso é feito de forma semelhante ao que foi descrito no Capítulo 4 para a estimativa da máscara EBM, mas dessa vez a estimação da SNR é feita a partir de cada banda de frequência dos filtros gammatone apresentados na seção anterior. O estimador IMCRA é dividido em duas etapas, onde cada uma possui duas fases, uma de suavização do espectro de potência do sinal ruidoso e outra de localização por estatísticas mínimas, que tem o objetivo de estimar o espectro de potência do ruído acústico presente no sinal de voz. Detalhes sobre o algoritmo IMCRA podem ser encontrado em [71].

Após de obter a SNR das unidades T-F, é realizado o procedimento de mascaramento INM proposto no capítulo 4, representado pela seguinte equação

$$INM[k] = \begin{cases} \dfrac{\gamma[k-1] + 2\gamma[k] + \gamma[k+1]}{4} & para LBP[k] = 0 \\ \sqrt{\dfrac{\gamma[k]}{1+\gamma[k]}} & para LBP[k] = 1 \; or \; 2 \\ 1 & para LBP[k] = 3 \end{cases} \tag{6-6}$$

onde $\gamma[k]$ é a SNR estimada com o algoritmo IMCRA na *k-ésima* unidade T-F. O algoritmo 3 resume o procedimento realizado para mascarar o sinal através do método INM no caso de $p = 2$ vizinhos. Nesse caso a máscara INM será denominada ENM.

Finalmente, quando se usa o estimador IMCRA, é aplicada uma variante na reconstrução do sinal depois de aplicar o mascaramento. São aproveitadas as características dos atributos *Power normalized filter bank* (PNFB) (que são a etapa prévia aos coeficientes cepstrais PNCC [220], ou seja, a etapa anterior a aplicação da DCT) que simulam o sistema auditivo humano. Esses surgiram como um conjunto de características para o reconhecimento de voz que são mais robustos em relação à variabilidade acústica, e não apresentam perda de desempenho quando o sinal de fala não é degradado. Os PNFB são obtidos a partir de uma matriz de unidades T-F com as mesmas configurações de decomposição do *cochleogram*, seguindo o seguinte procedimento:

– O sinal de entrada é passado por um filtro de pré-enfase a fim aumentar a magnitude de algumas frequências, para compensar os efeitos dos pulsos glotais e a ressonância dos lábios.

Algorithm 3 Cálculo da máscara ENM para p=2

Input: Sinal de entrada.
Output: $INM[k]$.

1: Aplicar filtro pré-enfase
2: Passar o sinal por um banco de filtros gammatone de 64 canais, com frequências igualmente espaçadas entre 50Hz e 800Hz.
3: Segmentar a voz em quadros de 20-ms com 10-ms de superposição entre quadros.
4: Calcular $SNR[k]$ *a priori*, ou seja, $\gamma[k]$ de cada segmento usando o algoritmo IMCRA [71] .
5: **while** $\gamma[k]$ True (para $k = [0 : N-1]$) **do**
6: Calcular códigos LBP na janela de análise deslisante W de comprimento p.
7: **if** $\gamma[i \pm 1] \geq \gamma[i]$ **then**
8: p=0.
9: **return** $W_p = 1$; incrementar p
10: **else**
11: **return** $W_p = 0$; incrementar p
12: **end if**
13: **if** $p = 2$ **then**
14: **return** $\gamma[k] = 2^{W_p} + 2^{W_{p+1}}$
15: **end if**
16: **end while**
17: Separar todos os segmentos com diferentes valores LBP
18: Calcular INM[k] de acordo com (6-6)

- O sinal filtrado é transformado ao domínio da frequência usando a *short-time Fourier transform* (STFT) com quadros de 20 ms superpostos com 10 ms.
- A potência espectral em N bandas de análise é obtida ponderando a magnitude quadrada das saídas da STFT pela resposta em frequência associada com as N bandas do banco de filtros gammatone.
- É realizada uma estimativa e remoção de ruído em cada banda através de uma série de operações não-lineares, variáveis no tempo, que são executadas usando uma análise temporal de longo tempo que faz subtração de ruído, acrescentando um grau de robustez em relação ao coeficientes tradicionais.
- Finalmente, a energia de cada banda é calculada aplicando uma não-linearidade denominada *power-law* com expoente 1/15

Informação detalhada do procedimento acima exposto encontra-se em [220]. Multiplicando a matriz que se obteve com a decomposição em PNFB pela máscara INM obtida na etapa anterior busca-se dar maior robustez ao sinal já que os coeficientes PNFB incluem uma etapa de remoção de ruído

chamada integração temporal para a análise de ambiente. Esse procedimento também é feito para as máscaras IBM e IRM.

6.2.2.1 Taxa de erro de palavra (WER) em reconhecimento de voz baseado em DNN

A Tabela 6.1 apresenta uma comparação entre as máscaras ideais INM[1] IBM[1] e IRM[1] e estimadas ENM[1] ,EBM[1] e ERM[1] do capítulo 4 e as máscaras ideais INM[2] IBM[2] e IRM[2] e estimadas ENM[2] ,EBM[2] e ERM[2] deste capítulo.

Os resultados que serão mostrados a seguir referem-se ao mascaramento ENM[2] sobre o banco de filtros gammatone tomando o mesmo subconjunto ruidoso das tarefas AURORA-4 com as mesmas condições descritas em capítulos anteriores. A medida de desempenho é dada em termos da taxa de erro de palavra WER. A configuração do sistema RAV é a mesma descrita no capítulo 4. Na Tabela. 6.1 são apresentados os resultados de simulações do algoritmo proposto usando o estimador IMCRA. Os resultados foram comparados com os métodos tradicionais de mascaramento IRM[2] e IBM[2] e o mascaramento proposto no capitulo 4, tanto em sua forma ideal (condição oráculo em que se supõe que os sinais são conhecidos) como em sua forma usando o estimador IMCRA. Observa-se que obviamente, os melhores resultados são para as máscaras ideais (condição oráculo). Nesse caso, o INM[2] apresentam os melhores resultados dentre todas as máscaras.

Dos resultados de desempenho em termos de WER mostrados na Tabela 6.1 pode-se ver como o mascaramento ENM[2] baseado em códigos LBP fornece melhores resultados em todos os ambientes, melhorando o desempenho dos sistemas RAV em comparação com os métodos tradicionais. Os resultados mostram a importância de utilizar esse mascaramento sobre os bancos de filtros que simulam o sistema auditivo humano. As taxas de erro médias do ENM[2] em relação ao sistema mais robusto da literatura ERM[2] são reduzidas em média de $45,87\%$ para $44,62\%$ no ruído *babble*, de $45,67\%$ para $45,02\%$ no ruido *airport*, de $52,07\%$ para $51,96\%$ no ruído *restaurant*, de $36,33\%$ para $35,55\%$ no ruído *street*, de $44,60\%$ para $40,73\%$ no ruído *car*, e de $25,44\%$ para $21,69\%$ no ruído *train*.

No entanto, com relação à técnica proposta no capítulo 4 (ENM[1]) pode-se ver que a nova abordagem ENM[2] fornece resultados significativamente piores em termos de taxas de erro de palavra.

[1]máscara estimada a partir do sinal de voz original
[2]máscara estimada após a passagem pelo banco de filtros

Tabela 6.1: Resultados de reconhecimento usando o estimador IMCRA, obtidos para o banco de dados AURORA-4. Tomando-se a média sobre as diferentes condições de SNR. RI significa melhoria relativa em relação ao sistema ruidoso

system	**babble**	**airport**	**restaurant**	**street**	**car**	**train**	**avg.**
Noisy	59,24	56,27	58,88	46,18	56,27	34,90	52,04
IBM^2	11,18	11,81	14,11	9,88	14,24	8,37	11,60
IRM^2	4,85	4,60	5,22	4,36	5,03	4,60	4,63
INM^2	**3,38**	**3,39**	**3,51**	**3,32**	**3,36**	**3,39**	**3,34**
IBM^1	39,30	37,00	38,68	35,14	40,67	33,59	37,40
IRM^1	29,07	28,17	29,60	27,27	29,48	26,53	28,35
INM^1	**20,93**	**27,23**	**29,93**	**26,42**	**26,42**	**25,37**	**25,93**
EBM^2	53,20	52,95	59,23	43,69	53,39	30,16	48,77
ERM^2	45,87	45,67	52,07	36,33	44,60	25,44	41,66
ENM^2	**44,62**	**45,02**	**51,96**	**35,55**	**40,73**	**21,69**	**39,93**
EBM^1	56,37	54,28	57,34	44,67	54,02	32,86	49,92
ERM^1	51,16	48,52	53,22	39,47	45,16	24,11	43,60
ENM^1	**43,86**	**41,79**	**46,68**	**32,45**	**37,48**	**16,91**	**36,52**

Com base nesses resultados foi proposta uma nova abordagem, dessa vez através de um algoritmo de aprendizagem supervisionado baseado em DNN usado para estimar técnicas de mascaramento.

6.2.2.2
B. Máscara INM sobre o banco de filtros com estimador DNN

Nessa seção é proposta uma nova forma de separar as unidade T-F predominantemente contaminadas com ruído das unidades predominantemente dominadas pelo sinal de voz, através de um processo de aprendizado supervisionado baseado em DNNs. A ideia principal deste tipo de técnica é criar uma DNN que aprenda o alvo de treinamento (SNR local associada à máscara ideal) através de um algoritmo de aprendizagem supervisionado, com o objetivo de mapear o conjunto de características extraídas de um sinal corrompido para esse alvo. Em [221] a separação da voz usando *Ideal Binary Mask* (IBM) como alvo das DNN apresentou grandes melhorias na inteligibilidade e qualidade da voz comparado com técnicas de aprendizado supervisionado como *support vector machine* (SVM). No entanto, como foi explicado no capítulo 4 a separação da voz usando técnicas de supressão como as máscaras binárias geralmente produz ruído musical residual. Devido a isso, em [127] a *ideal ratio mask* (IRM) foi proposta como alvo de treinamento mostrando melhoras significativas que superam a estimativa da IBM em termos de desempenho de sistemas RAV robustos.

No novo enfoque aqui proposto usamos o algoritmo supervisionado proposto em [221] para predizer a INM, buscando melhorar a estimação das

máscaras tradicionais IBM e IRM em termos de desempenho robusto de sistemas RAV. A Fig. 6.5 mostra o diagrama de blocos da estimação de uma máscara baseada em DNN.

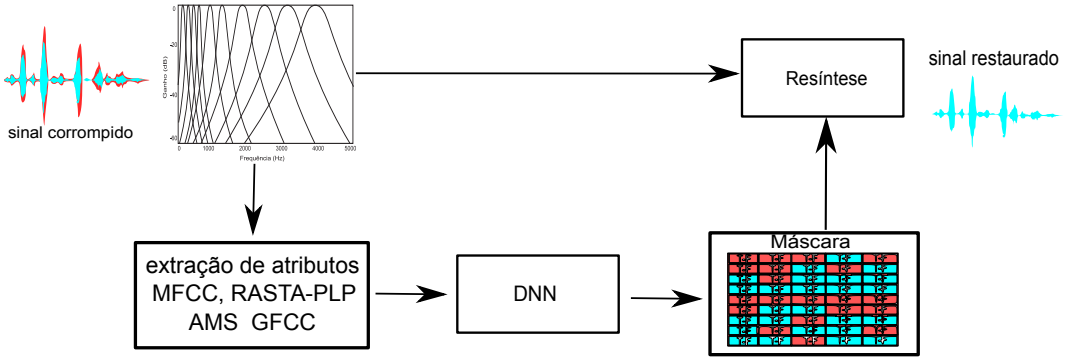

Figura 6.5: Diagrama em blocos de ENM sobre o banco de filtros com estimador baseado em DNN.

6.2.2.2 Extração de atributos para treinamento das DNNs

Nesta etapa são extraídas de cada unidade T-F atributos que ajudarão a decidir o quão confiável é cada elemento do espectro para ser etiquetado como voz. Essas características são usadas no treinamento das DNNs como entradas, gerando na saída uma estimativa da SNR ideal para as máscaras INM, IRM e IBM. Uma abordagem de extração de atributos que tem sido estudada nos últimos anos para mascaramento auditivo é a proposta por *Wang* [120] e está baseada na mistura de um conjunto de características obtidas da saída de cada banco de filtros. Estas características são:

- 31 coeficientes MFCC usando filtro de pré-enfase seguido por uma transformada STFT com quadros janelados de 20 ms e 10 ms de superposição.
- 13 coeficientes RASTA-PLP usando quadros de 20 ms com 10 ms de superposição.
- 15 atributos AMS *(Amplitude Modulation Spectrograms)* usando quadros de 32 ms com 22 ms de superposição, os atributos AMS são extraídos do banco do de filtros depois de sub amostrar o sinal para 4000 Hz
- 64 atributos GFCC (*Gammatone Frequency Cepstral Coefficients*) usando filtro de pré-enfase e um banco filtros gammatone, seguido por uma transformada STFT com quadros janelados de 20 ms e 10 ms de superposição que são dizimados para aplicar uma potenciação de 1/15.

Misturando essas características tem-se um vetor de atributos com suas componentes delta de dimensão $((31 + 13 + 15 + 64) \times 2) = 246$. Utilizando as representações acústicas ou conjunto de características, o próximo passo é identificar as unidades que contêm informação dominante em relação ao ruído para agrupá-las e etiquetá-las como unidades confiáveis pertencentes ao mesmo som. Este procedimento pode ser realizado com máscaras baseadas na estimativa da SNR local [128] [197], máscaras baseadas na classificação Bayesiana do espectro [116][222], entre outras. Neste capítulo continuaremos com as máscaras baseadas na estimativa da SNR local, da mesma forma que nos capítulos 4 e 5.

6.2.2.2 Alvo de treinamento e configuração da DNN

A *Ideal Neighbourhood mask* (INM) é definida como a codificação da energia das unidades T-F vizinhas com relação à unidade central de análise, de modo que a informação codificada será do nível mais alto. Considera-se aqui valores entre 0 e 3 com os que se etiquetará segmentos do sinal com voz dominante ou ruído dominante. Esta máscara, da mesma forma que a IBM e IRM, assume que o sinal de voz e o ruído são decorrelatados e que o sinal corrompido pode ser aproximado como a suma entre a energia do sinal e a energia do ruído. Como foi mencionado anteriormente, os alvos tradicionais para os mascaramentos são o IBM e o IRM. Nesta seção propomos como alvo de treinamento o nosso algoritmo de mascaramento proposto no capítulo 4, dado pela equação

$$INM[k] = \begin{cases} \dfrac{\gamma[k-1] + 2\gamma[k] + \gamma[k+1]}{4} & para LBP[k] = 0 \\ \sqrt{\dfrac{\gamma[k]}{1+\gamma[k]}} & para LBP[k] = 1 \; or \; 2 \\ 1 & para LBP[k] = 3 \end{cases} \tag{6-7}$$

onde $\gamma[k]$ representa a SNR instantânea (equação 6-4) estimada para cada unidade T-F em dB diretamente das unidades T-F corrompidas em um dado quadro. A estimação da INM é realizada em dois importantes passos. O primeiro treinando uma DNN com o conjunto de características, e uma segunda rede *multi layer perceptron* (MLP) que suavizará as saídas da DNN. Para o primeiro passo, neste trabalho usamos o algoritmo supervisionado de estimação da IBM proposto em [120] apresentado a seguir:

– A DNN usa 4 camadas ocultas, cada uma composta por 1024 neurônios,

a camada de saída é composta de 64 neurônios correspondente aos 64 filtros que compõem o banco de filtros de entrada.

– Na entrada da DNN foram agrupados 5 quadros de atributos com o objetivo de incorporar informação do contexto. Desse modo a dimensão do conjunto de atributos de entrada é de 1230 ((31 MFCC × 64 GFCC × 13 RASTA-PLP × 15 AMS) × 2 delta × 5 quadros)

– Uma normalização de média, variância e filtragem temporal (MVA) é aplicada no conjunto de atributos, ajudando a minimizar a variabilidade dos dados [223].

– O treinamento da DNN inclui uma fase de pré-treinamento não supervisionada baseada nas *restricted boltzmann machine* RBM [224] onde é modelada a dependência entre um grupo de variáveis aleatórias usando uma arquitetura de duas camadas uma visível V e uma oculta H, com as entradas de todos os nós visíveis V sendo passadas para todos os nós ocultos H como apresentado na Fig. 6.6. Cada RBM treina-se consecutivamente com o conjunto de atributos de entrada sem supervisão (sem alvo). A camada de entrada visível H recebe os atributos de entrada e em cada nó oculto, cada entrada v é multiplicada pelo seu respectivo peso w. Cada nó escondido recebe as entradas multiplicadas pelos respectivos pesos. A soma desses produtos é passada através do algoritmo de ativação produzindo uma saída para cada nó oculto. Isso é feito para cada par de camadas na DNN. Detalhes sobre o algoritmo de pré-treinamento podem ser encontrado em [225].

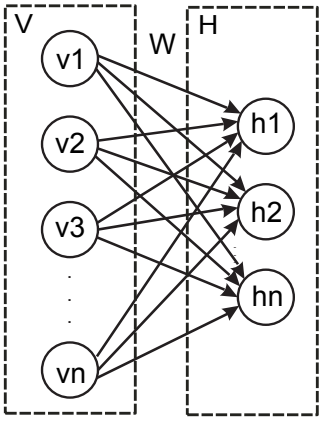

Figura 6.6: Arquitetura para uma RBM.

O método de pré-treinamento foi usado para melhorar a inicialização dos parâmetros antes da usar o treinamento supervisionado *back-propagation*. Os pesos das unidades ocultas são inicializados camada a camada usando

o pré-treinamento (RBM) acima exposto. Utiliza-se uma taxa de aprendizagem de 0,1 e momentum de 0,9 com uma degradação de peso de 0,0001. Cada camada é pré-treinada com 20 épocas.

– Com os pesos inicializados no pré-treinamento uma segunda etapa de ajuste de pesos é realizada através de um aprendizagem supervisionado *back-propagation*[226]. Nesta etapa, o método de gradiente descendente adaptativo [227] é usado ao longo do treinamento com uma taxa de aprendizagem geral de 0,05. O *mumentum* é ajustado inicialmente para 0,5 e aumentando para 0,9 depois que a rede treina a metade do número total de épocas, que é definida como 100.

– A DNN é treinada para prever a saída desejada em todas as bandas de frequência, para isso usa-se o função de erro *cross-entropy* como critério de aprendizagem, no qual o sinal de erro associado à camada de saída é diretamente proporcional à diferença entre os valores de saída desejados e reais (alvo). A função *cross-entropy* acelera o algoritmo de *back-propagation*

– A função de ativação *Sigmoid* é usada tanto nas camadas ocultas quanto na camada de saída.

Em [127] uma segunda rede MLP com uma camada oculta de 100 neurônios é usada à saída da DNN com o objetivo de suavizar a estimação realizada pela DNN. Esta rede recebe a saída da DNN e toma como entrada 5 quadros atrás e 5 quadros na frente do quadro atual para o qual a saída deve ser estimada tomando como alvo o mesmo usado nas DNN. Isso resulta em uma entrada de atributos de dimensão 704 (64×11). Porém, isto complicaria o treinamento, já que precisaria da restauração de muitos atributos de uma só vez, tornando a rede complexa no seu desenvolvimento. Em [216] propõe-se dividir a rede em várias redes com as mesmas características, com o objetivo que a entrada de cada rede tenha os 704 atributos, mas na saída cada rede representa a SNR associada à saída de cada filtro do banco de filtros. Tem-se, portanto, 64 redes de uma camada. A Fig. 6.7 mostra o diagrama geral da estimação da INM com a etapa final de suavização dividida em várias redes neurais, conforme apresentado em [216].

6.2.2.2 PESQ e Taxa de erro de palavra (WER) em reconhecimento de voz baseado em DNN

Nesta seção são apresentados os resultados de simulações em termos de qualidade PESQ e taxa de erro de palavra WER, obtidos para o conjunto

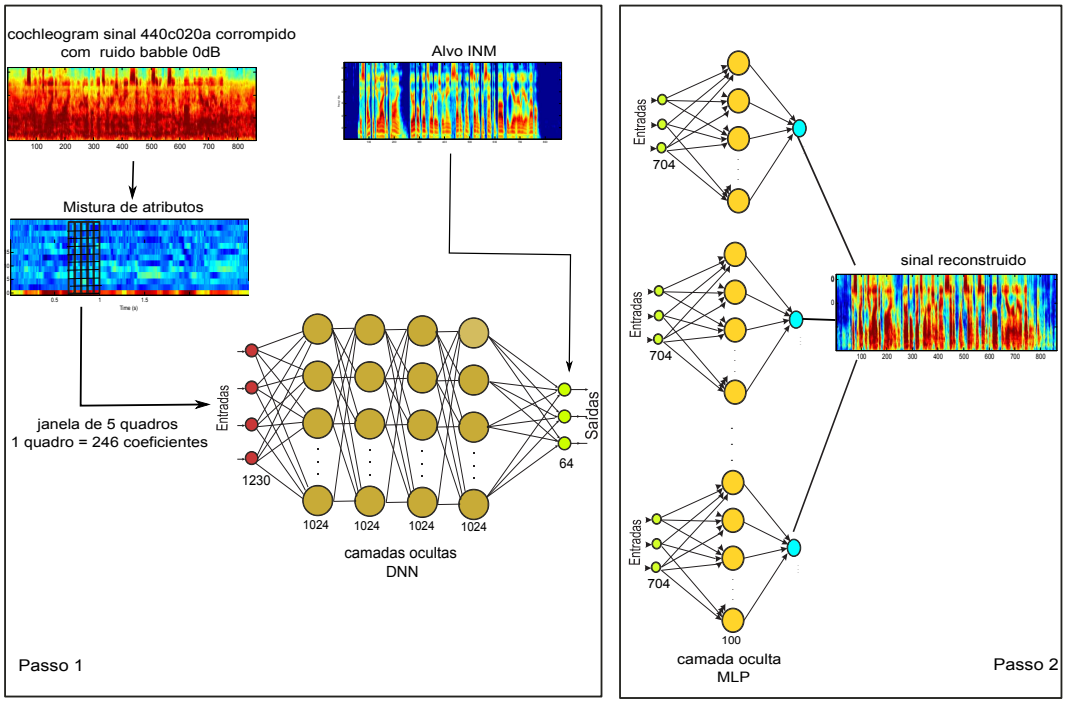

Figura 6.7: Diagrama em blocos do método de realce proposto baseado na estimação de uma máscara INM[18] com DNN.

de tarefas da base de dados AURORA-4. As configurações de cada uma das métricas de medição de desempenho assim como a configuração do sistema RAV encontram-se no capítulo 4.

Na Tabela 6.2 são mostrados os resultados da medida PESQ obtida com o mascaramento baseado em DNN[3] em comparação com o PESQ obtido no capítulo 4 para os métodos EBM[1], ERM[1] e ENM[1]. Os resultados são fornecidos para SNR de 0, 5, 10 e 15 dB, tomando-se a média sobre 6 tipos diferentes de ruído. O objetivo desta comparação é observar se o mascaramento sobre o banco de filtros melhora a inteligibilidade e qualidade do sinal em comparação a técnica INM proposta sobre o sinal da voz. De acordo com os resultados podemos dizer que o mascaramento INM baseado em DNN comporta-se melhor em todos os cenários com relação às técnicas tradicionais. Por outro lado, pode-se ver da Tabela 6.2 como mascarar através de DNN sobre o banco de filtros (ENM[3]) oferece um desempenho superior em relação à técnica proposta no capítulo 4 (ENM[1]) com taxas de qualidade acima a 0,4 para SNR de 0 e 5 dB e 0,5 para 10 e 15 dB.

A Tabela 6.3 apresenta as taxas de erro de palavra dos sistemas RAV baseado em DNN obtidas com o mascaramento INM proposto nesta seção em comparação com a técnica proposta no capítulo 4. Os resultados são

[3]Máscara INM sobre o banco de filtros com estimador DNN

Tabela 6.2: Média do *PESQ* sobre os diferentes tipos de ruído usando o estimador baseado em DNN.

SNR	**noisy**	**EBM**^{1}	**ERM**^{1}	**ENM**^{1}	**EBM**^{3}	**ERM**^{3}	**ENM**^{3}
0	1,06	1,11	1,14	1,24	1,37	1,59	1,66
5	1,19	1,22	1,30	1,46	1,53	1,88	1,91
10	1,45	1,47	1,59	1,81	1,75	2,31	2,40
15	1,87	1,89	2,03	2,27	2,03	2,79	2,86

fornecidos para 6 tipos diferentes de ruído, tomando-se a média sobre as diferentes condições de SNR. Como pode-se observar, o mascaramento ENM^3 com DNN oferece melhoras significativas em todos os cenários com relação aos métodos tradicionais ERM^3 e EBM^3. Também pode-se verificar que as taxas de erro de palavra em comparação com o método proposto no capítulo 4 são significativamente altas com uma diferença em média de $25,88\%$, o que confirma que os algoritmos de aprendizagem supervisionado são estimadores muito superiores ao método proposto na seção anterior com o estimador IMCRA e o proposto no capítulo 4.

Tabela 6.3: Resultados de reconhecimento obtidos para o banco de dados AURORA-4. Tomando-se a média sobre as diferentes condições de SNR. avg. significa uma média geral sobre todos os tipos de ruído

system	**babble**	**airport**	**restaurant**	**street**	**car**	**train**	**avg.**
Noisy	59,24	56,27	58,88	46,18	56,27	34,90	52,04
EBM^1	56,37	54,28	57,34	44,67	54,02	32,86	49,92
ERM^1	51,16	48,52	53,22	39,47	45,16	24,11	43,60
ENM^1	**43,86**	**41,79**	**46,68**	**32,45**	**37,48**	**16,91**	**36,52**
EBM^3	41,00	39,74	41,95	34,65	35,51	26,77	36,60
ERM^3	18,15	13,63	19,87	14,14	12,86	8,66	14,55
ENM^3	**14,96**	**12,42**	**10,57**	**8,55**	**11,87**	**5,44**	**10,64**

Em vista desses resultados, é importante destacar a grande capacidade que esse tipo de máscara tem para mitigar a degradação causada pelo ruído nas características da voz, separando eficientemente a voz dominante do ruído de fundo. O sistema que usa como alvo de treinamento a máscara INM com o estimador DNN (ENM^3) fornece, em média melhoras significativas em relação ao sistema ERM^3 na redução das taxas de erro, que passam de $18,15\%$ para $14,96\%$ no ruído *babble*, de $13,63\%$ para $12,42\%$ no ruido *airport*, de $19,87\%$ para $10,57\%$ no ruído *restaurant*, de $14,14\%$ para $8,55\%$ no ruído *street*, de $12,86\%$ para $11,87\%$ no ruído *car*, de $8,66\%$ para $5,44\%$ no ruído *train*. Da mesma forma que na proposta baseada em estimador IMCRA as melhorias atingidas pelos mascaramento INM^3 sobre o banco de filtros são significativamente notáveis.

A ultima coluna da Tabela 6.3 (*avg.*) fornece uma média geral sobre todas os tipos de ruído. Os valores de *avg.* são calculados como as médias gerais dobre as diferentes condições de SNR e os diferentes tipos de ruído. A melhoria de desempenho do método proposto neste capítulo é corroborada, apresentando menor valor de *avg.* dentre todos os sistemas avaliados.

6.2.3 Comparação Geral

Nesta seção se faz uma comparação geral dos resultados em termos de qualidade PESQ e taxas de erro WER dos resultados obtidos para a técnica de mascaramento INM baseada em LBPs em condições reais apresentada nos capítulos 4, 5 e 6. Ou seja não serão apresentados resultados ideais (oráculo). As Tabelas 6.4 e 6.5 comparam os métodos ENM^1, WLBP e ENM^3, com o objetivo de verificar o desempenho atingido nos diferentes domínios onde foi aplicado.

Na Tabela 6.4 a coluna *noisy* representa os resultados de qualidade PESQ quando nenhum tipo de mascaramento é aplicado ao sinal. Pode-se ver como o uso das técnica ENM melhora significativamente a qualidade do sinal. No primeiro caso ENM^1 o mascaramento feito sobre o sinal da voz usando o espectro através da transformada de Fourier (capítulo 4) apresentando resultados ligeiramente melhores em todos os cenários, mostrando ser uma boa alternativa para realçar o sinal. Buscando melhorar a qualidade obtida no domínio de Fourier se propôs no capítulo 5 uma nova abordagem, dessa vez sobre o domínio *wavelet* (WLBP). Pode-se ver como os resultados desta nova abordagem são superiores e produziram uma melhora média de 0,19 para SNR de 0 dB, de 0,34 para SNR de 5dB, de 0,47 para SNR de 10 dB, e de 0,51 para SNR de 15 dB com relação ao sinal mascarado no domínio de Fourier. Isso mostra como através da decomposição em *wavelet* consegue-se uma melhor qualidade do sinal quando aplicada nossa técnica de mascaramento nas componentes de decomposição de alta frequência. Finalmente, seguindo a metodologia dos sistemas CASA e usando filtros que imitam o comportamento do sistema auditivo humano, se treinou através de algoritmos não supervisionados uma máscara ideal INM^3 baseada em DNN. O objetivo é separar melhor as informações que são predominantemente ruído das que são predominantemente voz. É claramente visto da Tabela 6.4 que o algoritmo proposto supera amplamente os outros algoritmos em todos os ambientes ruidosos. Além disso, nosso algoritmo pode garantir uma melhora média de 0,36 para 0dB de 0,35 para 5 dB e de 0,46 para 10 e 15 dB em relação aos resultados dados pela técnica WLBP do capítulo 5.

Tabela 6.4: Comparação da média do $PESQ$ sobre os diferentes tipos de ruído aplicando o mascaramento INM em diferentes domínios.

SNR	**noisy**	**ENM**^{1}	**WLBP**	**ENM**^{3}
0	1,06	1,11	1,30	1,66
5	1,19	1,22	1,56	1,91
10	1,45	1,47	1,94	2,40
15	1,87	1,89	2,40	2,86

Na Tabela6.5 apresenta-se os resultados das diferentes abordagens da máscara ENM, avaliadas a partir de um sistema de reconhecimento baseado em DNN, tomando como medida de desempenho a taxa de erro de palavra WER junto com seu respetivo intervalo de confiança (95%). Em primeiro lugar, observa-se que efetivamente as técnicas de separação de voz baseadas em máscaras de segregação reduzem os erros de descasamento entre condições de treinamento e teste. Pode-se ver como o mascaramento através de ENM^1 melhora significativamente o sistema RAV reduzindo as taxas de erro em média $15,52\%$ em relação ao sinal corrompido sem nenhuma técnica de realce. Continuando com os experimentos no capítulo 5, foram obtidos resultados significativamente melhores em comparação com a técnica aplicada no capítulo 4, superando-o em média $4,05\%$. Finalmente, para concluir esta seção, se faz uma comparação entre a técnica do capitulo 5 WLBP e a técnica proposta neste capítulo ENM^3, outra vez consegue-se ver o ganho altamente positivo do mascaramento proposto, onde os resultados revelam a superioridade do mascaramento baseado em DNN, demonstrando claramente a robustez inerente do DNN à variabilidade indesejada do ruído, reduzindo as taxas de erro do sistema de $36,09\%$ para $14,96\%$ no ruído *babble*, de $34,29\%$ para $12,42\%$ no ruido *airport*, de $40,24\%$ para $10,57\%$ no ruído *restaurant*, de $31,14\%$ para $8,55\%$ no ruído *street*, de $34,39\%$ para $11,87\%$ no ruído *car* e de $18,77\%$ para $5,44\%$ no ruído *train*.

Tabela 6.5: Comparação dos resultados de reconhecimento obtidos para o banco de dados AURORA-4. Tomando-se a média sobre as diferentes condições de SNR. Usando o mascaramento INM em diferentes domínios

system	babble	airport	restaurant	street	car	train
Noisy	59,24 ± 0,96	56,27 ± 0,97	58,88 ± 0,96	46,18 ± 0,98	56,27 ± 0,97	34,90 ± 0,93
ENM^1	**43,86** ± 0,97	**41,79** ± 0,97	**46,68** ± 0,98	**32,45** ± 0,92	**37,48** ± 0,95	**16,91** ± 0,73
WLBP	**36,09** ± 0,94	**34,29** ± 0,93	**40,24** ± 0,96	**31,14** ± 0,91	**34,29** ± 0,93	**18,77** ± 0,76
ENM^3	**14,96** ± 0,70	**12,42** ± 0,64	**10,57** ± 0,60	**8,55** ± 0,55	**11,87** ± 0,63	**5,44** ± 0,44

6.3 Conclusões

Este capítulo apresenta duas formas diferentes de melhorar a qualidade da voz e o desempenho dos sistemas RAV, baseados em DNN-HMM, na presença do ruído aditivo, através do mascaramento INM sobre os bancos de filtros gammatone que imitam o sistema auditivo humano.

Inicialmente, avaliou-se o sistema de mascaramento baseado na estimação da SNR no caso ideal em que se conhece a *priori* os sinais de voz e ruído. A máscara INM quando comparada com as técnicas tradicionais IBM e IRM obtidas nas mesmas condições, apresenta resultados com melhoras consideráveis no desempenho dos sistemas RAV reduzindo substancialmente as taxas de erro do sistema. Porém, note-se que isso acontece apenas no caso ideal (oráculo). No entanto, quando as máscaras utilizam o estimador IMCRA, os resultados obtidos a partir do sinal de voz original são superiores aos obtidos a partir do banco de filtros. Devido a isso, em seguida foi proposta uma nova estimação da máscara, mas dessa vez usando o algoritmo de aprendizado supervisionado das DNN. Que, conforme foi visto ao longo do capítulo, consegue aprender de forma eficiente um alvo desejado (máscara ideal INM). Esse algoritmo é treinado com base em um conjunto amplo de características que levam à obtenção de um modelo estimado. Esta abordagem permite o mapeamento direto entre os espaços dos atributos do sinal degradado e do sinal limpo, o que constitui um sistema de segregação da voz com elevada robustez em ambientes adversos.

O mais notável dos experimentos foi comprovar que independentemente do domínio onde a técnica de mascaramento seja utilizada, o mascaramento INM mostrou diferenças significativas nas taxas de reconhecimento e na melhoria da qualidade da voz em comparação com as técnicas tradicionais.

É importante destacar com os resultados finais obtidos com o mascaramento da nova proposta em que a SNR estimada é baseada em DNN e aplicada sobre os bancos de filtros tem maior poder de realçar o sinal eliminando o ruído presente no sinal devido às condições do ambiente acústico. Consequentemente, consegue-se resultados notáveis em termos de WER em sistemas de reconhecimento de voz baseados em DNN-HMM

7
Conclusões e Sugestões para Trabalhos Futuros

Buscando atingir a precisão que o sistema auditivo humano tem para se adaptar às condições adversas de forma instantânea, conseguindo diferenciar com precisão o ruído ou interferências das informações que são de interesse para o ouvinte, a comunidade de processamento de voz ao longo das ultimas décadas focou seus esforços na área de robustez da voz. Dentro desse enfoque foram propostas nesta tese técnicas que melhoraram a qualidade da voz e consequentemente o desempenho de sistemas de reconhecimento de voz contínua baseados em DNN-HMM. Concentram-se aqui nas áreas de compensação de atributos e realce de voz. Estas abordagens foram baseadas em métodos convencionais de processamento de imagens que foram adaptados a fim de oferecer, em conjunto com as técnicas tradicionais do processamento de voz, melhoras em relação à qualidade da voz, removendo as distorções que a degradam e que geram o descasamento entre as condições de treinamento e teste dos sistemas RAV.

Foi avaliado o desempenho das técnicas propostas a partir de diferentes experimentos de análise quantitativos e qualitativos das diferentes representações do sinal de voz obtidas através de métricas como ganho de SNR, PESQ e WER (taxa de erro de palavra). Foi demostrado em cada uma delas a efetividade dos métodos propostos nesta tese sobre as tarefas do banco de dados AURORA-4. Testes de escuta informais foram feitos no ambiente do laboratório.

De forma geral pode-se dizer que o foco deste trabalho é certamente inovador. As abordagens desenvolvidas, assim como suas propriedades e vantagens apresentadas ao longo desta tese demostram que a adaptação destas técnicas para o realce e reconhecimento de voz fornecem qualidades favoráveis diante das técnicas tradicionais.

Nas seguintes seções serão apresentadas as conclusões especificas obtidas a partir das técnicas de robustez propostas nesta tese, assim como as contribuições mais importantes obtidas ao longo deste trabalho. Finalmente, serão discutidas sugestões para trabalhos futuros.

7.1 Conclusões específicas

No capitulo 2, se fez uma revisão bibliográfica das diferentes estratégias propostas para melhorar a qualidade da voz e consequentemente melhorar a robustez dos sistemas RAV, quando agem em condições adversas. Foi analisado como o ruído aditivo afeta os atributos da voz usados nos sistemas RAV, fazendo com que os modelos acústicos treinados previamente modelem incorretamente a voz que se quer reconhecer. Assim, ocorre o conhecido descasamento entre as condições de treinamento e teste do sistemas.

No capítulo 3, foi apresentada a primeira contribuição desta tese, o método de compensação de características baseado em uma filtragem não linear das distribuições de probabilidade dos coeficientes cepstrais seguido de uma equalização de histogramas (MED-HMAP). Esta abordagem utilizou a conhecida técnica de equalização de histogramas com o objetivos de mapear os atributos do sinal corrompido para um sinal limpo de referência diminuindo as distorções causadas pelo ruído aditivo. No entanto, foi reportado na literatura que o processo de mapeamento para o novo domínio não reduz completamente as oscilações das funções de distribuição que caracterizam cada atributo. Devido a isso, neste capítulo se propôs aplicar uma filtragem de mediana sobre cada função de distribuição dos coeficientes cepstrais com o objetivo de suavizar as distorções causadas pelo ruído aditivo. Uma das razões mais importantes de usar a filtragem por mediana é sua baixa sensibilidade a elevadas intensidades de ruído local em comparação com o filtro da média apresentado em [151] o qual é mais sensível a alterações locais.

Com as distribuições de probabilidade filtradas através da mediana, se fez uma normalização de cada atributo através de um mapeamento de histogramas que a diferencia de técnicas como MVN, pois não só normaliza os primeiros momentos das funções de distribuição dos coeficientes cepstrais, como também os momentos de ordem superior. Através desta proposta, consegue-se modificar os vetores de atributos antes da normalização reduzindo assim o descasamento entre condições de treinamento e teste.

Esta técnica tem como principal atrativo ser de baixa carga computacional, ser independente do *back-end* do sistema RAV e o ponto mais relevante, o fato de não usar dados nem suposições sobre o tipo de ruído ou SNRs que esperam-se durante o reconhecimento. Através desta abordagem conseguiram-se diminuir as taxas de erro de palavras do sistema de reconhecimento passando de $29,52\%$ para $26,97\%$ com ruído *white*, de $31,44\%$ para $25,34\%$ com ruído *babble*, de $33,55\%$ para $31,11\%$ com ruído *f16* e de $46,56\%$ para $45,88\%$ com ruído *factory*, utilizando o banco de dados AURORA-4 e a base de ruído

noisex-92.

No capítulo 4 apresentou-se uma das principais contribuições desta tese, a técnica de mascaramento baseada em *Local Binary Pattern* (LBP). O método foi desenvolvido para melhorar a qualidade do sinal de voz, estimando uma máscara de vizinhança ideal (INM) através de uma matriz de unidades T-F que representam a SNR instantânea de cada quadro. A ideia principal desta nova proposta foi codificar em uma unidade T-F de análise, as informações presentes nas unidades T-F vizinhas usando as propriedades dos LBPs.

A contribuição do nosso trabalho centrou-se em adaptar a técnica LBP usada em processamento de imagens para matrizes bidimensionais ao processamento de voz para trabalhar sobre vetores de uma dimensão, buscando a melhor forma de obter os códigos LBP da matriz de mascaramento, que melhor discrimine as unidades T-F dominadas pelo ruído das unidades T-F dominadas pela voz.

Uma característica importante no projeto desta nova máscara foi o limiar de decisão da codificação binária, o qual foi desenvolvido levando em consideração o método de eliminação de ruído *Visushrink*, onde se representa o limar como uma estimativa de limiar universal, que é importante porque evita a influencia de ruídos muito pequenos no processo de codificação.

Os resultados conseguidos em termos de qualidade da voz assim como taxas de erro de palavra mostraram que com a nova proposta de mascaramento os resultados superaram significativamente os métodos tradicionais como IBM e IRM, obtendo-se um ganho na qualidade da voz maior a 0,4 em condições ideais e de 0,2 em condições reais, assim como uma melhora significativa de $13,6\%$ nas taxas de erro de palavra com relação ao sistema melhor sucedido da literatura o IRM.

Em conclusão, esta técnica mistura a simplicidade dos modelos de mascaramento nos que se baseiam os sistemas tradicionais IBM e IRM, com o uso dos descritores discriminativos LBP. Com isso obtém-se uma máscara com unidades T-F codificadas, que a diferencia dos métodos de mascaramento tradicionais, onde o valor de cada unidade T-F é diretamente o valor da SNR capturada nesse instante de tempo. O valor LBP codifica a relação entre o valor da unidade T-F de análise com os valores de SNR das unidades T-F vizinhas, de modo que a informação codificada será realizada em um nível mais alto.

No capítulo 5, dando continuidade a nossa proposta de mascaramento INM, se explorou o domínio das transformadas *wavelet* como alternativa para o mascaramento e separação das unidades dominantemente contaminadas por ruído. O esquema proposto foi denominado WLBP. O principal objetivo desta nova abordagem foi aproveitar a decomposição do sinal em diferentes sub-

bandas oferecida pela transformada *wavelet*, a fim de aplicar a máscara INM sobre os coeficientes wavelet das sub-bandas de alta frequência, substituindo desta forma o tradicional limiar usado para realçar o sinal da técnica *wavelet-denoising*.

Observações interessantes podem ser feitas a partir dos resultados apresentados neste capítulo. Em primeiro lugar, os desempenhos alcançados mostram claramente que a técnica WLBP proposta como *front-end* conseguiu diminuir a incompatibilidade entre as condições de treinamento e teste. Em todos os casos, quando o método proposto foi usado, o desempenho obtido em condições reais foi superior ao método tradicional *wavelet-denoising* para aprimoramento da voz.

O algoritmo WLBP proposto neste capítulo para a estimação da máscara INM mostrou ser eficiente na caracterização de unidades dominadas pelo ruído e dominadas pela voz, em todos os ambientes avaliados. Esse algoritmo mostrou ser uma boa alternativa aos métodos tradicionais de realce de fala como *wavelet-denoising* ou substração espectral, conseguindo melhorar a qualidade da voz em mais de 0,20 para SNRs entre 0 e 5 dB e superior a 0,5 para níveis de ruído acima de 10dB, com relação ao sistema proposto no capítulo 4. e atingindo taxas de erro de palavra com uma melhora relativa média de $17,45\%$.

Finalmente, observando os bons resultados tanto em qualidade como em taxas de erro de palavra da técnica INM proposta nos capítulos 4 e 5, foram propostas no capítulo 6, duas novas abordagens, dessa vez, fazendo uso da teoria dos sistemas CASA.

Na primeira abordagem se fez a decomposição do sinal usando o banco de filtros gammatone que simulam o comportamento auditivo humano e se fez a estimativa do ruído (ou da SNR) através do estimador IMCRA, obtendo uma matriz de unidades T-F sobre a qual aplicou-se a máscara INM. A partir dos resultados obtidos conseguiu-se observar que da mesma forma que nos capítulos anteriores a máscara INM fornece melhores resultados em comparação com as técnicas tradicionais. No entanto, em comparação à técnica INM usada em capítulos anteriores o uso do estimador IMCRA em conjunto com o banco de filtros não oferece melhoras, devido à correlação dos bancos de filtros Portanto, constatou-se que o método proposto precisaria ser melhorado para poder lidar com as distorções dos atributos cepstrais.

Visando melhorar os resultados obtidos com o estimador IMCRA foi proposta uma segunda abordagem de segregação de voz a partir de banco de filtros, porém utilizando como estimador do ruído (ou da SNR) uma aprendizagem supervisionada da INM como alvo de treinamento usando as *deep neural*

networks (DNNs). Foi apresentada a arquitetura das DNNs em conjunto com uma alternativa à solução do problema de segregação de voz, submetendo o modelo a um processo de pre-treinamento visando evitar problemas de sobre adaptação aos dados (*overfitting*).

A matriz de unidades T-F do alvo de treinamento foi construída idealmente a partir das unidades T-F dos bancos de filtros gammatone do sinal limpo e do ruído de fundo. Por outro lado, foram estudados o conjunto de atributos a serem inseridos na entrada da rede, o que é de fundamental importância para a aprendizagem supervisionada. Isso foi motivado a partir de resultados da literatura onde diferentes configurações de atributos foram utilizados para a etapa de treinamento da segregação de voz supervisionada, onde principalmente dois tipos de recursos, como AMS e recursos baseados em filtros gammatone, foram usados em várias metodologias.

Comparando os alvos de treinamento se demonstrou que o mascaramento INM baseado em DNN supera significativamente o sistema de mascaramento IRM, considerado como estado da arte em separação e aprimoramento de voz.

Finalmente, de acordo com os resultados obtidos com esta nova abordagem, confirmou-se a eficácia do uso de DNN para solucionar o problema de segregação da voz dominante do ruído dominante. Um ponto chave para ressaltar de esta nova abordagem foi os ótimos resultados tanto em qualidade quanto em taxas de erro de palavra, que superaram de forma notável os apresentados como estado da arte usando IRM. Foi demostrado claramente a robustez inerente do DNN à variabilidade indesejada do ruído, reduzindo as taxas de erro do sistema de reconhecimento de $18,15\%$ para $14,96\%$ no ruído *babble*, de $13,63\%$ para $12,42\%$ no ruido *airport*, de $19,87\%$ para $10,57\%$ no ruído *restaurant* de $14,14\%$ para $8,55\%$ no ruído *street* de $12,86\%$ para $11,87\%$ no ruído *car* e de $8,66\%$ para $5,44\%$ no ruído *train*, conseguindo uma melhora relativa media de $3,91\%$ com relação ao IRM.

7.2 Sugestões para trabalhos futuros

Finalmente, com base nas observações realizadas ao longo deste trabalho, esta tese deixa as portas abertas a futuras revisões e melhorias, a fim de aperfeiçoar ainda mais o desempenho do sistema.

– No capítulo 3, o mapeamento de histogramas dos coeficientes cepstrais introduz um descasamento devido ao fato de nem todos os coeficientes cepstrias serem igualmente discriminativos nem serem afetados da mesma forma pelo ambiente acústico. Sugere-se experimentar o mapeamento de

histogramas nos coeficientes onde o sinal de ruído afeta mais severamente, que geralmente é nos primeiros coeficientes cepstrais, incluindo o coeficientes de energia.

– No mesmo capítulo considera-se que os coeficientes são estatisticamente independentes. No entanto, existe uma pequena correlação entre eles. Sugere-se levar em conta esta correlação no momento do mapeamento de histogramas o que poderia incrementar o desempenho do mapeamento.

– Levando em consideração a correlação do item anterior, utilizar coeficientes MELFB ou PNFB para analisar o efeito do mapeamento de histogramas para diferentes técnicas de parametrização.

– A maior parte das abordagens propostas foram derivadas do modelo de mascaramento proposto baseado nos *(Local Binary Patterns)* LBPs e apresentado no capítulo 4. Como primeira recomendação dentro desse tema se propõe usar um número maior de unidades T-F vizinhas para codificar os dados, já que nesta tese utilizou-se apenas 2 vizinhos.

– Devido ao fato de trabalhar com espectrogramas e cochleogramas propõe-se usar o LBP original em duas dimensões que considera uma vizinhança 3x3 em torno de cada unidade T-F de análise. Assim o código LBP de cada unidade T-F seria calculado como um número de 8 bits, ou seja, tratando o espectrograma ou cochleograma como uma imagem.

– Nesta tese foram estudados os códigos LBP normais que são os que usam apenas as unidades T-F vizinhas sem levar em conta nenhum padrão de repetição entre transições. Considerando-se o item anterior é considerado sugere-se usar a técnica LBP uniforme, que é uma variante amplamente utilizada de LBPs que seleciona apenas padrões uniformes reduzindo o comprimento do vetor característico melhorando o sistema. Um LBP é "uniforme"quando contém um máximo de duas transições de '1 para 0' e/ou de '0 para 1', por exemplo: 000100 tem duas transições (é LBP-Uniforme), 010101001 tem sete transições (não é LBP-Uniforme).

– Testar diferentes algoritmos para fazer uma melhor estimação do espectro do ruído (SNR das unidades T-F), levando em consideração as variações temporais do mesmo.

– No capítulo 6 as DNN, foram treinadas usando uma busca exaustiva em relação à arquitetura e configuração de treinamento. Com o objetivo de melhorar ainda mais seu desempenho se propõe usar diferentes configurações da DNN, diminuindo o número de camadas e testando o treinamento sem usar as RBM, as quais requerem uma alta carga computacional. Adicionalmente, seria de interesse empregar o algoritmo

de treinamento *dropout* que se fundamenta na eliminação aleatória de neurônios durante o processo de aprendizagem, para evitar a sobre adaptação aos dados.

– Devido ao fato da DNN não utilizar informação relativa à correlação dos coeficientes que representam o sinal se propõe usar diferentes configurações de atributos de treinamento usando os coeficientes baseados na quefrencia, ao invés de usar os coeficientes cepstrais. Isto é, usar coeficientes MELFB, PNFB, entre outros. Além disso, pode ser um tópico de pesquisa interessante o uso de atributos adicionais, como o *pitch*, para melhorar ainda mais a capacidade de generalização das DNNs,

– Finalmente, se propõe usar treinamento multicondição no sistema de reconhecimento, já que esse tipo de treinamento é um elemento essencial para ser integrado em um sistema RAV sempre que possível, a fim de proporcionar um bom ponto de partida em termos de robustez contra o ruído.

Referências bibliográficas

[1] HUANG, X.; ACERO, A.; HON, H.-W. ; REDDY, R.. **Spoken language processing: A guide to theory, algorithm, and system development**, volumen 95. Prentice hall PTR Upper Saddle River, 2001.

[2] PERICÁS, H.. **Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidosos**. Universitat Politècnica de Catalunya, 1993.

[3] MOORE, R. K.. **Spoken language processing: Piecing together the puzzle**. Speech communication, 49(5):418–435, 2007.

[4] LIPPMANN, R. P.. **Speech recognition by machines and humans**. Speech communication, 22(1):1–15, 1997.

[5] MORENO, P.. **Speech recognition in noisy environments**. PhD thesis, Carnegie Mellon University, 1996.

[6] HINTON, G.; DENG, L.; YU, D.; DAHL, G. E.; MOHAMED, A.-R.; JAITLY, N.; SENIOR, A.; VANHOUCKE, V.; NGUYEN, P.; SAINATH, T. N. ; OTHERS. **Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups**. IEEE Signal Processing Magazine, 29(6):82–97, 2012.

[7] XIONG, W.; L, W.; F, A.; DROPPO, J.; HUANG, X. ; STOLCKE, A.. **The microsoft 2017 conversational speech recognition system**. Technical Report MSR-TR-2017-39, 2017.

[8] WAMBACQ, P.; STOUTEN, V.. **Robust Automatic Speech Recognition in the Time-Varying environments**. PhD thesis, Kathaolieke Universiteit Leuven, 2006.

[9] RAJ, B.; SELTZER, M. L. ; STERN, R. M.. **Reconstruction of missing features for robust speech recognition**. Speech communication, 43(4):275–296, 2004.

[10] HARTMANN, W.; FOSLER-LUSSIER, E.. **Investigations into the incorporation of the ideal binary mask in asr**. In: INTERNATIONAL

CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-11, p. 4804–4807. IEEE, 2011.

[11] BENESTY, J.; SONDHI, M. M. ; HUANG, Y.. **Springer handbook of speech processing**. Springer Science & Business Media, 2007.

[12] PEINADO, A.; SEGURA, J.. **Speech recognition over digital channels: Robustness and Standards**. John Wiley & Sons, 2006.

[13] DE LA TORRE VEGA, A.; HERREROS, A. M. P. ; AYUSO, A. J. R.. **Reconocimiento automático de voz en condiciones de ruido**. Master's thesis, Monografías del Departamento de Electrónica, No 47. Universidad de Granada, 2001.

[14] ALCAIM, A.; OLIVEIRA, C.. **Fundamentos do Processamento de Sinais de Voz e Imagem**. Interciência e PUC Rio, 2012.

[15] NOLAN, F.. **The phonetic bases of speaker recognition**. PhD thesis, University of Cambridge, 1980.

[16] SEPÚLVEDA SEPÚLVEDA, F.; OTHERS. **Extracción de parámetros de señales de voz usando técnicas de análisis en tiempo-frecuencia**. PhD thesis, Universidad Nacional de Colombia-Sede Manizales, 2004.

[17] MITRA, S.; KUO, Y.. **Digital signal processing: a computer-based approach**, volumen 2. McGraw-Hill New York, 2006.

[18] QUATIERI, T. F.. **Discrete-time speech signal processing: principles and practice**. Pearson Education India, 2006.

[19] OPPENHEIM, A.; SCHAFER, R.; BUCK, J. ; OTHERS. **Discrete-time signal processing**, volumen 2. Prentice hall Englewood Cliffs, 1999.

[20] PERDIGÃO, F. M. D. S.. **Modelos do sistema auditivo periférico no reconhecimento automático da fala**. PhD thesis, Fac. de Ciências e Tecnologia de Coimbra, 1998.

[21] FLETCHER, H.. **Auditory patterns**. Reviews of modern physics, 12(1):47, 1940.

[22] ZWICKER, E.. **Subdivision of the audible frequency range into critical bands (frequenzgruppen)**. The Journal of the Acoustical Society of America, 33(2):248–248, 1961.

[23] FREDES, J.; NOVOA, J.; KING, S.; STERN, R. M. ; YOMA, N. B.. **Locally normalized filter banks applied to deep neural-network-based robust speech recognition**. IEEE Signal Processing Letters, 24(4):377–381, 2017.

[24] POBLETE, V.; ESPIC, F.; KING, S.; STERN, R. M.; HUENUPÁN, F.; FREDES, J. ; YOMA, N. B.. **A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification**. Computer Speech & Language, 31(1):1–27, 2015.

[25] FURUI, S.. **Cepstral analysis technique for automatic speaker verification**. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(2):254–272, 1981.

[26] BREGMAN, A. S.; OTHERS. **Auditory scene analysis**, volumen 10. Cambridge, ma: mit press, 1990.

[27] HERMANSKY, H.. **Perceptual linear predictive (plp) analysis of speech**. the Journal of the Acoustical Society of America, 87(4):1738–1752, 1990.

[28] LIM, J.; OPPENHEIM, A.. **All-pole modeling of degraded speech**. IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(3):197–210, 1978.

[29] KIM, D.-S.; LEE, S.-Y. ; KIL, R. M.. **Auditory processing of speech signals for robust speech recognition in real-world noisy environments**. IEEE Transactions on speech and audio processing, 7(1):55–69, 1999.

[30] HICKOK, G.; POEPPEL, D.. **The cortical organization of speech processing**. Nature Reviews Neuroscience, 8(5):393, 2007.

[31] ZWICKER, E.; FLOTTORP, G. ; STEVENS, S. S.. **Critical band width in loudness summation**. The Journal of the Acoustical Society of America, 29(5):548–557, 1957.

[32] WANG, D.; BROWN, G. J.. **Computational auditory scene analysis: Principles, algorithms, and applications**. 2006.

[33] HOHMANN, V.. **Frequency analysis and synthesis using a gammatone filterbank**. Acta Acustica united with Acustica, 88(3):433–442, 2002.

[34] EZZAT, T.; BOUVRIE, J. ; POGGIO, T.. **Spectro-temporal analysis of speech using 2-d gabor filters**. In: EIGHTH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 2007.

[35] SENEFF, S.. **A computational model for the peripheral auditory system: Application of speech recognition research**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-86, volumen 11, p. 1983–1986. IEEE, 1986.

[36] ATAL, B.; REMDE, J.. **A new model of lpc excitation for producing natural-sounding speech at low bit rates**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-82., volumen 7, p. 614–617. IEEE, 1982.

[37] ZHEN, B.; WU, X.; LIU, Z. ; CHI, H.. **On the use of bandpass liftering in speaker recognition**. In: SIXTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, 2000.

[38] TEVAH, R. T.. **implementação de um sistema de reconhecimento de fala contínua com amplo vocabulário para o português brasileiro**. Master's thesis, Universidade Federal do Rio de Janeiro, 2006.

[39] DAVIS, S. B.; MERMELSTEIN, P.. **Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences**. In: READINGS IN SPEECH RECOGNITION, p. 65–74. Elsevier, 1990.

[40] KIM, C.; STERN, R.. **Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction**. INTERSPEECH-2009, 1:28–31, 2009.

[41] PATTERSON, R.; ROBINSON, K.; HOLDSWORTH, J.; MCKEOWN, D.; ZHANG, C. ; ALLERHAND, M.. **Complex sounds and auditory images**. Auditory physiology and perception, 83:429–446, 1992.

[42] FUENTES, L. J. R.; BARAÑANO, M. I. T.. **Estudio y modelización acústica del habla espontánea en diálogos hombre-máquina y entre personas**. PhD thesis, Facultad de Ciencia y Tecnología Universidad del País Vasco, 2004.

[43] LEE, K.. **Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition**. Transactions on Acoustics Speech and Signal Processing, 38(4):599–609, 1990.

[44] BAKER, J.. **The dragon system–an overview**. Transactions on Acoustics Speech and Signal Processing, 23(1):24–29, 1975.

[45] JELINEK, F.; BAHL, L. ; MERCER, R.. **Design of a linguistic statistical decoder for the recognition of continuous speech**. Transactions on Information Theory, 21(3):250–256, 1975.

[46] RABINER, L.. **A tutorial on hidden markov models and selected applications in speech recognition**. Proceedings of the IEEE, 77(2):257–286, 1989.

[47] GALES, M. J. F.. **Model-based techniques for noise robust speech recognition**. PhD thesis, University of Cambridge Cambridge, 1995.

[48] HUANG, X.; ACERO, A.; HON, H. ; OTHERS. **Spoken language processing**, volumen 15. Prentice Hall PTR New Jersey, 2001.

[49] SANTOS, S. C. B. D.. **Reconhecimento de Voz Contínua para o Português Utilizando Modelos de Markov Escondidos**. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro, 1997.

[50] FORNEY JR, G.. **The viterbi algorithm**. Proceedings of the IEEE, 61(3):268–278, 1973.

[51] BAHL, L. R.; JELINEK, F. ; MERCER, R. L.. **A maximum likelihood approach to continuous speech recognition**. In: READINGS IN SPEECH RECOGNITION, p. 308–319. Elsevier, 1990.

[52] FURUI, S.; SONDHI, M. M.. **Advances in speech signal processing**. Marcel Dekker, Inc. New York, NY, USA, 1991.

[53] NEY, H.; MERGEL, D.; NOLL, A. ; PAESELER, A.. **Data driven search organization for continuous speech recognition**. IEEE Transactions on signal processing, 40(2):272–281, 1992.

[54] BAHL, L. R.; DE GENNARO, S. V.; GOPALAKRISHNAN, P. ; MERCER, R. L.. **A fast approximate acoustic match for large vocabulary speech recognition**. IEEE Transactions on Speech and Audio Processing, 1(1):59–67, 1993.

[55] YOUNG, S.. **A review of large-vocabulary continuous-speech**. IEEE signal processing magazine, 13(5):45, 1996.

[56] GALES, M.; YOUNG, S.. **The application of hidden markov models in speech recognition**. Foundations and trends in signal processing, 1(3):195–304, 2008.

[57] SAON, G.; CHIEN, J.-T.. **Large-vocabulary continuous speech recognition systems: A look at some recent advances**. IEEE Signal Processing Magazine, 29(6):18–33, 2012.

[58] BOURLARD, H. A.; MORGAN, N.. **Connectionist speech recognition: a hybrid approach**, volumen 247. Springer Science & Business Media, 2012.

[59] GRAVES, A.; JAITLY, N. ; MOHAMED, A.-R.. **Hybrid speech recognition with deep bidirectional lstm**. In: WORKSHOP ON AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING ASRU-13, p. 273–278. IEEE, 2013.

[60] SARMA, M.; SARMA, K. K.. **Acoustic modeling of speech signal using artificial neural network: A review of techniques and current trends**. In: INTELLIGENT APPLICATIONS FOR HETEROGENEOUS SYSTEM MODELING AND DESIGN, p. 282–299. IGI Global, 2015.

[61] LI, J.; DENG, L.; HAEB-UMBACH, R. ; GONG, Y.. **Robust Automatic Speech Recognition: A Bridge to Practical Applications**. Academic Press, 2015.

[62] MAAS, A. L.; QI, P.; XIE, Z.; HANNUN, A. Y.; LENGERICH, C. T.; JURAFSKY, D. ; NG, A. Y.. **Building dnn acoustic models for large vocabulary speech recognition**. Computer Speech & Language, 41:195–213, 2017.

[63] PIERACCINI, R.; RABINER, L.. **The voice in the machine: building computers that understand speech**. MIT Press Cambridge, MA, USA, 2012.

[64] SERCU, T.; PUHRSCH, C.; KINGSBURY, B. ; LECUN, Y.. **Very deep multilingual convolutional neural networks for lvcsr**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-16, p. 4955–4959. IEEE, 2016.

[65] ZHANG, L.; WU, X. ; LUO, D.. **Human activity recognition with hmm-dnn model**. In: 14TH INTERNATIONAL CONFERENCE ON COGNITIVE INFORMATICS & COGNITIVE COMPUTING (ICCI* CC), p. 192–197. IEEE, 2015.

[66] SAGAYAMA, S.; AIKAWA, K.. **Issues relating to the future of asr for telecommunications applications**. In: ROBUST SPEECH RECOGNITION FOR UNKNOWN COMMUNICATION CHANNELS, 1997.

[67] MEEKER, M.. **Kp internet trends 2017 code conference**. p. 48, 2017.

[68] BELLEGARDA, J.. **Statistical techniques for robust asr: review and perspectives**. In: FIFTH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY, p. KN 33–36, 1997.

[69] CHEN, J.; BENESTY, J.; HUANG, Y. A. ; DIETHORN, E. J.. **Fundamentals of noise reduction**. In: SPRINGER HANDBOOK OF SPEECH PROCESSING, p. 843–872. Springer, 2008.

[70] ACERO, A.. **Acoustical and environmental robustness in automatic speech recognition**, volumen 201. Springer Science & Business Media, 2012.

[71] LOIZOU, P. C.. **Speech enhancement: theory and practice**. CRC press, Boca Raton, London, 2013.

[72] SHANNON, B. J.; PALIWAL, K. K.. **Effect of speech and noise cross correlation on amfcc speech recognition features**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING, ICASSP-07, volumen 4, p. IV–1033. IEEE, 2007.

[73] TAN, Z.-H.; LINDBERG, B.. **Automatic speech recognition on mobile devices and over communication networks**. Springer Science & Business Media, 2008.

[74] GALLARDO ANTOLÍN, A.. **Reconocimiento de habla robusto frente a condiciones de ruido aditivo y convolutivo**. PhD thesis, Universidad Politécnica de Madrid, 2002.

[75] HERMANSKY, H.; MORGAN, N.. **Rasta processing of speech**. IEEE transactions on speech and audio processing, 2(4):578–589, 1994.

[76] DENG, L.; DROPPO, J. ; ACERO, A.. **Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features**. IEEE Transactions on Speech and Audio Processing, 12(3):218–233, 2004.

[77] COHEN, I.; BERDUGO, B.. **Noise estimation by minima controlled recursive averaging for robust speech enhancement**. IEEE signal processing letters, 9(1):12–15, 2002.

[78] HANSEN, J. H.. **Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition**. Speech communication, 20(1-2):151–173, 1996.

[79] MOORE, B. C.. **Hearing (Handbook of Perception and Cognition)**. Academic Press, New York, 1995.

[80] JOHN, M.; GALES, F. ; OTHERS. **Model-Based Techniques For Noise Robust Speech Recognition**. PhD thesis, Europe PubMed Central, 1995.

[81] OPENSHAW, J. P.; MASAN, J.. **On the limitations of cepstral features in noise**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING, ICASSP-94, volumen 2, p. II–49. IEEE, 1994.

[82] LIU, F.. **Environmental adaptation for robust speech recognition**. PhD thesis, ECE Department CMU, 1994.

[83] EPHRAIM, Y.. **Statistical-model-based speech enhancement systems**. Proceedings of the IEEE, 80(10):1526–1555, 1992.

[84] WARREN, R. M.; OTHERS. **Perceptual restoration of missing speech sounds**. Science, 167(3917):392–393, 1970.

[85] BHATT, K.; VINITHA, C. ; GUPTA, R.. **Secure speech enhancement using lpc based fem in wiener filter**. In: DATA ENGINEERING AND INTELLIGENT COMPUTING, p. 657–665. Springer, 2018.

[86] UPADHYAY, N.; KARMAKAR, A.. **An improved multi-band spectral subtraction algorithm for enhancing speech in various noise environments**. Procedia Engineering, 64:312–321, 2013.

[87] ZHANG, Y.; ZHAO, Y.. **Real and imaginary modulation spectral subtraction for speech enhancement**. Speech Communication, 55(4):509–522, 2013.

[88] BOLL, S.. **Suppression of acoustic noise in speech using spectral subtraction**. Transactions on Acoustics Speech and Signal Processing, 27(2):113–120, 1979.

[89] VASEGHI, S.. **Advanced digital signal processing and noise reduction**. John Wiley and Sons, 2008.

[90] XIA, Y.; WANG, P.. **Speech enhancement in presence of colored noise using an improved least square estimation**. In: PROCEEDINGS OF THE 3RD INTERNATIONAL CONFERENCE ON MULTIMEDIA TECHNOLOGY, p. 779–786, 2013.

[91] BECKER, R.; CORSETTI, G.; SILVEIRA, J. G.; BALBINOT, R. ; CASTELLO, F.. **A silence detection and suppression technique design for voice over ip systems**. In: PACIFIC RIM CONFERENCE ON COMMUNICATIONS COMPUTERS AND SIGNAL PROCESSING. PACRIM-05, p. 173–176. IEEE, 2005.

[92] CHAUDHARI, A.; DHONDE, S.. **A review on speech enhancement techniques**. In: INTERNATIONAL CONFERENCE ON PERVASIVE COMPUTING ICPC-15, p. 1–3. IEEE, 2015.

[93] ARAKAWA, T.; TSUJIKAWA, M. ; ISOTANI, R.. **Model-basedwiener filter for noise robust speech recognition**. In: 2006 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING PROCEEDINGS, volumen 1, p. I–I. IEEE, 2006.

[94] DOCLO, S.; MOONEN, M.. **On the output snr of the speech-distortion weighted multichannel wiener filter**. IEEE Signal Processing Letters, 12(12):809–811, 2005.

[95] BENESTY, J.; CHEN, J.; HUANG, Y. A. ; DOCLO, S.. **Study of the wiener filter for noise reduction**. In: SPEECH ENHANCEMENT, p. 9–41. Springer, 2005.

[96] VIRAG, N.. **Single channel speech enhancement based on masking properties of the human auditory system**. IEEE Transactions on speech and audio processing, 7(2):126–137, 1999.

[97] EPHRAIM, Y.; MALAH, D.. **Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator**. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(6):1109–1121, 1984.

[98] THÜNE, P.; ENZNER, G.. **Maximum-likelihood approach with bayesian refinement for multichannel-wiener postfiltering**. IEEE Transactions on Signal Processing, 65(13):3399–3413, 2017.

[99] MARTIN, R.. **Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-02, volumen 1, p. I–253. IEEE, 2002.

[100] LU, Y.; LOIZOU, P. C.. **Estimators of the magnitude-squared spectrum and methods for incorporating snr uncertainty**. IEEE transactions on audio, speech, and language processing, 19(5):1123–1137, 2011.

[101] BENESTY, J.; HUANG, Y.. **Adaptive signal processing: applications to real-world problems**. Springer Science & Business Media, 2013.

[102] DONOHO, D.; JOHNSTONE, I.. **Threshold selection for wavelet shrinkage of noisy data**. In: 16TH ANNUAL INTERNATIONAL CONFERENCE OF THE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY. ENGINEERING ADVANCES: NEW OPPORTUNITIES FOR BIOMEDICAL ENGINEERS., p. A24–A25, 1994.

[103] SEOK, J. W.; BAE, K. S.. **Speech enhancement with reduction of noise components in the wavelet domain**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING, ICASSP-97., volumen 2, p. 1323–1326. IEEE, 1997.

[104] HU, Y.; LOIZOU, P. C.. **Speech enhancement based on wavelet thresholding the multitaper spectrum**. IEEE transactions on Speech and Audio processing, 12(1):59–67, 2004.

[105] DONOHO, D. L.. **De-noising by soft-thresholding**. IEEE transactions on information theory, 41(3):613–627, 1995.

[106] SALIMPOUR, Y.; ABOLHASSANI, M.. **Auditory wavelet transform based on auditory wavelet families**. In: ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY, EMBS'06., p. 1731–1734. IEEE, 2006.

[107] BHOWMICK, A.; CHANDRA, M.. **Speech enhancement using voiced speech probability based wavelet decomposition**. Computers & Electrical Engineering, 2017.

[108] SINGH, S.; TRIPATHY, M. ; ANAND, R.. **A wavelet packet based approach for speech enhancement using modulation channel selection**. Wireless Personal Communications, 95(4):4441–4456, 2017.

[109] HUANG, N. E.; SHEN, Z.; LONG, S. R.; WU, M. C.; SHIH, H. H.; ZHENG, Q.; YEN, N.-C.; TUNG, C. C. ; LIU, H. H.. **The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis**. In: PROCEEDINGS OF THE ROYAL SOCIETY OF LONDON A: MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES, volumen 454, p. 903–995. The Royal Society, 1998.

[110] HASAN, T.; HASAN, M. K.. **Suppression of residual noise from speech signals using empirical mode decomposition**. IEEE Signal Processing Letters, 16(1):2–5, 2009.

[111] CHATLANI, N.; SORAGHAN, J. J.. **Emd-based filtering (emdf) of low-frequency noise for speech enhancement**. IEEE Transactions on Audio, Speech, and Language Processing, 20(4):1158–1166, 2012.

[112] HAO, H.; WANG, H. ; REHMAN, N.. **A joint framework for multivariate signal denoising using multivariate empirical mode decomposition**. Signal Processing, 135:263–273, 2017.

[113] ROMAN, N.; WANG, D. ; BROWN, G. J.. **Speech segregation based on sound localization**. The Journal of the Acoustical Society of America, 114(4):2236–2252, 2003.

[114] KIM, G.; LU, Y.; HU, Y. ; LOIZOU, P. C.. **An algorithm that improves speech intelligibility in noise for normal-hearing listeners**. The Journal of the Acoustical Society of America, 126(3):1486–1494, 2009.

[115] RAJ, B.; STERN, R. M.. **Missing-feature approaches in speech recognition**. IEEE Signal Processing Magazine, 22(5):101–116, 2005.

[116] KIM, W.; STERN, R. M.. **Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise**. Speech Communication, 53(1):1–11, 2011.

[117] HOLMES, J.; SEDGWICK, N.. **Noise compensation for speech recognition using probabilistic models**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING, ICASSP-86, volumen 11, p. 741–744. IEEE, 1986.

[118] COOKE, M.; GREEN, P.; JOSIFOVSKI, L. ; VIZINHO, A.. **Robust automatic speech recognition with missing and unreliable acoustic data**. Speech communication, 34(3):267–285, 2001.

[119] COOKE, M.; GREEN, P.; JOSIFOVSKI, L. ; VIZINHO, A.. **Robust asr with unreliable data and minimal assumptions**. In: PROC. OF WORKSHOP ON ROBUST METHODS FOR SPEECH RECOGNITION IN ADVERSE CONDITIONS, volumen 99, 1999.

[120] WANG, Y.; HAN, K. ; WANG, D.. **Exploring monaural features for classification-based speech segregation**. IEEE Transactions on Audio, Speech, and Language Processing, 21(2):270–279, 2013.

[121] WANG, D.. **On ideal binary mask as the computational goal of auditory scene analysis**. In: SPEECH SEPARATION BY HUMANS AND MACHINES, p. 181–197. Springer, 2005.

[122] JIANG, Y.; ZHOU, H. ; FENG, Z.. **Performance analysis of ideal binary masks in speech enhancement**. In: 4TH INTERNATIONAL CONGRESS ON IMAGE AND SIGNAL PROCESSING (CISP), volumen 5, p. 2422–2425. IEEE, 2011.

[123] SRINIVASAN, S.; ROMAN, N. ; WANG, D.. **Binary and ratio time-frequency masks for robust speech recognition**. Speech Communication, 48(11):1486–1501, 2006.

[124] WANG, Y.; NARAYANAN, A. ; WANG, D.. **On training targets for supervised speech separation**. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22(12):1849–1858, 2014.

[125] JIN, Z.; WANG, D.. **A supervised learning approach to monaural segregation of reverberant speech**. IEEE Transactions on Audio, Speech, and Language Processing, 17(4):625–638, 2009.

[126] HAN, K.; WANG, D.. **A classification based approach to speech segregation**. The Journal of the Acoustical Society of America, 132(5):3475–3483, 2012.

[127] NARAYANAN, A.; WANG, D.. **Ideal ratio mask estimation using deep neural networks for robust speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), p. 7092–7096. IEEE, 2013.

[128] LI, B.; SIM, K. C.. **An ideal hidden-activation mask for deep neural networks based noise-robust speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP-14), p. 200–204. IEEE, 2014.

[129] MELLOR, B.; VARGA, A.. **Noise masking in a transform domain**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, ICASSP-93, volumen 2, p. 87–90. IEEE, 1993.

[130] CERISARA, C.; DEMANGE, S. ; HATON, J.-P.. **On noise masking for automatic missing data speech recognition: A survey and discussion**. Computer Speech & Language, 21(3):443–457, 2007.

[131] NARAYANAN, A.; WANG, D.. **Investigation of speech separation as a front-end for noise robust speech recognition**. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22(4):826–835, 2014.

[132] **RASTA extensions: Robustness to additive and convolutional noise**, 1992.

[133] KANEDERA, N.; HERMANSKY, H. ; ARAI, T.. **On properties of modulation spectrum for robust automatic speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, volumen 2, p. 613–616. IEEE, 1998.

[134] HERMANSKY, H.; MORGAN, N.; BAYYA, A. ; KOHN, P.. **The challenge of inverse-e: the rasta-plp method**. In: CONFERENCE RECORD OF THE TWENTY-FIFTH ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS AND COMPUTERS, 91, p. 800–804. IEEE, 1991.

[135] MOKBEL, C.; JOUVET, D. ; MONNÉ, J.. **Deconvolution of telephone line effects for speech recognition**. Speech Communication, 19(3):185–196, 1996.

[136] STERN, R. M.; LIU, F.-H.; OHSHIMA, Y.; SULLIVAN, T. M. ; ACERO, A.. **Multiple approaches to robust speech recognition**. In: PROCEEDINGS OF THE WORKSHOP ON SPEECH AND NATURAL LANGUAGE, p. 274–279, 1992.

[137] MORENO, P. J.; STERN, R. M.. **Sources of degradation of speech recognition in the telephone network**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, ICASSP-94, volumen 1, p. I–109. IEEE, 1994.

[138] XIAO, X.; CHNG, E. S. ; LI, H.. **Temporal structure normalization of speech feature for robust speech recognition**. IEEE signal processing letters, 14(7):500–503, 2007.

[139] LIU, F.-H.; STERN, R. M.; HUANG, X. ; ACERO, A.. **Efficient cepstral normalization for robust speech recognition**. In: PROCEEDINGS OF THE WORKSHOP ON HUMAN LANGUAGE TECHNOLOGY, p. 69–74, 1993.

[140] VIIKKI, O.; LAURILA, K.. **Noise robust hmm-based speech recognition using segmental cepstral feature vector normalization**. In: ROBUST SPEECH RECOGNITION FOR UNKNOWN COMMUNICATION CHANNELS, 1997.

[141] VIIKKI, O.; LAURILA, K.. **Cepstral domain segmental feature vector normalization for noise robust speech recognition**. Speech Communication, 25(1):133–147, 1998.

[142] PUJOL, P.; MACHO, D. ; NADEU, C.. **On real-time mean-and-variance normalization of speech recognition features**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP-06, volumen 1, p. I–I. IEEE, 2006.

[143] GROZDIC, D.; JOVICIC, S.; SUMARAC PAVLOVIC, D.; GALIC, J. ; MARKOVIC, B.. **Comparison of cepstral normalization techniques in whispered speech recognition**. Advances in Electrical and Computer Engineering, 17(1):21–26, 2017.

[144] SEGURA, J.; BENITEZ, C.; DE LA TORRE, A. ; RUBIO, A.. **Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust asr**. In: PROC. ICSLP 02, p. 225–228, 2002.

[145] DE LA TORRE, A.; PEINADO, A.; SEGURA, J.; PÉREZ-CÓRDOBA, J.; BENÍTEZ, M. ; RUBIO, A.. **Histogram equalization of speech representation for robust speech recognition**. Transactions on Speech and Audio Processing, 13(3):355–366, 2005.

[146] SAON, G.; DHARANIPRAGADA, S. ; POVEY, D.. **Feature space gaussianization**. In: NTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, ICASSP-04, volumen 1, p. I–329. IEEE, 2004.

[147] SEGURA, J. C.; BENÍTEZ, C.; DE LA TORRE, Á.; RUBIO, A. J. ; RAMÍREZ, J.. **Cepstral domain segmental nonlinear feature transformations for robust speech recognition**. IEEE Signal Processing Letters, 11(5):517–520, 2004.

[148] HILGER, F.; NEY, H.. **Quantile based histogram equalization for noise robust large vocabulary speech recognition**. IEEE Transactions on Audio, Speech, and Language Processing, 14(3):845–854, 2006.

[149] GARCÍA, L.; ORTÚZAR, C. B.; DE LA TORRE, A. ; SEGURA, J. C.. **Class-based parametric approximation to histogram equalization for asr**. IEEE Signal Processing Letters, 19(7):415–418, 2012.

[150] SEGURA, J. C.; BENÍTEZ, C.; DE LA TORRE, Á.; RUBIO, A. J. ; RAMÍREZ, J.. **Cepstral domain segmental nonlinear feature transformations for robust speech recognition**. IEEE Signal Processing Letters, 11(5):517–520, 2004.

[151] WANG, S.-S.; TSAO, Y. ; HUNG, J.-W.. **Filtering on the temporal probability sequence in histogram equalization for robust speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING ICASSP-13, p. 7112–7116. IEEE, 2013.

[152] STERN, R. M.; RAJ, B. ; MORENO, P. J.. **Compensation for environmental degradation in automatic speech recognition**. In: ROBUST SPEECH RECOGNITION FOR UNKNOWN COMMUNICATION CHANNELS, 1997.

[153] UN, C. K.; KIM, N. S. ; OTHERS. **Speech recognition in noisy environments using first-order vector taylor series**. Speech Communication, 24(1):39–49, 1998.

[154] MORENO, P. J.; RAJ, B. ; STERN, R. M.. **A vector taylor series approach for environment-independent speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, ICASSP-96, volumen 2, p. 733–736. IEEE, 1996.

[155] OBUCHI, Y.; STERN, R. M.. **Normalization of time-derivative parameters using histogram equalization.** In: EIGHTH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY, 2003.

[156] LEE, C.-H.; HUO, Q.. **On adaptive decision rules and decision parameter adaptation for automatic speech recognition**. Proceedings of the IEEE, 88(8):1241–1269, 2000.

[157] GALES, M.. **Nice model-based compensation schemes for robust speech recognition**. In: ROBUST SPEECH RECOGNITION FOR UNKNOWN COMMUNICATION CHANNELS, 1997.

[158] LIPPMANN, R.; MARTIN, E. ; PAUL, D.. **Multi-style training for robust isolated-word speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING, ICASSP'87., volumen 12, p. 705–708. IEEE, 1987.

[159] KALINLI, O.; SELTZER, M. L.; DROPPO, J. ; ACERO, A.. **Noise adaptive training for robust automatic speech recognition**. IEEE Transactions on Audio, Speech, and Language Processing, 18(8):1889–1901, 2010.

[160] ACERO, A.; DENG, L.; KRISTJANSSON, T. T. ; ZHANG, J.. **Hmm adaptation using vector taylor series for noisy speech recognition.** In: SIXTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, p. 869–872, 2000.

[161] GAUVAIN, J.-L.; LEE, C.-H.. **Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains**. IEEE transactions on speech and audio processing, 2(2):291–298, 1994.

[162] LEGGETTER, C. J.; WOODLAND, P. C.. **Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models**. Computer Speech & Language, 9(2):171–185, 1995.

[163] WOODLAND, P.; GALES, M.; PYE, D. ; VALTCHEV, V.. **The htk large vocabulary recognition system for the 1995 arpa h3 task**. In: PROC. ARPA SPEECH RECOGNITION WORKSHOP, volumen 99, p. 104. Haniman New York, USA, 1996.

[164] DIGALAKIS, V. V.; RTISCHEV, D. ; NEUMEYER, L. G.. **Speaker adaptation using constrained estimation of gaussian mixtures**. IEEE Transactions on speech and Audio Processing, 3(5):357–366, 1995.

[165] GALES, M. J.. **Maximum likelihood linear transformations for hmm-based speech recognition**. Computer speech & language, 12(2):75–98, 1998.

[166] GALES, M. J. F.; YOUNG, S. J.. **A fast and flexible implementation of parallel model combination**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING,ICASSP-95, volumen 1, p. 133–136. IEEE, 1995.

[167] HUNG, J.-W.; SHEN, J.-L. ; LEE, L.-S.. **New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (pmc) techniques**. IEEE Transactions on Speech and Audio Processing, 9(8):842–855, 2001.

[168] TUFEKCI, Z.; GOWDY, J. N.; GURBUZ, S. ; PATTERSON, E.. **Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition**. Speech communication, 48(10):1294–1307, 2006.

[169] HIZLISOY, S.; TUFEKCI, Z.. **Noise robust speech recognition using parallel model compensation and voice activity detection methods**. In: 5TH INTERNATIONAL CONFERENCE ON ELECTRONIC DEVICES SYSTEMS AND APPLICATIONS ICEDSA-16, p. 1–4. IEEE, 2016.

[170] LU, L.; GHOSHAL, A. ; RENALS, S.. **Regularized subspace gaussian mixture models for speech recognition**. IEEE Signal Processing Letters, 18(7):419–422, 2011.

[171] DAHL, G. E.; YU, D.; DENG, L. ; ACERO, A.. **Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition**. IEEE Transactions on audio, speech, and language processing, 20(1):30–42, 2012.

[172] SELTZER, M. L.; YU, D. ; WANG, Y.. **An investigation of deep neural networks for noise robust speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-13, p. 7398–7402. IEEE, 2013.

[173] YAO, K.; YU, D.; SEIDE, F.; SU, H.; DENG, L. ; GONG, Y.. **Adaptation of context-dependent deep neural networks for automatic speech recognition**. In: SPOKEN LANGUAGE TECHNOLOGY WORKSHOP (SLT), p. 366–369. IEEE, 2012.

[174] LEE, H.-Y.; CHO, J.-W.; KIM, M. ; PARK, H.-M.. **Dnn-based feature enhancement using doa-constrained ica for robust speech recognition**. IEEE Signal Processing Letters, 23(8):1091–1095, 2016.

[175] PARTHASARATHI, S. H. K.; HOFFMEISTER, B.; MATSOUKAS, S.; MANDAL, A.; STROM, N. ; GARIMELLA, S.. **fmllr based feature-space speaker adaptation of dnn acoustic models**. In: SIXTEENTH

ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 2015.

[176] VIRTANEN, T.; SINGH, R. ; RAJ, B.. **Techniques for noise robustness in automatic speech recognition**. John Wiley & Sons, 2012.

[177] LI, J.; DENG, L.; GONG, Y. ; HAEB-UMBACH, R.. **An overview of noise-robust automatic speech recognition**. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4):745–777, 2014.

[178] SENIOR, A.; LOPEZ-MORENO, I.. **Improving dnn speaker independence with i-vector inputs**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-14, p. 225–229. IEEE, 2014.

[179] RUSS, J. C.; MATEY, J. R.; MALLINCKRODT, A. J.; MCKAY, S. ; OTHERS. **The image processing handbook**. Computers in Physics, 8(2):177–178, 1994.

[180] BALCHANDRAN, R.; MAMMONE, R.. **Non-parametric estimation and correction of non-linear distortion in speech systems**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-98, volumen 2, p. 749–752. IEEE, 1998.

[181] HARVILLA, M.; STERN, R.. **Histogram-based subband powerwarping and spectral averaging for robust speech recognition under matched and multistyle training**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-12, p. 4697–4700. IEEE, 2012.

[182] MOLAU, S.; PITZ, M. ; NEY, H.. **Histogram based normalization in the acoustic feature space**. In: WORKSHOP ON AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING,ASRU'01, p. 21–24. IEEE, 2001.

[183] DE LA TORRE, A.; SEGURA, J. C.; BENITEZ, C.; PEINADO, A. M. ; RUBIO, A. J.. **Non-linear transformations of the feature space for robust speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-02, volumen 1, p. I–401. IEEE, 2002.

[184] GORDILLO, C. D. A.. **Reconhecimento de voz contínua combinando os atributos mfcc e pncc com métodos de robustez ss,**

wd, map e frn. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio, 2013.

[185] CHEN, C.-P.; FILALI, K. ; BILMES, J. A.. **Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases.** In: SEVENTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, 2002.

[186] **The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) Training and Test Data**, 1993.

[187] HIRSCH, H.; PEARCE, D.. **The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions**. In: ASR2000-AUTOMATIC SPEECH RECOGNITION: CHALLENGES FOR THE NEW MILLENIUM ISCA TUTORIAL AND RESEARCH WORKSHOP (ITRW), 2000.

[188] GELFAND, S. A.. **Hearing: An introduction to psychological and physiological acoustics**, volumen 5. CRC Press, New York, USA, 2016.

[189] HAYKIN, S.; CHEN, Z.. **The cocktail party problem**. Neural computation, 17(9):1875–1902, 2005.

[190] DARWIN, C.. **Computational auditory scene analysis: Principles, algorithms and applications**. The Journal of the Acoustical Society of America, 124(1):13–13, 2008.

[191] PRINCEN, J.; BRADLEY, A.. **Analysis/synthesis filter bank design based on time domain aliasing cancellation**. IEEE Transactions on Acoustics, Speech, and Signal Processing, 34(5):1153–1161, 1986.

[192] BRUNGART, D. S.; CHANG, P. S.; SIMPSON, B. D. ; WANG, D.. **Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation**. The Journal of the Acoustical Society of America, 120(6):4007–4018, 2006.

[193] SRINIVASAN, S.; WANG, D.. **Robust speech recognition by integrating speech separation and hypothesis testing**. Speech Communication, 52(1):72–81, 2010.

[194] BROWN, G. J.; COOKE, M.. **Computational auditory scene analysis**. Computer Speech & Language, 8(4):297–336, 1994.

[195] SHAO, Y.; SRINIVASAN, S.; JIN, Z. ; WANG, D.. **A computational auditory scene analysis system for speech segregation and robust speech recognition**. Computer Speech & Language, 24(1):77–93, 2010.

[196] LI, Y.; WANG, D.. **On the optimality of ideal binary time–frequency masks**. Speech Communication, 51(3):230–239, 2009.

[197] HUMMERSONE, C.; STOKES, T. ; BROOKES, T.. **On the ideal ratio mask as the goal of computational auditory scene analysis**. In: BLIND SOURCE SEPARATION, p. 349–368. Springer, 2014.

[198] OJALA, T.; PIETIKÄINEN, M. ; HARWOOD, D.. **A comparative study of texture measures with classification based on featured distributions**. Pattern recognition, 29(1):51–59, 1996.

[199] HE, S.; SORAGHAN, J. J.; O'REILLY, B. F. ; XING, D.. **Quantitative analysis of facial paralysis using local binary patterns in biomedical videos**. IEEE Transactions on Biomedical Engineering, 56(7):1864–1870, 2009.

[200] LIAO, S.; LAW, M. W. ; CHUNG, A. C.. **Dominant local binary patterns for texture classification**. IEEE transactions on image processing, 18(5):1107–1118, 2009.

[201] CHEN, J.; SHAN, S.; HE, C.; ZHAO, G.; PIETIKAINEN, M.; CHEN, X. ; GAO, W.. **Wld: A robust local image descriptor**. IEEE transactions on pattern analysis and machine intelligence, 32(9):1705–1720, 2010.

[202] GUPTA, D.; JINDAL, A.. **Content based image retrieval using enhanced local tetra patterns**. International journal of innovative research in science and engineering, 2017.

[203] CHATLANI, N.; SORAGHAN, J. J.. **Local binary patterns for 1-d signal processing**. In: 18TH EUROPEAN SIGNAL PROCESSING CONFERENCE EUSIPCO-10, p. 95–99. IEEE, 2010.

[204] KIM, G.; LOIZOU, P. C.. **A new binary mask based on noise constraints for improved speech intelligibility**. In: ELEVENTH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 2010.

[205] COHEN, I.. **Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging**. IEEE Transactions on speech and audio processing, 11(5):466–475, 2003.

[206] RANGACHARI, S.; LOIZOU, P. C.. **A noise-estimation algorithm for highly non-stationary environments**. Speech communication, 48(2):220–231, 2006.

[207] HUYNH-THU, Q.; GARCIA, M.-N.; SPERANZA, F.; CORRIVEAU, P. ; RAAKE, A.. **Study of rating scales for subjective quality assessment of high-definition video**. IEEE Transactions on Broadcasting, 57(1):1–14, 2011.

[208] POVEY, D.; GHOSHAL, A.; BOULIANNE, G.; BURGET, L.; GLEMBEK, O.; GOEL, N.; HANNEMANN, M.; MOTLICEK, P.; QIAN, Y.; SCHWARZ, P. ; OTHERS. **The kaldi speech recognition toolkit**. In: IEEE 2011 WORKSHOP ON AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING, número EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[209] FISCHER, A.; IGEL, C.. **An introduction to restricted boltzmann machines**. In: IBEROAMERICAN CONGRESS ON PATTERN RECOGNITION, p. 14–36. Springer, 2012.

[210] ALLEN, J. B.; RABINER, L. R.. **A unified approach to short-time fourier analysis and synthesis**. Proceedings of the IEEE, 65(11):1558–1564, 1977.

[211] RICKER, N.. **Wavelet contraction, wavelet expansion, and the control of seismic resolution**. Geophysics, 18(4):769–792, 1953.

[212] LONG, Y.; GANG, L. ; JUN, G.. **Selection of the best wavelet base for speech signal**. In: INTERNATIONAL SYMPOSIUM ON INTELLIGENT MULTIMEDIA VIDEO AND SPEECH PROCESSING, p. 218–221. IEEE, 2004.

[213] SAFA, S.; MOUHAMED, B. ; ADNEN, C.. **The real time implementation on dsp of speech enhancement based on kalman filter and wavelet thresholding**. Indian Journal of Science and Technology, 10(24), 2017.

[214] MALLAT, S. G.. **A theory for multiresolution signal decomposition: the wavelet representation**. IEEE transactions on pattern analysis and machine intelligence, 11(7):674–693, 1989.

[215] VAIDYANATHAN, P.. **Quadrature mirror filter banks, m-band extensions and perfect-reconstruction techniques**. IEEE Assp Magazine, 4(3):4–20, 1987.

[216] SIQUIERA, J. K.. **Reconhecimento de voz contínua com atributos mfcc, ssch e pncc, wavelet denoisign e redes neurais**. Master's thesis, Pontífica Universidade Católica do Rio de Janiero, 2011.

[217] WANG, D.; BROWN, G.. **Computational auditory scene analysis: Principles, algorithms, and applications**. IEEE Press, 2006.

[218] HOLDSWORTH, J.; NIMMO-SMITH, I.; PATTERSON, R. ; RICE, P.. **Implementing a gammatone filter bank**. Annex C of the SVOS Final Report: Part A: The Auditory Filterbank, 1:1–5, 1988.

[219] SHARAN, R. V.; MOIR, T. J.. **Cochleagram image feature for improved robustness in sound recognition**. In: INTERNATIONAL CONFERENCE ON DIGITAL SIGNAL PROCESSING DSP-15, p. 441–444. IEEE, 2015.

[220] KIM, C.; STERN, R. M.. **Power-normalized cepstral coefficients (pncc) for robust speech recognition**. In: INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP-12, p. 4101–4104. IEEE, 2012.

[221] WANG, Y.; WANG, D.. **Towards scaling up classification-based speech separation**. IEEE Transactions on Audio, Speech, and Language Processing, 21(7):1381–1390, 2013.

[222] SELTZER, M. L.; RAJ, B. ; STERN, R. M.. **A bayesian classifier for spectrographic mask estimation for missing feature speech recognition**. Speech Communication, 43(4):379–393, 2004.

[223] CHEN, C.-P.; BILMES, J. A.. **Mva processing of speech features**. IEEE Transactions on Audio, Speech, and Language Processing, 15(1):257–270, 2007.

[224] LING, Z.-H.; KANG, S.-Y.; ZEN, H.; SENIOR, A.; SCHUSTER, M.; QIAN, X.-J.; MENG, H. M. ; DENG, L.. **Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends**. IEEE Signal Processing Magazine, 32(3):35–52, 2015.

[225] HINTON, G. E.. **A practical guide to training restricted boltzmann machines**. In: NEURAL NETWORKS: TRICKS OF THE TRADE, p. 599–619. Springer, 2012.

[226] MAZUMDAR, J.; HARLEY, R. G.. **Recurrent neural networks trained with backpropagation through time algorithm to estimate nonlinear load harmonic currents**. IEEE Transactions on Industrial Electronics, 55(9):3484–3491, 2008.

[227] HINTON, G. E.; SRIVASTAVA, N.; KRIZHEVSKY, A.; SUTSKEVER, I. ; SALAKHUTDINOV, R. R.. **Improving neural networks by preventing co-adaptation of feature detectors**. arXiv:1207.0580, 2012.

[228] HAN, W.; CHAN, C.; CHOY, C. ; PUN, K.. **An efficient mfcc extraction method in speech recognition**. In: INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS ISCAS-06, p. 4–pp. IEEE, 2006.

[229] SLANEY, M.; OTHERS. **An efficient implementation of the patterson-holdsworth auditory filter bank**. Apple Computer, Perception Group, Tech. Rep, 35:8, 1993.

[230] HOLDSWORTH, J.; NIMMO-SMITH, I.; PATTERSON, R. ; RICE, P.. **Implementing a gammatone filter bank**. Annex C of the SVOS Final Report: Part A: The Auditory Filterbank, 1:1–5, 1988.

[231] SÁROSI, G.; MOZSÁRY, M.; MIHAJLIK, P. ; FEGYÓ, T.. **Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise environment**. In: 6TH CONFERENCE ON SPEECH TECHNOLOGY AND HUMAN-COMPUTER DIALOGUE SPED-11, p. 1–8. IEEE, 2011.

[232] KIM, C.; STERN, R.. **Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring**. In: ACOUSTICS SPEECH AND SIGNAL PROCESSING (ICASSP), 2010 IEEE INTERNATIONAL CONFERENCE ON, p. 4574–4577. IEEE, 2010.

A
Atributos MFCC e PNCC

Na atualidade existem diferentes metodologias e procedimentos de análise para extração de atributos do sinal de voz, que se centram em diferentes aspectos representativos. Neste apêndice apresentam-se e analisam-se as técnicas de parametrização utilizadas nos experimentos desta tese, as quais são:

- Mel-Frequency Cepstral Coefficients (MFCC).
- Power Normalized Cepstral Coefficients (PNCC).

Essas duas técnicas possuem etapas em comum como amostra a Fig. A.1, tais como: a pré-ênfases, análise espectral de curto prazo (divisão em quadros), transformada discreta do cosseno, coeficientes delta e os coeficientes de aceleração. Diferenciando-se no bloco chamado de *informação do espectro*. Esse bloco consiste em dividir cada quadro fornecido pela DFT em B bandas de frequência e extrair um valor de cada um deles separadamente. Esse procedimento será explicado para cada um dos métodos nas seções a seguir.

A.1
Pré-ênfase

Prévio à segmentação do sinal, é aplicado um filtro digital passa-alta de primeira ordem, a fim de compensar os efeitos dos pulsos glotais e ressaltar as frequências dos formantes [5]. Esse procedimento justifica-se por duas razões:

- Evitar a perda de dados durante o processo de segmentação, já que a maior parte da informação está contida nas frequências baixas.
- Remover a componente DC do sinal, aplainando-o espectralmente.

A função de transferência do filtro de pré-ênfase é dada por

$$H(z) = 1 - \alpha_{pre} z^{-1} \qquad 0 \leq \alpha_{pre} \leq 1 \tag{A-1}$$

onde α_{pre} determina a frequência de corte, com valores tipicamente variando entre 0,95 e 0,98.

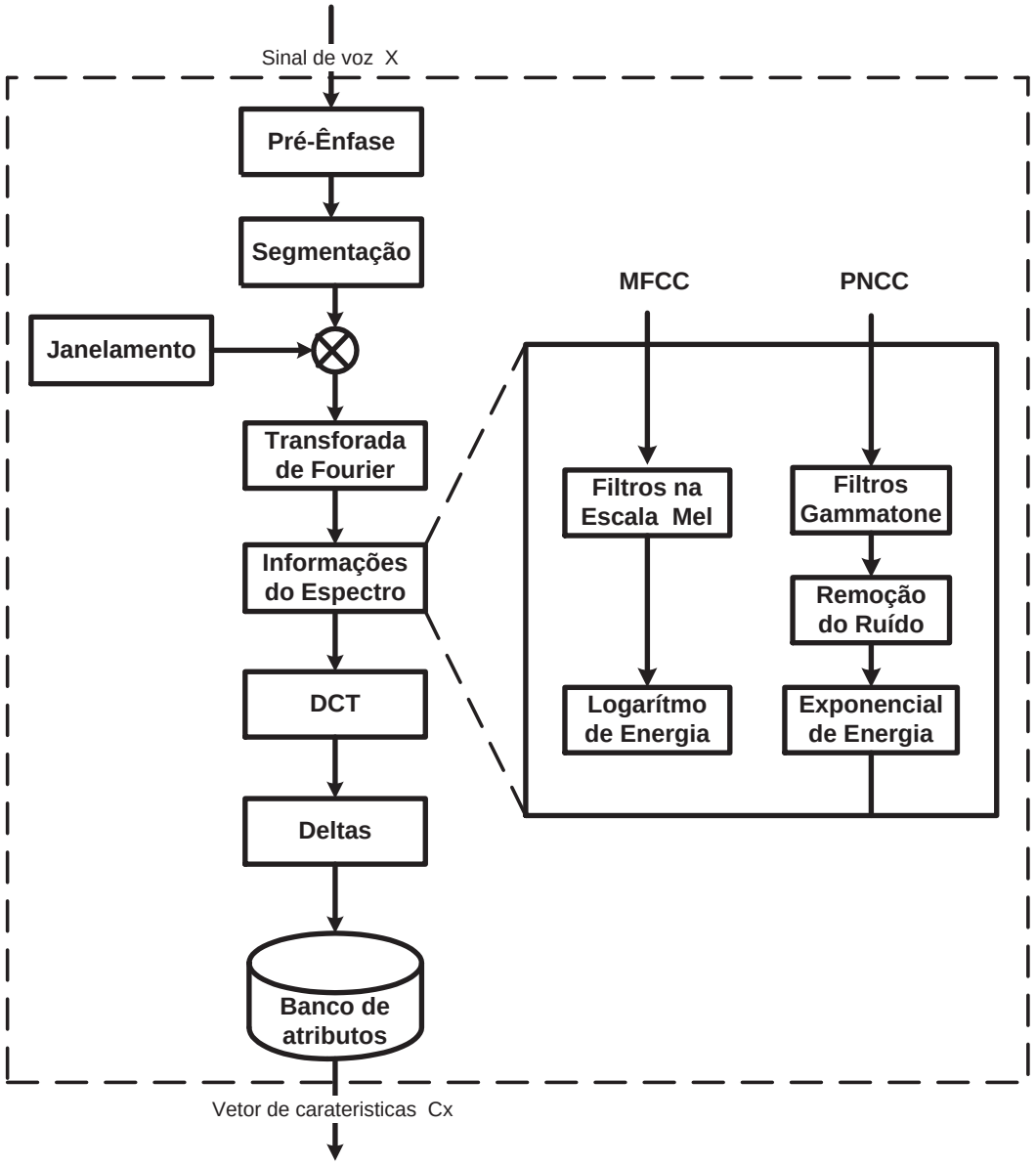

Figura A.1: Comparação dos métodos de extração de atributos

A.2 Segmentação

Nos sistemas de reconhecimento, um dos grandes inconvenientes é acompanhar as mudanças ao longo do sinal causadas pelas características próprias do sinal de voz que se alteram de acordo com as vogais, consoantes e os momentos de ausência de voz. Uma forma de conseguir acompanhar essas mudanças é segmentando o sinal de voz em intervalos consecutivos relativamente pequenos e do mesmo tamanho, nos quais assume-se características de quase estacionariedade [16]. A segmentação é feita levando em consideração a continuidade do sinal em tempo e frequência. No entanto, essa segmentação provoca perda de

informação devido ao fato de quebrar em quadros vizinhos. Por esse motivo os quadros vizinhos são superpostos mantendo assim toda a informação relevante que foi cortada no fim de um quadro e no quadro seguinte. O tamanho do segmento escolhido determina a resolução de frequência da representação. com janelas curtas, obtém-se uma boa resolução temporal (maior capacidade de discriminação entre eventos temporários), porém, obtém-se uma resolução de frequência mais baixa e vice-versa. Para calcular o numero de amostras que compõem cada segmento, multiplica-se a duração do segmento L_t (em segundos) pela frequência de amostragem F_s, ou seja,

$$N = F_s(amostras/segundo) * L_t(segundos) \tag{A-2}$$

Tendo em conta a duração dos fones, o tamanho do quadro geralmente é de 20 a 30ms, com um deslocamento típico de 10 ms entre quadros. Isso impede a perda de representação de um segmento.

A.3 Janelamento

Segmentar o sinal de voz traz o problema de descontinuidade ao início e ao final de cada quadro, devido ao fato de cada um começar e terminar bruscamente. Se simplesmente toma-se as amostras segmentadas no passo anterior, quando aplica-se uma técnica de análise espectral como a transformada de Fourier, ela agiria como se estivesse operando em um sinal que é zero antes do início do segmento e, em seguida, saltaria bruscamente para o sinal durante o segmento é depois voltaria a zero quando o segmento termina. O que introduziria uma distorção significativa do sinal, fazendo com que pareça haver ruído de alta frequência no início e nos pontos finais de cada segmento.

É necessário, então, diminuir este efeito, multiplicando cada quadro por uma janela que seja adequada, visando suavizar as bordas do quadro até chegar a zero, e realçando a parte central para acentuar as propriedades características do segmento, como amostra a Fig. A.2.

No reconhecimento de voz, existem diferentes tipos de janelas. No entanto, a mais utilizada é a janela de Hamming [17]. Representada matematicamente por:

$$W(t) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi t}{T-1}\right) & \text{para} \quad 0 \leq t \leq T-1, \\ 0 \quad \text{para} & \text{caso contrário.} \end{cases} \tag{A-3}$$

O novo sinal de voz segmentado e janelado, com características de quase

estacionariedade, é definido pela multiplicação das amostras de cada quadro pela janela de Hamming, ou seja,

$$x'(t) = x(t) * W(t) \tag{A-4}$$

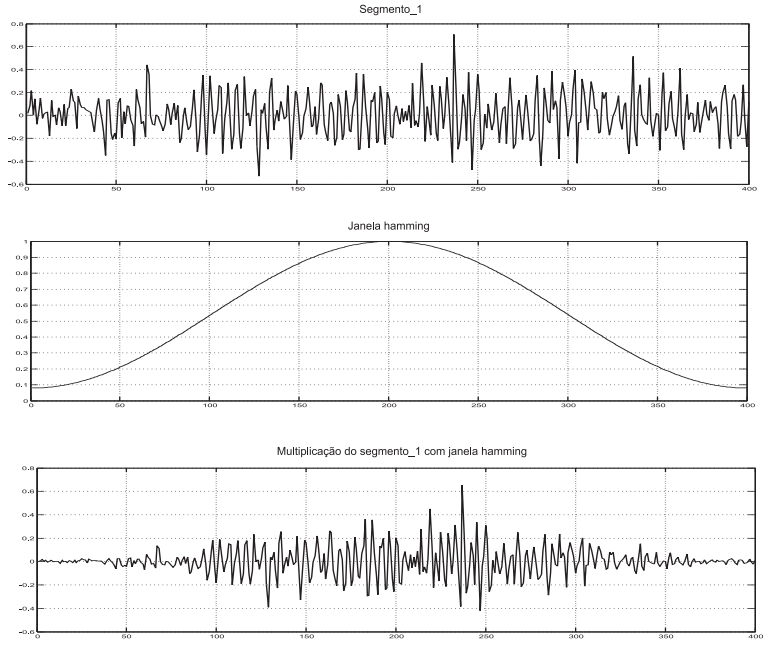

Figura A.2: Segmento janelado com Hamming

A.4 Transformada de Fourier

Enquanto uma função no domínio do tempo indica como a amplitude do sinal muda no tempo, sua representação no domínio da frequência permite saber quantas vezes essas mudanças ocorrem, já que é mais fácil ver que os fones são formados por vibrações que se repetem ao longo do tempo, como foi visto na Fig. 2.4 do capítulo 1. Devido a essa mudanças nase a ideia de extrair atributos dobre essas oscilações, através da transformada discreta de Fourier [19], matematicamente dada por

$$X(k) = \sum_{k=0}^{T-1} x(t) e^{\frac{-j2\Pi kt}{T}} \tag{A-5}$$

onde N é o total de amostras do quadro.

A.5
Informação do espectro

Como apresentado no capítulo 1. É nessa etapa que diferem as duas técnicas de parametrização usadas nesta tese (MFCC e PNCC). Já que a extração de informação numérica do espectro de cada quadro do sinal é feita de formas diferentes como sera apresentado a seguir.

A.5.1
Mel-Frequency Cepstral Coefficients (MFCC)

A técnica de extração de atributos *Mel-Frequency Cepstral Coefficients* (MFCC)[228] faz uma análise de características espectrais de tempo curto, baseando-se no uso do espectro da voz convertido para uma escala de frequências denominada MEL que é uma escala que visa imitar as características únicas perceptíveis pelo ouvido humano. Estes coeficientes são uma representação definida como o cepstrum de um sinal janelado no tempo, que tem sido derivado da aplicação da DFT, em escalas de frequência não lineares.

Para a extração dos vetores de características MFCC, são necessárias as etapas mencionadas nas seções anteriores, sendo configuradas da seguinte forma.

- O sinal de voz *x(t)* a parametrizar é passado através do filtro de pre-ênfase da equação A-1 com $\alpha_{pre} = 0.97$. Essa etapa é recomendável para compensar a atenuação das componentes de alta frequência causadas pelo mecanismo da produção de voz.
- Depois do sinal ser filtrado, é necessário atenuar as descontinuidades causadas no início e no final do sinal de cada segmento, aplicando uma janela Hamming de 25 ms de comprimento, com deslocamento entre janelas de 10 ms, obtendo-se assim vetores MFCC a cada 10 ms (equações A-3 e A-4.
- Após a etapa de janelamento do sinal, aplica-se a DFT da equação A-5 para obter o espectro.
- Uma vez calculada a DFT obtém-se a potência espectral, utilizando a equação

$$S[k] =\mid X[k] \mid^2 = (real(X[k]))^2 + (imag(X[k]))^2 \qquad \text{(A-6)}$$

– A etapa a seguir é a chamada **informações de espectro**, que faz a distinção entre as técnicas de extração, na qual aplica-se um banco de M filtros à potência espectral.

O banco de filtros está formado por filtros triangulares, espaçados de acordo com a escala de frequência MEL, representada pela equação

$$Mel(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{A-7}$$

que, como foi mencionado, imita a resposta em frequência do sistema auditivo humano.

Matematicamente, os filtros MEL, são definidos pela seguinte resposta em frequência

$$H_m[k] = \begin{cases} 0 & k < k[m-1] \\ \frac{2(k-k[m-1])}{(k[m+1]-k[m-1])(k[m]-k[m-1])} , & k[m-1] \leq k \leq k[m] \\ \frac{2(k[m+1]-k)}{(k[m+1]-k[m-1])(k[m+1]-k[m])} , & k[m] \leq k \leq k[m+1] \\ 0 & k > k[m+1] \end{cases} \tag{A-8}$$

Cada filtro calcula a média do espectro em torno da frequência central, e têm diferentes larguras de banda. Quanto maior é a frequência maior é a largura de banda, como mostra a Fig. A.3.

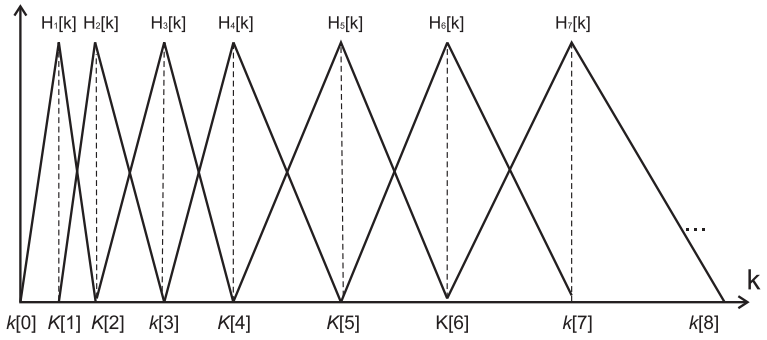

Figura A.3: Banco de filtros usado na técnica MFCC

Para determinar matematicamente os segmentos, parte-se das frequências extremas f_l e f_h que são as frequências de corte do banco de filtros em Hz. Esses valores são usados para dividir o intervalo em $B+1$ partes iguais. Para obter os valores em Hz, basta aplicar a função inversa

$$k[m] = \left(\frac{N}{F_s} \right) Mel^{-1} \left(Mel(f_l) + m \frac{Mel(f_h) - Mel(f_l)}{M+1} \right) \qquad \text{(A-9)}$$

onde F_s é a frequência de amostragem em Hz, M é o número de filtros e N o número de amostras da DFT. $k[m]$ são as frequências digitais e Mel^{-1} determina a largura do banco de filtros e é dado por

$$Mel^{-1}(m) = 700 \left(e^{\frac{m}{1125}} - 1 \right) \qquad \text{(A-10)}$$

– Em seguida, obtém-se a log-energia da saída de cada um dos filtros MEL.

$$\widehat{S}(m) = \ln \left(\sum_{m=0}^{\frac{N}{2}-1} S[k] H_m[k] \right), \qquad 1 < m < M \qquad \text{(A-11)}$$

– Finalmente, os coeficientes MFCC são obtidos aplicando a transformada inversa do cosseno (DCT^{-1}) ao logaritmo dos coeficientes de energia obtidos no item anterior

$$c[n] = \sum_{m=0}^{M-1} \widehat{S}[m] \cos \left(\frac{\pi n (m + 0,5)}{M} \right), \qquad 0 < n < M - 1 \qquad \text{(A-12)}$$

Por exemplo, se M = 13 tem-se um vetor como é mostrado a seguir:

$$C_{mel} = c_0, c_1, c_2, ..., c_{12}.$$

O primeiro coeficiente do vetor C_{mel}, denotado por c_0, pode carregar muita informação do sinal transmitido. Este coeficiente por vezes é considerado e por vezes não; isto vai depender do tipo de reconhecimento desejado, que pode ser reconhecimento de voz, ou reconhecimento de locutor.

A avantagem de utilizar DCT no lugar da IFFT (transformada inversa de Fourier),é que a DCT reduz o número de coeficientes gerados após utilizar as técnicas de parametrização especificadas (MFCC ou PNCC). Esta redução é feita através de uma propriedade da DCT conhecida como compactação da energia, concentrando os valores mais significativos nos

primeiros termos do vetor, e descartando os últimos, melhorando assim a eficiência computacional.

– A ideia principal da extração de atributos é captar as mudanças temporais bruscas presentes no espectro. Devido a isto, utilizam-se além, dos coeficientes extraídos até agora, chamados coeficientes "estáticos", os coeficientes delta e de aceleração, chamados coeficientes "dinâmicos", que capturam essas mudanças e incorporam informação relativa à transição dos coeficientes estáticos entre quadros vizinhos.

  O cálculo dos coeficientes dinâmicos faz-se através de regressão linear sobre uma janela, cobrindo dois vetores antes e dois após o vetor calculado [38], ou seja,

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{A-13}$$

  onde d_t é o coeficiente (diferencial) delta (Δ) computado no tempo t, calculado em termos dos correspondentes coeficientes estáticos $C_{t-\theta}$ até $C_{t+\theta}$. O parâmetro Θ indica o tamanho da janela de regressão, usualmente igual a 2.

  Os parâmetros de segunda ordem chamados delta-delta, são obtidos replicando a derivada sobre os resultados obtidos na primeira derivação [187].

A.5.2 Power-Normalized Cepstral Coefficients (PNCC)

Diferentes experimentos ao longo do tempo têm mostrado que os tons não são representados em escalas lineares de frequência. Por isso, tenta-se aproximar, através de escalas de frequências não lineares, o comportamento auditivo humano, tal como acontece com os atributos MFCC. Porém, nestes atributos a eficácia do reconhecimento cai rapidamente com a presença do ruído.

Recentemente, [40] introduziu um método mais eficiente para extração de atributos, chamado *Power-Normalized Cepstral Coefficients* (PNCC). Sua eficiência é devida à adição de uma nova etapa de remoção de ruído, a qual, através da média das energias de uma banda ao longo de alguns quadros consecutivos, consegue remover a adição do ruído do sinal. Esse procedimento é feito após a divisão do sinal em bandas de frequência superpostas, similar ao utilizado nos MFCC. A diferença é o uso de um novo tipo de escala que imita

a resposta em frequência do sistema auditivo. De acordo com isso, os atributos PNCC são considerados uma evolução dos atributos MFCC.

Os PNCC utilizam a mesma metodologia de análise de tempo curto que os MFCC, visando desenvolver conjuntos de atributos baseados em critérios perceptuais. A estrutura do método PNCC é mostrada na Fig A.1, onde pode-se ver que é similar à estrutura MFCC descrita na seção anterior, com algumas variações, especialmente na etapa de *informações de espectro*.

O pré-processamento para a extração de atributos PNCC é o mesmo para os MFCC, que consiste em um filtro de pré-ênfase e a análise de Fourier utilizando o mesmo janelamento Hamming de 25 ms de comprimento com deslocamento entre janelas de 10 ms.

Uma vez obtida esta informação, procede-se à análise espectral constituída por três partes explicadas a seguir.

– A primeira parte consiste na utilização de filtros Gammatome baseados na escala de Bandas Retangulares Equivalentes (ERB) [41].

  Esses filtros possuem bandas de passo não uniformes e sobrepostos, como é mostrado na Fig. A.4, onde cada filtro representa a resposta em frequência relacionada com um ponto particular da membrana basilar [217].

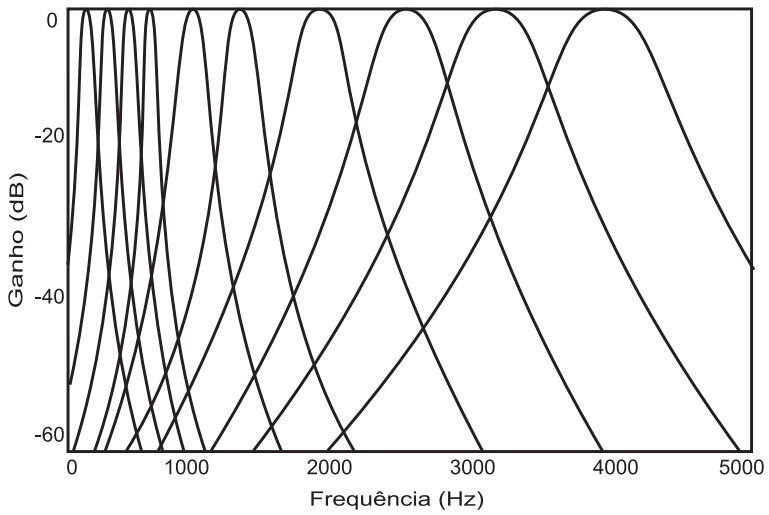

Figura A.4: Banco de filtros Gammatone.

O modelo utilizado para a construção dos filtros é o proposto em [229]. A resposta ao impulso de cada filtro é dada por

$$g(t) = t^{l-1}e^{-2\pi t 1{,}019ERB}\cos(2\pi f_c t) \qquad com \qquad t \geq 0 \qquad \text{(A-14)}$$

onde l é a ordem do filtro ERB e f_c é a frequência central associada a ela. Assim, a largura de banda de cada filtro é ajustada conforme as medidas da largura de ERB dos filtros auditivos humanos dados pela equação

$$ERB(f_c) = 24,7\left(4,37\frac{f_c}{1000}+1\right)Hz \tag{A-15}$$

A largura da anda ERB corresponde aproximadamente 11% de sua frequência central, pelo que os filtros auditivos equivalentes têm uma largura de banda inferior à apresentadas pelas bandas criticas (20% de f_c em 500 Hz). Devido a isso, é necessário uma maior quantidade de filtros ERB, a fim de caracterizar a faixa completa de frequências do sistema auditivo humano (de 20 Hz a 22.050 Hz)

Estas frequências centrais são distribuídas uniformemente em uma escala auditiva de frequências ERB A-16, representadas por uma função quase logarítmica que relaciona a frequência com o número de canais do banco de filtros.

$$ERB_N = 21,4\log_{10}(0,00437f+1) \tag{A-16}$$

onde f é a frequência em Hz e ERB_N é o número ERB(razão ERB)

A implementação completa dos filtros Gammatone pode ser encontrada em [230].

– A segunda modificação é a implementação da etapa de remoção de ruído acima referida, afim de estimar a redução da qualidade da fala causada pelo ruído [231], já que este costuma ser mais estacionário que o sinal de voz.

  Este procedimento é motivado pelo fato de que o sistema auditivo humano é mais sensível a alterações na frequência ao longo do tempo, do que a excitação relativamente constante de fundo [40].

  A implementação detalhada é descrita em [232].

– A terceira modificação está relacionada com a mudança da função logarítmica utilizada na saída dos bancos MEL, por uma função de potenciação aplicada na operação não linear sobre a energia de cada banda. A utilização desta nova função tenta evitar que os valores de saída das bandas estejam perto de zero, já que as regiões onde o sinal

possui pequenas energias serão mas vulneráveis à adição de ruído aditivo, como acontece com a função logarítmica alterando os atributos MFCC. É por isso que se utiliza uma função de potenciação, que vai crescer mais suave, reduzindo assim a distorção espectral.

Uma vez obtida a informação do espectro, o cepstrum em escala ERB é a transformada discreta do cosseno das saídas dos bancos de filtros, similar ao utilizado para os MFCC, (equação A-12).

Finalmente, o vetor de atributos é constituído pelos n coeficientes determinados da DCT, além da adição dos correspondentes coeficientes de regressão (delta e delta-delta) obtidos através da equação A-13.