5
Projeto de um Sistema de Reconhecimento de Voz Contínua através de Técnicas de pré-extração de Atributos

Nos capítulos anteriores foi explicada a teoria dos sistemas de reconhecimento de voz e os problemas que se apresentam quando o sinal de voz é adquirido em condições adversas. Foi verificado que a melhor maneira de melhorar o desempenho do reconhecedor e reduzir o custo computacional é utilizar métodos de robustez antes e depois da extração de atributos, (Fig. 4.4) e dessa forma fornecer ao reconhecedor dados mais parecidos com os de um sinal limpo.

O objetivo deste capítulo é descrever os métodos *Substração Espectral* e *Wavelet Denoising*, baseados na categoria *realce de fala*, e avaliar experimentalmente sua eficiência para combater o ruído aditivo.

5.1
Substração Espectral

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 [50]. Este método está baseado na premissa de que o sinal de voz e o ruído estão decorrelatados e são aditivos no domínio do tempo, pelo qual o espectro em potência do sinal degradado é a soma dos espectros da potência da voz e o ruído.

O princípio básico desta técnica é resumido no diagrama em blocos da Fig. 5.1.

Segundo o modelo de ambiente acústico apresentado em 4.1, e supondo desprezível a presença do ruído convolutivo e outras distorções devidas à variedade intra e interlocutor, a expressão analítica do efeito do ruído aditivo $r(t)$ sobre o sinal de voz $x(t)$ pode ser expressa como uma distorção aditiva no domínio do tempo:

$$y(t) = x(t) + r(t) \tag{5-1}$$

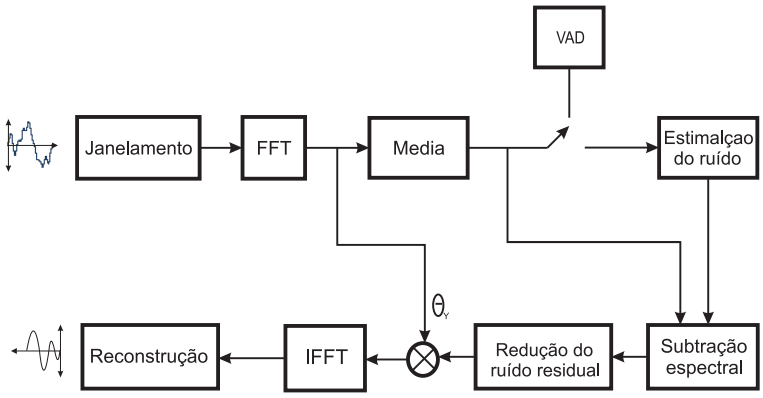

Figura 5.1: Diagrama de blocos do processo de substração espectral.

Sob a suposição de que o ruído é só aditivo e estacionário, o processo de análise é feito de forma localizada utilizando a janela de Hamming definida na seção 2.4.1, a fim de transformar a entrada do sistema (equação 5-1) para domínio da frequência:

$$Y(e^{jw}) = X(e^{(jw)}) + R(e^{(jw)}) \tag{5-2}$$

A ideia principal da subtração espectral é a estimação do espectro do sinal limpo a partir do espectro do sinal ruidoso e uma estimação do espectro do ruído [51]

O espectro do sinal de voz não degradado estima-se como a diferença entre o espectro do sinal com ruído e o espectro do ruído, ou seja

$$\mid \widehat{X}(e^{(jw)}) \mid = \mid Y(e^{jw}) \mid - \mid R(e^{jw}) \mid \tag{5-3}$$

Finalmente, obtém-se o espectro estimado do sinal restaurado, tomando a magnitude do espectro da equação 5-3 e a fase correspondente ao sinal de voz com ruído. Isto fornece como resultado o estimador da subtração espectral

$$\widehat{X}(e^{jw}) = \left[\mid Y(e^{jw}) \mid - E\left\{\mid R(e^{jw}) \mid\right\}\right] e^{j\theta_y} \tag{5-4}$$

onde o valor $E\mid R(e^{jw}) \mid$ é a estimativa do ruído e é representado matematicamente pela equação 5-5, calculada nos instantes onde a atividade vocal é nula, através de um detector de atividade de voz / não voz (VAD), 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.

$$E\left\{\mid R(e^{jw}) \mid\right\} = \alpha E\left\{\mid R(e^{jw}) \mid\right\} + (1-\alpha) \mid Y(e^{jw}) \mid e^{j\theta_x(e^{jw})} \tag{5-5}$$

onde α é o fator de memória e varia entre $0 \leq \alpha \leq 1$. No entanto, esta estimativa produz um erro espectral, resultado da diferença entre o espectro do sinal restaurado e o espectro do sinal limpo

$$\varepsilon(e^{jw}) = \widehat{X}(e^{jw}) - X(e^{jw}) \tag{5-6}$$

Segundo [52], para diminuir este erro é preciso fazer adaptações à equação 5-4, as quais consistem em:

– **Média da magnitude do sinal degradado**

Esta etapa consiste em substituir, na 5-4 a magnitude do espectro de cada quadro do sinal com ruído $\mid Y(e^{jw}) \mid$, por um média de um determinado número de quadros do sinal de voz com ruído, $\mid \overline{Y(e^{jw})} \mid$ onde:

$$\mid \overline{Y(e^{jw})} \mid = \frac{1}{M} \sum_{1=0}^{M-1} \mid Y_i(e^{jw}) \mid \tag{5-7}$$

Deve-se observar que como a voz não é um processo estacionário, só pode-se tomar tempos curtos, usualmente no máximo três quadros por análise, a fim de não perder claridade do discurso.

– **Retificação de onda completa**

A diferença da retificação de meia onda utilizada nos métodos de subtração de Boll [53] e a retificação de onda completa consiste em que na primeira os valores do sinal de voz estimados X que são menores que a estimativa do ruído são considerados nulos, enquanto na segunda, propõe-se estabelecer um limiar de ruído, dado pela magnitude do sinal original $\mid Y(e^{jw}) \mid$ multiplicado por uma constante de atenuação β.

Assim, os valores que estejam debaixo do limiar β. $\mid Y(e^{jw}) \mid$ serão substituídos por este valor, então a retificação é definida por

$$| X_{retificada}(e^{jw}) |= \begin{cases} | X(e^{jw}) | & \text{se } | X(e^{jw}) | \geqslant \beta. | Y(e^{jw}) | \\ \beta. | Y(e^{jw}) | & \text{caso contrário.} \end{cases} \tag{5-8}$$

onde β é um fator usado para minimizar as distorções do sinal processado e está entre 0 e 1.

– **Redução do ruído residual**

Na ausência da voz, a diferença entre o sinal de ruído e a estimativa do ruído pode-se denominar ruído residual. Este ruído tomará valores entre 0 e um valor máximo da estimativa do ruído, medido nos momentos sem voz.

$$R_r = R - E\left\{| R(e^{jw}) |\right\} e^{j\theta_x} \tag{5-9}$$

onde R_r é o ruído residual e R é o ruído de fundo. Este ruído residual pode ser interpretado como uma soma de geradores de tons com frequências fundamentais aleatórias e pode ser percebido nos momentos de atividade vocal.

Este efeito indesejável pode ser reduzido, substituindo o seu valor atual pelo valor mínimo das estimativas dos quadros adjacentes, isto é, tomar o valor mínimo só quando o valor da estimação de $X(e^{jw})$ for menor do que o máximo ruído residual R_r, calculado no instante de não atividade de voz.

A razão para esta substituição é dada em função da amplitude da estimação de $X(e^{jw})$ e é representada matematicamente por:

$$\begin{cases} | \widehat{X}_i(e^{jw}) |=| \widehat{X}_i(e^{jw}) |, & \text{para } | \widehat{X}_i(e^{jw}) | \geq \max | R_r(e^{jw}) | \\ | \widehat{\mathrm{X}}_i(e^{jw}) |= \min\left\{| \widehat{X}_j(e^{jw}) |_{j=i-1,i,i+1}\right\}, & \text{para } | \widehat{\mathrm{X}}_\mathrm{i}(e^{jw}) | < \max | R_r(e^{jw}) | \end{cases} \tag{5-10}$$

onde $\widehat{X}_i(e^{jw})$ é a estimativa do espectro do sinal de voz não degradado e $| R_r(e^{jw}) |$ é o máximo valor do ruído residual estimado nos períodos sem atividade vocal.

Detetor de atividade de voz VAD

Estudos realizados em discursos de fala têm mostrado que o período ativo de uma conversação abrange aproximadamente 40% do tempo, enquanto o 60% restante é constituído por silêncios os quais contêm informação relacionada ao ruído. Assim, para conseguir uma melhor aproximação da estimativa do ruído, é preciso utilizar o VAD para atualizar as amostras tomadas como ruído nesses 60% de instantes onde a voz desaparece.

Existem muitos métodos na literatura para detecção de voz. No entanto, procurando simplicidade, eficiência e baixo custo computacional, o método de Boll foi considerado o mais rápido, diminuindo o tempo total de cálculo em comparação com outros métodos. O princípio básico deste método é resumido na Fig. 5.2

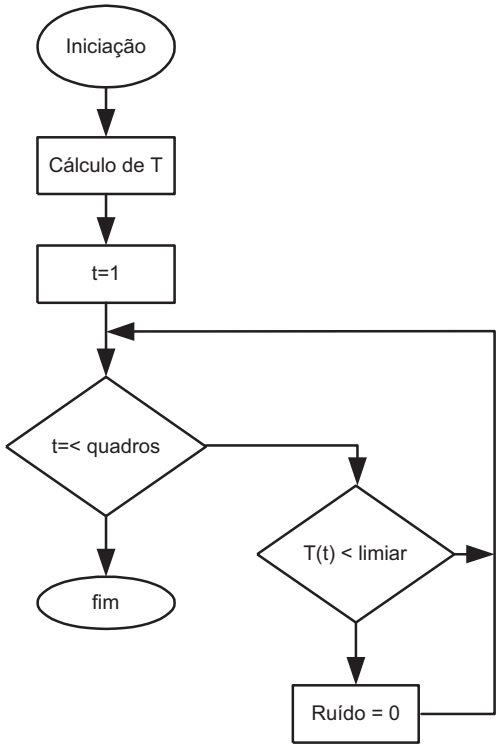

Figura 5.2: Diagrama de fluxo do método VAD.

Seu princípio geral consiste em calcular o nível do sinal de entrada **T** em (dB) para cada quadro e definir o limiar de decisão, em função do qual define se o quadro de análise denominado como quadro ativo marcando-o com 1, ou se o quadro é só ruído marcando-o com 0.

5.2 Wavelet Denoising

Através da transformada wavelet [54] é possível modelar o comportamento do canal auditivo humano, representando a informação no domínio tempo-frequência. Isso permite melhorar algumas das limitações da transformada de Fourier, como a largura da janela de tempo, que uma vez escolhida é mantida fixa durante roda a análise. Ao contrário, a transformada wavelet que varia, ou as pequenas descontinuidades que Fourier não apresenta enquanto que wavelet mostra exatamente a localização dessa descontinuidade 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 estas características, a transformada wavelet tem sido usada, em vez da transformada de Fourier janelada clássica, para extração de características espectrais do sinal de voz [55].

Segundo [56], 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. 5.3.

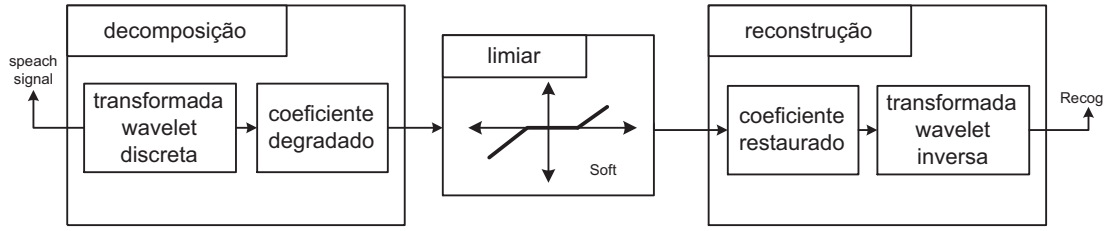

Figura 5.3: Diagrama de blocos da técnica wavelet denoising.

A seguir serão detalhadas as funções que compõem os blocos do processo de *denoising*.

5.2.1 Transformada Wavelet Discreta

As transformadas *wavelet* discreta e inversa definidas nas equações 5-11 e 5-12, respectivamente, são usadas para decomposição e filtragem de qualquer série temporal, tendo como característica principal a eliminação da redundância de coeficientes, através da discretização dos parâmetros de escala a e deslocamento b.

$$W(a,b) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} g(n)\psi_{a,b}(n) \tag{5-11}$$

$$g(n) = W^{-1}(a,b) = \frac{1}{\sqrt{N}} \sum_{a} \sum_{b} W(a,b)\psi_{a,b}(n) \tag{5-12}$$

onde $W(a,b)$ são os coeficientes da transformada *wavelet*, $g(n)$ é a trasnformada inversa de *wavelet* e $\psi_{a,b}(n)$ é a família de funções *wavelet* com escala a e deslocamento b. A discretização do parâmetro b deve ser tal que *wavelets* estreitas (de alta frequência) sejam deslocadas por passos pequenos e *wavelets* largas (baixa frequência) sejam deslocadas por passos maiores, como apresenta a Fig 5.4.

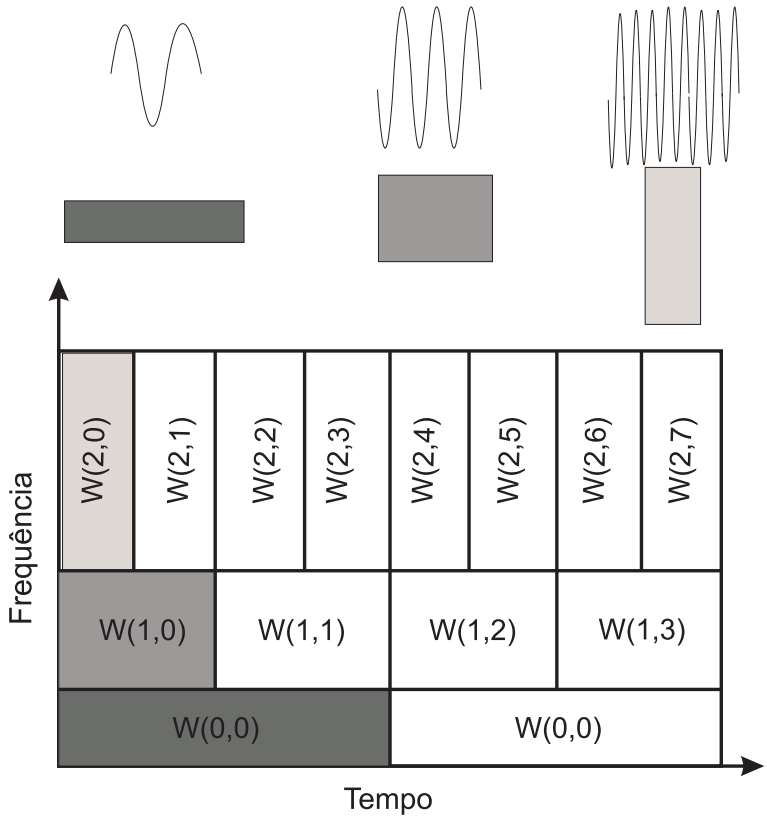

Figura 5.4: Transformada *Wavelet* no domínio tempo- frequência.

A diferença entre as funções senoidais, que são a base da análise de Fourier (equação 2-6), e as wavelets $\psi_{a,b}(n)$ (equação 5-13), é que as segundas são funções obtidas a partir de uma função protótipo denominada *wavelet* mãe $\psi(n)$, a qual pode ser escolhida dependendo do sinal analisado.

$$\psi_{a,b}(n) = \sqrt{2^a}\psi(2^a n - b) \tag{5-13}$$

A seguir, serão apresentadas as propriedades da função *wavelet* mãe utilizada neste projeto.

Wavelet Daubechies

Na maioria de sinais são as componentes de baixa frequência as que entregam ao sinal a maior parte de sua informação, enquanto as componentes de alta frequência incorporam características mais detalhadas. É por isso que as componentes do sinal são divididas em duas categorias

- Aproximações (baixa frequência)
- Detalhes (alta frequência)

Separando estas duas componentes através de filtros, como mostra a Fig. 5.5, exemplifica-se a decomposição do sinal de voz no domínio da frequência como aplicações de filtros passa-baixas e passa-altas.

Uma função que separa essas componentes e que são comummente utilizadas em reconhecimento de voz são as wavelet Daubechies, as quais estão ligadas a famílias de filtros com propriedades especiais, possuindo maior regularidade (suavidade) e aproximando melhor as funções que apresentem descontinuidade.

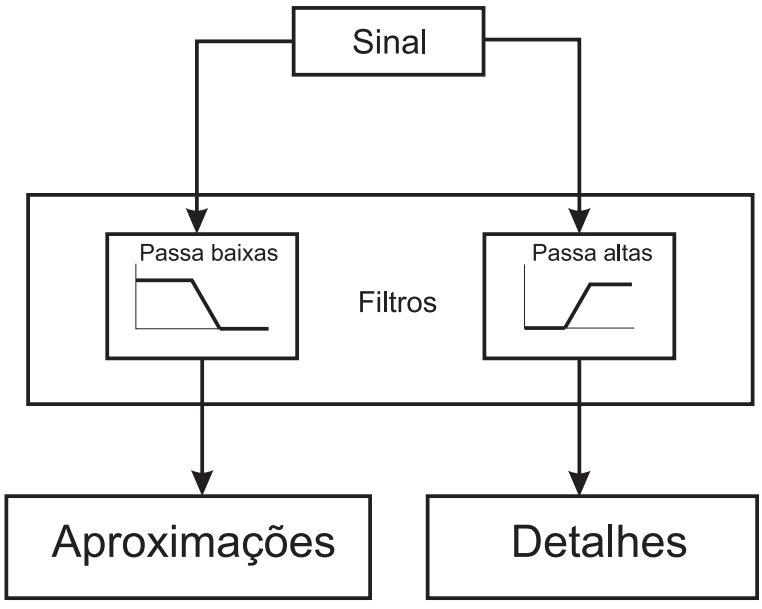

Figura 5.5: Diagrama de decomposição de sinais através de transformadas *wavelet*.

A Fig. 5.6 faz uma comparação da função wavelet daubechies com as funções senoidais (base da análise de Fourier). Pode-se ver que a principal diferença é que o sinal senoidal não tem duração limitada, estendendo-se desde $-\infty$ até ∞.

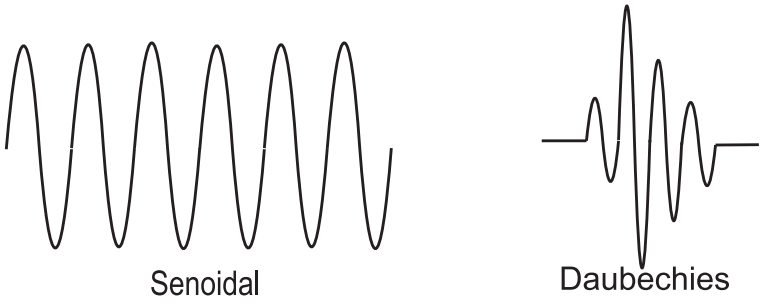

Figura 5.6: Comparação do sinal seniodal e sinal *wavelet*.

Além disso, as *wavelets* têm comportamento irregular e assimétrico, ao contrário dos sinais senoidais. Assim, os sinais com mudanças bruscas serão melhor analisados com transformadas *wavelet* irregulares, do que com suaves senoides.

5.2.2 Limiar (thresholding)

O ponto chave do método *wavelet denoising* está no bloco central da Fig. 5.3, onde, através da função thresholding, tenta-se eliminar os valores menores da transformada $W(a, b)$, que provavelmente representam ruído.

As funções de thresholding mais utilizadas são conhecidas como *hard* e *soft thresholding* [57] e são representadas, graficamente, na Figura 5.7, onde as linhas tracejadas correspondem aos coeficientes originais e as linhas cheias correspondem aoss coeficientes ajustados, após o *thresholding* mantidos ou ajustados.

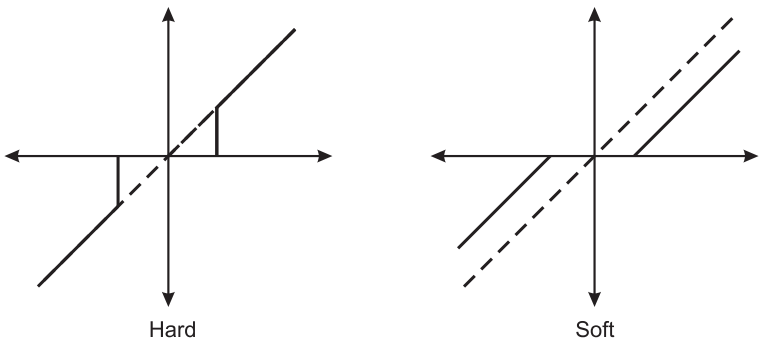

Figura 5.7: Comparação dos gráficos da função *hard thresholding* e a função *soft thresholding*.

O procedimento de *hard thresholding* consiste em "conservar ou eliminar" os coeficientes *wavelet*. Isto é, valores com módulo maior que o limiar λ serão mantidos, caso contrário serão anulados. Este procedimento é dado pela expressão

$$W_h = \begin{cases} W, & \mid W \mid \geq \lambda \\ 0, & \text{caso contrário} \end{cases} \tag{5-14}$$

O procedimento *soft thresholding* visa ajustar os coeficientes *wavelet* a novos valores quando seu módulo é maior que o limiar λ, ao contrário ao método *hard*, que os conserva. A função de thresholding que define o soft thresholding é dada por

$$W_s = \begin{cases} sgn(W)(\mid W \mid -\lambda), & \mid W \mid \geq \lambda \\ 0, & \text{caso contrário} \end{cases} \tag{5-15}$$

Os trabalhos [56] [58] mostram que a função com melhor desempenho para eliminação do ruído é *soft thresholding*, devido a que faz uma transição suave entre valores. Portanto, ela foi escolhida para este trabalho.

Estimação do Limiar

Existem diferentes métodos de estimação do limiar λ, como o desvio padrão, o desvio médio absoluto, etc.

De acordo com os resultados mostrados em [59], [56] e [57], este trabalho escolheu o método que utiliza um processo de estimação dos coeficientes baseado no desvio médio absoluto dos mesmos, o qual também utiliza o *thresholding* universal, sendo dado por

$$\lambda = s\sqrt{2\log(N)} \tag{5-16}$$

onde N é a quantidade de amostras em cada nível e s é a estimação do ruído dada por

$$s = \frac{med(\mid w \mid)}{0.6745} \tag{5-17}$$

onde w é o coeficiente wavelet e med é o desvio médio absoluto.

5.3 Avaliação das técnicas pré-extração de atributos

Até este ponto foram apresentadas os fundamentos do reconhecimento de voz contínua, os problemas que estes sistemas apresentam quando o sinal é adquirido em condições adversas e as técnicas de pré-extração de atributos que visam dar solução à degradação do sinal de voz.

O foco desta seção é avaliar o desempenho dessas técnicas através de experimentos práticos, realizando testes com sinais contaminados a diferentes relações sinal-ruído, a fim de analisar e verificar se as técnicas propostas cumprem os objetivos expostos.

A seguir são apresentados os bancos de dados utilizados para treinar e testar os sistemas propostos, as condições e configurações de teste e os resultados dos experimentos realizados, evoluindo-se gradualmente entre um experimento e o seguinte.

5.3.1 Bancos de dados de voz e de ruído

Para medir as taxas de acerto do reconhecedor de voz e comprovar a forma na qual cada uma das técnicas de aumento de robustez propostas afeta ao mesmo, dispõe-se dos seguintes bancos de dados:

Bancos de dados de voz TIMIT

O banco de dados TIMIT [60] 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. Ela foi criada com ajuda de diferentes grupos de pesquisa, respaldados pelo *Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO)*

Suas siglas vêm 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

Bancos de dados de ruído NOISEX-92

A análise das técnicas de robustez frente a condições adversas precisa de dados degradados. Uma forma de obtê-los é acrescentar aos dados limpos do banco de dados TIMIT o ruído de forma artificial e controlada .

As fontes de ruído foram tomadas do banco de dados NOISEX-92[61], o qual contém sons variados, como falatório e ruído branco, entre outros.

Neste trabalho foram formados dois conjuntos de validação degradados (um por cada nível de ruído) mediante a adição de ruído branco gaussiano. Esta contaminação foi gerada utilizando uma relação sinal-ruído de 15 dB e 10 dB. Uma descrição mais detalhada deste tipo de ruído e a geração do sinal ruidoso é apresentada no apêndice E.

5.3.2 Configurações de prova

A configuração básica utilizada a partir de agora no treinamento e teste do sistema foi a utilizada na maioria dos trabalhos vistos na área, citados ao longo desta dissertação 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**
  - MFCC: B= 26 bandas, somente os 12 primeiros valores da DCT foram considerados, e coeficientes delta e de aceleração foram incluídos.

  - 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).

- **Parâmetros do Wavelet Denoising**

  - Função wavelet mãe: Daubechies 10
  - Número de níveis : 5

- **Parâmetros de Substração Espectral**

  - Constante de Atenuação β :0.3
  - Fator de memória α: 0,8

5.3.3 Resultados Experimentais

Nesta seção serão apresentados finalmente os resultados experimentais obtidos após a aplicação das técnicas de robustez referidas neste capítulo contra degradação gerada pelas condições adversas no sinal de voz.

A avaliação do reconhecedor é dada pela taxa de acerto das palavras 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

$$R(\%) = 100\frac{N-(S+D+I)}{N} \tag{5-18}$$

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.

Teste sistema referência (Ref)

O ponto de partida dos experimentos realizados neste trabalho é testar o reconhecedor sem incorporar nenhum método de robustez, simplesmente avaliá-lo apenas com os métodos de extração de atributos MFCC e PNCC, em condições limpas e posteriormente corrompidas com ruído branco a 10dB e 15dB.

Os resultados de desempenho são apresentados na Tabela 5.1. Eles serão tomados como referência (daqui em diante *Ref*) para os testes posteriores, a

fim de verificar como as técnicas de robustez propostas melhoram as taxas de acerto.

SNR[dB]	**MFCC**	**PNCC**
	Ref	**Ref**
Limpo	89.42%	86,58%
15 dB	34,47%	80,89%
10dB	7,85%	74,52%

Tabela 5.1: Taxas de acerto do sistema de referência.

Os dados da Tabela 5.1 apresentam uma diferença significativa no desempenho do reconhecedor entre as técnicas MFCC e PNCC, mostra também claramente a tendência das taxas de acerto do sistema ser consideravelmente inferiores quanto maior seja a adição de ruído, como vê se nos sistemas com extração de atributos MFCC, onde a degradação do sinal causa no reconhecedor uma queda considerável nas taxas de acerto. Já com os atributos PNCC, pode-se ver como o reconhecedor atinge uma maior robustez em todos os cenários analisados, especialmente quando os ambientes são mais ruidosos. Isso pode ser explicado tendo em conta que este método conta com uma etapa de remoção de ruído, como foi mostrado no capítulo 4, o que faz com que compense a degradação do sinal, conseguindo assim uma forma mais eficiente de extrair informação do sistema.

Teste das técnicas de pré-extração de atributos

Um segundo teste foi realizado visando obter um sistema mais robusto ao efeito da adição do ruído. Para isso foram acrescentados ao sistema *Ref* os métodos *substração espectral* e *Wavelet denoising*, testados nas mesmas condições anteriores. Os testes foram feitos da seguinte maneira

- *Ref + Substração Espectral* (SS).
- *Ref + Wavelet Denoising* (WD).

Os resultados obtidos são apresentados na Tabela 5.2.

SNR[dB]	**MFCC**			**PNCC**		
	Ref	**Ref + SS**	**Ref + WD**	**Ref**	**Ref + SS**	**Ref + WD**
limpo	89,42%	83.62%	74.29%	86,58%	72,13%	70,19%
15db	34,47%	68,03%	63,14%	80,89%	77,59%	81,80%
10db	7,85%	44,25%	27,42%	74,52%	67,35%	75,65%

Tabela 5.2: Taxas de acerto utilizando as técnicas pré-extração de atributos.

Observando os dados da Tabela 5.2, deduz-se que os resultados obtidos depois de utilizar as técnicas *pré-extração de atributos* melhoram significativamente as taxas de acerto para cada um dos cenários dados nos sistemas com atributos MFCC, exceto com o sinal limpo.

Porém, para o sistema com atributos PNCC, nem todas as técnicas de realce são favoráveis. No caso da subtração espectral, as taxas de acerto foram inferiores em todos os cenários. Uma possível razão está ligada ao ruído residual, o qual continua presente nos momentos de atividade vocal, concentrado principalmente nos primeiros coeficientes cepstrais, aparecendo e desaparecendo em intervalos de tempo de $20ms$ aproximadamente. Isso produz uma amplificação deste ruído durante o procedimento de remoção do ruído intrínseco no método PNCC, degradando a inteligibilidade da voz.

Por outro lado, considerando as taxas de acerto obtidas utilizando o método de robustez *Wavelet Denoising* e comparando-as com o sistema *Ref*, pode-se observar que o método de realce apresenta um reconhecimento ligeiramente maior, exceto no cenário limpo, onde apenas tem-se dados sobre a voz. Comprovando assim, tanto para MFCC como PNCC, que qualquer método de eliminação do ruído aplicado a um sinal limpo, elimina informações necessárias para o reconhecimento. observando principalmente que o desempenho do reconhecedor utilizando PNCC supera de maneira considerável o desempenho do reconhecedor utilizando MFCC.

Teste misturando as técnicas de pré-extração de atributos

Um último teste foi realizado concatenando as duas técnicas de robustez na seguinte ordem:

– *Ref* + *substração espectral* (SS) + *Wavelet denoising* (WD)

Os resultados para este teste são apresentados na tabela 5.3.

SNR[dB]	**MFCC**		**PNCC**	
	Ref	**Ref + SS + WD**	**Ref**	**Ref + SS + WD**
limpo	89,42%	72,58%	86,58%	53,24%
15db	34,47%	72,47%	80,89%	78.38%
10db	7,85%	49,26%	74,52%	69.28%

Tabela 5.3: Taxas de acerto utilizando a mistura de técnicas pré-extração de atributos

Pode-se ver que a mistura dos métodos reduz as taxas de acerto em todos os cenários dos sistemas com atributos PNCC, como esperado. Isso

é devido ao efeito do ruído residual da subtração espectral. Já para os sistemas com atributos MFCC, a mistura dos métodos gera um aumento nas taxas de acerto consideravelmente maior em comparação com o sistema *Ref*, principalmente nos cenários mais ruidosos, além de trazer maiores benefícios quando comparado com as taxas de acerto do que a combinação parcial dos testes da 5.2.

Conclusões dos testes

A fim de concluir com a análise global dos resultados apresentados neste capítulo, pode-se dizer o seguinte: Os métodos de pré-extração de atributos melhoram os resultados de reconhecimento dependendo do tipo de extração e das misturas de métodos utilizadas no sistema.

Para o caso dos sistemas com atributos MFCC, o uso dos métodos isoladamente mostrou um incremento considerável no desempenho do reconhecedor, considerando a subtração espectral o melhor método para eliminar o ruído aditivo do sinal. Porém, apesar de sua utilização, pode-se ver que o ruído ainda é forte no sinal de voz, embora menor do que sem a utilização dos mesmos. Além disso, comprovou-se que as melhoras são maiores ao usar a mistura dos métodos de robustez só para atributos MFCC, só que para efeitos da eficiência computacional, esta mistura traz um retardo maior na hora do reconhecimento

Para o sistema com atributos PNCC, a utilização do método wavelet denoising melhora discretamente o desempenho do reconhecedor, resultando numa maior redução do ruído e diminuindo as distorções no sinal de voz provocadas pelo surgimento de componentes em alta frequência, mostrando ser eficientemente mais robusto que a substração espectral. Já para a mistura dos métodos, os resultados foram inferiores em todos os casos, apresentando uma ligeira degradação devido à permanência do ruído residual depois de aplicar a subtração espectral.

Por último, e devido aos resultados, pode-se concluir que a melhor mistura para utilizar a técnica de *pré-extração de atributos* é

Wavelet denoising + PNNC