Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Departamento de Engenharia Elétrica
Programa de Pós-Graduação em Engenharia Elétrica

Dissertação de Mestrado

Redução de Ruído Sonoro Aplicada ao Reconhecimento Automático de Voz

Ísis de Andrade Lima

Marcelo Sampaio de Alencar
Orientador

Campina Grande
Março de 2014

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Departamento de Engenharia Elétrica
Programa de Pós-Graduação em Engenharia Elétrica

Redução de Ruído Sonoro Aplicada ao Reconhecimento Automático de Voz

Ísis de Andrade Lima

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para obtenção do grau de Mestre em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação/Comunicações.

Marcelo Sampaio de Alencar
Orientador

Campina Grande

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

L732r Lima, Ísis de Andrade.

Redução de ruído sonoro aplicada ao reconhecimento automático de voz / Ísis de Andrade Lima. – Campina Grande, 2014.

60 f.: color.

Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2014.

"Orientação: Prof. Dr. Marcelo Sampaio de Alencar".

Referências.

1.Processamento Digital de sinais. 2. Redução de Ruído 3. Reconhecimento de voz I. Alencar, Marcelo Sampaio. II. Título.

CDU 621.391(043)

"REDUÇÃO DE RUÍDO SONORO APLICADO AO RECONHECIMENTO AUTOMÁTICO DE VOZ"

ISIS DE ANDRADE LIMA

DISSERTAÇÃO APROVADA EM 28/03/2014

MARCELO SAMPAIO DE ALENCAR, Ph.D., UFCG
Orientador(a)

LUCIANA RIBEIRO VELOSO, D.Sc., UFCG
Examinador(a)

WASLON TERLLIZZIE ARAÚJO LOPES, D.Sc., UFCG
Examinador(a)

WAMBERTO JOSÉ LIRA DE QUEIROZ, D.Sc
Examinador(a)

CAMPINA GRANDE - PB

Para Dona Tiquinha (mainha).

Agradecimentos

"Por trás de todo feito visível, existem pessoas invisíveis, sem os quais a vida profissional não seria possível" (*The Door*. Direção: Istvàn Szabó, 2012). Eu gostaria de agradecer a alguns daqueles que foram essas pessoas para mim ao longo dos últimos dois anos.

Muito obrigada a painho, pela ajuda em todos os momentos críticos, pelas conversas, conselhos e trocas de experiências. A mainha, pelo apoio, pelos carinhos, pela presença, por sua amizade e amor incondicional. A minha irmã Tha, pela companhia, pelo ombro, pelas muitas conversas sobre os mesmos assuntos.

Agradeço imensamente a todos os meus professores, por terem sempre se importado com meu crescimento profissional e pessoal. Em especial, ao professor Marcelo Sampaio de Alencar, que aceitou orientar este trabalho e o fez com dedicação, respeito e atenção. Aos professores Waslon, Edmar e Ewerton, que estiveram presentes, sempre dispostos a me ajudar.

A todos os meus colegas do Iecom, pelas trocas de informação e pela companhia. Especialmente a Raphael e Raissa, por terem sido tão atenciosos e prestativos.

A todo o pessoal da COPELE e aos demais funcionários de todo o departamento de Engenharia Elétrica, especialmente a Angela, Pedro, Adail e Tchai, aos quais sempre pude recorrer.

A todos os muitos amigos, sem vocês esses dois anos não teriam sido tão felizes e este trabalho seria impossível. Em especial a Milla, Ítalo, Jana, Sarinha, Talita, Ricardo, Bobby, Túlio, Caramelo, Laís, Beth e Tiago, por terem permanecido ao meu lado nos momentos difíceis.

A Lopes (Camila), minha melhor amiga e companheira de 7,5 anos acadêmicos. Nessa linha os agradecimentos são tantos que nem sei como dizer. Muito obrigada!

A todo o pessoal da empresa Alpargatas com quem tive oportunidade de evoluir profissionalmente.

Ao apoio da CAPES, da Universidade Federal de Campina Grande (UFCG) e do Instituto de Estudos Avançados em Telecomunicações (Iecom).

No que se apresenta
O triste se ausenta
Fez-se a alegria
Corra e olhe o céu
Que o sol vem trazer
Bom dia

—CARTOLA (Corra e Olhe o Céu)

Resumo

Um dos principais problemas no desenvolvimento de filtros para sinais de voz é a avaliação do seu desempenho. Não é possível determinar o desempenho de uma técnica de tratamento de ruído sonoro apenas pela análise da SNR obtida, pois a qualidade do sinal filtrado está ligada à sua inteligibilidade. As avaliações subjetivas também não são conclusivas.

Esta dissertação apresenta uma avaliação comparativa dos filtros com resposta finita ao impulso de Wiener ótimo e sub-ótimo, que permite a ponderação entre redução de ruído obtida e distorção inserida a partir do ajuste de um parâmetro α, por meio da observação da taxa de acertos de um sistema de reconhecimento automático de voz (RAV).

Os filtros implementados possuem ordem 20 e janela de análise de 20 ms (intervalo no qual o sinal de voz pode ser considerado estacionário). Para o filtro sub-ótimo foram usados $\alpha = 0,5$, $\alpha = 0,7$ e $\alpha = 0,8$. Para o reconhecedor foi utilizado o decodificador de amplo vocabulário Julius, modelo acústico baseado em cadeias de Markov (*Hidden Markov Models* – HMMs) e modelo linguístico N-grama para o português brasileiro.

Os testes foram realizados com 20 frases de locutores distintos, totalizando 146 palavras. Foram obtidos os percentuais de palavras reconhecidas corretamente para os sinais sem adição de ruído, e para ruído aditivo gaussiano branco com SNR de 20 dB, 15 dB, 10 dB, 5 dB, 3 dB e 0 dB.

Para avaliar o efeito de distorção nos filtros implementados, os sinais obtidos pela filtragem dos arquivos de voz sem ruído são processados pelo reconhecedor, observando que a percentagem de acerto aumenta com a diminuição do parâmetro α (o filtro de Wiener corresponde a $\alpha = 1$).

A partir da análise dos resultados de reconhecimento para os diferentes valores de SNR se conclui que a aplicação do filtro sub-ótimo com $\alpha = 0,7$ resulta na melhor taxa de acertos para o reconhecedor utilizado dentre os quatro filtros desenvolvidos quando o ruído é aditivo gaussiano branco. A melhoria observada foi de 10% para a menor SNR avaliada e de 14% para a maior SNR avaliada.

Palavras-chave: Filtro de Wiener, reconhecimento de voz, redução de ruído.

Abstract

One of the main problems in the development of filters for speech signals is performance evaluation. It is not possible to evaluate the technique only by the obtained SNR analysis, because the quality of the filtered signal is related to its intelligibility. Subjective evaluations are also not conclusive.

This dissertation presents a comparative evaluation of finite impulse response Wiener optimal and sub-optimal filters, which allows weighting between noise reduction and distortion insertion by setting a parameter α, through the observation of an automatic speech recognition (ASR) system error rate.

The 20 order filters were implemented with analysis window of 20 ms (for which the speech signal can be considered stationary). A sub-optimal filter was tested, for $\alpha = 0.5$, $alpha = 0.7$ and $\alpha = 0.8$. The large vocabulary decoder Julius was chosen for the ASR system. Hidden Markov Models (HMMs) and N-gram language model for Brazilian Portuguese were used for acoustic and linguistic training.

The tests were performed with 20 sentences from different speakers, totaling 146 words. The percentage of correctly recognized words for the clean speech signals, additive white Gaussian noise (AWGN) was obtained, for a SNR of 20 dB, 15 dB, 10 dB, 5 dB, 3 dB, 0 dB, and filtered signals.

To evaluate the distortion effect caused by filtering, the filtered version of clean speech signals were processed by the recognizer, and it was observed that the error rate decreases with the reduction of the parameter α (the Wiener filter corresponds to $\alpha = 1$).

Based on the analysis of recognition results for different values of SNR, the application of sub-optimal filter, with $\alpha = 0.7$, produces the best recognition rate for a specified AWGN among the four designed filters. The observed improvement was 10% for the lowest SNR and 14% for the highest SNR evaluated.

Keywords: Wiener filter, speech recognition, noise reduction.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A fala é um dos principais meios de comunicação humana e, por isso, a evolução tecnológica foi acompanhada do desenvolvimento de sistemas de comunicação baseados na fala. Além disso, a automatização dos processos torna necessária a existência de uma interface de comunicação entre humanos e máquinas.

Com a possibilidade do emprego de processamento digital de sinais, permitindo a implementação de funções mais complexas que as empregadas no processamento analógico, surgiu o conceito de processamento digital de voz, que consiste no tratamento de sinais de voz em níveis discretos de intensidade e tempo [3]. Algumas áreas de aplicação de sistemas de comunicação por voz são transmissão, armazenamento, reconhecimento automático de voz, síntese de voz, verificação e identificação automática de locutor, melhoria da qualidade do sinal e aplicações ligadas à acessibilidade conforme ilustrado no diagrama da Figura 1.1 [1].

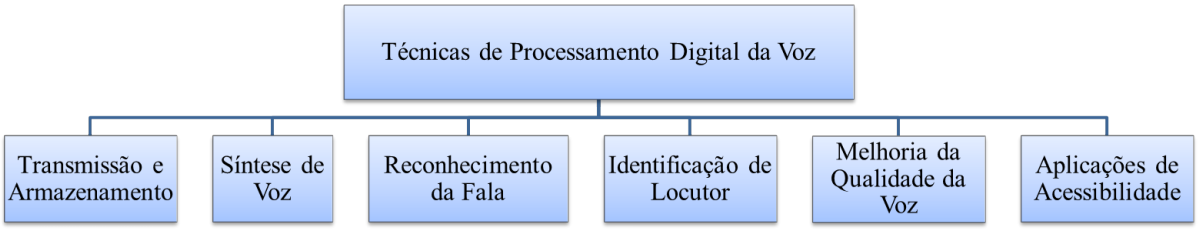

Figura 1.1 Diagrama de blocos representando as aplicações das técnicas de processamento digital da fala [1].

Assim, as pesquisas sobre processamento da fala, ou voz, envolvem especialmente os problemas ligados à facilitação da comunicação humana à distância e comunicação ser humano-máquina utilizando a fala. Um dos principais exemplos de aplicação do processamento digital da voz para a comunicação humana é a telefonia digital.

Este capítulo apresenta uma introdução aos conceitos de comunicação vocal ser humano-máquina na Seção 1.1. Na Seção 1.2 são discutidas as questões que motivam esta dissertação.

Os objetivos geral e específicos são listados na Seção 1.3 e na Seção 1.4 é apresentado um resumo da estrutura do texto.

1.1 Comunicação Ser Humano-Máquina por Fala

A complexidade de operação de sistemas automatizados está diretamente ligada ao tipo de comunicação utilizada em sua interface com o usuário. Como a fala é uma das principais formas de comunicação humana, o desenvolvimento de tecnologias que permitam seu uso nessas interfaces é um processo crescente [3–6].

A utilização da fala na interface ser humano-máquina também oferece a vantagem da velocidade na troca de informação, pois a taxa média de comunicação por voz é de 200 palavras por minuto, enquanto a de digitação é em torno de 60 palavras por minuto [7].

Na década de 1950 surgiram os primeiros trabalhos descrevendo a possibilidade de comunicação vocal ser humano-máquina [8], especificamente para o reconhecimento de dígitos falados. Na década de 1960 as pesquisas em torno do espectro de voz, permitindo o levantamento das características do sinal de voz, e a evolução dos computadores digitais proporcionaram grandes avanços na área [7]. O desenvolvimento dos sistemas de comunicação vocal ser humano-máquina seguiu considerando também a possibilidade de sintetização da voz pela máquina, permitindo a produção de respostas vocais.

Algumas aplicações possíveis para interfaces de comunicação vocal entre ser humano e máquina são comando por voz, soluções para medicina na área de identificação de patologias por voz [9], acessibilidade para pessoas com deficiência visual, auditiva, vocal e física [10], troca de informações com sistemas automatizados, como guichês de estacionamentos e caixas bancários eletrônicos, reconhecimento de ditado, e segurança, por meio da identificação pessoal.

Os sistemas de comunicação vocal ser humano-máquina podem então ser divididos em sistemas de síntese de voz, sistemas de reconhecimento de voz e sistemas de reconhecimento de locutor [7]. Nas seções seguintes são apresentadas as características e exemplos de aplicação para esses tipos de sistemas, bem como as considerações acerca de sua utilização em ambientes que possuem ruído sonoro.

1.1.1 Sistemas de Síntese Vocal

Sistemas de síntese de voz têm como objetivo a conversão de mensagens na forma de texto para voz sintética inteligível e com sonoridade natural [1], fazendo a transmissão da mensagem da máquina para o usuário humano.

Os dois processos fundamentais dos sistemas de síntese de voz por texto são a análise de texto e a síntese de voz. A estrutura desses sistemas é ilustrada na Figura 1.2. A análise de texto determina uma descrição linguística abstrata da mensagem, enquanto a síntese da voz produz os sons correspondentes a essa mensagem [1].

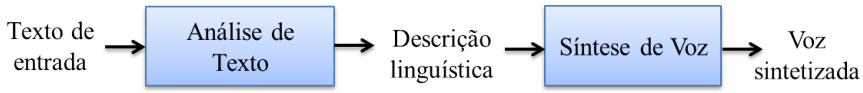

Figura 1.2 Diagrama de blocos representando sistemas de síntese de voz.

Para produzir a voz sintetizada é necessário utilizar um conjunto de elementos de texto (fonemas, palavras, frases ou parâmetros no caso do uso de codificação paramétrica) armazenado no formato digital. Esses elementos podem ser combinados para gerar a mensagem desejada.

Os métodos de codificação de forma de onda – modulação por codificação de pulsos (PCM), PCM diferencial, entre outros [11] – e métodos de análise-síntese – codificação por predição linear – podem ser usados no armazenamento dos elementos textuais. A qualidade do sinal sonoro sintetizado depende, essencialmente, do método de codificação utilizado [7].

A descrição dos fundamentos envolvidos no desenvolvimento dos sistemas de síntese da fala é feita em [1].

1.1.2 Sistemas de Reconhecimento de Voz

Os sistemas de reconhecimento automático de voz (RAV), ou da fala, têm como objetivo a interpretação automática da informação dada por um locutor, o que pode ser visto como o processo inverso da síntese vocal [4]. O principal fator que impulsiona as pesquisas na área de conhecimento da fala é a potencial redução de custos na substituição dos serviços de interação entre humanos por interação ser humano-máquina [1].

Sistemas de interface para dispositivos eletrônicos pessoais, atendimento automático, controle de equipamentos e segurança baseados em voz são algumas aplicações diretas dos sistemas de reconhecimento automático de voz. Esses serviços oferecem facilidade no acesso a informações e serviços para os usuários [1]. Dependendo de sua aplicação, o reconhecimento de fala possui variações no tamanho do vocabulário de reconhecimento, natureza da fala e população de locutores.

Em geral, os sistemas de reconhecimento de fala podem ser subdivididos em sistemas de interface e sistemas transcritores [12]. Os sistemas de interface utilizam técnicas de reconhecimento de voz para comando de ações ou navegação por menus de sistemas, facilitando ou acelerando a operação de dispositivos diversos. Os sistemas transcritores têm como objetivo a captação e transcrição do texto falado por um locutor [7].

O desempenho dos sistemas de reconhecimento de voz é afetado pelo conjunto de parâmetros escolhidos para representar a voz, o modelo linguístico que descreve as características linguísticas de comparação, o modelo acústico baseado nas informações dos sinais de voz e o codificador de voz utilizado [11].

As características e a estrutura dos sistemas de reconhecimento de fala são tratadas mais detalhadamente no Capítulo 2. Mais informações e a descrição dos conceitos envolvidos no desenvolvimento desses sistemas podem ser encontrados em [4] e [1].

1.1.3 Sistemas de Reconhecimento do Locutor

Os sistemas de reconhecimento de locutor têm como objetivo a identificação robusta da identidade de locutores. As principais aplicações desses sistemas são na área de segurança e criminalística, constituindo em uma das principais áreas da comunicação vocal ser humano-máquina [3]. Nesses sistemas a comunicação é feita do usuário humano para a máquina.

O processo de reconhecimento da identidade vocal de locutores consiste essencialmente na extração de parâmetros vocais do locutor, a partir dos quais é possível definir um modelo que preserva suas características vocais que o diferenciam dos demais indivíduos [7].

Os sistemas de reconhecimento de locutor podem ser classificados em duas categorias: identificação de locutor e verificação de locutor. A identificação de locutor se aplica aos casos nos quais o objetivo é atribuir uma identidade ao locutor, enquanto a verificação de locutor se aplica quando o objetivo é confirmar a identidade do locutor [7].

A divisão geral para as aplicações relacionadas à comunicação ser humano-máquina por voz é ilustrada no diagrama de blocos da Figura 1.1.3.

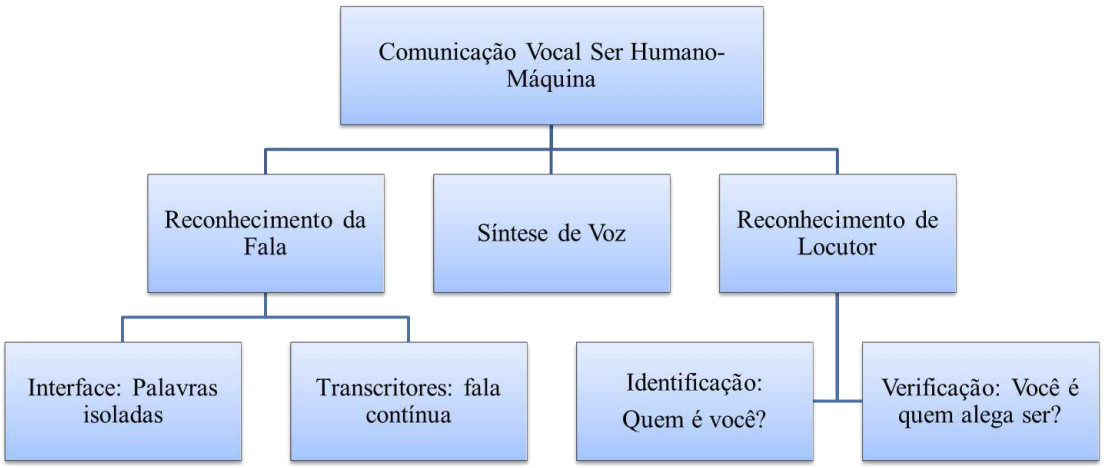

Figura 1.3 Diagrama de blocos representando a divisão das aplicações para comunicação vocal ser humano-máquina.

Mais informações sobre os sistemas de reconhecimento de locutor podem ser encontradas em [7] e [13].

1.1.4 Considerações para Ambientes com Ruído Sonoro

Os sistemas de comunicação vocal ser humano-máquina são normalmente projetados em laboratório e treinados com um dicionário de frases com baixo ou nenhum ruído sonoro.

Como os ambientes típicos de utilização das aplicações desses sistemas não possuem isolamento acústico, o efeito de ruído sonoro pode degradar seu desempenho.

A presença de ruído afeta especialmente as aplicações de comunicação do usuário humano para a máquina, ou seja, sistemas de reconhecimento de voz e de reconhecimento de locutor. Os sistemas de reconhecimento de voz tendem a ter seu desempenho mais degradado que os sistemas de reconhecimento de locutor [14]. Em geral, duas medidas mitigadoras podem ser consideradas para reduzir o efeito da presença de ruído: melhoria do modelo de reconhecimento e pré-tratamento do sinal [6].

A robustez do sistema de reconhecimento pode ser melhorada com uso de técnicas de representação paramétrica da voz (estágio paramétrico) e modelagem, considerando que os sinais de voz estão misturados ao ruído (estágio de modelagem). O pré-tratamento do sinal pode ser usado no estágio acústico para diminuir o nível de ruído ou aumentar o nível do sinal de voz (melhoria da relação sinal-ruído, ou *Signal-to-Noise Ratio* – SNR).

As soluções aplicadas, como pré-processamento, são mais modulares, pois o sistema de reconhecimento pode ser mantido independente do ambiente no qual será utilizado [5]. Essas soluções consistem no cálculo de uma estimativa para o sinal de voz a partir de um sinal de observação correspondente ao áudio alterado pelas condições do ambiente.

Existem diversas técnicas para o tratamento de sinais contaminados com ruído, incluindo filtros adaptativos, ótimos, lineares ou não lineares [15]. As técnicas de filtragem ótima buscam a minimização do erro médio quadrático entre a estimativa obtida para o sinal e o sinal original, enquanto as técnicas de filtragem adaptativa variam os parâmetros de filtragem de acordo com o sinal observado. Considerando que em grande parte das situações as características do ruído variam com o tempo, também foram desenvolvidas técnicas que combinam filtragem ótima e filtragem adaptativa [5, 6, 16, 17].

Especificamente para o problema de redução de ruído sonoro na fala, algumas técnicas se destacam [17]: subtração espectral [18], subespaço de sinais [19], modelagem estatística [20] e filtragem de Wiener [5]. A subtração espectral e a filtragem de Wiener são mais utilizadas, porque apresentam pouca complexidade de implementação e podem ser empregadas com o uso de apenas um canal (um microfone de captura) [6]. No entanto, o uso de subtração espectral e filtro de Wiener não são indicados para situações de baixa SNR [5].

O reconhecimento automático de voz em ambientes de fábrica é uma abordagem específica do problema, pois esses ambientes possuem nível de ruído sonoro bastante elevado. Diversas aplicações para o reconhecimento de voz são propostas para esses ambientes, como a utilização de comando por voz em sistemas ligados à segurança e à operação de máquinas em situações nas quais as mãos dos operários executam outras funções.

1.2 Motivação

O reconhecimento de voz possibilita a passagem de informação do usuário humano para máquinas, podendo ser aplicado diretamente na comunicação ser humano-máquina, na comunicação humana e também em interfaces para melhorar a acessibilidade para pessoas com deficiência.

Os sinais de voz utilizados em sistemas de comunicação por fala são normalmente originados em ambientes sem isolamento acústico. Assim, o sinal recebido pelo sistema corresponde a uma versão corrompida do sinal original, que contém o ruído sonoro aditivo do ambiente e o eco produzido pela fala [6]. Os sistemas de reconhecimento automático da fala são especialmente sensíveis ao efeito de ruído sonoro, o que torna necessária a utilização de medidas mitigadoras.

Um dos filtros mais utilizados na redução de ruído sonoro quando existe apenas um sinal de observação é o filtro ótimo de Wiener [21]. No entanto, o aumento do nível de ruído no sinal leva ao aumento da distorção no sinal filtrado. A necessidade de tratamento para ambientes com ruído muito elevado é justificada nas aplicações dos sistemas de comunicação vocal ser humano-máquina em fábricas, aeroportos, rodoviárias, estações de metrô e etc.

Em [5] é proposta uma alteração no filtro de Wiener com resposta finita ao impulso, desenvolvendo um filtro sub-ótimo em que é possível diminuir a distorção inserida por meio da redução da SNR do sinal filtrado a partir da alteração de um parâmetro de distorção α.

Um dos principais problemas no desenvolvimento de filtros para voz é a avaliação do seu desempenho. Não é possível determinar o desempenho de uma técnica de tratamento de ruído sonoro apenas pela análise da SNR obtida, pois a qualidade do sinal filtrado está ligada à sua inteligibilidade. As avaliações subjetivas também não são conclusivas. A análise comparativa do uso de técnicas de tratamento de ruído a partir de sua aplicação no pré-tratamento de sistemas de RAV pode ser usada para avaliar o desempenho dessas técnicas.

1.3 Objetivos

O objetivo geral deste trabalho é a análise de desempenho dos filtros de Wiener ótimo e sub-ótimo, com resposta finita ao impulso, em termos da taxa de acertos de um sistema de reconhecimento automático de voz baseado em Modelos Escondidos de Markov (*Hidden Markov Models* – HMMs).

Os objetivos específicos que subdividem o objetivo principal são listados a seguir.

1.3.1 Objetivos Específicos

- Estudo de algumas técnicas de redução de ruído sonoro e avaliação teórica de seu desempenho para ambientes com baixa SNR;

- Implementação dos filtros de Wiener ótimo e sub-ótimo;
- Comparação do desempenho das técnicas de filtragem utilizadas por meio de sua aplicação a um sistema de RAV para português brasileiro baseado em HMMs;
- Variação dos parâmetros dos filtros implementados de forma a obter o melhor percentual de palavras reconhecidas corretamente para SNR compatível a ambientes de fábrica.

1.4 Estrutura da Dissertação

Esta dissertação está organizado em seis capítulos. O presente capítulo contém as informações introdutórias acerca das aplicações do processamento digital da voz em sistemas de comunicação ser humano-máquina e os problemas envolvidos no uso dessas aplicações em ambientes com presença de ruído sonoro. A partir desses conceitos é apresentada na Seção 1.2 a motivação e na Seção 1.3 os objetivos para o trabalho.

No Capítulo 2 é descrito o funcionamento dos sistemas de reconhecimento automático de voz (RAV). Para isso, inicialmente são apresentadas as características do sinal de voz e sua representação paramétrica, e em seguida as propriedades e a estrutura dos sistemas de RAV. Por fim, é feita uma breve discussão acerca das ferramentas de implementação utilizadas e os desafios encontrados no desenvolvimento desses sistemas.

No Capítulo 3 são discutidas duas técnicas muito usadas na redução de ruído sonoro em voz (subtração espectral e filtragem de Wiener). Além disso, é realizada uma discussão acerca do desempenho dessas técnicas para ambientes com baixa relação sinal-ruído, e são apresentadas alternativas propostas em trabalhos anteriores, sendo detalhado o filtro sub-ótimo baseado no filtro de Wiener.

No Capítulo 4 é apresentado o sistema utilizado, sendo discutidos os passos usados para a obtenção dos sinais de voz misturados a diferentes níveis de ruído, desenvolvimento dos filtros de Wiener e sub-ótimo, e aplicação dos sinais obtidos no sistema de reconhecimento.

No Capítulo 5 os resultados obtidos pela filtragem são mostrados e é feita a análise de desempenho do sistema de RAV para os sinais filtrados em termos do percentual de palavras reconhecidas corretamente pelo sistema desenvolvido, considerando o uso dos diferentes filtros para diversos valores de SNR.

Por fim, no Capítulo 6 são apresentadas as considerações finais deste trabalho e feitas as propostas para trabalhos futuros.

CAPÍTULO 2

Reconhecimento Automático de Voz (RAV)

O objetivo dos sistemas de reconhecimento automático de voz (RAV), ou da fala, é fazer a conversão de um sinal de voz em uma mensagem transcrita identificável pela máquina, de forma eficiente e acurada, independentemente do dispositivo usado na aquisição do sinal (transdutor, microfone), do sotaque do locutor e do ambiente acústico [1]. Isto é, um sistema de RAV ideal deveria atuar como um ouvinte humano.

O processamento a ser realizado após a identificação pode ser a sintetização (repetição), transmissão, escrita em forma de texto ou execução de um comando. Esses sistemas tiveram a primeira implementação documentada pelos Laboratórios Bell na década de 1950 [8], empregando métodos baseados em reconhecimento de padrões. Desde então, os estudos buscam melhorar seu desempenho e a interface entre homens e máquinas, pois a fala é uma das principais formas de comunicação humana.

Um modelo conceitual simples para os processos de geração do sinal de voz e reconhecimento de voz é mostrado na Figura 2.1. O primeiro passo para a produção da fala assumido como parte do processo da comunicação homem-máquina é a intenção do locutor de expressar algum pensamento. Então o locutor deve compor uma sentença que possua significado linguístico ($\mathbf{W}$) na forma de uma sequência de palavras. Tendo sido escolhidas as palavras, são enviados através do sistema nervoso sinais de controle para os órgãos de articulação da fala, originando a forma de onda do sinal de voz $x(n)$. Esse processo de formação do sinal da fala a partir da intenção do locutor é chamado de Modelo de Locução, pois reflete o sotaque e escolha de palavras para expressar a mensagem desejada [1].

O processo de reconhecimento de voz consiste em um processamento acústico, que faz a análise do sinal de voz e converte em um conjunto de características acústicas X (espectrais e/ou temporais), que representa de forma eficiente os sons da fala. Essas características são processadas por um decodificador linguístico que faz a estimativa das palavras, utilizando uma medida de verossimilhança (*likelihood*), para gerar a sentença reconhecida $\hat{\mathbf{W}}$.

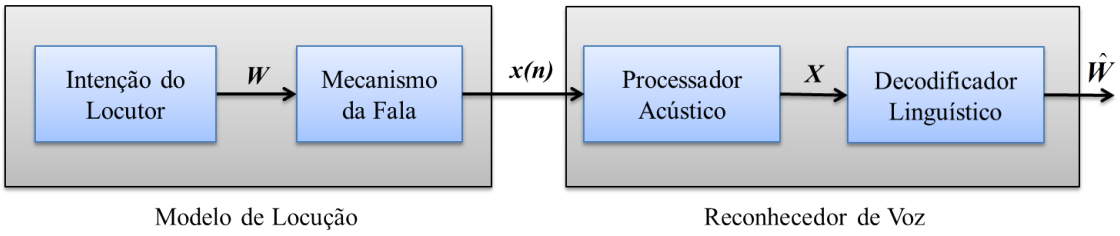

Figura 2.1 Modelo conceitual para processos de produção e reconhecimento de voz [1].

Neste capítulo são apresentados os fundamentos envolvidos na produção e as características do sinal da fala, a fim de justificar o tratamento realizado no processo de Reconhecimento da Fala. Em seguida é feito um resumo das características dos sistemas de Reconhecimento da Fala, que dependem da aplicação desejada.

Na Seção 2.3 o processo de Reconhecimento da Fala descrito resumidamente a partir da Figura 2.1 é apresentado de forma mais detalhada. As ferramentas de implementação utilizadas para o desenvolvimento das diferentes etapas dos sistemas de RAV são apresentadas na Seção 2.4. Por fim, são destacados os desafios para o desenvolvimento desses sistemas e algumas considerações finais.

2.1 Características do Sinal de Voz

O sinal de voz pode ser definido como a onda acústica que corresponde à forma fundamental da mensagem contida na fala [21]. Embora sua representação fundamental seja uma onda contínua, a fala pode ser representada foneticamente por um conjunto finito de símbolos, chamados de *fonemas* da linguagem [4]. O número de fonemas depende do idioma, sendo 38 para o português brasileiro [22, 23]. Assim, a informação contida na fala é de natureza discreta. De acordo com a Teoria da Informação, apresentada por Shannon em [24], uma mensagem composta de símbolos discretos pode ser quantificada em *bits* por seu conteúdo de informação e a taxa de transmissão pode ser medida em *bits* por segundo ($bits/s$) [1].

Um conjunto de regras da linguagem determina a formação dos sons da fala. A linguística é a área científica dedicada ao estudo da linguagem e da utilização de suas regras na comunicação humana, enquanto a fonética é a área de estudos das características da produção do som da fala humana [4].

O sinal da fala pode ser convertido em uma forma de onda elétrica pelo uso de um microfone ou um transdutor, em seguida pode ser processado na sua forma analógica ou, após uma conversão analógico/digital, ser tratado por um processador digital. É possível converter o sinal novamente para uma onda acústica (conversão digital/analógico) com o uso de auto-falantes. Apesar das aplicações iniciais para o processamento de voz não levarem em consideração os fundamentos de Teoria da Informação, os estudos mais recentes utilizam cada vez mais esses conceitos. Assim, existem duas abordagens possíveis para o tratamento do sinal de voz:

o tratamento da forma de onda e representação paramétrica do sinal, sendo essa abordagem considerada clássica, e o tratamento da informação codificada no sinal de voz [1]. Os estudos apresentados neste trabalho consideram a abordagem clássica.

O processo de produção e percepção da fala brevemente descrito na Figura 2.1 é apresentado de forma detalhada na cadeia da fala ilustrada na Figura 2.2.

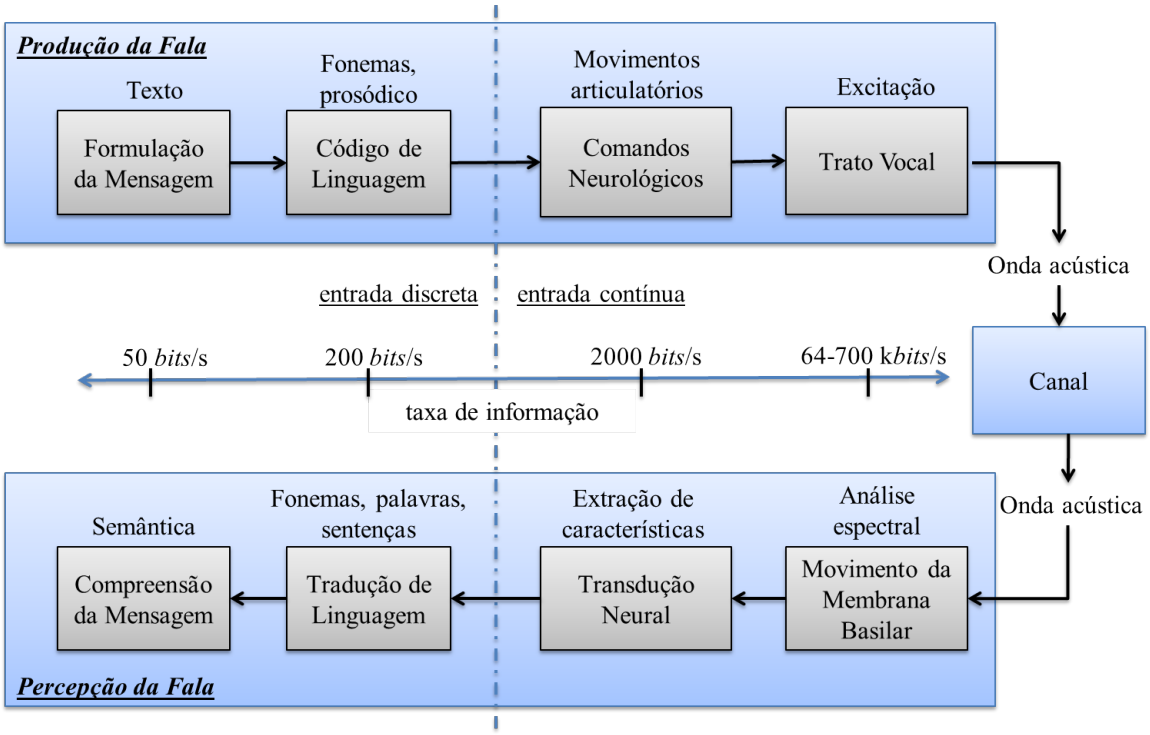

Figura 2.2 Cadeia da fala ilustrando as etapas envolvidas na produção e percepção da fala [1].

A representação da mensagem transmitida pela fala assume diferentes formatos ao longo da produção e percepção, sendo esses formatos informados na parte superior de cada bloco na Figura 2.2. O início da cadeia da fala se dá na origem da mensagem, representada de alguma forma no cérebro do locutor. Para transformar a mensagem em uma onda acústica, o locutor implicitamente faz a conversão para uma representação simbólica da sequência de sons correspondente. Esse passo é representado pelo código de linguagem, que converte os símbolos da forma que podem ser representados textualmente para símbolos fonéticos descrevendo as unidades sonoras da mensagem.

O terceiro passo representa a conversão da informação para uma sequência de controles neuro-musculares, que representam os comandos de movimentos para o sistema articulador da fala. A última etapa do processo de produção da fala é o sistema de trato vocal, responsável pela produção da onda acústica.

O primeiro passo no modelo de percepção da voz, mostrado na parte inferior da Figura 2.2, é a conversão da onda acústica em uma representação espectral, que é feita no interior do ouvido pela membrana basilar, que atua como um analisador de espectro não-uniforme sepa-

rando as componentes espectrais do sinal e fazendo sua análise em um sistema que atua como um banco de filtros não-uniforme [1].

Após a obtenção dos parâmetros espectrais, o sistema de percepção realiza uma transdução neural desses parâmetros em um conjunto de características do áudio, que podem ser decodificadas e processadas pelo cérebro. Em seguida, é feita a conversão desse conjunto de características audíveis obtido em um conjunto de fonemas, palavras e sentenças associadas pelo processo de translação de linguagem no cérebro humano à mensagem recebida. O último passo do modelo é a conversão dos fonemas, palavras e sentenças da mensagem em uma compreensão do significado da mensagem.

2.1.1 Modelagem do Sinal de Voz

A fonte de um sinal de voz pode ser modelada como um sistema linear discreto variante no tempo, conforme ilustrado na Figura 2.3. O gerador do sinal de excitação simula os diferentes modelos de produção do som, podendo ser, por exemplo, um trem de impulsos para sinais sonoros (que possuem periodicidade, como as vogais) e um ruído aleatório para sinais surdos. O sistema linear variante no tempo simula a modulação do som no trato vocal. A saída do sistema corresponde a amostras do sinal de voz [1].

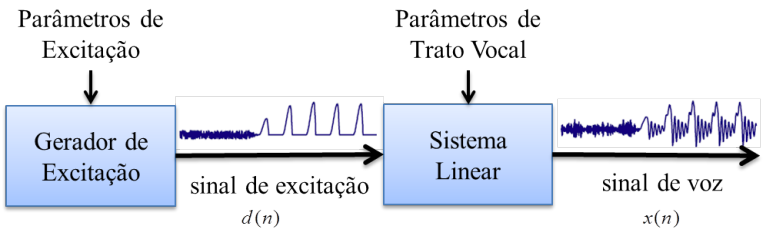

Figura 2.3 Modelo de fonte para o sinal de voz [1].

Como o trato vocal humano possui natureza lentamente variante com o tempo [4], o sistema linear usado na sua modelagem também possui variação lenta. Isto é, as variações nos parâmetros do trato vocal ocorrem a intervalos em torno de 20 ms [21]. Essa é uma das características mais importantes do sinal de voz, pois permite que o processamento seja feito considerando o sinal como estacionário dentro de janelas de tempo de aproximadamente 20 ms (geralmente utilizam-se janelas entre 18 ms e 30 ms [21]).

O sinal periódico de trem de pulsos usado como sinal de excitação para geração de sinais sonoros possui o espaçamento dado pelo valor do período de *pitch* da amostra mais próxima. Esses pulsos são combinados à resposta ao impulso do sistema linear invariante no tempo originado a partir do janelamento por meio de uma convolução.

2.2 Características dos Sistemas de Reconhecimento Automático de Voz

As características dos sistemas de RAV dependem da sua aplicação. Em alguns casos o sistema tem atuação mais específica e pode ser menos complexo. Algumas das principais características dos sistemas de RAV que podem variar são resumidas nas próximas seções.

2.2.1 Dependência de Locutor

Dependendo do banco de dados usado no treinamento do sistema, ele pode ser classificado como dependente ou independente de locutor. Quando o sistema é dependente de locutor, ele é treinado com um banco de dados originados pelo locutor específico. Embora a limitação ao locutor seja um fator que reduz bastante a aplicação do sistema, essa característica facilita o treinamento e otimiza o reconhecimento para o locutor específico.

Além disso, o sistema dependente de locutor pode passar pela etapa de treinamento sempre que for utilizado por um novo usuário. Aplicações comuns para esses sistemas são as interfaces de automóveis e telefones celulares.

Quando são usados sinais de fala originados por muitos locutores, o sistema pode ser capaz de reconhecer a fala independente do locutor. No entanto, além de tornar mais complexo o treinamento, esse procedimento tende a reduzir a taxa de acertos do sistema, pois o torna sensível a variações na forma da oratória.

2.2.2 Vocabulário de Reconhecimento

O vocabulário de reconhecimento é o conjunto de palavras que o sistema de RAV é capaz de reconhecer. O sistema compara o sinal recebido com as palavras do vocabulário e determina quais palavras estão mais próximas da sequência de palavras do sinal recebido.

Quando é feita a comparação entre as características do sinal recebido e as características das palavras disponíveis, quanto maior o número de palavras com características semelhantes no vocabulário de reconhecimento, maior a probabilidade de ocorrência de erros na sentença reconhecida.

Sistemas com vocabulário curto costumam ser usados em aplicações com comandos. Nesses casos, o reconhecedor possui no vocabulário apenas os comandos disponíveis na aplicação. No entanto, vocabulários pequenos restringem as aplicações do sistema e não podem ser usados para reconhecimento automático de texto ou ditados.

Por outro lado, os sistemas com amplo vocabulário permitem o reconhecimento de textos longos, permitindo o reconhecimento de ditado e também a utilização do reconhecedor em contextos diferentes.

2.2.3 *Corpus* de Treinamento

O *corpus* de treinamento corresponde à base de dados usada para treinar os modelos do sistema, consistindo em um conjunto de frases originadas pelos locutores específicos (no caso de dependência de locutor) ou por diferentes locutores. O tamanho e a natureza do *corpus* tem importância essencial no processo de reconhecimento.

A escolha do *corpus* depende do vocabulário de reconhecimento e varia com a dependência de locutor, possuindo amostras correspondentes ao locutor específico ou locutores diversos.

Quanto maior e mais diversificado (maior quantidade de variações linguísticas) o *corpus*, maior será a abragência do reconhecimento.

2.2.4 Natureza da Fala

O banco de dados utilizado pode ser composto por palavras isoladas ou fala contínua. Como existe um pequeno intervalo de silêncio entre as palavras na fala contínua, os vocabulários de palavras isoladas tendem a apresentar melhor resultado no treinamento do modelo [25]. O reconhecimento da fala contínua é mais complexo devido às variações nos tamanhos dos intervalos de silêncio, que podem não ser detectados.

Os reconhecedores para palavras isoladas são indicados para aplicações de comandos, enquanto os de fala contínua são mais usados no reconhecimento de frases e ditado.

2.2.5 Dicionário Fonético

O dicionário fonético de um sistema de reconhecimento consiste em um conjunto de palavras associadas às suas transcrições fonéticas. Os dicionários fonéticos são normalmente desenvolvidos e disponibilizados por grupos de estudos da área [26]. Uma das maiores dificuldades associada à implementação do sistema de RAV em português é a escassez de dicionários para o português brasileiro quando comparado, por exemplo, aos de língua inglesa [27].

As palavras usadas no dicionário fonético determina a quantidade de fonemas isolados e combinações fonéticas que podem ser identificadas pelo reconhecedor. Assim, quanto maior o dicionário fonético, melhor o desempenho no reconhecimento de palavras compostas por fonemas diferentes.

2.3 Estrutura dos Sistemas de Reconhecimento Automático de Voz

Um dos desafios no desenvolvimento de sistemas de RAV é a natureza multidisciplinar do seu conceito, que envolve processamento de sinais, acústica, reconhecimento de padrões,

teoria da informação e comunicação, linguística, fisiologia, ciência da computação e psicologia [4]. Então, é desejável que esses sistemas sejam desenvolvidos de forma modular, ou seja, permitindo a alteração de parâmetros que modifiquem seu desempenho nos diferentes aspectos relacionados a cada uma dessas áreas.

Uma forma de desenvolver sistemas de processamento com subsistemas menos dependentes entre si é tratá-los como sequências de blocos e um fluxo troca de dados entre eles. A estrutura em forma de blocos para sistemas de RAV pode ser descrita como um *front-end*, um sistema de modelagem acústica e um sistema de modelagem linguística, conforme mostrado na Figura 2.4.

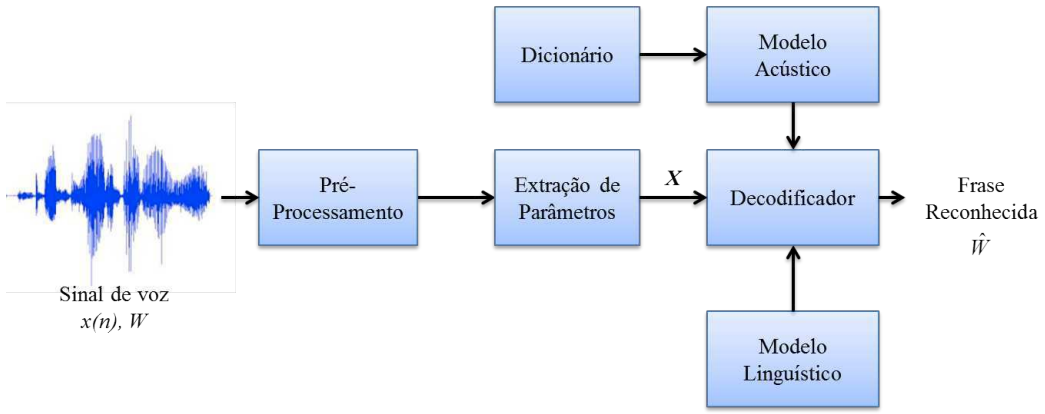

Figura 2.4 Diagrama de blocos para sistema de reconhecimento automático de voz.

O sinal de voz de entrada x é convertido na sequência de vetores de características $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_F\}$, em que F representa o número de quadros do sinal de voz, pelo bloco de extração de parâmetros. Em particular, os coeficientes cepstrais (*Mel-Frequency Cepstrum Coefficients* – MFCCs), são comumente utilizados para representar as características espectrais em curto período [1]. A partir dos vetores característicos é obtida uma representação simbólica que é a sequência de máxima verossimilhança, $\hat{W}$.

O decodificador de um sistema de reconhecimento de voz utiliza um conjunto de modelos acústicos (comumente os modelos escondidos de Markov) associado a um dicionário (léxico) de palavras, que fornece uma probabilidade para cada combinação acústica correspondente a cada sequência proposta. Além disso, um modelo linguístico (N-grama) é usado para calcular a probabilidade associada cada sequência de palavras proposta.

O processo de reconhecimento automático da fala é análogo ao processo de percepção humana da fala, detalhado na Seção 2.1. As funções de cada um dos blocos envolvidos no processo de RAV, apresentados na Figura 2.4, são detalhadas nas subseções seguintes.

2.3.1 Pré-Processamento

A etapa de pré processamento compreende o tratamento do sinal de voz capturado, para produzir o sinal na forma desejável ao sistema de RAV. A primeira operação é a conversão

analógico-digital, responsável pela transformação do sinal de som analógico em seu equivalente digital, por meio dos processos de amostragem, quantização e codificação.

A taxa de amostragem e o número de níveis utilizados nesse processo determinam a qualidade do sinal digitalizado em relação ao seu equivalente analógico. O formato utilizado na digitalização deve ser compatível com o formato aceito na extração de parâmetros.

Durante a conversão, o sinal também é filtrado e as componentes de frequência indesejáveis são descartadas. O ruído presente no sinal capturado pode sobrepor o sinal de voz gerado, dependendo do ambiente de captura. Assim, além da filtragem para extração das componentes de frequência indesejáveis, muitas vezes é necessário aplicar outra técnica de tratamento de ruído sonoro.

2.3.2 Extração de Parâmetros

Como as aplicações dos sistemas de processamento digital de voz estão diretamente ligadas ao mecanismo de produção e percepção humano da fala, esses sistemas levam em consideração as limitações e o funcionamento desse mecanismo. O estudo detalhado do trato vocal e percepção auditiva humanos é apresentado em [4].

É desejável que o sinal de voz seja representado em uma forma paramétrica, que facilite as operações sobre o sinal, como a separação dos fones e a síntese de voz. O processo de extração dos parâmetros não é reversível devido à perda de informação [11].

Não existe uma classe de características da fala que seja definida como padrão para o reconhecimento de voz, sendo utilizadas várias combinações de características acústicas, articulatórias e auditivas [1]. O modelo paramétrico desenvolvido mais popular na literatura é o MFCC, derivado da codificação preditiva linear (*Linear Predictive Coding* – LPC) [1]. Também são usados os modelos de derivada de primeira e segunda ordem dos MFCC.

O cálculo dos MFCCs é feito a partir do cálculo do cepstro, que pode ser definido como transformada inversa de Fourier da magnitude espectral logarítmica de um sinal [1]. Assim, a variável independente do cepstro é o tempo. Esse conceito é usado porque as distâncias cepstrais possuem interpretação equivalente em termos de distância no domínio da frequência, facilitando a modelagem do sistema auditivo humano, que é baseado na análise em frequência. O cálculo do cepstro e mais detalhes sobre seu significado físico são apresentados em [1].

O processamento para extração dos parâmetros é feito a partir de uma operação de janelamento, que segmenta o sinal de voz digitalizado e analisa cada janela separadamente, considerando a possibilidade de sobreposição entre janelas. A operação para cálculo dos parâmetros é realizada em cada janela, ou quadro, sendo aplicada a transformada de Fourier e transformada discreta do cosseno para extração de parâmetros. Esse processo é detalhado em [25], [1] e [28].

Ao final desse processo é obtido um conjunto de MFCCs que caracteriza o sinal de voz. Esses coeficientes são usados na decodificação do sistema de RAV.

2.3.3 Modelagem Acústica

Os primeiros sistemas de reconhecimento funcionavam com base na geração de padrões para fones, palavras ou sequências de palavras, permitindo o reconhecimento por meio do cálculo de distância espectral entre o sinal a ser reconhecido e os padrões. Métodos baseados em programação dinâmica também foram empregados para alinhar os modelos de padrões acústicos com as informações acústicas que seriam reconhecidas [11].

A função da modelagem acústica é associar probabilidades às ocorrências acústicas de uma sequência de palavras, dado o vetor acústico observado. Ou seja, o cálculo da probabilidade de que sequência de vetores acústicos $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_F]$ foi originado pela sequência de palavras $\mathbf{W} = [w_1, w_2, \cdots, w_P]$, em que P é o número total de palavras na sentença.

Atualmente, grande parte dos modelos acústicos utiliza as informações estatísticas dos sinais de voz e a representação das sequências de palavras é normalmente feita por modelos de estados, em que cada unidade fonética representa um estado com características estatísticas da voz, e os estados estão associados uns aos outros pelas transições de estados [1].

Os Modelos Escondidos de Markov (*Hidden Markov Models* – HMMs), que são baseados em cadeias de estados e permitem o uso da relação entre elementos vizinhos [4, 11, 29], são os mais usados na representação dos modelos de unidades acústicas.

Para o treinamento do modelo acústico é necessário fazer diversas gravações de cada unidade do modelo (palavras, fonemas) nos mais variados contextos, para que o método de aprendizado estatístico possa criar distribuições com boa precisão para cada modelo de estado. O treinamento acústico se baseia em sequências de fala que possuem um identificador preciso, sendo essas sequências segmentadas de acordo com suas transcrições fonéticas. Dessa forma, o treinamento envolve segmentação da fala em unidades do modelo de reconhecimento e também a utilização dessa sequência segmentada para construção das distribuições estatísticas do modelo para cada estado das unidades do vocabulário.

2.3.4 Modelagem Linguística

Para que um reconhecedor seja capaz de identificar o sinal de voz, é necessário que disponha de um conjunto de características linguísticas como parâmetro de comparação. Essas informações podem estar na forma de uma gramática ou modelo de linguagem.

A gramática livre de contexto consiste em um conjunto de regras que define o que pode ser reconhecido pelo sistema. Essas regras estão na forma de variáveis seguidas de expressões regulares que descrevem as palavras que podem ser reconhecidas e sua ordenação. Quando o vocabulário é muito extenso, a construção dessas gramáticas é inviável e são utilizados os modelos de linguagem [11].

Os modelos de linguagem têm como objetivo estimar a probabilidade de ocorrência de uma sequência de palavras. Para isso, os modelos se baseiam na medida de verossimilhança da ocorrência de uma dada palavra dentro do contexto no qual o sistema de RAV está aplicado.

Assim, a probabilidade de ocorrência de uma dada sequência de palavras W é nula desde que a sentença expressa seja identificada como incoerente ao contexto.

O treinamento do modelo de linguagem é feito a partir de uma sequência de sentenças textuais representando a sintaxe da fala. Em geral, o texto de treinamento pode ser obtido automaticamente (baseado em um modelo de gramática para a aplicação de reconhecimento) ou utilizando fontes de texto pré-existentes, como artigos de jornais e revistas, legendas televisivas (*closed caption*), entre outros [1]. Em outros casos, o treinamento pode ser criado a partir de bancos de dados. Assim, existem diferentes formas de desenvolver modelos de linguagem para aplicações específicas (dependentes de contexto) [1]:

- Treinamento estatístico a partir de bancos de dados contendo textos transcritos de diálogos pertencentes à aplicação específica (procedimento de aprendizagem);
- Aprendizado a partir das regras formais da gramática associada à aplicação;
- Listagem manual de todas as sequências de texto válidas e associação dos valores de probabilidade a cada sequência.

O método de construção do modelo de linguagem mais utilizado parte de uma gramática estatística N-grama, que é estimada com uso de uma longa sequência de treinamento formada de expressões textuais (pertencentes a um banco de dados genérico que pode ter dependência de contexto ou não) [1].

2.3.5 Decodificador

A função principal do decodificador é a transcrição da amostra de voz para a informação reconhecida. Isto é, para forma de texto ou tomada de decisão, a depender da aplicação. A decodificação é baseada em uma rede de palavras construída a partir dos modelos acústico e linguístico.

O reconhecimento do padrão utiliza um conjunto de modelos acústicos e o dicionário fonético para calcular um escore de comparação para cada sequência proposta. Quando são utilizados HMMs na modelagem acústica, o decodificador calcula o caminho mais provável para as transições de estados. Essa etapa é detalhadamente descrita em [11]. O modelo de linguagem é usado para calcular a probabilidade de ocorrência de cada sequência de palavras.

Ou seja, o decodificador realiza a comparação das características acústicas do sinal com o modelo acústico para obter as probabilidades de ocorrência das palavras associadas ao dicionário. Além disso, faz o cálculo de probabilidade de ocorrência das sentenças descritas no modelo linguístico, obtendo uma estimativa para a sentença de entrada.

2.4 Ferramentas de Implementação

As principais plataformas utilizadas na implementação de sistemas de RAV disponibilizam ferramentas para construção do modelo acústico, do modelo linguístico e da decodificação.

Para construção de modelos acústicos baseados em HMMs, uma ferramenta de livre uso é o HTK (*The Hidden Markov Model Toolkit*) [29], que disponibiliza facilidades para análise do sinal da fala, treinamento dos HMMs e análise de resultados.

O SRILM (*SRI Language Modeling Toolkit*) [30] é uma ferramenta para construção de modelos linguísticos.

O decodificador Julius [27] tem suporte para modelos N-grama e HMMs dependentes de contexto.

Neste trabalho são utilizados modelos acústico e linguístico disponibilizados pelo grupo de pesquisa FalaBrasil [26], que são implementados com HTK e SRILM. Esses modelos são combinados ao decodificador Julius para formar o sistema de reconhecimento de voz para português brasileiro.

2.5 Desafios

O reconhecimento automático de voz ainda é considerado complexo, devido a alguns fatores inerentes às aplicações. Alguns dos problemas envolvidos são listados a seguir [5, 22, 23]:

- Existência de palavras homófonas, ou seja, que possuem a mesma pronúncia, a exemplo de "mais/mas"e "trás/traz".
- As variações linguísticas, ou seja, a pronúncia das palavras difere de acordo com o meio vivenciado pelo locutor. Alguns exemplos são as variações de ordem social e regional.
- Dependendo da velocidade da fala, alguns fonemas podem não ser reconhecidos.
- A determinação automática da transição de fonemas consecutivos (segmentação).
- Como os ambientes nos quais os sistemas de RAV são aplicados não possuem isolamento acústico, a presença de ruído sonoro pode comprometer o funcionamento do sistema.

2.6 Considerações Finais

O desenvolvimento de um sistema de RAV requer o fornecimento de um modelo acústico associado a um dicionário, um modelo linguístico e um decodificador. Esses recursos podem ser desenvolvidos utilizando as ferramentas HTK, SRILM e Julius, respectivamente, desde que sejam fornecidos o dicionário e o *corpus* de treinamento.

Como o *corpus* de treinamento é obtido usualmente em ambientes com SNR elevada, o desempenho do reconhecimento em ambientes com ruído sonoro pode ser degradado. Para ambientes com SNR baixa, como ambientes de fábrica, torna-se necessário aplicar tratamento de ruído no pré-processamento do sinal.

No pré-processamento, diferentes técnicas de redução de ruído sonoro podem ser aplicadas. Os sistemas de RAV podem ser usados como ferramenta de avaliação dos métodos de redução de ruído sonoro, pois sua implementação considera a modelagem do sistema auditivo humano. Assim, eles podem ser utilizados como parâmetro de análise da inteligibilidade do sinal.

CAPÍTULO 3

Técnicas de Tratamento de Ruído

A redução de ruído sonoro com apenas um canal de observação (uma fonte de observação do sinal) se baseia nas diferenças entre as características do sinal de voz e do ruído. Em alguns casos essas diferenças simplificam o tratamento de ruído. Por exemplo, a maior parte da energia do sinal de voz se distribui entre 300 Hz e 3300 Hz [21], então para telefonia qualquer ruído presente em outras faixas de frequência pode ser removido pela aplicação de um filtro passa-faixa. No entanto, na maior parte dos casos, o ruído se apresenta em larga faixa de frequência, sobrepondo-se ao sinal e possui natureza aleatória [6].

Os sinais sonoros, como as vogais, são componentes da fala produzidas pela vibração periódica das cordas vocais. Por outro lado, os sons surdos, produzidos pela modulação de um fluxo de ar, como sussurros, apresentam energia espalhada por uma faixa larga do espectro de frequências [23]. Assim, a separação do sinal de voz e do ruído se torna um problema mais complexo que a simples filtragem em frequência.

As técnicas de redução de ruído sonoro baseadas nas características estatísticas do sinal de voz têm se mostrado mais eficientes [6]. A propriedade estatística mais utilizada é a autocorrelação, ou seu equivalente no domínio da frequência, a densidade espectral de potência [21]. Esses métodos são apresentados e desenvolvidos em [31].

A maioria das técnicas de redução de ruído sonoro podem ser classificadas em quatro tipos básicos [17]: subtração espectral [18], subespaço [19], modelagem estatística [20] e filtragem de Wiener [5].

Os sistemas de reconhecimento de voz são projetados com base nas características dos fonemas e o seu desempenho pode ser comprometido pela presença de ruído sonoro, inerente a todos os ambientes naturais, que não possuem isolamento acústico. Alguns desses ambientes apresentam nível elevado de ruído aditivo, que se sobrepõe ao sinal da fala. Assim, os sistemas de RAV tanto podem ser usados na avaliação de desempenho de técnicas de redução de ruído quanto necessitam do emprego dessa técnicas para garantir seu funcionamento em presença de ruído sonoro elevado.

Neste capítulo são discutidas as técnicas de subtração espectral e filtro de Wiener, que consideram a disponibilidade de apenas um canal. Essas técnicas se baseiam na autocorrelação e densidade espectral de potência do sinal de voz e do ruído.

A princípio é feita a formulação do problema de redução de ruído e a técnica de subtração espectral é então apresentada [18]. Em seguida, são abordadas a subtração espectral de potência [31], que realiza estimativa da magnitude espectral quadrática, e a subtração espectral de potência com compensação [6], na qual é feita redução do efeito sonoro musical introduzido pela subtração.

A técnica de filtragem de Wiener é apresentada considerando o filtro com resposta finita ao impulso e o filtro com resposta infinita ao impulso não causal [15].

Devido às limitações apresentadas pela subtração espectral e o filtro de Wiener quando o nível de ruído é elevado [5], diversas outras técnicas para tratamento de ruído foram desenvolvidas, nesse sentido, alguns trabalhos anteriores são citados e brevemente discutidos. A técnica sub-ótima desenvolvida em [5], que propõe a redução da distorção inserida em detrimento da redução do erro médio quadrático, é detalhada.

3.1 Notação Utilizada e Formulação do Problema

Considerando que o sinal de voz é afetado somente por ruído sonoro aditivo, a n -ésima amostra do sinal resultante é

$$y(n) = x(n) + v(n), \tag{3.1}$$

em que $x(n)$ é a amostra do sinal de voz puro e $v(n)$ é a amostra do ruído no instante de tempo n. Um quadro com K amostras do sinal observado é dado por

$$\mathbf{y} = \mathbf{x} + \mathbf{v}, \tag{3.2}$$

em que $\mathbf{x}$ é um quadro com K amostras do sinal de voz e $\mathbf{v}$ é um quadro com K amostras do ruído sonoro. Ou seja,

$$\mathbf{y} = [y(0) \quad y(2) \cdots y(K-1)], \tag{3.3}$$

$$\mathbf{x} = [x(0) \quad x(2) \cdots x(K-1)] \quad \text{e} \tag{3.4}$$

$$\mathbf{v} = [v(0) \quad v(2) \cdots v(K-1)]. \tag{3.5}$$

O problema de tratamento de ruído consiste no cálculo de uma estimativa $\hat{\mathbf{x}}$ para o quadro com K amostras do sinal de voz a partir do quadro $\mathbf{y}$ observado.

3.2 Subtração Espectral de Potência

Na técnica de subtração espectral de potência [18] a estimativa do sinal de voz é feita com base no conhecimento *a priori* de uma estimativa de potência espectral do ruído. Como o sinal de voz não é estacionário, considera-se a multiplicação do sinal observado por um sinal tipo janela limitante no tempo, originando o sinal dado na forma vetorial pela Equação 3.2. Os quadros obtidos no janelamento devem possuir cerca de 20 ms do sinal observado, garantindo assim que o sinal de voz seja aproximadamente estacionário, conforme discutido no Capítulo 2.

Aplicando a transformada de Fourier de tempo discreto (*discrete-time Fourier transform* – DTFT) ao quadro $\mathbf{y}$, que possui K amostras do sinal observado, o espectro do sinal de saída correspondente é obtido em termos dos espectros do sinal de voz $X(\Omega)$ e de ruído $V(\Omega)$, como

$$Y(\Omega) = X(\Omega) + V(\Omega), \tag{3.6}$$

em que

$$\mathbf{y} \longleftrightarrow Y(\Omega), \text{ sendo } Y(\Omega) = \sum_{l=0}^{K-1} y(l)e^{-j\Omega l}. \tag{3.7}$$

Assim, a magnitude espectral quadrática do sinal observado é [21]

$$\begin{aligned} |Y(\Omega)|^2 &= Y(\Omega)Y^*(\Omega) && (3.8)\\ &= |X(\Omega)|^2 + |V(\Omega)|^2 + X^*(\Omega)V(\Omega) + X(\Omega)V^*(\Omega). && (3.9) \end{aligned}$$

Aplicando o operador valor esperado $\mathrm{E}[\cdot]$,

$$\mathrm{E}[|Y(\Omega)|^2] = \mathrm{E}[|X(\Omega)|^2] + \mathrm{E}[|V(\Omega)|^2] + \mathrm{E}[X^*(\Omega)V(\Omega)] + \mathrm{E}[X(\Omega)V^*(\Omega)]. \tag{3.10}$$

Considerando ausência de eco, o sinal de voz é descorrelacionado do ruído. Como a média do ruído é nula, tem-se

$$\mathrm{E}[X^*(\Omega)V(\Omega)] = \mathrm{E}[X(\Omega)V^*(\Omega)] = 0. \tag{3.11}$$

Substituindo na Equação 3.10,

$$\mathrm{E}[|Y(\Omega)|^2] = \mathrm{E}[|X(\Omega)|^2] + \mathrm{E}[|V(\Omega)|^2]. \tag{3.12}$$

Assim, é possível calcular a estimativa para a magnitude quadrática do sinal de voz

$$|\hat{X}(\Omega)|^2 = \begin{cases} |Y(\Omega)|^2 - \mathrm{E}[|V(\Omega)|^2], & \text{se } |Y(\Omega)|^2 - \mathrm{E}[|V(\Omega)|^2] \geq 0, \\ 0, & \text{caso contrário.} \end{cases} \tag{3.13}$$

É possível obter $\mathrm{E}[|V(\Omega)|^2]$ a partir dos quadros de sinal observado para os quais seja detectada ausência de voz.

A estimativa para a fase de $X(\Omega)$ é tomada como a fase do sinal observado, ou seja,

$$e^{j\hat{\theta}_X} = \frac{Y(\Omega)}{|Y(\Omega)|}. \tag{3.14}$$

Assim, a estimativa para a transformada discreta de Fourier do sinal de voz é

$$\hat{X}(\Omega) = |\hat{X}(\Omega)|\frac{Y(\Omega)}{|Y(\Omega)|}. \tag{3.15}$$

As estimativas de amostras do sinal de voz para o quadro podem ser calculadas pela transformada inversa

$$\hat{x}(i) = \frac{1}{N_{dtft}} \sum_{b=0}^{N_{dtft}-1} X(\Omega) \cdot e^{j\Omega b}. \tag{3.16}$$

Esse desenvolvimento considera que o ruído do ambiente permanece estacionário durante um intervalo de tempo suficiente para cálculo da média de magnitude espectral quadrática do ruído no período de silêncio e utilização dessa média como estimativa de magnitude espectral quadrática no intervalo subsequente [6].

Conforme pode ser observado na Equação 3.13, a estimativa de magnitude espectral obtida com a subtração espectral pode apresentar muitos valores nulos quando a magnitude espectral do ruído é elevada. Esse processo origina muitos valores nulos na magnitude espectral do sinal de voz estimado, gerando picos que são responsáveis pelo surgimento de um efeito sonoro desagradável chamado ruído musical [6].

3.2.1 Subtração Espectral com Compensação

Para o modelo proposto na Equação 3.13, como a magnitude não pode assumir valores negativos, a ocorrência de muitos valores nulos no sinal obtido produz um efeito sonoro musical indesejado. Um modelo alternativo é proposto pela inserção de dois parâmetros para reduzir esse efeito sonoro [6], sendo $\beta_1 > 1$ responsável pela minimização da ocorrência de valores negativos e $0 < \beta_2 \ll 1$ o limitante inferior que reduz a percepção dos pontos que seriam igualados a zero. A estimativa da função espectral de potência do sinal é então escrita como [6]

$$|\hat{X}^2(\Omega)| = \begin{cases} |Y^2(\Omega)| - \beta_1 \mathrm{E}[|V(\Omega)|^2], & \text{se} \quad |Y^2(\Omega)| - \mathrm{E}[|V(\Omega)|^2] \geq \beta_2 |\mathrm{E}[|V(\Omega)|^2], \\ \beta_2 \mathrm{E}[|V(\Omega)|^2], & \text{caso contrário.} \end{cases} \tag{3.17}$$

Novamente a fase é aproximada pela fase do sinal observado. Então,

$$\hat{X}(\Omega) = |\hat{X}(\Omega)|\frac{Y(\Omega)}{|Y(\Omega)|}, \tag{3.18}$$

e o sinal de voz é aproximado pela transformada inversa de Fourier.

Esse método reduz o ruído musical, mas as técnicas de subtração espectral degradam o sinal sob ruído com magnitude espectral elevada [6]. A fase do sinal de voz é aproximada pela fase do sinal observado, e como a fase do espectro de ruído é diferente da fase do sinal de voz, o erro na estimativa de fase se eleva bastante quando o ruído possui magnitude de ordem próxima à magnitude do sinal de voz.

3.3 Filtro de Wiener

O filtro de Wiener foi desenvolvido por Norbert Wiener na década de 1940 como uma solução para o problema de estimativa de um sinal aleatório desejado por meio do processamento de outro sinal aleatório associado. O trabalho de Norbert Wiener descrevendo o tratamento de sinais no domínio contínuo no tempo foi publicado em 1949 [32]. O equivalente para sinais no domínio discreto foi desenvolvido por Andrey Kolmogorov e publicado em 1941 [33], por isso a teoria também é conhecida como filtragem de Wiener-Kolmogorov. No presente capítulo são tratados os conceitos aplicáveis aos sinais discretos [5].

O filtro de Wiener é considerado ótimo do ponto de vista do erro médio quadrático (*Minimum Mean Squared Error* – MMSE) relativo ao sinal estimado e o sinal de voz original, tendo sido o primeiro filtro desenvolvido em termos estatísticos. O desenvolvimento considera que os sinais de entrada são estacionários em sentido amplo e, por isso, trata-se de um filtro não adaptativo. O problema consiste em calcular uma estimativa $\hat{x}(n)$ para a amostra do sinal desejado com base no critério de MMSE. Três abordagens podem ser feitas [15]:

a) Filtro causal com resposta finita ao impulso (*Finite Impulse Response* – FIR);

b) Filtro não-causal com resposta infinita ao impulso (*Infinite Impulse Response* – IIR);

c) Filtro causal com resposta infinita ao impulso.

O desenvolvimento para o terceiro caso é omitido neste trabalho porque apresenta solução por fatoração, que é tratada em [15]. Nos tópicos a seguir é mostrada a solução para os dois primeiros casos. Um estudo detalhado sobre o filtro de Wiener é feito em [34] e [15].

3.3.1 Filtro de Wiener com Resposta Finita ao Impulso (FIR)

Considerando a n-ésima amostra para o sinal de voz com ruído dada na Equação 3.1, a amostra estimada $\hat{x}(n)$ para o sinal de voz, obtida a partir do filtro FIR de comprimento L, definido pelos coeficientes h_i, é dada por

$$\hat{x}(n) = \sum_{i=0}^{L-1} h_i y(n-i). \tag{3.19}$$

Na forma vetorial, $\hat{x}(n) = \mathbf{h}^T\mathbf{y}(n)$ e o sinal de erro é definido como

$$e_X(n) \triangleq x(n) - \hat{x}(n) = x(n) - \mathbf{h}^T\mathbf{y}(n), \tag{3.20}$$

em que o sobrescrito ^{T} denota o operador transposto, o vetor $\mathbf{h}$ representa o filtro FIR, dado por

$$\mathbf{h} = [h_0 \quad h_1 \quad \cdots \quad h_{L-1}]^T \tag{3.21}$$

e $\mathbf{y}(n)$ é o vetor contendo a amostra atual e as $L-1$ amostras anteriores de $y(n)$

$$\mathbf{y}(n) = [y(n) \quad y(n-1) \quad \cdots \quad y(n-L+1)]^T. \tag{3.22}$$

O erro médio quadrático (*Mean Squared Error* – MSE) associado à amostra do sinal de voz $x(n)$ e sua estimativa $\hat{x}(n)$ é definido por meio do operador valor esperado $\mathrm{E}[\cdot]$ como

$$\begin{aligned} \mathrm{E}[e_X^2(n)] &= \mathrm{E}[\big(x(n) - \hat{x}(n)\big)^2] \\ &= \mathrm{E}[x^2(n)] + \mathrm{E}[\hat{x}^2(n)] - 2\mathrm{E}[x(n)\hat{x}(n)] \\ &= \mathrm{E}[x^2(n)] + \mathrm{E}[(\mathbf{h}^T\mathbf{y}(n))^2] - 2\mathrm{E}[x(n)\mathbf{h}^T\mathbf{y}(n)]. \end{aligned} \tag{3.23}$$

Assim, o erro médio quadrático (MSE) é uma função da amostra n e do filtro $\mathbf{h}$ escolhido. O critério MSE é definido em função do filtro e denotado por

$$J_X(\mathbf{h}) \triangleq \mathrm{E}[e_X^2(n)]. \tag{3.24}$$

Para o filtro particular $\mathbf{u}_1 = [1 \quad 0 \quad \cdots \quad 0]^T$, a estimativa obtida na saída é

$$\hat{x}(n) = [1 \quad 0 \quad \cdots \quad 0] \cdot \mathbf{y}(n) = y(n). \tag{3.25}$$

Assim, o filtro $\mathbf{u}_1$, que corresponde a um impulso unitário discreto, não altera a entrada e o MSE correspondente é

$$J_X(\mathbf{u}_1) = \mathrm{E}\big[\big(x(n) - \mathbf{u}_1^T\mathbf{y}(n)\big)^2\big] = \mathrm{E}\big[\big(x(n) - y(n)\big)^2\big] \tag{3.26}$$

$$= \mathrm{E}[v^2(n)]. \tag{3.27}$$

Considerando que a média do ruído é nula, $\mathrm{E}[v^2(n)]$ é igual à variância do ruído σ_V^2 e corresponde ao limite superior para o comportamento erro médio quadrático mínimo, ou seja,

$$J_X(\mathbf{h_o}) < J_X(\mathbf{u_1}) = \sigma_V^2. \tag{3.28}$$

Os coeficientes do filtro de Wiener $\mathbf{h_o}$ produzem minimização do MSE e levam à obtenção da estimativa ótima $\hat{x}_o(n)$. Eles são obtidos a partir do argumento que minimiza o critério

$$\mathbf{h_o} = \texttt{arg} \quad \texttt{min}_{\mathbf{h}} J_X(\mathbf{h}). \tag{3.29}$$

A derivada parcial do erro médio quadrático em relação ao filtro na forma escalar é

$$\frac{\partial}{\partial h_i}\mathrm{E}[e^2(n)] = 2\sum_{j=0}^{L-1}\mathrm{E}[y(n-j)y(n-i)]h_j - 2\mathrm{E}[y(n-i)x(n)], \tag{3.30}$$

sendo $i = 0, 1, 2, \cdots, L$. Na forma vetorial,

$$\frac{d}{d\mathbf{h}}J_X(\mathbf{h}) = 2\mathrm{E}[\mathbf{y}(n)\mathbf{y}^T(n)]\mathbf{h} - 2\mathrm{E}[\mathbf{y}(n)x(n)]. \tag{3.31}$$

Para a minimização, a derivada é nula, e então o filtro ótimo $\mathbf{h_o}$ satisfaz

$$\sum_{j=0}^{L}\mathrm{E}[y(n-j)y(n-i)]h_j = \mathrm{E}[y(n-i)x(n)], \tag{3.32}$$

ou seja,

$$\mathrm{E}[\mathbf{y}(n)\mathbf{y}^T(n)]\mathbf{h_o} = \mathrm{E}[\mathbf{y}(n)x(n)]. \tag{3.33}$$

A matriz de autocorrelação do sinal observado é dada por

$$\mathbf{R}_Y = \mathrm{E}[\mathbf{y}(n)\mathbf{y}^T(n)], \tag{3.34}$$

e o vetor de correlação entre o vetor das amostras observadas e a amostra de voz por

$$\mathbf{p} = \mathrm{E}[\mathbf{y}(n)x(n)]. \tag{3.35}$$

A equação usada para desenvolver a solução da Equação 3.29, conhecida como Wiener-Hopf, pode ser escrita na forma vetorial como

$$\mathbf{R}_Y\mathbf{h_o} = \mathbf{p}. \tag{3.36}$$

Como o sinal de voz no modelo $y(n) = x(n) + v(n)$ não é observável, não é possível calcular a correlação $\mathbf{p}$, que deve ser desenvolvida em termos do sinal observado e do ruído,

$$\begin{aligned}
\mathbf{p} &= \mathrm{E}[\mathbf{y}(n)x(n)] \\
&= \mathrm{E}\big[\mathbf{y}(n)\big(y(n) - v(n)\big)\big] \\
&= \mathrm{E}[\mathbf{y}(n)y(n)] - \mathrm{E}[\mathbf{y}(n)v(n)] \\
&= \mathrm{E}[\mathbf{y}(n)y(n)] - \mathrm{E}\big[\big(\mathbf{x}(n) + \mathbf{v}(n)\big)v(n)\big] \\
&= \mathrm{E}[\mathbf{y}(n)y(n)] - \mathrm{E}[\mathbf{x}(n)v(n)] - \mathrm{E}[\mathbf{v}(n)v(n)]. \qquad (3.37)
\end{aligned}$$

Sendo $\mathbf{x}(n) = [x(n) \quad x(n-1) \cdots x(n-L+1)]$ e $\mathbf{v}(n) = [v(n) \quad v(n-1) \cdots v(n-L+1)]$.

Considerando que o sinal de voz e o ruído são independentes e que a média do ruído é nula, $\mathrm{E}[\mathbf{x}(n)v(n)] = 0$. Assim,

$$\mathbf{p} = \mathbf{r}_Y - \mathbf{r}_V, \qquad (3.38)$$

em que $\mathbf{r}_Y$ e $\mathbf{r}_V$ são os vetores de autocorrelação do sinal observado e do ruído

$$\mathbf{r}_Y = \mathrm{E}[\mathbf{y}(n)y(n)], \quad \mathbf{r}_V = \mathrm{E}[\mathbf{v}(n)v(n)]. \qquad (3.39)$$

Substituindo na Equação 3.36,

$$\begin{aligned}
\mathbf{h_o} = \mathbf{R}_Y^{-1}\mathbf{p} &= \mathbf{R}_Y^{-1}(\mathbf{r}_Y - \mathbf{r}_V) \\
&= \mathbf{R}_Y^{-1}\mathbf{r}_Y - \mathbf{R}_Y^{-1}\mathbf{r}_V. \qquad (3.40)
\end{aligned}$$

Como $\mathbf{r}_Y$ corresponde à primeira coluna de $\mathbf{R}_Y$, $\mathbf{R}_Y^{-1}\mathbf{r}_Y = \mathbf{u}_1$ e

$$\begin{aligned}
\mathbf{h_o} &= \mathbf{u}_1 - \mathbf{R}_Y^{-1}\mathbf{r}_V \qquad (3.41) \\
&= [\mathbf{I} - \mathbf{R}_Y^{-1}\mathbf{R}_V]\mathbf{u}_1. \qquad (3.42)
\end{aligned}$$

A matriz de autocorrelação $\mathbf{R}_Y$ pode ser escrita como

$$\begin{aligned}
\mathbf{R}_Y &= \mathrm{E}[\mathbf{y}(n)\mathbf{y}^T(n)] \\
&= \mathrm{E}\big[\big(\mathbf{x}(n) + \mathbf{v}(n)\big)\big(\mathbf{x}(n) + \mathbf{v}(n)\big)^T\big] \\
&= \mathrm{E}[\mathbf{x}(n)\mathbf{x}^T(n)] + \mathrm{E}[\mathbf{x}(n)\mathbf{v}^T(n)] = \mathrm{E}[\mathbf{v}(n)\mathbf{x}^T(n)] + \mathrm{E}[\mathbf{v}(n)\mathbf{v}^T(n)] \\
&= \mathrm{E}[\mathbf{x}(n)\mathbf{x}^T(n)] + \mathrm{E}[\mathbf{v}(n)\mathbf{v}^T(n)] \\
&= \mathbf{R}_X + \mathbf{R}_V. \qquad (3.43)
\end{aligned}$$

A relação sinal-ruído é dada em termos das variâncias do sinal de voz σ_X^2 e de ruído σ_V^2 como

$$\gamma \triangleq \frac{\sigma_X^2}{\sigma_V^2}. \tag{3.44}$$

O MMSE e o MMSE normalizado são dados por

$$J_X(\mathbf{h_o}) = \sigma_X^2 - \mathbf{p}^T\mathbf{h_o} = \sigma_V^2 - \mathbf{r}_V^T\mathbf{h_o}, \tag{3.45}$$

$$\tilde{J}_X(\mathbf{h_o}) = \frac{J_X(\mathbf{h_o})}{\sigma_V^2}. \tag{3.46}$$

É possível observar que quando a SNR aumenta muito, ou seja, quando o sinal observado é muito próximo do sinal original, o filtro $\mathbf{h_o}$ tende para o filtro impulsivo $\mathbf{u}_1$. A distorção inserida pelo filtro é analisada na Subseção 3.4.1.

3.3.2 Filtro de Wiener Não Causal IIR

Assumindo que a estimativa $\hat{x}(n)$ para a n-ésima amostra o sinal de voz pode ser dependente de todas as amostras do sinal de observação e não apenas de um vetor finito de amostras, o filtro tem resposta infinita ao impulso (*Infinite Impulse Response* – IIR) com coeficientes h_i e a saída $\hat{x}(n)$ é dada em termos da entrada $y(n)$ como

$$\hat{x}(n) = \sum_{i=-\infty}^{\infty} h_i y(n-i). \tag{3.47}$$

Analogamente ao filtro de Wiener FIR, na forma vetorial, $\hat{x}(n) = \mathbf{h}^T\mathbf{y}(n)$ e o sinal de erro é definido como

$$e_X(n) \triangleq x(n) - \hat{x}(n) = x(n) - \mathbf{h}^T\mathbf{y}(n), \tag{3.48}$$

em que o sobrescrito ^{T} denota o operador transposto, o vetor $\mathbf{h}$ é um filtro com resposta infinita ao impulso (IIR) e $\mathbf{y}(n)$ e o vetor de todas as amostras do sinal de observação.

O MSE é definido com uso do operador valor esperado $\mathrm{E}[\cdot]$ como

$$\mathrm{E}[e_X^2(n)] = \mathrm{E}[x^2(n)] + \mathrm{E}[(\mathbf{h}^T\mathbf{y}(n))^2] - 2\mathrm{E}[x(n)\mathbf{h}^T\mathbf{y}(n)]. \tag{3.49}$$

O erro médio quadrático é uma função de n e depende do filtro $\mathbf{h}$ escolhido e o critério MSE é definido em função do filtro

$$J_X(\mathbf{h}) \triangleq \mathrm{E}[e_X^2(n)]. \tag{3.50}$$

A minimização do erro médio quadrático em termos do filtro $\mathbf{h}$ corresponde ao filtro ótimo $\mathbf{h_o}$. Conforme discutido no tópico anterior, a minimização resulta nas equações de Wiener-Hopf

$$\sum_{l=-\infty}^{\infty} \mathrm{E}[y(n-l)y(n-i)]h_l = \mathrm{E}[y(n-i)x(n)], \tag{3.51}$$

em que $i \in \mathbb{Z}$.

Na forma vetorial,

$$\mathrm{E}[\mathbf{y}(n)\mathbf{y}^T(n)]\mathbf{h} = \mathrm{E}[\mathbf{y}(n)x(n)]. \tag{3.52}$$

Aplicando a transformada de Fourier de tempo discreto [35],

$$S_Y(\Omega)H_o(\Omega) = S_{XY}(\Omega), \tag{3.53}$$

em que S_Y é a densidade espectral de potência de y e S_{XY} é a densidade espectral de potência cruzada, dada pela DTFT de $\mathbf{p}$.

Assim, o filtro de Wiener possui uma resposta em frequência

$$H_o(\Omega) = \frac{S_{XY}(\Omega)}{S_Y(\Omega)} \tag{3.54}$$

e o sinal estimado é dado por

$$\hat{X}(\Omega) = Y(\Omega)H_o(\Omega), \tag{3.55}$$

Para $\mathbf{y} = \mathbf{x} + \mathbf{v}$, em que os processos $\mathbf{x}$ e $\mathbf{y}$ são independentes, essa função de transferência é [36]

$$H_o(\Omega) = \frac{S_X(\Omega)}{S_X(\Omega) + S_V(\Omega)}, \tag{3.56}$$

em que $S_X(\Omega)$ e $S_V(\Omega)$ são as funções densidade espectral de potência do sinal de voz isolado e do ruído, respectivamente, que são obtidas a partir da transformada de Fourier das funções de autocorrelação. Assim, a implementação do filtro de Wiener requer conhecimento do comportamento estatístico dos sinais de voz e de ruído, que são processos aleatórios.

A resposta em frequência do filtro de Wiener não causal é degradada quando a magnitude do ruído se torna muito elevada em relação à magnitude do sinal de voz. Observa-se que

$$S_X(\Omega) \gg S_V(\Omega) \longrightarrow H_o(\Omega) \approx 1, \quad \text{e} \tag{3.57}$$

$$S_X(\Omega) \ll S_V(\Omega) \longrightarrow H_o(\Omega) \approx 0. \tag{3.58}$$

Assim, a degradação do sinal para SNR baixa pode ser justificada pela análise da magnitude da resposta em frequência do filtro dada na Equação 3.56. Além disso, a fase do sinal obtido como estimativa é aproximada pela fase do sinal observado. Quando o ruído aumenta,

a fase do sinal observado se distancia do sinal de voz original e a informação de fase não pode ser recuperada.

3.4 Trabalhos Anteriores

Como as características e a natureza do ruído sonoro dependem do ambiente, é muito difícil desenvolver um algoritmo universal para tratamento de ruído sonoro [5]. Além disso, os sistemas de processamento digital de voz podem alterar de diversas formas o sinal originado pelo locutor. Para o caso das técnicas de redução de ruído, os principais efeitos observados são a distorção e o surgimento de efeitos sonoros musicais.

As limitações das técnicas consideradas clássicas levou a diversos estudos para solucionar o problema de tratamento de ruído sonoro. Os principais esforços são no sentido de desenvolver um método de filtragem adaptativa, ou seja, que seja aplicável a casos em que os sinais envolvidos não são estacionários, como o sinal de voz. Além disso, verificou-se a necessidade de reduzir a distorção inserida pelo processo de filtragem.

Em [37] é proposta uma técnica de redução de ruído sonoro baseada na combinação da subtração espectral e filtro de Wiener, a fim de reduzir os efeitos de distorção inseridos. O filtro implementado possui dois estágios, consistindo na combinação dessas duas técnicas em série. Primeiramente é feita a estimativa da densidade espectral de potência do ruído e do sinal de voz pelo uso da subtração espectral e, em seguida, essas estimativas são usadas para obtenção da resposta em frequência do filtro de Wiener. O desempenho do filtro é avaliado pela razão logarítmica de verossimilhança (*Log-Likelihood Ratio* – LLR), que é uma medida de avaliação objetiva baseada na técnica de predição linear, indicando melhoria com relação aos filtros de subtração espectral e de Wiener isolados. A avaliação subjetiva também indicou melhoria na qualidade da voz obtida.

Em [16] é desenvolvida uma técnica de filtragem adaptativa baseada no filtro de Wiener IIR não causal. Nessa técnica as mudanças estatísticas do sinal de voz (média e variância) são levadas em consideração, e a função de transferência do filtro é calculada a cada amostra. Os resultados obtidos são comparados com o filtro de subtração espectral simples e com o filtro de Wiener. No entanto, apenas a SNR é observada, não havendo avaliação da inteligibilidade dos sinais filtrados.

Em [38] é feita a implementação do Filtro de Kalman [15], que utiliza modelo de equações de estados, aplicado à redução de ruído sonoro em voz. O desempenho é comparado ao das técnicas de filtragem de Wiener e subtração espectral, indicando que a SNR obtida é superior para casos em que o ruído não é estacionário para os intervalos considerados no janelamento do sinal. No entanto, a implementação do filtro de Kalman é mais complexa, por se tratar de um filtro iterativo.

Alguns trabalhos também foram desenvolvidos com o emprego da transformada *wavelet* [39, 40].

Na subseção seguinte é detalhada a técnica desenvolvida em [5], que propõe uma modificação no filtro Wiener FIR para redução da distorção.

3.4.1 Filtro de Wiener FIR Modificado

O filtro de Wiener é projetado para possibilitar a minimização do erro médio quadrático entre o sinal estimado e o sinal original. No entanto, pode degradar o sinal de voz, afetando sua qualidade e inteligibilidade devido ao efeito de inserção de distorção. Em [5] é desenvolvida uma técnica de filtragem baseada no filtro de Wiener FIR (apresentado na Seção 3.3.1), chamado de filtro sub-ótimo porque não minimiza o MSE, mas permite a redução da distorção do sinal. O MSE pode ser observado em termos da relação sinal-ruído, enquanto a distorção é medida por um índice de distorção de voz.

Estimativa do Ruído

O sinal de erro entre as amostras de ruído e de estimativa de ruído é dado por

$$e_V(n) \triangleq v(n) - \hat{v}(n) = v(n) - \mathbf{g}^T\mathbf{y}(n), \tag{3.59}$$

em que

$$\mathbf{g} = [g_0 \quad g_1 \quad \cdots \quad g_{L-1}]^T \tag{3.60}$$

é o filtro FIR de Wiener para estimativa do ruído a partir do vetor de L amostras observadas $\mathbf{y}(n)$. O critério MSE associado é

$$J_V(\mathbf{g}) \triangleq \mathrm{E}[e_V^2(n)]. \tag{3.61}$$

A minimização do erro médio quadrático, que deve levar à atenuação do sinal de voz, é feita a partir das equações de Wiener-Hopf dadas na Equação 3.36, conforme desenvolvido na Seção 3.3.1,

$$\mathbf{R}_Y\mathbf{g_o} = \mathbf{q}, \tag{3.62}$$

em que $\mathbf{q} = \mathrm{E}[\mathbf{y}(n)v(n)]$. Desenvolvendo,

$$\begin{aligned} \mathbf{q} = \mathrm{E}[\mathbf{y}(n)v(n)] &= \mathrm{E}[(\mathbf{x}(n) + \mathbf{v}(n))v(n)] \\ &= \mathrm{E}[\mathbf{x}(n)v(n)] + \mathrm{E}[\mathbf{v}(n)v(n)] \\ &= \mathbf{r}_V. \end{aligned} \tag{3.63}$$

O filtro ótimo é então obtido

$$\begin{aligned}\mathbf{g_o} = \mathbf{R}_Y^{-1}\mathbf{r}_V &= \mathbf{R}_Y^{-1}\mathbf{R}_V\mathbf{u}_1 \\ &= \left[\gamma\cdot\mathbf{I} + \tilde{\mathbf{R}}_X^{-1}\tilde{\mathbf{R}}_V\right]^{-1}\tilde{\mathbf{R}}_X^{-1}\tilde{\mathbf{R}}_V\mathbf{u}_1. \end{aligned} \tag{3.64}$$

O MMSE e o MMSE normalizado são

$$J_V(\mathbf{g_o}) = \sigma_V^2 - \mathbf{q}^T\mathbf{g_o} = \sigma_V^2 - \mathbf{r}_V^T\mathbf{g_o}, \tag{3.65}$$

$$\tilde{J}_V(\mathbf{g_o}) = \frac{J_V(\mathbf{g_o})}{\sigma_X^2}. \tag{3.66}$$

Relação entre Redução de Ruído e Distorção

É possível relacionar a estimativa do sinal de voz com a do sinal de ruído,

$$\mathbf{h_o} = \mathbf{u}_1 - \mathbf{g_o}. \tag{3.67}$$

A minimização de $J_X(\mathbf{h})$ ou $J_V(\mathbf{h}-\mathbf{u}_1)$ com respeito a $\mathbf{h}$ é equivalente. Analogamente, a minimização de $J_V(\mathbf{g})$ ou $J_X(\mathbf{u}_1-\mathbf{g})$ com respeito a $\mathbf{g}$ também é equivalente. Isso significa que a filtragem para obter a estimativa do sinal de voz pode ser feita pelo uso direto do filtro com coeficientes $\mathbf{h_o}$, para o qual é necessário cálculo da estimativa do vetor de autocorrelação do ruído e matriz de autocorrelação do sinal observado, ou pelo uso do filtro com coeficientes $\mathbf{g_o}$, que permite a obtenção de uma estimativa para o ruído a partir de uma estimativa para o vetor de autocorrelação do sinal de voz e matriz de autocorrelação do sinal observado.

Para o caso ótimo, o sinal de erro entre a estimativa calculada e o sinal original é [5]

$$\begin{aligned} e_{X,o}(n) = x(n) - \hat{x}(n) &= x(n) - \mathbf{h_o}^T\mathbf{y}(n) \\ &= x(n) - [\mathbf{u}_1 - \mathbf{g_o}]^T[\mathbf{x}(n) + \mathbf{v}(n)] \\ &= x(n) - \mathbf{u}_1^T\mathbf{x}(n) - \mathbf{u}_1^T\mathbf{v}(n) + \mathbf{g_o}^T\mathbf{x}(n) + \mathbf{g_o}^T\mathbf{v}(n) \\ &= x(n) - x(n) - v(n) + \mathbf{g_o}^T[\mathbf{x}(n) + \mathbf{v}(n)] \\ &= -v(n) + \mathbf{g_o}^T\mathbf{y}(n) = -e_{V,o}(n). \end{aligned} \tag{3.68}$$

A partir da Equação 3.45 e da Equação 3.65 se observa que os MMSEs são iguais, mas a relação entre os MMSEs normalizados é

$$\tilde{J}_V(\mathbf{g_o}) = \frac{J_V(\mathbf{g_o})}{\sigma_X^2} = \frac{\tilde{J}_X(\mathbf{h_o})}{\gamma}. \tag{3.69}$$

A potência do sinal estimado com filtro de Wiener é

$$\begin{aligned} \mathrm{E}[\hat{x}_o^2(n)] &= \mathrm{E}[\mathbf{h_o}^T\mathbf{y}(n)\mathbf{y}^T(n)\mathbf{h_o}] \\ &= \mathbf{h_o}^T\mathbf{R}_Y\mathbf{h_o} = \sigma_X^2 - J_X(\mathbf{h_o}) \\ &= \mathbf{h_o}^T\mathbf{R}_X\mathbf{h_o} + \mathbf{h_o}^T\mathbf{R}_V\mathbf{h_o}, \end{aligned} \tag{3.70}$$

que corresponde à soma da potência do sinal de voz atenuado (primeiro termo) e do ruído residual (segundo termo). Observa-se que a redução do ruído implica também a redução do sinal de voz por $J_X(\mathbf{h_o}) + \mathbf{h_o}^T\mathbf{R}_v\mathbf{h_o}$ e consequente distorção inevitável, pois partes do sinal de voz são atenuadas.

O índice de distorção de voz devido à filtragem de Wiener é definido como [5]

$$\begin{aligned} v_D(\mathbf{g_o}) &\triangleq \frac{\mathrm{E}\left[\left(x(n) - \mathbf{h_o}^T\mathbf{x}(n)\right)^2\right]}{\sigma_X^2} \\ &= \frac{\mathbf{g_o}^T\mathbf{R}_X\mathbf{g_o}}{\sigma_X^2} = \frac{1}{\gamma}\left[\tilde{J}_X(\mathbf{h_o}) - \mathbf{h_o}^T\tilde{\mathbf{R}}_V\mathbf{h_o}\right] \\ &< \tilde{J}_X(\mathbf{g_o}). \end{aligned} \tag{3.71}$$

Para o filtro ótimo, esse índice varia entre 0 e 1 [5], sendo a distorção maior quando $v_D(\mathbf{g_o})$ está próximo de 1 e menor quando está próximo de 0. Além disso,

$$\lim_{\gamma\to 0} v_D(\mathbf{g_o}) = 1, \tag{3.72}$$

$$\lim_{\gamma\to \infty} v_D(\mathbf{g_o}) = 0, \tag{3.73}$$

ou seja, para valores baixos de SNR, o filtro de Wiener insere muita distorção no sinal, podendo não ser aplicável.

O fator de redução de ruído é definido como [5]

$$\begin{aligned} \xi_R(\mathbf{h_o}) &\triangleq \frac{\sigma_V^2}{\mathrm{E}\left[\left(\mathbf{h_o}^T\mathbf{R}_V\mathbf{h_o}\right)^2\right]} \\ &= \frac{\sigma_V^2}{\mathbf{h_o}^T\mathbf{R}_V\mathbf{h_o}} = \frac{1}{\gamma\left[\tilde{J}_V(\mathbf{g_o}) - \mathbf{g_o}^T\tilde{\mathbf{R}}_X\mathbf{g_o}\right]} \\ &> \frac{1}{\tilde{J}_X(\mathbf{h_o})}, \end{aligned} \tag{3.74}$$

sendo $\xi_R(\mathbf{h_o}) > 1$, e quanto maior seu valor, maior a redução de ruido. Além disso,

$$\lim_{\gamma \to 0} \xi_R(\mathbf{h_o}) = \infty, \tag{3.75}$$

$$\lim_{\gamma \to \infty} \xi_R(\mathbf{h_o}) = 1. \tag{3.76}$$

Outro índice para avaliação da redução do ruído pode ser definido como

$$\zeta_R \triangleq 1 - \tilde{J}_X(\mathbf{h_o}) < 1, \tag{3.77}$$

que se aproxima de 1 quando a redução de ruído aumenta.

Assim, o filtro de Wiener insere menor distorção e menor redução de ruído quando a SNR é alta. Com o decaimento da SNR, o filtro insere maior distorção e a redução de ruído aumenta. Para melhorar o desempenho do filtro em termos de distorção, torna-se necessário modificar sua estrutura, permitindo o aumento do erro médio quadrático e construção de um filtro que não é ótimo nesse aspecto, conforme descrito no tópico seguinte.

Filtro sub-ótimo FIR

Para redução do efeito de distorção é introduzido o parâmetro α (um número real) e o filtro sub-ótimo é dado em termos do filtro de Wiener para estimativa de ruído desenvolvido no tópico anterior ($\mathbf{g_o}$) como

$$\mathbf{h_s} = \mathbf{u}_1 - \mathbf{g_s} = \mathbf{u}_1 - \alpha \mathbf{g_o}. \tag{3.78}$$

O filtro de Wiener corresponde ao caso especial em que $\alpha = 1$. O sinal de voz estimado com o filtro sub-ótimo é $\hat{x}_s(n) = \mathbf{h_s}^T \mathbf{y}(n)$ e o erro médio quadrático (MSE) associado é

$$\begin{aligned} J_X(\mathbf{h_s}) &= \mathrm{E}\big[\big(x(n) - \mathbf{h_s}^T \mathbf{y}(n)\big)^2\big] \\ &= \sigma_V^2 - \alpha(2-\alpha)\mathbf{r}_V^T \mathbf{R}_Y^{-1} \mathbf{r}_V. \end{aligned} \tag{3.79}$$

Analogamente ao caso do filtro ótimo de Wiener, a estimativa do sinal de voz é equivalente à estimativa do ruído, então

$$J_V(\mathbf{g_s}) = \mathrm{E}\big[\big(v(n) - \alpha \mathbf{g_o}^T \mathbf{y}(n)\big)^2\big] = J_X(\mathbf{h_s}). \tag{3.80}$$

Quando $\alpha = 1$, $J_X(\mathbf{h_s}) = J_X(\mathbf{h_o})$, para qualquer valor de α escolhido, $J_X(\mathbf{h_s}) \geq J_X(\mathbf{h_o})$. Para o filtro $\mathbf{u}_1$ não há redução de ruído, então α deve ser escolhido de forma que $J_X(\mathbf{h_s}) < J_X(\mathbf{u}_1)$, ou seja,

$$0 < \alpha < 2. \tag{3.81}$$

A potência do sinal de voz estimado é

$$\begin{aligned}
\mathrm{E}[\hat{x}_s^2(n)] &= \mathrm{E}[\mathbf{h_s}^T\mathbf{y}(n)\mathbf{y}(n)^T\mathbf{h_s}] \\
&= \mathbf{h_s}^T\mathbf{R}_Y\mathbf{h_s} = [\mathbf{u}_1 - \alpha\mathbf{g_o}]^T\mathbf{R}_Y[\mathbf{u}_1 - \alpha\mathbf{g_o}] \\
&= [\mathbf{u}_1 - \alpha\mathbf{R}_Y^{-1}\mathbf{r}_Y]^T[\mathbf{r}_Y - \alpha\mathbf{r}_V] \\
&= \sigma_X^2 + \sigma_X^2 - 2\alpha\sigma_V^2 + \alpha^2\mathbf{r}_V^T\mathbf{R}_Y^{-1}\mathbf{r}_V \\
&= \mathbf{h_s}^T\mathbf{R}_X\mathbf{h}_s + \mathbf{h_s}^T\mathbf{R}_V\mathbf{h}_s. \qquad (3.82)
\end{aligned}$$

A relação entre a distorção inserida pelo filtro de Wiener e pelo filtro sub-ótimo é

$$\begin{aligned}
v_D(\mathbf{g_s}) &= \frac{\mathrm{E}\left[\left(x(n) - \mathbf{h_s}^T\mathbf{x}(n)\right)^2\right]}{\sigma_X^2} \\
&= \alpha^2\mathbf{g_o}^T\tilde{R}_X\mathbf{g_o} = \alpha^2 v_D(\mathbf{g_o}), \qquad (3.83)
\end{aligned}$$

dependente unicamente de α. Isso significa que o valor de α escolhido determina a alteração na distorção com relação ao filtro de Wiener. Para que haja redução de distorção, $v_{rr}(\mathbf{g_s}) < v_{rr}(\mathbf{g_o})$, respeitando os limites dados na Equação 3.81, tem-se

$$0 < \alpha < 1. \qquad (3.84)$$

Para $\alpha = 0$ não há redução de ruído ou distorção, para $\alpha = 1$ a redução de ruido e a distorção são máximas.

Apesar da razão entre os índices de distorção para o filtro de Wiener e o filtro sub-ótimo ser função apenas de α, a razão entre os fatores de redução de ruído depende das características estatísticas do sinal de voz e do ruído. No entanto, a razão entre a redução de ruído do filtro sub-ótimo $\mathbf{h_s}$ e o filtro de Wiener $\mathbf{h_o}$, para o fator de redução de ruído definido na Equação 3.77 é

$$\frac{\zeta_R(\mathbf{h_s})}{\zeta_R(\mathbf{h_o})} = \alpha(2 - \alpha). \qquad (3.85)$$

Assim, definindo a função de custo entre distorção e redução de ruído

$$\begin{aligned}
J_{\zeta v}(\alpha) &\triangleq \frac{\zeta_R(\mathbf{h_s})}{\zeta_R(\mathbf{h_o})} - \frac{v_D(\mathbf{g_s})}{v_D(\mathbf{g_o})} \\
&= 2\alpha(1 - \alpha), \qquad (3.86)
\end{aligned}$$

aplicando a derivada em função de α e igualando a zero, calcula-se o valor de α que maximiza essa função de custo, dado por

$$\alpha_o = \frac{1}{2}. \qquad (3.87)$$

A função de custo escolhida não otimiza o desempenho para todas as aplicações. Assim, muitas vezes é possível ajustar o valor de α experimentalmente. Para sistemas de RAV, por exemplo, os valores podem ser testados por meio da observação do percentual de palavras reconhecidas corretamente.

3.5 Considerações Finais

Os filtros de subtração espectral e de Wiener FIR apresentam maior facilidade de implementação e podem ser usados para tratamento de ruído sonoro de voz desde que o ruído e o sinal de voz possam ser considerados estacionários. Nesse caso, os filtros podem ser aplicados a quadros com duração em torno de 20 ms, intervalo para o qual a voz é aproximadamente estacionária [21].

No entanto, a análise teórica dos filtros e estudos anteriores mostram que, apesar do aumento de SNR proporcionado, a distorção inserida no sinal aumenta bastante com a elevação do nível de ruído sonoro. Nesses casos, conforme descrito pelos trabalhos anteriores analisados, o sinal da fala obtido perde qualidade e inteligibilidade.

Assim, um dos principais problemas no desenvolvimento de filtros para voz é a avaliação do seu desempenho, pois a observação da SNR de entrada e saída do filtro é o método usado para justificar o uso das técnicas desenvolvidas nos trabalhos citados na Seção 3.4. As avaliações subjetivas também não são conclusivas. O uso de um sistema de reconhecimento automático de voz pode determinar qual o desempenho efetivo de técnicas de redução de ruído sonoro [5].

CAPÍTULO 4

Descrição do Sistema Desenvolvido

As técnicas de filtragem escolhidas (filtro de Wiener FIR e filtro FIR sub-ótimo) foram implementadas utilizando o Matlab. O sinal de entrada do filtro consiste na palavra a ser reconhecida acrescida do ruído sonoro. O nível de ruído inserido depende da SNR (relação sinal-ruído) considerada para o sinal com ruído incorporado.

Para fazer a mistura do sinal de voz ao ruído com a SNR desejada é necessário adotar um método de cálculo de SNR e utilizar um algoritmo para fazer a soma do sinal de voz e ruído com nível adequado para obter essa SNR.

O sistema de reconhecimento escolhido utiliza o decodificador Julius [27], baseado em HMMs e que necessita de modelos acústico e linguístico. Foram utilizados os modelos disponibilizados pelo grupo Fala Brasil [26]. Além disso, como diferentes ferramentas são usadas, é utilizado o *sox* [41] como programa conversor de tipos de arquivo de áudio.

Os detalhes de implementação do sistema de tratamento de ruído, o método de cálculo de SNR escolhido e as configurações utilizadas no reconhecimento de voz são descritos nas Seções 4.1 e 4.2. Na Seção 4.3 é feito um resumo da descrição de todo o sistema implementado para filtragem e reconhecimento da fala.

4.1 Tratamento de Ruído

Conforme discutido no Capítulo 3, as técnicas de filtragem de Wiener e sub-ótima são desenvolvidas considerando que o sinal de voz e o ruído são estacionários e independentes. A fala é um sinal com comportamento não-estacionário, mas suas características estocásticas não variam para intervalos de tempo pequenos (entre 15 ms e 30 ms) [21].

Por esse motivo o processamento de sinais de voz normalmente inicia com a divisão do sinal em blocos de dados, ou quadros, dentro dos quais as propriedades estatísticas do sinal (momentos e autocorrelação) podem ser consideradas inalteradas [1]. Sinais com esse comportamento podem ser tratados pela análise em curto período, que é representada na forma geral

$$S_{\hat{n}} = \sum_{m=-\infty}^{\infty} P\{s(m)w(\hat{n}-m)\}, \tag{4.1}$$

em que $S_{\hat{n}}$ representa o parâmetro de análise em curto período no instante de análise $\hat{n}$, o operador $P\{\quad\}$ define a natureza da função de análise e $w[\hat{n}-m]$ representa a sequência de janelas.

O tratamento de ruído é feito com análise em curto período. A primeira etapa do processamento consiste na obtenção dos parâmetros necessários para aplicação dos filtros. Em seguida é feito o cálculo dos coeficientes, a partir da autocorrelação do sinal observado e do ruído, para os filtros FIR de Wiener e sub-ótimo.

As especificações para o janelamento, o método usado para cálculo das matrizes de autocorrelação necessárias para obtenção dos filtros e a operação de filtragem são apresentadas nas Seções 4.1.1, 4.1.2 e 4.1.3. Na Seção 4.1.4 é apresentado o método usado para cálculo de relação sinal-ruído para os sinais de entrada e de saída dos filtros.

4.1.1 Janelamento

O sinal de voz pode ser considerado estacionário em intervalos de tempo em torno de 20 ms [21]. Considerando que o ruído também seja estacionário nesse intervalo, sinal observado é dividido em quadros de tamanho $L_a = 20 \times 10^{-3} \times F_s$, em que F_s representa a taxa de amostragem, que tipicamente assume os valores 8000 amostras/s, 16000 amostras/s, 22050 amostras/s e 44100 amostras/s. O tratamento do ruído é feito para cada quadro separadamente.

Considerando a notação usada na Equação 4.1, a janela usada para separação dos quadros é definida por

$$w_R(m) = \begin{cases} 1, & -M \leq m \leq M \\ 0, & \text{caso contrário,} \end{cases}$$

que corresponde à janela retangular. Para quadros de 20 ms, $M = L_a/2$.

4.1.2 Obtenção das Matrizes de Autocorrelação

A matriz de autocorrelação de um sinal discreto real representado na forma vetorial por $\mathbf{s}(n)$, em que n representa o instante (ou amostra) observado, pode ser definida como [36]

$$\mathbf{R}_S = \mathrm{E}[\mathbf{s}(n)\mathbf{s}^T(n)]. \tag{4.2}$$

Para sinais com variações estatísticas lentas, a função de autocorrelação de curto período é definida como a função de autocorrelação da sequencia $s_{\hat{n}}(m) = s(m)w(\hat{n}-m)$, selecionada pela janela deslocada para o instante $\hat{n}$, ou seja [1],

$$\begin{aligned} R_{S,\hat{n}}(l) &= \sum_{m=-\infty}^{\infty} s_{\hat{n}}(m)s_{\hat{n}}(m+l) \\ &= \sum_{m=-\infty}^{\infty} s(m)w(\hat{n}-m)s(m+l)w(\hat{n}-m-l). \end{aligned} \tag{4.3}$$

Dependendo do tamanho da janela usada na análise de curto período, essa definição pode ter a inserção de um parâmetro de ponderação como, por exemplo, uma média aritmética.

A estimativa para a matriz de autocorrelação do sinal observado e o vetor de autocorrelação do ruído, necessários para o cálculo dos coeficientes dos filtros FIR de Wiener e sub-ótimo, é descrita a seguir.

Autocorrelação do Sinal Observado

Para o tratamento de ruído, o processamento do sinal é feito para cada quadro separadamente. Portanto, a autocorrelação do sinal observado é estimada para cada quadro. O cálculo dos coeficientes para o filtro de ordem L projetado para cada quadro depende da obtenção da estimativa de $\mathbf{R}_Y = E\big[\mathbf{y}(n)\mathbf{y}^T(n)\big]$, que representa a matriz de autocorrelação do vetor de observação $\mathbf{y}(n)$. Como $\mathbf{y}(n) = [y(0) \quad y(1) \cdots y(L-1)]^T$ é estacionário, conforme desenvolvido em [15], $\mathbf{R}_Y$ possui uma estrutura de matriz Toeplitz, ou seja,

$$\mathbf{R}_Y = \begin{bmatrix} R_Y(0) & R_Y(1) & \cdots & R_Y(L-1) \\ R_Y(1) & R_Y(0) & \cdots & R_Y(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_Y(L-1) & R_Y(L-2) & \cdots & R_Y(0) \end{bmatrix}.$$

Em matrizes com estrutura Toeplitz, cada diagonal descendente da esquerda para a direita possui valor constante [15]. Assim, é necessário calcular apenas a primeira linha da matriz, que corresponde ao vetor de autocorrelação do sinal observado.

Cada quadro obtido a partir do janelamento descrito na Subseção 4.1.1 possui L_a amostras do sinal de voz afetado pelo ruído, e a autocorrelação do sinal observado $\mathbf{y}(n)$ é estimada por [1]

$$\begin{aligned} R_{Y,\hat{n}}(l) &= \frac{1}{L_a} \sum_{m=-\infty}^{\infty} y(m)w_H(\hat{n}-m)y(m+l)w_H(\hat{n}-m-l) \\ &= \frac{1}{L_a} \sum_{m=m_0}^{m_0+L_a-1} y(m)y(m+l), \end{aligned} \tag{4.4}$$

em que m_0 representa o início do quadro.

Autocorrelação do Ruído

O cálculo da autocorrelação do ruído é feito a partir da suposição que o ruído é estacionário ou possui propriedades estocásticas que variam mais lentamente que as características da voz. Assim, é possível calcular uma estimativa para a sua autocorrelação utilizando um quadro do sinal observado em um instante de ausência de fala e fazer uso dessa estimativa nos quadros seguintes com presença de voz [5].

Assim, o algoritmo desenvolvido utiliza um sinal contendo informação sobre a presença de voz, denotado por $VAD_{\hat{n}}$ (*Voice Activity Detection*) para o quadro centrado na n-ésima amostra do sinal de observação $y(n)$. Esse sinal pode ser descrito como

$$VAD_{\hat{n}} = \begin{cases} 1, & \text{presença de voz detectada} \\ 0, & \text{ausência de voz detectada.} \end{cases}$$

O vetor de autocorrelação do ruído $\mathbf{r}_V$ é inicializado como um vetor nulo. Para cada quadro sem presença de voz, o vetor de autocorrelação do ruído é igualado ao vetor de autocorrelação do sinal observado. Para cada quadro com presença de voz, o vetor é mantido como o vetor atribuído anteriormente. Ou seja,

$$\begin{aligned} &\text{se } VAD_{\hat{n}} = 0 \text{ , } \mathbf{r}_{V,\hat{n}} = \mathbf{r}_{Y,\hat{n}} \\ &\text{se } VAD_{\hat{n}} = 1, \mathbf{r}_{V,\hat{n}} = \mathbf{r}_{V,\hat{n-1}}, \end{aligned}$$

em que $VAD_{\hat{n}}$ representa a informação de presença de voz no quadro centrado em $\hat{n}$ e $\mathbf{r}_{V,\hat{n-1}}$ representa o quadro anterior. O detector de atividade de voz utilizado é descrito na Seção 4.1.4.

4.1.3 Filtragem

A operação de filtragem é feita sequencialmente para cada quadro janelado. Para o filtro FIR de Wiener, o único parâmetro ajustável é a ordem L do filtro, enquanto para o filtro subótimo existe além da ordem L o parâmetro α responsável pela ponderação entre melhoria de SNR e inserção de distorção.

Filtro de Wiener FIR

O vetor de coeficientes do filtro de Wiener é estimado para cada quadro como

$$\mathbf{h_o} = \mathbf{u}_1 - \mathbf{R}_{Y,\hat{n}}^{-1}\mathbf{r}_{V,,\hat{n}}.$$

O processo de filtragem é feito para cada amostra dentro do quadro, resultando na estimativa do sinal de voz

$$\hat{x}(n) = \mathbf{h_o}^T\mathbf{y}(n).$$

Para as primeiras L amostras de cada quadro, o filtro utilizado é aquele projetado para o quadro atual, mas é necessário utilizar as $L-1$ amostras finais do sinal observação do quadro anterior.

Filtro FIR sub-ótimo

O processamento para o filtro sub-ótimo é feito da mesma forma que o filtro de Wiener, com exceção do cálculo dos coeficientes, dados por

$$\mathbf{h_s} = (1-\alpha)\mathbf{u}_1 + \mathbf{h_o}. \tag{4.5}$$

4.1.4 Cálculo da Relação Sinal-Ruído

O cálculo da relação sinal-ruído para sinais de voz deve ser feito considerando que existem intervalos de silêncio na fala. Assim, é necessário utilizar um parâmetro indicativo de presença de voz. As definições associadas à relação sinal-ruído para sinais de voz são resumidas a seguir [42].

A definição convencional de relação sinal-ruído, chamada de SNR global (GSNR) para aplicações de voz, é

$$\mathrm{GSNR} = 10\log\frac{\sigma_X^2}{\sigma_V^2}, \tag{4.6}$$

em que σ_Y^2 é a variância do sinal de voz e σ_V^2 é a variância do ruído, correspondentes a todas as amostras do sinal. A SNR otimizada para sinais de voz é denotada por SNR e se baseia na contabilização da SNR global para trechos com presença de voz, ou seja,

$$\mathrm{SNR} = 10\log\frac{\sum_{m=0}^{N-1} x^2(m)\cdot \mathit{vad}(m)}{\sum_{m=0}^{N-1} v^2(m)\cdot \mathit{vad}(m)}, \tag{4.7}$$

em que $x(m)$ e $v(m)$ correspondem a m-ésima amostra de voz e ruído respectivamente, $\mathit{vad}(m)$ é a informação sobre presença de voz na m-ésima amostra do sinal observado e N é o número total de amostras.

Como o sinal da fala possui comportamento quase estacionário para quadros em torno de 20 ms, a SNR calculada para cada um desses quadros é definida como SNR local, dada para a janela em torno da amostra $\hat{n}$ por

$$\mathrm{SNR}_{\hat{n}} = 10\log\frac{\sum_{m=0}^{M-1} x_{\hat{n}}^2(m)\mathit{vad}(m)}{\sum_{m=0}^{M-1} v_{\hat{n}}^2(m)\mathit{vad}(m)} = 10\log\frac{\sigma_{X,\hat{n}}^2}{\sigma_{V,\hat{n}}^2}. \tag{4.8}$$

A relação sinal-ruído segmental (SSNR) é definida como a média dos valores de SNR local para quadros com presença de voz. A SSNR é um indicador que se relaciona com a percepção humana do ruído na fala [42] e é calculada por

$$\mathrm{SSNR} = \frac{1}{K}\sum_{\hat{n}=n_0}^{n_f}\left(10\log\frac{\sum_{m=0}^{M-1}x_{\hat{n}}^2(m)}{\sum_{m=0}^{M-1}v_{\hat{n}}^2(m)}\cdot \mathit{VAD}_{\hat{n}}\right), \tag{4.9}$$

em que n_0 e n_f são as amostras centrais do primeiro e último quadro respectivamente e $VAD_{\hat{n}}$ é a informação de presença de voz no quadro centrado em $\hat{n}$.

Observa-se que o cálculo de SSNR envolve a média geométrica da relação sinal-ruído. Assim, é possível definir a relação sinal-ruído segmental aritmética (ASSNR) como um indicador alternativo e mais simples de calcular, isto é,

$$\mathrm{ASSNR} = 10\log\left(\frac{1}{K}\sum_{\hat{n}=n_0}^{n_f}\frac{\sum_{m=0}^{M-1}x_{\hat{n}}^2(m)}{\sum_{m=0}^{M-1}v_{\hat{n}}^2(m)}\cdot \mathit{VAD}_{\hat{n}}\right). \tag{4.10}$$

Os cálculos de relação sinal-ruído e ajuste da variância do ruído para obter a SNR desejada foram feitos utilizando a ferramenta *snr* [42] desenvolvida pelo Grupo de Processamento da Fala (*Speech Processing Group*) da *Czech Technical University* e disponibilizado juntamente com a documentação no *website* http://noel.feld.cvut.cz/speechlab .

A ferramenta *snr* também pode ser usada para fazer a estimativa de SNR a partir do sinal de voz misturado ao ruído, por meio da detecção da presença de voz e estimativa de variância do ruído. As funções disponíveis na ferramenta são ilustradas graficamente na Figura 4.1.4.

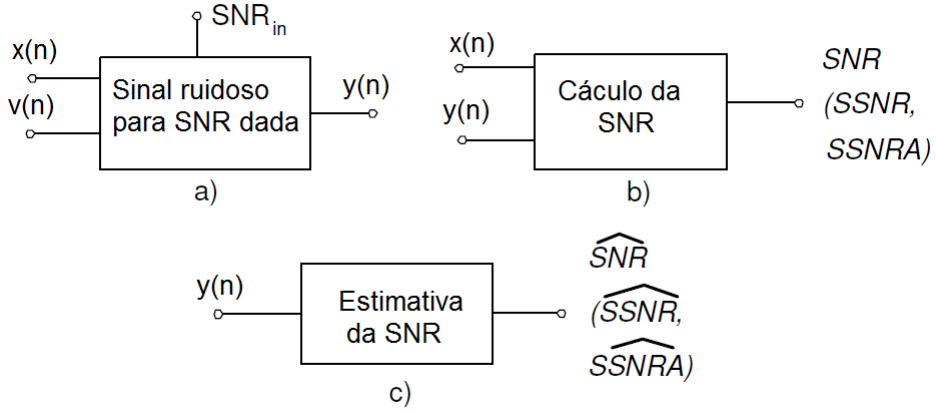

Figura 4.1 Diagrama de blocos da ferramenta *snr*, para as operações de (a) mistura de sinal de voz e ruído com SNR desejada, (b) cálculo de SNR do sinal misturado a partir do sinal de voz puro e (c) estimativa de SNR a partir do sinal observado.

Detector de Presença de Voz (*VAD*)

O indicador de presença de voz é disponibilizado pela ferramenta *snr*, podendo ser obtido pelos métodos de detecção de energia ou cepstral. O *vad* utilizado foi o de detecção cepstral, visto que o método da energia tem seu desempenho degradado com o aumento do nível de ruído.

Relação Sinal-Ruído na Saída dos Filtros

Para o cálculo da relação sinal-ruído após a filtragem é utilizado o ruído residual

$$\tilde{v}(n) = \hat{x}(n) - x(n). \tag{4.11}$$

Os valores de SNR podem então ser obtidos da mesma forma que o sinal de observação, substituindo as amostras do ruído original $v(n)$ pelas do ruído residual $\tilde{v}(n)$.

4.2 Sistema de Reconhecimento da Fala

Na Figura 4.2 é mostrado o diagrama de blocos para o sistema de reconhecimento usado.

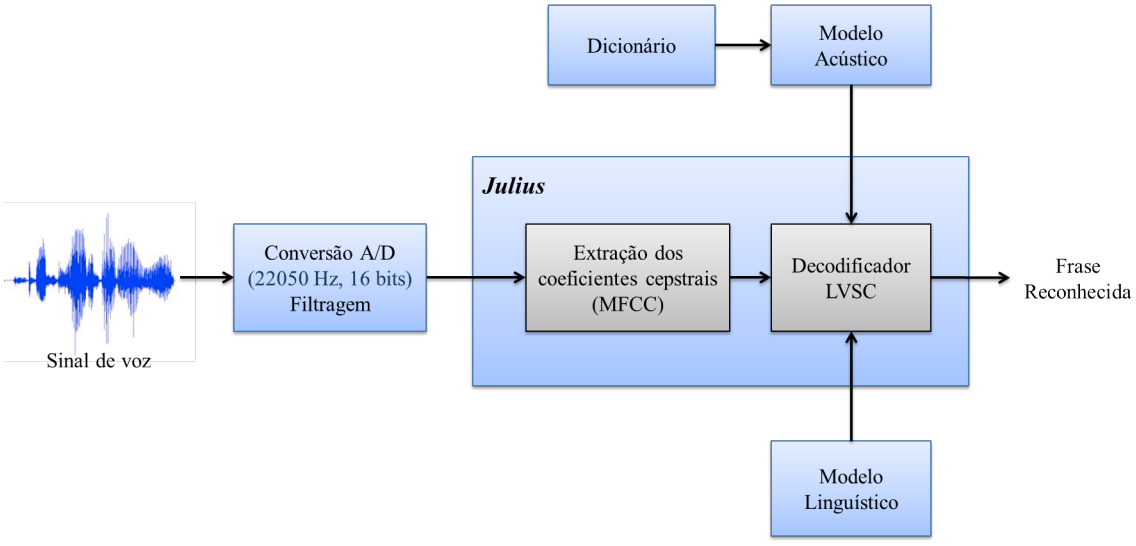

Figura 4.2 Diagrama de blocos representando o sistema de RAV utilizado.

O reconhecimento da fala é feito utilizando a ferramenta *Open-Source Large Vocabulary CSR Engine* Julius [27], Versão 4.3.1, disponível no *website* http://julius.sourceforge.jp. O Julius é um codificador de voz contínua desenvolvido para pesquisa e desenvolvimento. O codificador utiliza modelos de Markov (HMM) com dependência de contexto e modelo N-grama. As ferramentas incorporadas são descritas em [27].

O Julius disponibiliza a estrutura de codificação de voz. Assim, é possível utilizar a ferramenta para reconhecimento, desde que o desenvolvedor indique um modelo de linguagem e um modelo acústico no arquivo de configuração do Julius. Os modelos acústicos utilizados no Julius possuem o formato ASCII do HTK, descrito na Seção 2.4.

Para o sistema implementado, os modelos em português brasileiro utilizados para o treinamento foram obtidos a partir do banco de dados desenvolvido pelo Laboratório de Processamento de Sinais da UFPA e disponibilizado no site do grupo Fala Brasil [26]. Os modelos usados são:

- LaPSAM v1.3 - Modelo acústico criado com o *software* HTK. Para o treinamento foi utilizado o *corpus* LapsStory. O modelo acústico utiliza o dicionário UFPAdic3.0 com modelos trifones dependentes de contexto (*cross-word triphones*) com 14 gaussianas por mistura e taxa de amostragem de 22050 amostras/s. O tipo paramétrico utilizado se baseia nos coeficientes cepstrais e derivadas cepstrais (MFCC-E-D-A-Z).

- LaPSLM v1.0 - Modelo de linguagem N-grama construído com o *toolkit* SRILM. Para treinamento são utilizadas frases dos *corpora* CETENFolha, Spoltech, OGI-22, Westpoint, LapsStory e LapsNews (todos desenvolvidos pelo grupo Fala Brasil), totalizando 1,6 milhões de frases. O dicionário utilizado no treinamento é o UFPAdic3.0, com 64.972 palavras.

4.3 Considerações Finais

Na Figura 4.2 é mostrado um diagrama que representa a troca de dados entre os subsistemas que compõem todo o sistema desenvolvido. O texto falado pelo locutor é capturado pelo microfone e salvo em um arquivo no formato *WAVEform Audio Format* (.wav) com 16 *bits* e taxa de amostragem 22050 amostras/s. O arquivo é então convertido para o formato binário utilizando a ferramenta *sox*.

O ruído utilizado foi gerado no Matlab e a soma do sinal de voz com o ruído foi feita pela ferramenta *snr*, de acordo com a relação sinal-ruído desejada. Além do sinal de voz misturada ao ruído, a ferramenta também disponibiliza o sinal de informação de presença de voz. O arquivo contendo o sinal de voz misturado ao ruído é convertido para o formato .wav e processado pelo algoritmo implementado no MatLab, que recebe na entrada também o sinal *vad*.

As saídas do algoritmo são os sinais filtrados pelo filtro FIR de Wiener e FIR sub-ótimo considerando diferentes valores do parâmetro α.

Os arquivos contendo os sinais de voz filtrados são então passados para o reconhecedor, que fornece na saída um arquivo de texto com as frases reconhecidas.

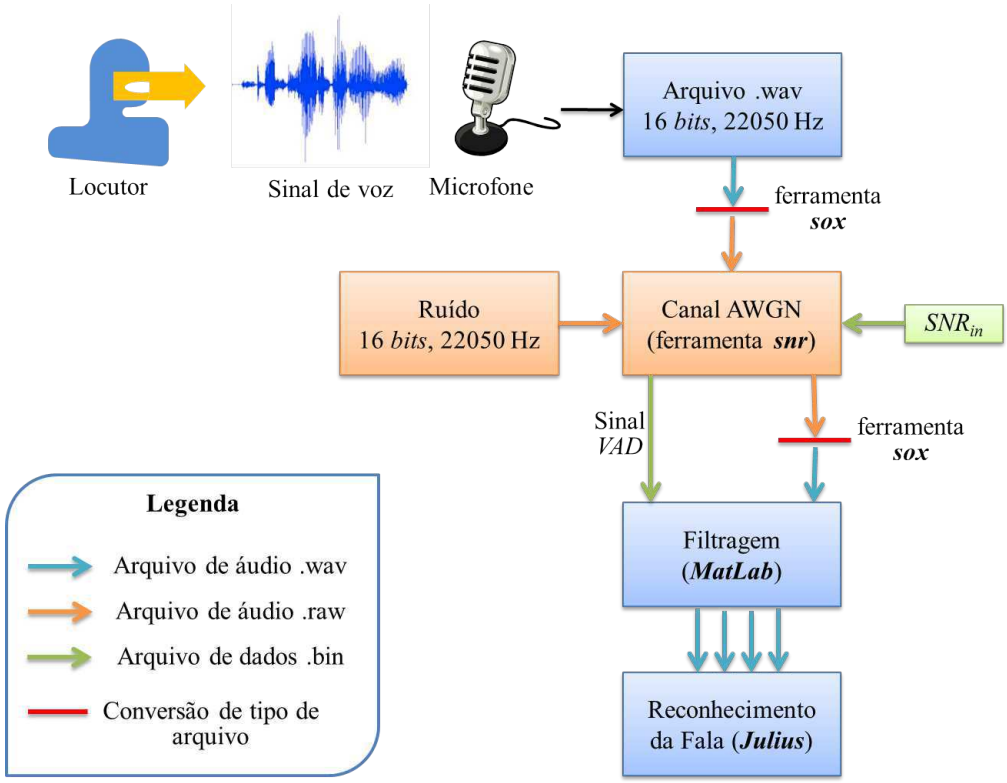

Figura 4.3 Diagrama de blocos representando o sistema implementado.

CAPÍTULO 5

Resultados

Para a simulação foram utilizados 20 arquivos de áudio, cada um contendo uma frase de um locutor distinto. A frequência de amostragem é 22050 amostras/s, com representação de 16 *bits* e codificação PCM. As frases foram originadas por 10 locutores do sexo feminino e 10 locutores do sexo masculino de diferentes regiões do Brasil. As sentenças correspondentes a cada frase são apresentadas no Apêndice B, totalizando 146 palavras.

Valores típicos de ruído sonoro para diferentes tipos de fontes sonoras presentes em ambientes com possíveis aplicações de sistemas de RAVs são mostrados na Tabela 5.1 [2]. Para voz, pode-se considerar que a SNR é elevada para valores acima de 30 dB e baixa para valores abaixo de 10 dB [43].

O ruído inserido nos sinais de voz possui distribuição gaussiana, com média nula e diferentes valores de variância, dependendo do valor de SNR desejado. Para cada sinal de voz foram gerados 6 sinais ruidosos, com SNR de aproximadamente 0 dB, 3 dB, 5 dB, 10 dB, 15 dB e 20 dB. A mistura do ruído com o sinal de voz é feita pela soma dos dois sinais, sem inserção de eco, configurando um ruído aditivo gaussiano branco (*Aditive White Gaussian Noise* – AWGN).

Os resultados iniciais apresentados são referentes aos testes para ajuste de parâmetros dos filtros. Foram realizadas simulações para cinco frases, considerando diferentes tamanhos de janela e ordem dos filtros. Os valores de SNR obtidos na saída e a quantidade de palavras

Fonte Sonora	Nível de Pressão Sonora (dB)
Avião a jato decolando	120
Trem de metrô	100
Trânsito intenso	80 dB
Fala	70 dB
Shopping Center	60 dB
Área residencial calma	50 dB
Estúdio de gravação	40 dB
Sussurro	40 dB

Tabela 5.1 Níveis típicos de ruído para diferentes ambientes [2].

		Sub-Ótimo		
Entrada	Wiener	$\alpha = 0,8$	$\alpha = 0,7$	$\alpha = 0,5$
21,6 dB	23,6 dB	25,9 dB	25,5 dB	24,0 dB
16,7 dB	20,4 dB	20,0 dB	19,5 dB	19,3 dB
12 dB	16,4 dB	16,3 dB	16,2 dB	15,6 dB
5,5 dB	12,7 dB	11,5 dB	12,1 dB	13,3 dB
1,3 dB	7,1 dB	6,8 dB	6,5 dB	5,8 dB
-1,2 dB	5,0 dB	4,6 dB	4,3 dB	3,2 dB

Tabela 5.2 SNR obtida após a filtragem para sentença "hoje pela manhã não haverá aula" para filtros com ordem igual a 20.

reconhecidas corretamente foram utilizadas para fixar os parâmetros. Os resultados obtidos na etapa de ajuste de parâmetros é descrita na Seção 5.1.

Os resultados finais correspondem aos percentuais de palavras reconhecidas corretamente nos sinais obtidos na filtragem com o filtro de Wiener e o filtro sub-ótimo com três diferente valores do parâmetro α. A análise desses resultados é apresentada na Seção 5.2.

5.1 Ajuste de Parâmetros dos Filtros

O filtro de Wiener FIR e o filtro sub-ótimo possuem dois parâmetros que podem ser ajustados: o comprimento do filtro L e o tamanho da janela de análise $winTime$ (tamanho em segundos) ou L_a (tamanho em número de amostras). Em [5] é demonstrado que, para aplicações com tratamento de voz, o valor indicado para L é 20.

O tamanho da janela escolhido foi 20 ms, pois nesse intervalo de tempo não ocorrem variações significativas na autocorrelação do sinal. Janelas de 18 ms e 25 ms foram testadas para 5 frases, não tendo sido observada melhoria significativa na SNR ou inteligibilidade dos sinais filtrados.

Para o filtro sub-ótimo, além do comprimento do filtro e tamanho da janela de análise, é possível ajustar o parâmetro α, que corresponde ao parâmetro de ponderação entre redução de ruído e inserção de SNR. De acordo com o estudo apresentado em [5] e discutido no Capítulo 3, a redução de ruído obtida pelo filtro é maior para valores de α próximos de 1. Para valores de $\alpha < 0,5$, a SNR obtida na saída é muito próxima a de entrada, indicando que não é aplicável situações de baixa SNR. Assim, os valores escolhidos para α foram $0,5$, $0,7$ e $0,8$.

As sentenças obtidas na saída do reconhecedor e as formas de onda dos sinais filtrados para filtros de comprimento igual a 10 e 20 são apresentados no Apêndice D para a sentença "hoje pela manhã não haverá aula", correspondente à Frase 1. Para essa mesma sentença os valores de SNR obtidos após a filtragem são mostrados na Tabela 5.2 e Tabela 5.3 para os filtros de ordem 20 e 10 .

Os ganhos de SNR obtidos na filtragem não variam para as diferentes frases testadas usadas na fase de testes. Observa-se que o ganho é maior para os filtros com ordem 10. No

		Sub-Ótimo		
Entrada	Wiener	$\alpha = 0,8$	$\alpha = 0,7$	$\alpha = 0,5$
21,6 dB	27,1 dB	26,8 dB	26,5 dB	25,2 dB
16,7 dB	20,9 dB	20,3 dB	19,9 dB	19,5 dB
12 dB	17,7 dB	17,5 dB	17,4 dB	16,3 dB
5,5 dB	13,5 dB	13,3 dB	11,9 dB	11,0 dB
1,3 dB	8,4 dB	8,1 dB	7,7 dB	7,2 dB
-1,2 dB	5,3 dB	4,8 dB	4,2 dB	2,4 dB

Tabela 5.3 SNR obtida após a filtragem para sentença "hoje pela manhã não haverá aula" para filtros com ordem igual a 10.

			Sub-Ótimo		
SNR	Sem Filtragem	Wiener	$\alpha = 0,8$	$\alpha = 0,7$	$\alpha = 0,5$
20 dB	4	5	4	5	4
15 dB	4	5	4	5	4
10 dB	4	5	4	5	4
5 dB	2	2	2	3	3
3 dB	0	1	2	2	2
0 dB	0	1	2	2	1

Tabela 5.4 Palavras reconhecidas corretamente na sentença "hoje pela manhã não haverá aula" para filtros com ordem igual a 20.

entanto, a inteligibilidade percebida para os sinais obtidos para SNR menor que 10 dB é consideravelmente melhor para ordem 20. Também foi avaliado o desempenho do reconhecedor para os dois casos.

As Tabelas 5.4 e 5.5 apresentam o número de palavras reconhecidas corretamente para a Frase 1, que possui total de seis palavras. O desempenho do reconhecedor é melhor para os filtros de ordem 20. Para as outras quatro frases o mesmo comportamento foi observado.

Após a análise de SNR, percentual de palavras reconhecidas corretamente e distorção percebida, o parâmetro de comprimento do filtro foi escolhido como $L = 20$. Para ilustrar o resultado da filtragem, são mostradas na Figura 5.1 as formas de onda do sinal de voz sem ruído, sinal com ruído e sinais filtrados para a Frase 1 e SNR de 0 dB.

			Sub-Ótimo		
SNR	Sem Filtragem	Wiener	$\alpha = 0,8$	$\alpha = 0,7$	$\alpha = 0,5$
20 dB	4	4	4	5	4
15 dB	4	4	4	4	4
10 dB	4	1	2	5	3
5 dB	2	2	2	2	2
3 dB	0	1	1	2	1
0 dB	0	1	1	1	0

Tabela 5.5 Palavras reconhecidas corretamente na sentença "hoje pela manhã não haverá aula" para filtros com ordem igual a 10.

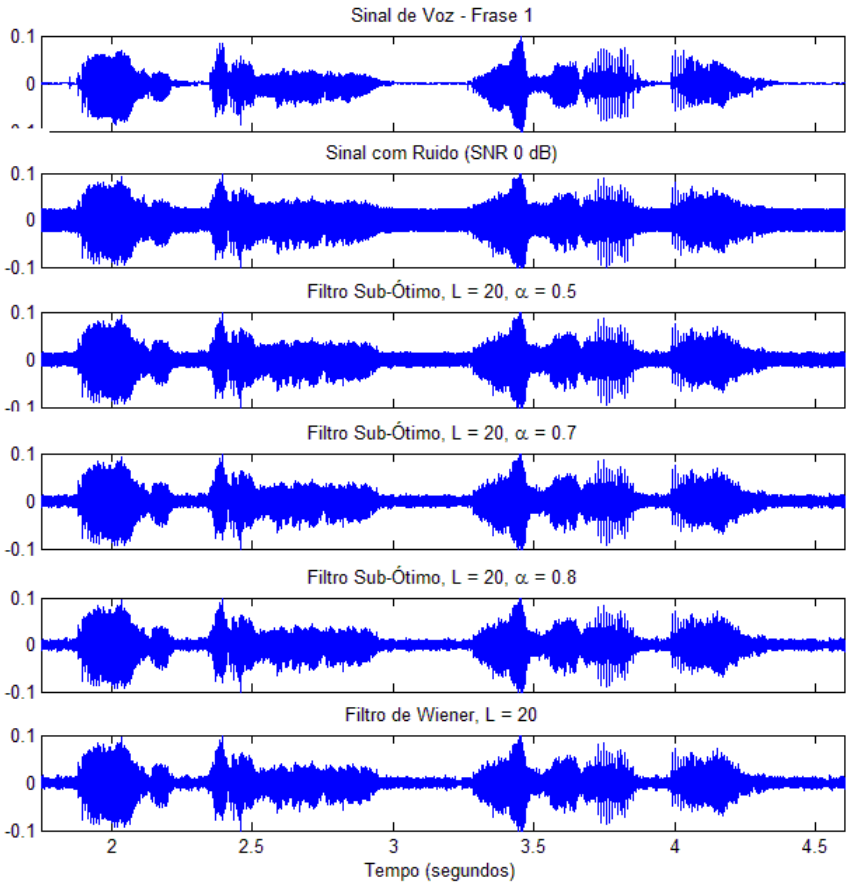

Figura 5.1 Formas de onda para o sinal de voz, sinal com ruído e sinais obtidos na saída dos filtros.

A observação das formas de onda obtidas indica que a SNR possui maior redução quando é aplicado o filtro de Wiener. Para os filtros do tipo sub-ótimo, a redução menor ocorre para o parâmetro α é igual a $0,5$, aumentando gradualmente para $0,7$ e $0,8$, conforme indicado também Tabela 5.2.

5.2 Resultados de Reconhecimento

Para cada frase de entrada (Apêndice B) foram obtidas seis frases somadas ao ruído gaussiano branco com as SNRs pré-definidas, compondo os sinais ruidosos. Cada uma desses sinais e também as frases originais (sem adição de ruído) foram aplicadas aos filtros implementados. Para cada sinal de entrada nos filtros são obtidos quatro sinais filtrados. Esse processo é ilustrado por um fluxo de arquivos na Figura 5.2 para uma frase.

Para cada frase de entrada são enviados para o reconhecedor 35 arquivos, totalizando 700 arquivos de áudio. As tabelas com o número de palavras reconhecidas por frase para os diferentes valores de SNR e para os sinais de voz sem ruído inserido são apresentadas no Apêndice D. Os resultados do sistema de reconhecimento são mostrados de forma resumida como o percentual de palavras reconhecidas corretamente na Tabela 5.6.

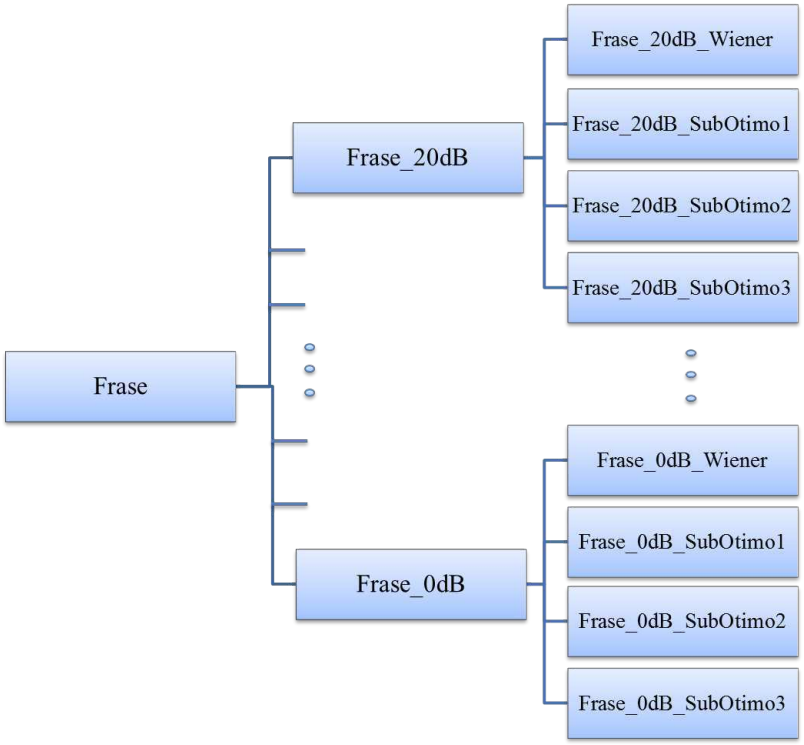

Figura 5.2 Fluxograma ilustrando todos os arquivos referentes a uma frase usados como entrada do sistema de reconhecimento.

O percentual de acerto para os sinais originais (sem filtragem e sem adição de ruído) é de 71%. Esse percentual diminui gradativamente com a inserção de ruído e diminuição da SNR, indicando que a operação dos sistemas de RAV em ambientes com SNR abaixo de 20 dB é comprometida. Para os valores de SNR abaixo de 10 dB o percentual de acerto do RAV é menor que 20%.

Quando um arquivo sem adição de ruído é processado por um filtro digital ocorre um efeito de distorção inerente ao processamento. Para avaliar esse efeito nos filtros implementados, os arquivos; obtidos pela filtragem dos sinais de voz sem adição ruído são processados pelo reconhecedor. O gráfico de barras ilustrativo para o percentual de palavras reconhecidas corretamente com aplicação de cada um dos filtros para sinal de entrada sem ruído é ilustrado na Figura 5.3.

Como o filtro de Wiener corresponde ao filtro sub-ótimo com $\alpha = 1$, observa-se que a percentagem de acerto aumenta com a diminuição do parâmetro α. Esse é o processo inverso ao observado para o comportamento da SNR na Tabela 5.2, que diminui. Esse resultado é coerente com a existência da relação entre diminuição de SNR e inserção de distorção, conforme discutido no Capítulo 3.

Com a adição de ruído, para melhorar o desempenho do reconhecedor é necessário utilizar um filtro que possibilite a redução do nível de ruído sem causar distorção significativa para o reconhecedor. Na Figura 5.4 é mostrado o gráfico de barras com percentual de palavras reconhecidas em função da SNR para os filtros desenvolvidos.

			Sub-Ótimo		
SNR	Sem Filtragem	Wiener	$\alpha = 0,8$	$\alpha = 0,7$	$\alpha = 0,5$
Sem Ruído	71%	62%	66%	67%	69%
20 dB	41%	48%	49%	54%	50%
15 dB	32%	38%	38%	44%	36%
10 dB	21%	23%	27%	35%	25%
5 dB	14%	16%	18%	23%	20%
3 dB	8%	12%	17%	20%	15%
0 dB	6%	10%	13%	16%	10%

Tabela 5.6 Percentual de palavras reconhecidas corretamente.

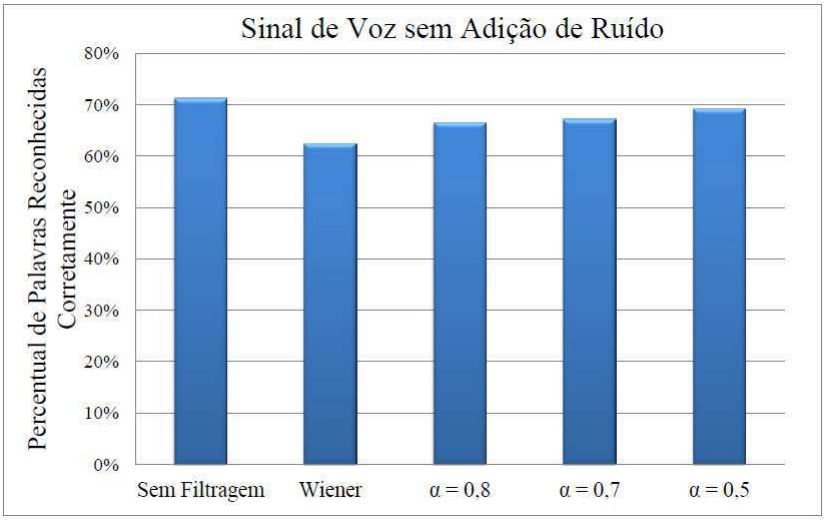

Figura 5.3 Percentual de palavras reconhecidas corretamente para sinais de entrada sem adição de ruído.

Para todos os valores de SNR analisados todos os filtros demonstraram melhoria no percentual de palavras reconhecidas corretamente com relação ao obtido para o sinal sem tratamento de ruído. No entanto, o desempenho do filtro de Wiener foi inferior, devido à inserção de distorção.

Os sinais processados pelo filtro sub-ótimo com parâmetro $\alpha = 0,7$ apresentam melhor percentual de palavras reconhecidas corretamente para todos os valores de SNR analisados. Para esse valor de parâmetro ocorre redução de ruído próxima a obtida com o filtro de Wiener, no entanto a distorção inserida é menor.

5.3 Considerações Finais

Com a redução de SNR o desempenho do sistema de RAV utilizado é degradado, tendo sua utilização comprometida especialmente para valores de SNR abaixo de 20 dB. A aplicação dos filtros implementados proporciona uma melhoria no percentual de palavras reconhecidas corretamente.

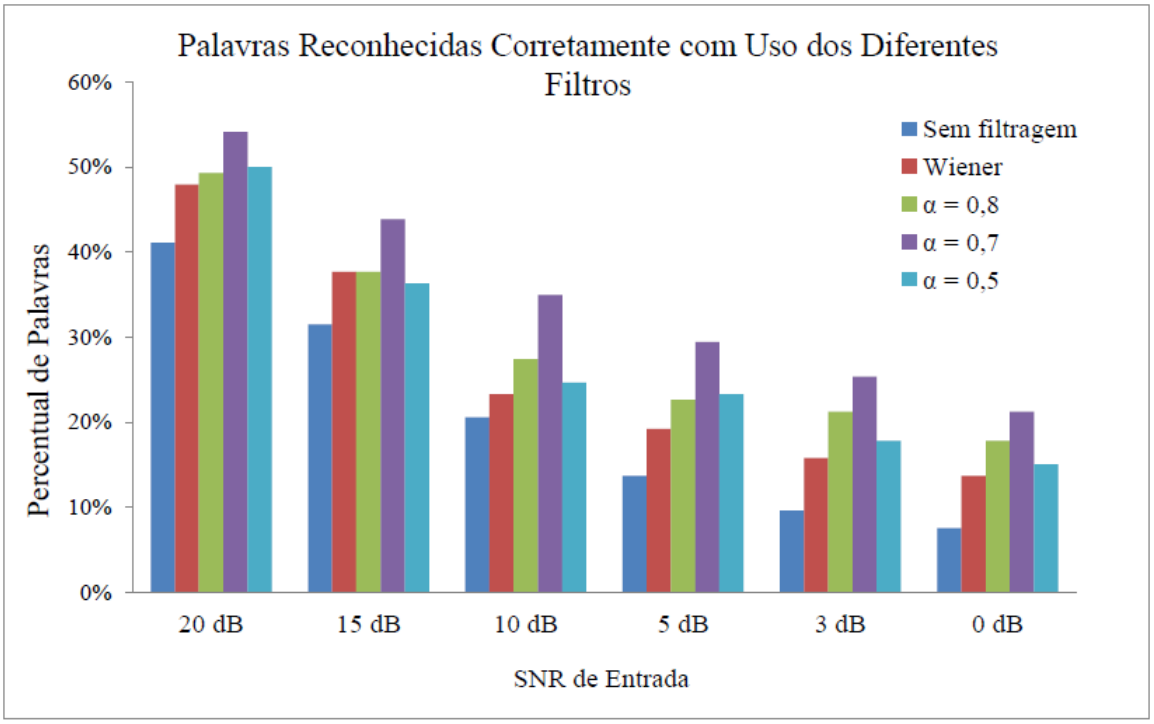

Figura 5.4 Percentual de palavras reconhecidas corretamente em função da SNR dos sinais de entrada.

No entanto, de acordo a observação do resultado de reconhecimento para o sinal sem adição de ruído em comparação com o resultado para os sinais obtidos a partir de sua aplicação aos filtros, a inserção de distorção também altera o desempenho do sistema de RAV.

Para entrada sem adição ruído, a distorção inserida pelos filtros reduz o percentual de acertos. No entanto, os percentuais de palavras reconhecidas corretamente obtidos indicam que a aplicação dos filtros implementados para os valores de SNR analisados e ruído aditivo gaussiano branco sempre implica em melhoria no desempenho do sistema de RAV. Esse resultado indica que o efeito do ruído gera alterações mais significativas nos coeficientes paramétricos que o efeito de distorção.

A partir da análise dos resultados de reconhecimento para os diferentes valores de SNR, conclui-se que a aplicação do filtro sub-ótimo com $\alpha = 0,7$ resulta na melhor taxa de acertos para o reconhecedor utilizado dentre os quatro filtros desenvolvidos quando o ruído é aditivo gaussiano branco.

CAPÍTULO 6

Conclusões e Trabalhos Futuros

Sistemas de reconhecimento automático de voz têm como objetivo converter a mensagem transmitida por meio da fala por um locutor em um formato compreensível pela máquina. Para realizar essa tarefa, os RAVs possuem uma estrutura análoga ao mecanismo de percepção humana da fala, baseada em coeficientes paramétricos que representam as informações do sinal de voz. Esses coeficientes são decodificados a partir de modelos acústico e linguístico.

O grande número de palavras, as variações linguísticas e existência de palavras com sonoridade semelhante são alguns dos complicadores no desenvolvimento dos sistemas de RAV. O efeito desses aspectos pode ser tratado na fase de treinamento dos sistemas, com inserção de análise de contexto da aplicação e uso de bancos de dados extensos da fala. Assim, para a fase de treinamento é necessário dispor de um banco de dados com grande número e diversidade de amostras.

As aplicações dos sistemas de reconhecimento de voz usualmente ocorrem em ambientes sem isolamento acústico, causando alterações no sinal de voz por meio da adição de ruído e ocorrência de eco. O tratamento desses efeitos pode ser feito utilizando um banco de dados de voz para treinamento que já possuam essas alterações. No entanto, as mudanças nas características do ambiente ou a utilização do reconhecedor em outros ambientes se tornam inviáveis.

Assim, o processamento do sinal de voz a fim de reduzir os efeitos de ruído é uma alternativa na aplicação de RAVs em diversos ambientes. De acordo com o estudo das técnicas de redução de ruído sonoro feita no Capítulo 3 desta dissertação, a utilização do filtro de Wiener minimiza o erro médio quadrático. Porém, com o aumento de redução de ruído ocorre o aumento de inserção de distorção.

Para aplicações com baixa SNR, a distorção inserida pelo filtro de Wiener é bastante elevada e, por isso, um filtro sub-ótimo permitindo o ajuste entre inserção de distorção e redução de ruído pode ser utilizado.

Neste trabalho de mestrado foram implementados o filtro de Wiener e filtro sub-ótimo para redução de ruído aditivo gaussiano branco. Os sinais filtrados foram usados como entrada

para um reconhecedor automático de voz para o português brasileiro para fala contínua com amplo vocabulário e sem orientação a locutor.

As contribuições e os possíveis trabalhos futuros e melhorias para o sistema desenvolvido são apresentadas na Seção 6.1 e Seção 6.2.

6.1 Contribuições

Duas contribuições são identificadas: a avaliação de desempenho dos filtros de Wiener e Sub-Ótimo para ruído aditivo gaussiano branco e a melhoria do reconhecimento automático de voz na presença desse ruído. Essas contribuições são analisadas nas sub-seções seguintes.

6.1.1 Avaliação de Desempenho dos Filtros de Wiener e Sub-Ótimo

A avaliação de desempenho de técnicas de redução de ruído sonoro em voz usualmente é feita pela análise de SNR e avaliação subjetiva. Neste trabalho é apresentada a avaliação de desempenho em termos do percentual de palavras reconhecidas corretamente por um reconhecedor automático de voz.

De acordo com os resultados obtidos para filtragem dos sinais de voz sem adição de ruído, é possível comparar a distorção inserida pelos filtros implementados. Nesse caso, o filtro de Wiener apresenta menor desempenho, enquanto o filtro sub-ótimo possui melhoria de desempenho para diminuição do parâmetro α. A partir desse resultado se conclui que a distorção inserida altera as características paramétricas do sinal de voz.

Com a diminuição da SNR é necessário avaliar o efeito de inserção combinado ao efeito de redução de ruído. A maior redução de ruído obtida foi para o filtro de Wiener, tendo sido observado que para o filtro sub-ótimo a SNR diminui com o parâmetro α.

A avaliação em termos do desempenho de reconhecimento indica que o filtro sub-ótimo com $\alpha = 0,7$ possui melhor relação entre redução de ruído e inserção de distorção para canal AWGN.

6.1.2 Melhoria do Reconhecimento Automático de Voz na Presença de Ruído

Os percentuais de palavras reconhecidas corretamente obtidos indicam que a aplicação dos filtros implementados para os valores de SNR analisados e ruído aditivo gaussiano branco sempre implica em melhoria no desempenho do sistema de RAV.

Para o filtro sub-ótimo com $\alpha = 0,7$, que apresentou maior percentual de palavras reconhecidas corretamente, a melhoria observada foi de 10% para a menor SNR avaliada e de 14% para a maior SNR avaliada. Esse resultado indica que o uso do filtro sub-ótimo pode facilitar a aplicação dos sistemas RAV em ambientes diversos.

6.2 Trabalhos Futuros

Como os sistemas de reconhecimento automático de voz possuem aplicações em diversas áreas e consequentemente diversos ambientes de uso, a natureza do ruído sonoro presente no sinal de voz capturado pode variar bastante. Vários aspectos referentes à melhoria da estimação de ruído podem ser aperfeiçoados para melhorar o desempenho dos filtros implementados. Alguns desses aspectos são apresentados a seguir.

- Implementação de uma função de detecção de presença de voz que seja mais robusta para baixo valor de SNR.
- Melhoria da estimativa da autocorrelação do ruído por meio do uso da média de estimativas em janelas sem presença de voz.
- Utilização de outro microfone para capturar o áudio em um ponto mais distante da origem de locução, ou seja, um segundo canal que poderá fornecer um sinal de observação com menor SNR mais baixa, possibilitando a estimativa de autocorrelação do ruído nos intervalos de presença de voz.

Além da melhoria de estimativa de ruído, a melhoria do sistema de redução de ruído para reconhecimento em ambientes com baixa SNR pode ser alcançada por:

- Análise do desempenho dos filtros por meio do percentual de palavras reconhecidas corretamente pelo sistema de RAV para sinais de voz contaminados com outros tipos de ruído a fim de ajustar o parâmetro α para diferentes ambientes;
- Ajuste de implementação dos filtros para operação em tempo real após a fase de testes, sem necessidade de utilizar arquivos de áudio, a fim de permitir o uso do reconhecedor de voz para sinais obtidos diretamente por um microfone.

Alguns dos trabalhos futuros possíveis para o sistema desenvolvido são listados a seguir.

- Análise de desempenho do sistema para aplicação de execução de comandos por voz em ambiente de fábrica. Nesse caso, o vocabulário de treinamento é bem mais restrito, e a probabilidade de reconhecimento correto de palavras é maior. Assim, espera-se que a taxa de reconhecimento para SNR abaixo de 5 dB seja maior que a apresentada para reconhecedor com amplo vocabulário;
- Análise do desempenho dos filtros usando análise subjetiva, permitindo a comparação da percepção humana e percepção da máquina mediante alterações de SNR e distorção.

Referências Bibliográficas

[1] L. R. Rabiner and R. W. Schafer. *Introduction to Digital Speech Processing (Foundations and Trends in Signal Processing)*. Now Publishers Inc., 2007.

[2] F. A. Everest. *Master Handbook of Acoustics*. McGraw-Hill, Reading, MA, fourth edition, 2001.

[3] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, New Jersey, first edition, 1978.

[4] L. R. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, first edition, 1996.

[5] J. Chen, J. Benesty, Y. Huang and S. Doclo. "New Insights Into the Noise Reduction Wiener Filter". *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 257 – 286, July 2006.

[6] J. Ortega-García and J. González-Rodríguez. "Overview of Speech Enhancement Techniques for Automatic Speaker Recognition". *International Conference on Spoken Language*, vol. 2, pp. 929 – 932, 1996.

[7] J. M. Fechine. "Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística". Tese de doutorado, Universidade Federal da Paraíba, Campina Grande, PB, Brasil.

[8] K. H. Davis, R. Biddulph and S. Balashek. "Automatic Recognition of Spoken Digits". *Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637 – 642, 1952.

[9] L. Gavidia-Ceballos and J. H. L. Hansen. "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection". *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 4, pp. 373 – 383, April 1996.

[10] M. Turunen, J. Hakulinen, K. Raiha, E. Salonen, A. Kainulainen and P. Prusi. "An architecture and applications for speech-based accessibility systems". *IBM Systems Journal*, vol. 44, no. 3, pp. 485 – 504, 2005.

[11] R. B. Rocha. “Desenvolvimento de um Codificador de Voz Pessoal de Baixa Taxa Baseado em Modelos de Markov Escondidos”. Dissertação de mestrado, Universidade Federal de Campina Grande, Campina Grande, PB, Brasil, Junho 2012.

[12] R. T. Tevah. “Implementação de um Sistema de Reconhecimento de Fala Contínua com Amplo Vocabulário para o Português Brasileiro”. Dissertação de mestrado, Universidade Federal do Rio de Jaineiro, Rio de Janeiro, RJ, Brasil, Junho 2006.

[13] O’Shaughnessy. *Speech Communication, Human and Machine*. Addison-Wesley, Reading, MA, 1987.

[14] R. Singh, R. M. Stern and B. Raj. *Noise Reduction in Speech Applications: Signal and Feature Compensation Methods for Robust Speech Recognition*, chapter 9. CRC Press, 2002.

[15] S. O. Haykin. *Adaptive Filter Theory*. Prentice Hall, fourth edition, 2002.

[16] I. M. A. A. El-Fattah, M. I. Dessouky, S. M. Diab and F. E. A. El-samie. “Speech Enhancement Using an Adaptive Wiener Filtering Approach”. *Progress in Electromagnetics Research*, vol. 4, pp. 167 – 184, 2008.

[17] L. F. da Silva and J. C. M. Bermudez. “Speech Enhancement using a Frame Adaptive Gain Function for Wiener Filtering”. *IEEE Statistical Signal Processing Workshop*, pp. 389 – 392, 2011.

[18] S. F. Boll. “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”. *IEEE Transactions ASSP*, vol. 27, no. 2, pp. 113 – 120, April 1979.

[19] Y. Ephraim and H. L. V. Trees. “A Signal Subspace Approach for Speech Enhancement”. *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[20] Y. Ephraim. “Statistical-Model-Based Speech Enhancement Systems”. *Proceedings of the IEEE*, pp. 1524 – 1555, 1992.

[21] G. P. Eatwell. *Noise Reduction in Speech Applications: Single-Channel Speech Enhancement*, chapter 6. CRC Press, 2002.

[22] M. R. D. Martins. *Ouvir Falar: Introdução à Fonética do Português*. Editora Caminho, 1988.

[23] M. R. D. Martins. *Fonética do Português: Trinta Anos de Investigação*. Editora Caminho, 2002.

[24] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. Illinois Books, Illinois, 1949.

[25] C. P. A. da Silva. “Um Software de Reconhecimento de Voz para Português Brasileiro”. Dissertação de mestrado, Universidade Federal do Pará, Belém, PR, Brasil, Setembro 2010.

[26] Fala Brasil. “Reconhecimento de Voz para o Português Brasileiro”. `http://www.laps.ufpa.br/falabrasil`.

[27] Julius Development Team. “Open-Source Large Vocabulary CSR Engine Julius”. `http://julius.sourceforge.jp/en`, 2012.

[28] S. Davis and P. Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357 – 366, August 1980.

[29] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland. *The HTK Book (for HTK Version 3.1)*. Microsoft Corporation, 2009.

[30] S. International. “SRILM - The SRI Language Modeling Toolkit”. `http://www.speech.sri.com/projects/srilm`, 2011.

[31] J. S. Lim and A. V. Oppenheim. “Enhancement and Bandwidth Compression of Noisy Speech”. *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586 – 1604, December 1979.

[32] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, New York, 1949.

[33] A. N. Kolmogorov. “Stationary Sequences in Hilbert Space (In Russian)”. *Moscow University*, vol. 2, no. 6, pp. 1 – 40, 1941. English translation in Kailath T. (ed.) Linear least squares estimation Dowden, Hutchinson and Ross 1977.

[34] A. V. Oppenheim and G. C. Verghese. *Signals, Systems and Inference – Class Notes for 6.011*. Massachusetts Institute of Technology, 2010.

[35] B. P. Lathi. *Modern Digital and Analog Communications Systems*. Oxford University Press, first edition, 1988.

[36] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, third edition, 1991.

[37] W. Guang-yan, Z. Xiao-qun and W. Xia. “Musical Noise Reduction Based on Spectral Subtraction Combined with Wiener Filtering for Speech Communication”. *IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2009)*, , no. December, pp. 726 – 729, 2009.

[38] M. Mathe, S. P. Nandyala and T. K. Kumar. "Speech Enhancement Using Kalman Filter for white, random and color noise". *2012 International Conference on Devices, Circuits and Systems (ICDCS)*, pp. 195 – 198, 2012.

[39] A. D. Malayeri. "Noise Speech Wavelet Analyzing in Special Time Ranges". *The 12th International Conference on Advanced Communication Technology*, pp. 525 – 528, February 2010.

[40] B. Carnero and A. Drygajlo. "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms". *IEEE Transactions on Signal Processing*, vol. 47, pp. 1622 – 1635, June 1999.

[41] SoX. "SoX – Sound eXchange". `http://sox.sourceforge.net`.

[42] M. Vondrasek and P. Pollak. "Methods for Speech SNR Estimation: Evaluation Tool and Analysis of VAD Dependency". *Radioengineering*, vol. 14, no. 1, April 2005.

[43] S. Weiss, R. W. Stewart and G. M. Davis. *Noise Reduction in Speech Applications: Noise and Digital Signal Processing*, chapter 1. CRC Press, 2002.

[44] Speech Processing Group - Czech Technical University. "Speech Processing Group". `http://noel.feld.cvut.cz/speechlab`, 2005.

[45] T. Raitio, M. Takanen, O. Santala, A. Suni, M. Vaini and P. Alku. "On Measuring the Intelligibility of Synthetic Speech in Noise – Do we Need a Realistic Noise Environment?" *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, , no. March, pp. 4025 – 4028, 2012.

[46] G. H. Lee, S. J. Kang, C. W. Han and N. S. Kim. "Feature Enhancement Error Compensation for Noise Robust Speech Recognition". *2012 9th International Multi-Conference on Systems, Signals and Devices*, pp. 1 – 4, March 2012.

[47] L. ying Sui, X. wei Zhang, J. jun Huang and B. Zhou. "An Improved Spectral Subtraction Speech Enhancement Algorithm under Non-Stationary Noise". *2011 International Conference on Wireless Communications and Signal Processing*, pp. 1 – 5, November 2011.

[48] P. Fardkhaleghi and M. H. Savoji. "New Approaches to Speech Enhancement Using Phase Correction in Wiener Filtering". *2011 International Conference onWireless Communications and Signal Processing*, pp. 895 – 899, December 2011.

[49] S. V. R. Rao, M. B. R. Murthy and K. S. Rao. "Speech Enhancement Using Perceptual Wiener Filter Combined with Unvoiced Speech – A New Scheme". *2011 IEEE Recent Advances in Intelligent Computational Systems*, pp. 688 – 691, September 2011.

[50] K. Ngo, M. Moonen, S. H. Jensen and J. Wouters. “A flexible Speech Distortion Weighted Multi-channel Wiener Filter for Noise Reduction in Hearing Aids”. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2528 – 2531, March 2011.

[51] J. Hou, Y. Liu, C. Zhang and S. Huang. “An In-car Chinese Noise Corpus for Speech Recognition”. *2011 International Conference on Asian Language Processing*, pp. 228 – 231, November 2011.

[52] A. R. Fukane and S. L. Sahare. “Role of Noise Estimation in Enhancement of Noisy Speech Signals for Hearing Aids”. *2011 International Conference on Computational Intelligence and Communication Networks*, vol. 1, pp. 648 – 652, October 2011.

[53] L. R. Rabiner. “Special Issue on Man-machine Communication by Voice”. *Proceedings of the IEEE*, vol. 64, pp. 403 – 404, 1976.

[54] M. S. de Alencar. *Probabilidade e Processos Estocásticos*. Érica, 2009.

[55] X. D. Huang, A. Ariki and M. A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, 1990.

[56] T. Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, The Netherlands, March 2001.

[57] F. A. Everest. *Handbook for Sound Engineers: Fundamentals of sound*. Howard W. Sams & Company, 1987.

Apêndice A
Código para Implementação dos Filtros

```
% Autor: Ísis de Andrade Lima
% Fevereiro de 2014

% Simulações para dissertação:
% Filtros de Wiener FIR e FIR modificado

% Características do algoritmo:
% – Janelamento retangular usando função "buffer"
% – Estimativa de das autocorrelações usando 'xcorr' 'biased'
% – Cálculo de Ry pela função 'toeplitz'

% Variáveis Importantes:
% L: ordem dos filtros FIR
% alpha: fator de multiplicação do filtro FIR modificado
% winTime: tamando da janela (quadro) em segundos
% treshold: limiar para considerar que existe algum sinal presente
% speechSignal (x): vetor de amostras de entrada (voz sem ruído)
% noisySignal (y): vetor de amostras do sinal observado (voz e ruido)
% VAD_n = sinal de verificação de presença (quadros com tamanho
    winTime)
% winLen: tamando da janela (quadro) em número de amostras

%OBS: todos os arquivos de áudio estão no mesmo formato, com mesma
%frequencia de amostragem e número de bits.

clear all
close all

%Arquivos de Entrada:
```

```
clean_speech = 'frase1.wav';                %Voz sem ruído
mixed_speech = 'frase1_0dB.wav';            %Voz com ruído
VAD = 'isis_casa_vad'                       %VAD para quadros com
    duração winTime

%Arquivos de Saída:
outputWiener = 'frase1_wiener_0dB.wav';     %Saída do filtro de Wiener
outputSubOpt1= 'frase1_subOpt1_0dB.wav';    %Saída do filtro Sub-Ótipo
    para alpha1
outputSubOpt2= 'frase1_subOpt1_0dB.wav';    %para alpha2
outputSubOpt3= 'frase1_subOpt1_0dB.wav';    %para alpha3

v_Wiener = 'res_frase1_wiener_0dB.wav';     %Ruídos residuais
v_SubOpt1 = 'res_frase1_subOpt1_0dB.wav';
v_SubOpt2 = 'res_frase1_subOpt2_0dB.wav';
v_SubOpt3 = 'res_frase1_subOpt3_0dB.wav';

% Parâmetros dos Filtros
L = 20;          %Ordem dos filtros
alpha1 = 0.8;
alpha2 = 0.7;
alpha3 = 0.5;

%Tamanho da Janela de Análise em Segundos
winTime = 20e-3;

%Nível mínimo (de potência) para presença de sinal
treshold = 6e-4;

% Sinal de Entrada: Arquivo de audio WAV
[speechSignal, Fs, nbits1] = wavread(cleanSpeech, 'double');

% Saida do Canal AWGN: Arquivo de audio WAV
[noisySignal, Fs, nbits1] = wavread(mixedSpeech, 'double');

% Sinal de detecção de presença de voz
VAD_n = open(VAD);

% Ajuste do tamanho dos sinais
L_x = min([length(speechSignal) length(noisySignal)]); % Comprimento
    do sinal em número de amostras
```

```
speechSignal = speechSignal(1:L_x,1)-mean(speechSignal(:,1));
noisySignal = noisySignal(1:L_x,1);

% Ruido
v = y-x;

% Janelas
winLen = ceil(Fs*winTime);
winOverlap = 0;

%wHamm = hamming(winLen);
wRect = rectwin(winLen);

% Framing/windowing
% sigFramed = buffer(noisySignal, winLen, winOverlap, 'nodelay');
% sigWindowed = diag(sparse(wHamm)) * sigFramed;
sigFramed = buffer(noisySignal, winLen, winOverlap, 'nodelay');
sigWindowed = diag(sparse(wRect)) * sigFramed;

u_1 = zeros([L,1]);  %filtro impulso
u_1(1)=1;

% Autocorrelação do ruido
r_v = zeros(L,1); % é inicializado como um vetor nulo

% Calculo Matriz Autocorrelação e Filtro
[winLen, N] = size(sigWindowed);

for k=1:N

    %Verificação de presença de sinal
    if (sum(sigWindowed(:,k)).^2)<treshold)  %se não existe sinal
       presente,
        r_v = zeros(L,1);                    %a autocorrelção do
           ruido é nula
    end

    %Verificação de presença de voz
    if (VAD_n(k) == 0)  %se não existe voz presente,
```

```
        r_v = xcorr(sigWindowed(:,k),L-1,'biased');  % a
            autocorrelação do
        r_v = r_v(L:end);                            % ruido é
            atualizada
    end

    %Matriz de autocorrelação de y
    r_y = xcorr(sigWindowed(:,k),L-1,'biased');
    r_y = r_y(L:end);
    R_y = toeplitz(r_y(1:L));

    %matriz inversa e filtros
    [Lu, U] = lu(R_y);                    %calculo das matrizes L-U

    h_0 = u_1 - (inv(Lu)*inv(U))*r_v;     %filtro de wiener FIR

    h_s1 = (1-alpha1)*u_1 + alpha1*h_0;      %filtro sub-otimo FIR
    h_s2 = (1-alpha2)*u_1 + alpha2*h_0;      %filtro sub-otimo FIR
    h_s3 = (1-alpha3)*u_1 + alpha3*h_0;      %filtro sub-otimo FIR

    % Buffer para calculo das primeiras L amostras
    if (k==1)
        buffer = zeros([2*L-1,1]);
    end
    if (k>1)
        buffer = cat (1,sigWindowed((winLen-L+1):winLen,k-1),
            sigWindowed(1:L-1,k));
    end

    %Aplicacao da filtragem: s para Wiener e sn para Sub-Otimo
    for l=1:winLen
        if (l<L) s_buffer(l) = h_0'*wrev(buffer(l:(l+L-1)));
                 s1_buffer(l) = h_s1'*wrev(buffer(l:(l+L-1)));
                 s2_buffer(l) = h_s2'*wrev(buffer(l:(l+L-1)));
                 s3_buffer(l) = h_s3'*wrev(buffer(l:(l+L-1)));
        end
        if(l>=L) s_buffer(l)=h_0'*wrev(sigWindowed((l-L+1):l,k));
                 s1_buffer(l)=h_s1'*wrev(sigWindowed((l-L+1):l,k));
                 s2_buffer(l)=h_s2'*wrev(sigWindowed((l-L+1):l,k));
                 s3_buffer(l)=h_s3'*wrev(sigWindowed((l-L+1):l,k));
        end
```

```
    end
    s(:,k) = s_buffer;
    s1(:,k) = s1_buffer;
    s2(:,k) = s2_buffer;
    s3(:,k) = s3_buffer;
end

% Ajuste do sinal de saida: blocos(janelas)->vetor
L_out = numel(s);
L_out = min([L_out L_x]);

outputSignalWiener = reshape(s,1,numel(s));
outputSignalWiener = outputSignalWiener(1:L_out)';

outputSignalSubOpt1 = reshape(s1,1,numel(s1));
outputSignalSubOpt1 = outputSignalSubOpt1(1:L_out)';

outputSignalSubOpt2 = reshape(s2,1,numel(s2));
outputSignalSubOpt2 = outputSignalSubOpt2(1:L_out)';

outputSignalSubOpt3 = reshape(s3,1,numel(s3));
outputSignalSubOpt3 = outputSignalSubOpt3(1:L_out)';

%%
wavwrite(outputSignalWiener, Fs, nbits1, outputWiener);
wavwrite(outputSignalSubOpt1, Fs, nbits1, outputSubOpt1);
wavwrite(outputSignalSubOpt2, Fs, nbits1, outputSubOpt2);
wavwrite(outputSignalSubOpt3, Fs, nbits1, outputSubOpt3);
%%

% Obtenção das SNR de entrada e saída
% Ruido residual
v_residualWiener = outputSignalWiener(1:L_out) - speechSignal(1:L_out
    );
v_residualSubOpt1 = outputSignalSubOpt1(1:L_out) - speechSignal(1:
    L_out);
v_residualSubOpt2 = outputSignalSubOpt2(1:L_out) - speechSignal(1:
    L_out);
v_residualSubOpt3 = outputSignalSubOpt3(1:L_out) - speechSignal(1:
    L_out);
```

```
% Calculo das SNR
SNR_in = mean(x(1:L_out).^2)/mean(v(1:L_out).^2);
SNR_outWiener = mean(x(1:L_out).^2)/mean(v_residualWiener(1:L_out)
    .^2);
SNR_outSubOpt1 = mean(x(1:L_out).^2)/mean(v_residualSubOpt1(1:L_out)
    .^2);
SNR_outSubOpt2 = mean(x(1:L_out).^2)/mean(v_residualSubOpt2(1:L_out)
    .^2);
SNR_outSubOpt3 = mean(x(1:L_out).^2)/mean(v_residualSubOpt3(1:L_out)
    .^2);

% SNR em dB
SNR_in_db = 10*log10(SNR_in)
SNR_out_dbWiener = 10*log10(SNR_outWiener)
SNR_out_dbSubOpt1 = 10*log10(SNR_outSubOpt1)
SNR_out_dbSubOpt2 = 10*log10(SNR_outSubOpt2)
SNR_out_dbSubOpt3 = 10*log10(SNR_outSubOpt3)

%GRÁFICOS
% Vetor de tempo
t = [0:L_out-1]/Fs;

xmin = 1.75;
xmax = 4.6;
ymin = -0.2;
ymax = 0.2;

figure (1)
subplot(6,1,1)
plot(t, speechSignal(1:L_out));
title('Sinal de Entrada - Frase 1');
axis ([xmin xmax ymin ymax]);

subplot(6,1,2)
plot(t, noisySignal(1:L_out))
title('Sinal com Ruido')
axis ([xmin xmax ymin ymax]);

subplot(6,1,3)
plot(t, outputSignalSubOpt1(1:L_out),'b');
title('Filtro Sub-Ótimo, \alpha = 0.5');
```

```
axis ([xmin xmax ymin ymax]);

subplot(6,1,4)
plot(t, outputSignalSubOpt2(1:L_out),'b');
title('Filtro Sub-Ótimo, \alpha = 0.7');
axis ([xmin xmax ymin ymax]);

subplot(6,1,5)
plot(t, outputSignalSubOpt3(1:L_out),'b');
title('Filtro Sub-Ótimo, \alpha = 0.8');
axis ([xmin xmax ymin ymax]);

subplot(6,1,6)
plot(t, outputSignalWiener(1:L_out),'b');
title('Filtro de Wiener');
xlabel ('Tempo (segundos)')
axis ([xmin xmax ymin ymax]);
figure (2)
plot(t, outputSignalWiener(1:L_out),'b');
hold on
plot(t, outputSignalSubOpt1(1:L_out), '-r');
xlim(xlims);
xlabel('Tempo (seg)');
legend({'Filtro Wiener','Filtro Sub-Otimo'});
grid on
```

Apêndice B
Frases Utilizadas nos Testes Realizados

Frases utilizadas para testes de reconhecimento

Locutores do sexo feminino

Frase 1. Hoje pela manhã não haverá aula.

Frase 2. A medida seria tomada caso o pacote fiscal fracassasse.

Frase 3. O telefone toca na delegacia e ouve-se uma voz desesperada.

Frase 4. Ele já foi vítima de dois graves atentados a bomba.

Frase 5. Quero mandar um abraço para Angela Maria.

Frase 6. Identifica uma consciência dupla que resiste.

Frase 7. O banco é um instrumento para isso.

Frase 8. Aposentadoria não estava em seus planos.

Frase 9. Polícia mostra que está na hora de acabar com a droga.

Frase 10. Áreas de maior potencial de geração de emprego.

Locutores do sexo masculino

Frase 11. Cada ilha possui seus diferentes pratos típicos

Frase 12. O valor do negócio não foi revelado.

Frase 13. Vilarejos próximos a fronteira também foram bombardeados.

Frase 14. Nossos clientes saem das boates.

Frase 15. Para isso recorreu à formação original do grupo.

Frase 16. Os dois bancos deram bons lucros.

Frase 17. O presidente também quer mudar.

Frase 18. Não se trata no imaginário desses artistas.

Frase 19. As duas redes nacionais agora estariam disputando vendas.

Frase 20. Meu marido parecia muito com ele.

Apêndice C
Resultados Detalhados para Frase 1

Formas de Onda Obtidas para Frase 1

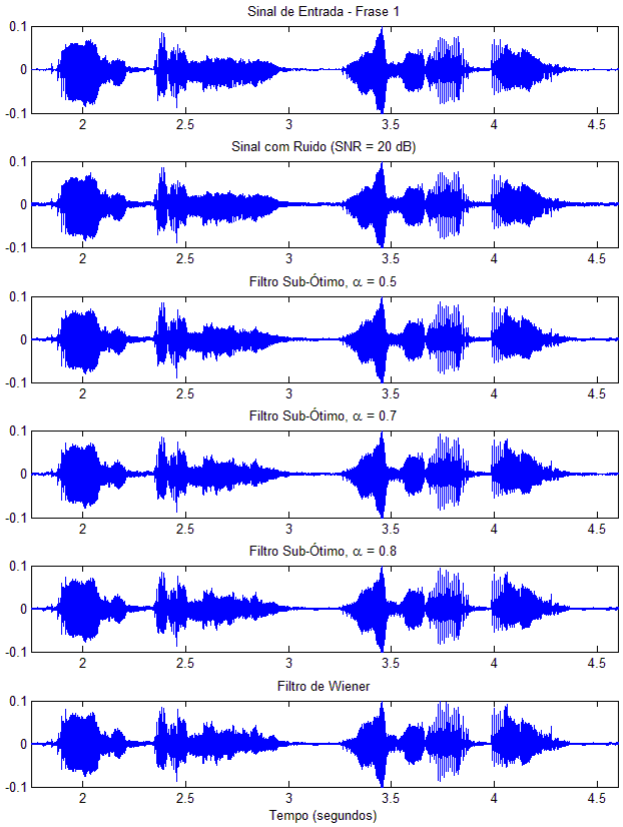

Figura C.1 Sinais de saída dos filtros obtidos para Frase 1 e SNR 20 dB para $L = 20$.

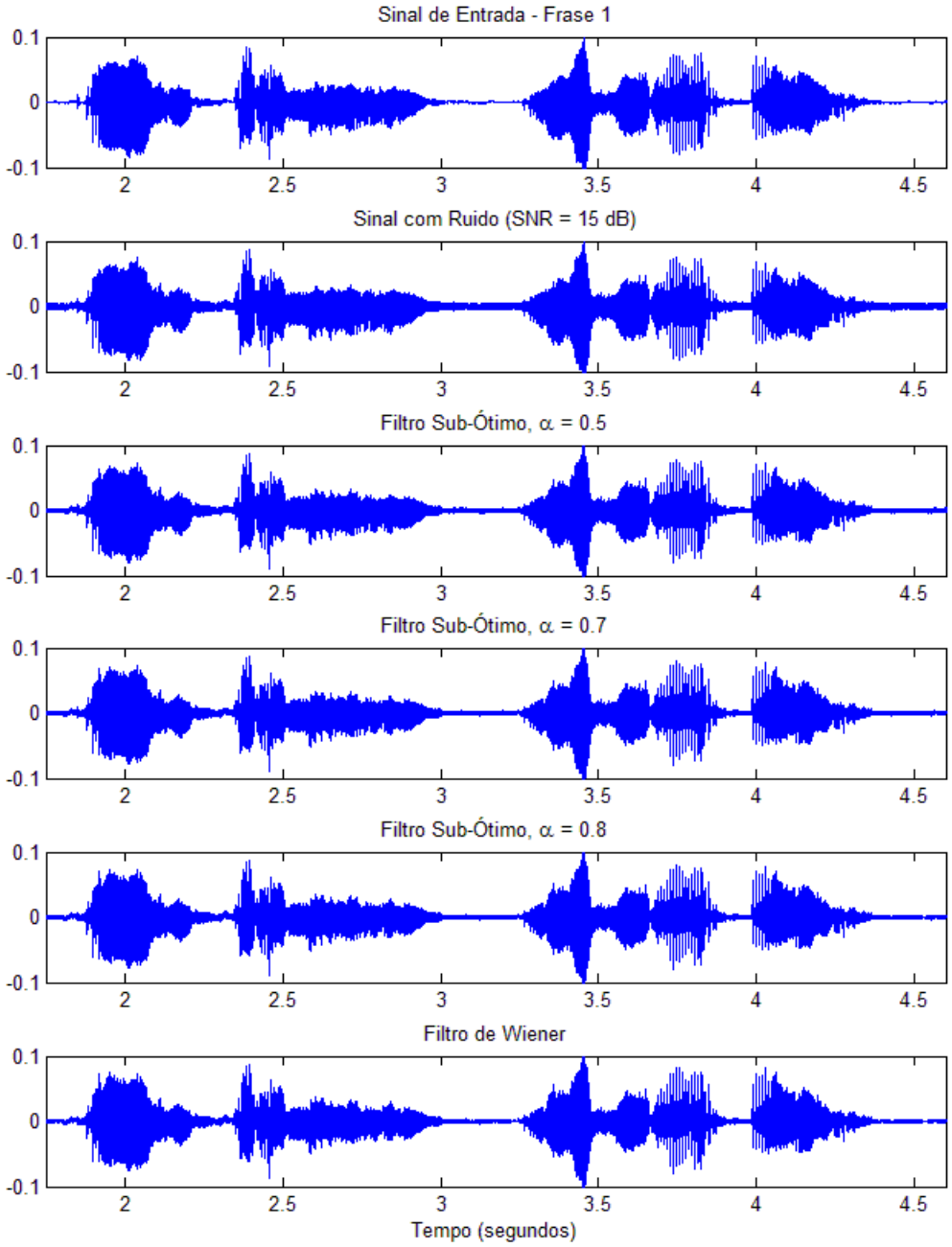

Figura C.2 Sinais de saída dos filtros obtidos para Frase 1 e SNR 15 dB para $L = 20$.

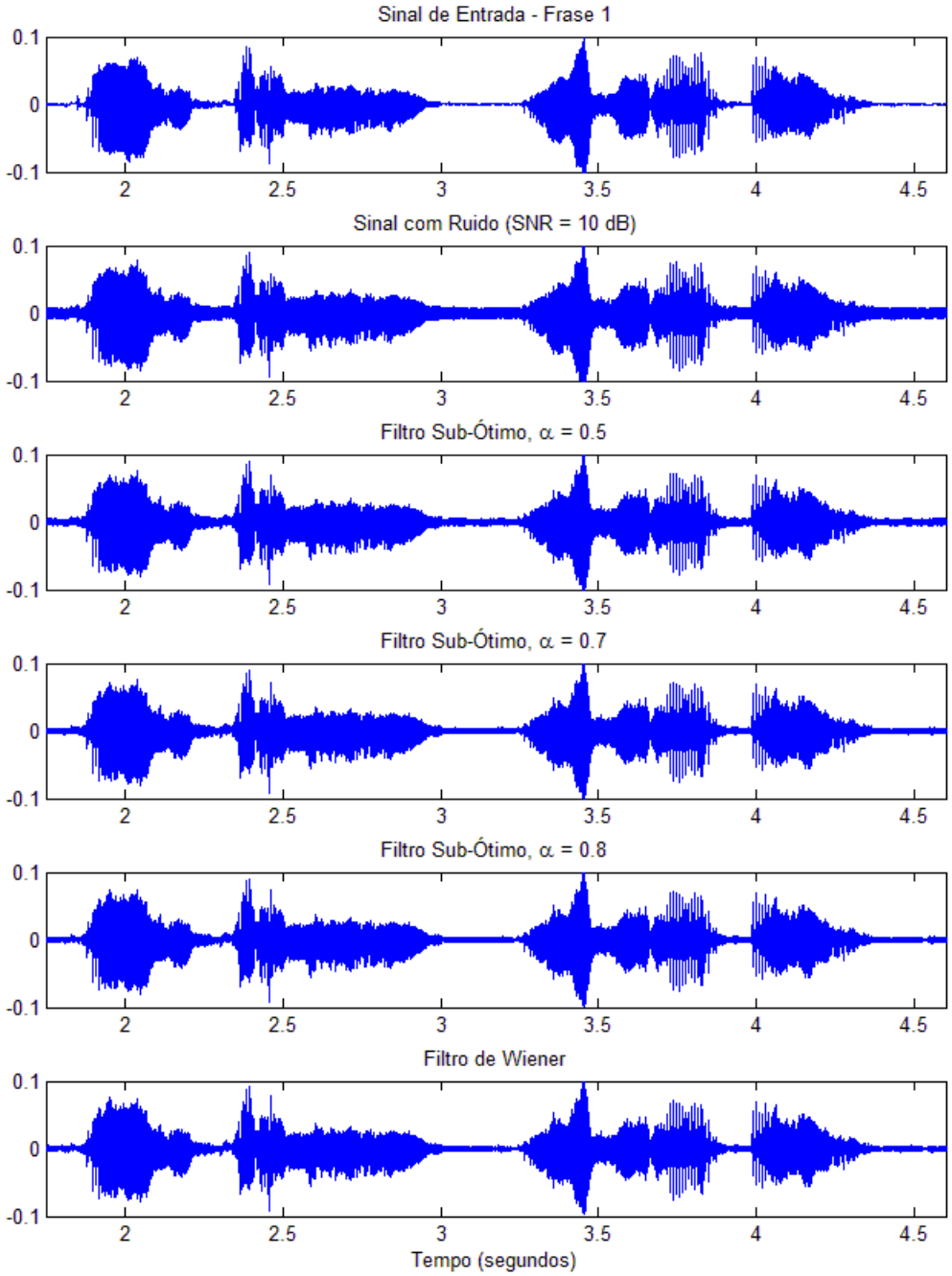

Figura C.3 Sinais de saída dos filtros obtidos para Frase 1 e SNR 10 dB para $L = 20$.

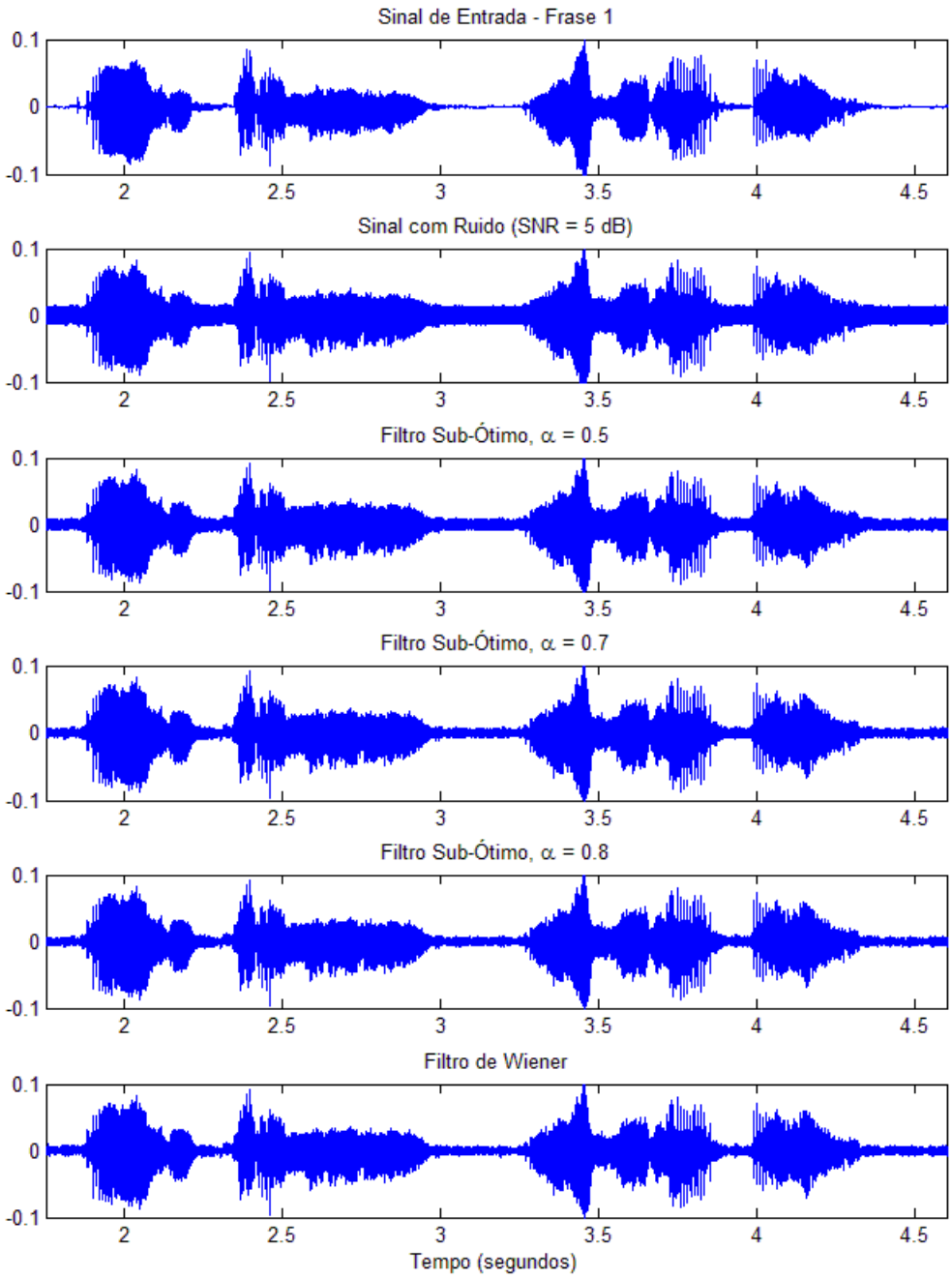

Figura C.4 Sinais de saída dos filtros obtidos para Frase 1 e SNR 5 dB para $L = 20$.

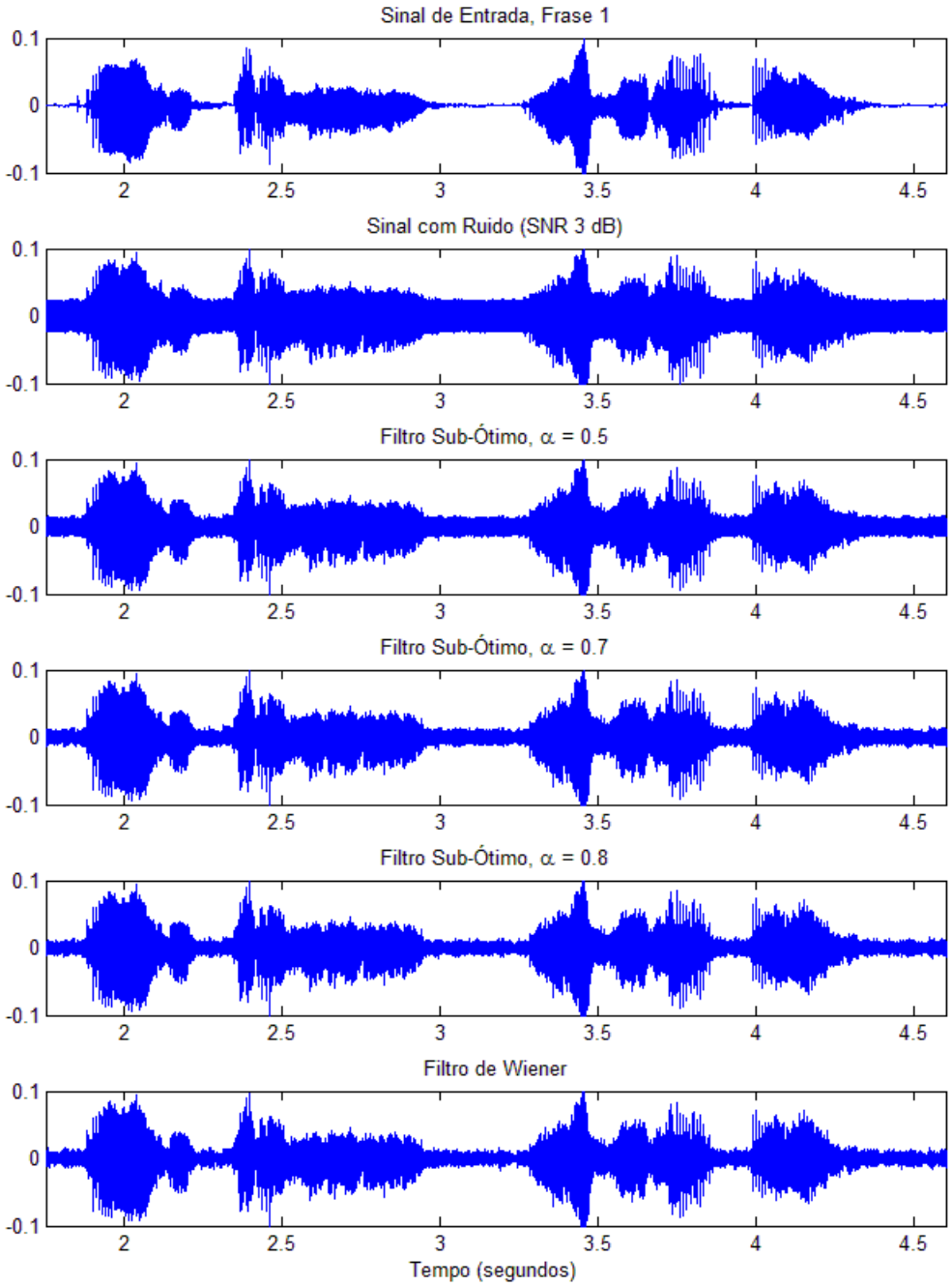

Figura C.5 Sinais de saída dos filtros obtidos para Frase 1 e SNR 3 dB para $L = 20$.

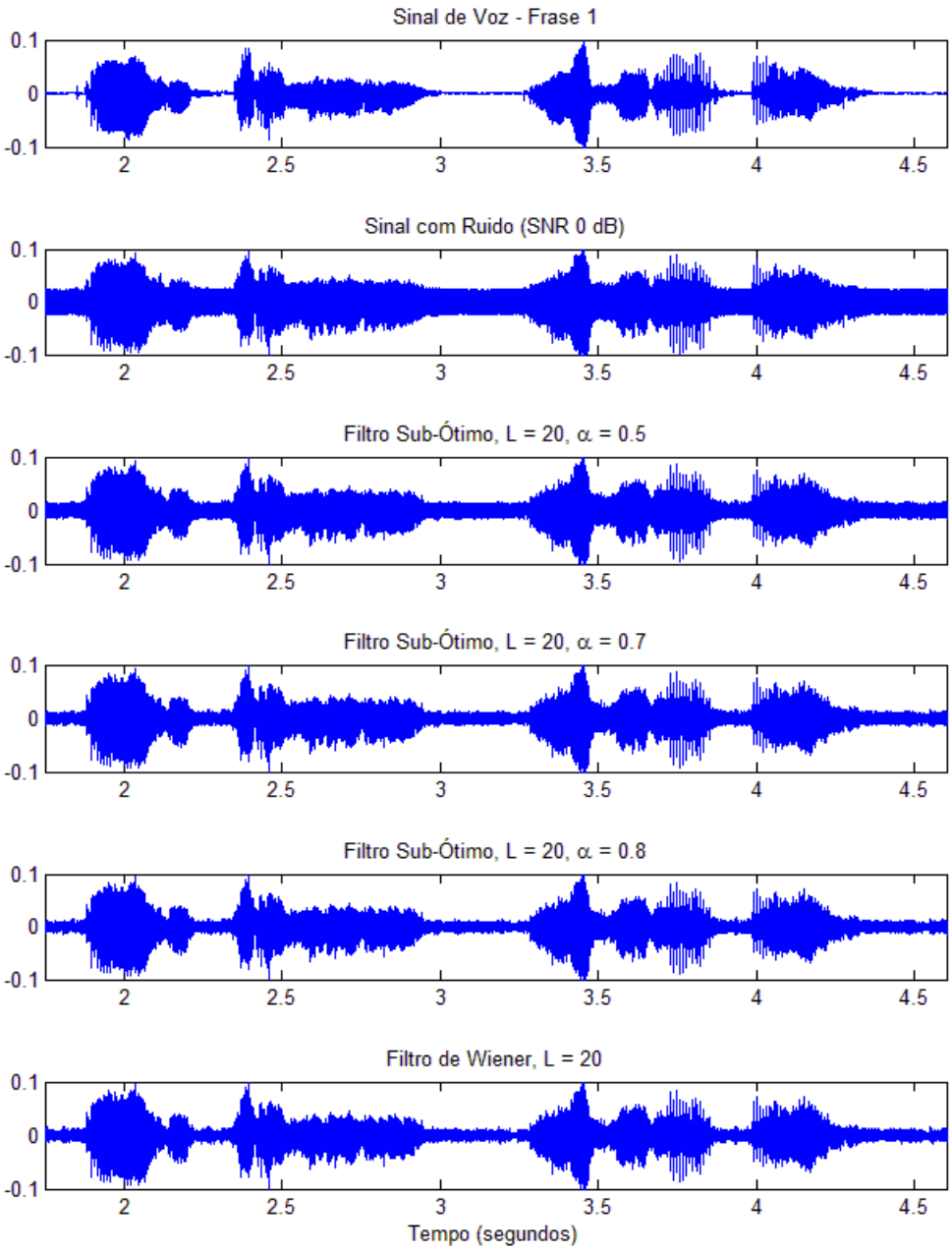

Figura C.6 Sinais de saída dos filtros obtidos para Frase 1 e SNR 0 dB para $L = 20$.

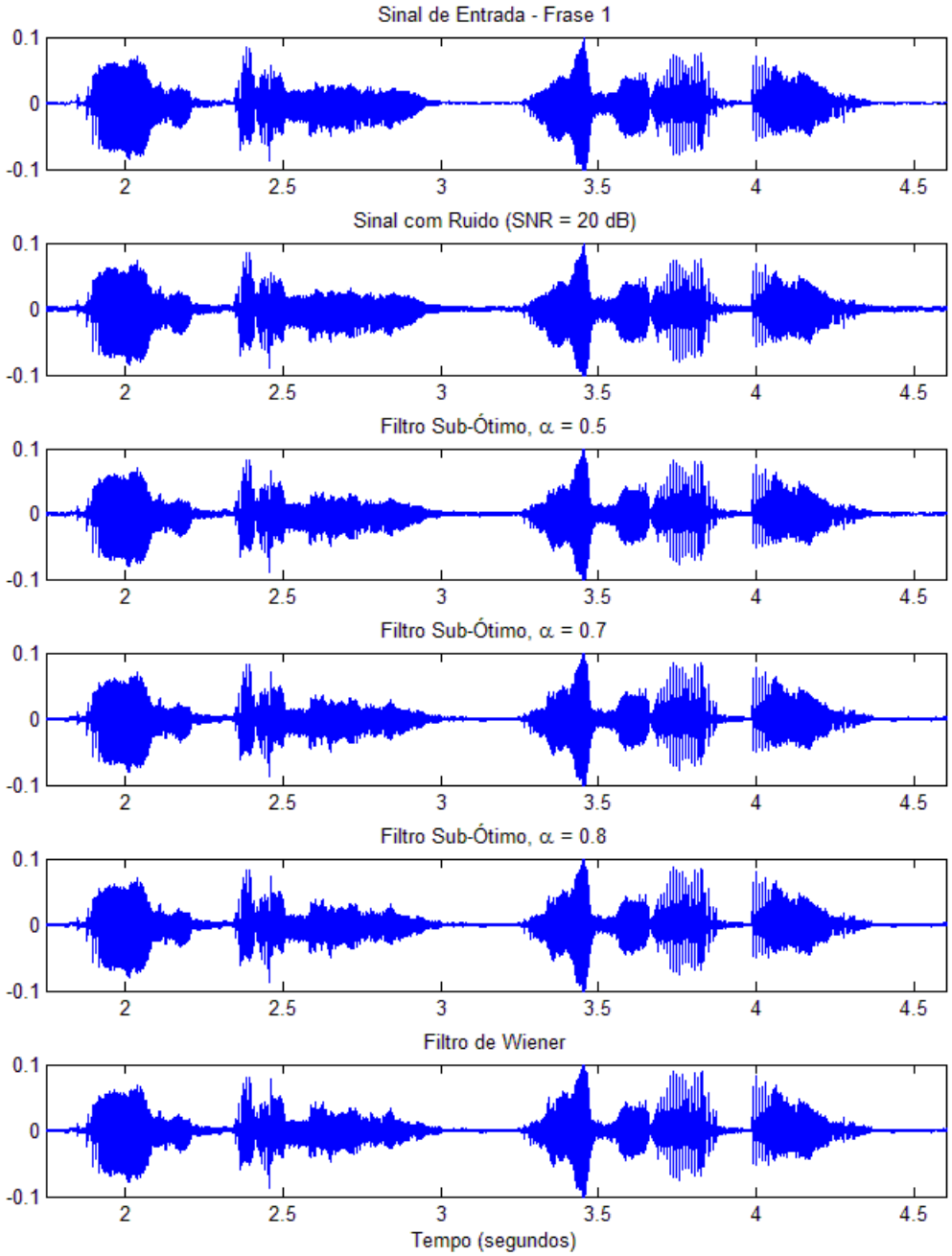

Figura C.7 Sinais de saída dos filtros obtidos para Frase 1 e SNR 20 dB para $L = 10$.

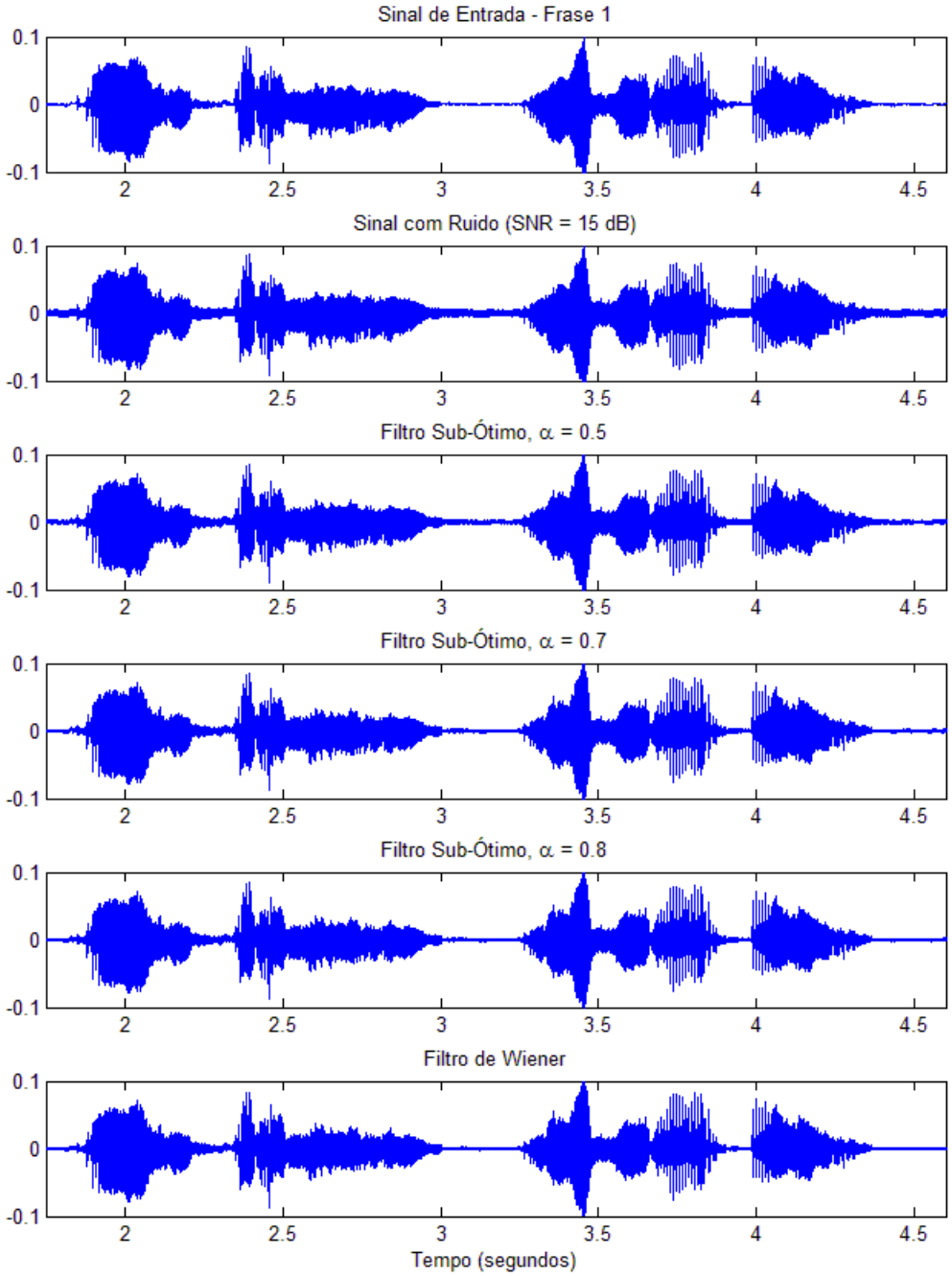

Figura C.8 Sinais de saída dos filtros obtidos para Frase 1 e SNR 15 dB para $L = 10$.

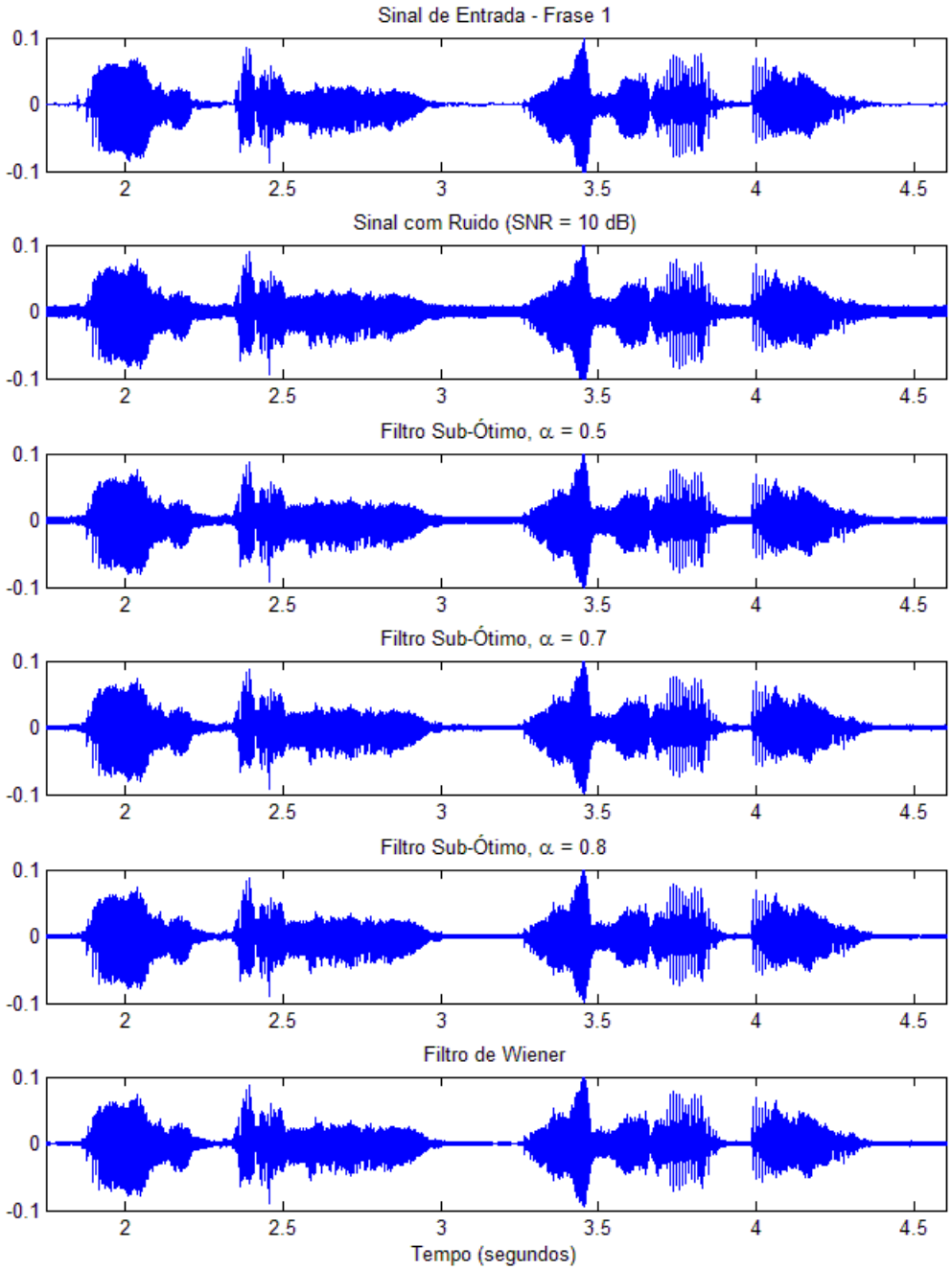

Figura C.9 Sinais de saída dos filtros obtidos para Frase 1 e SNR 10 dB para $L = 10$.

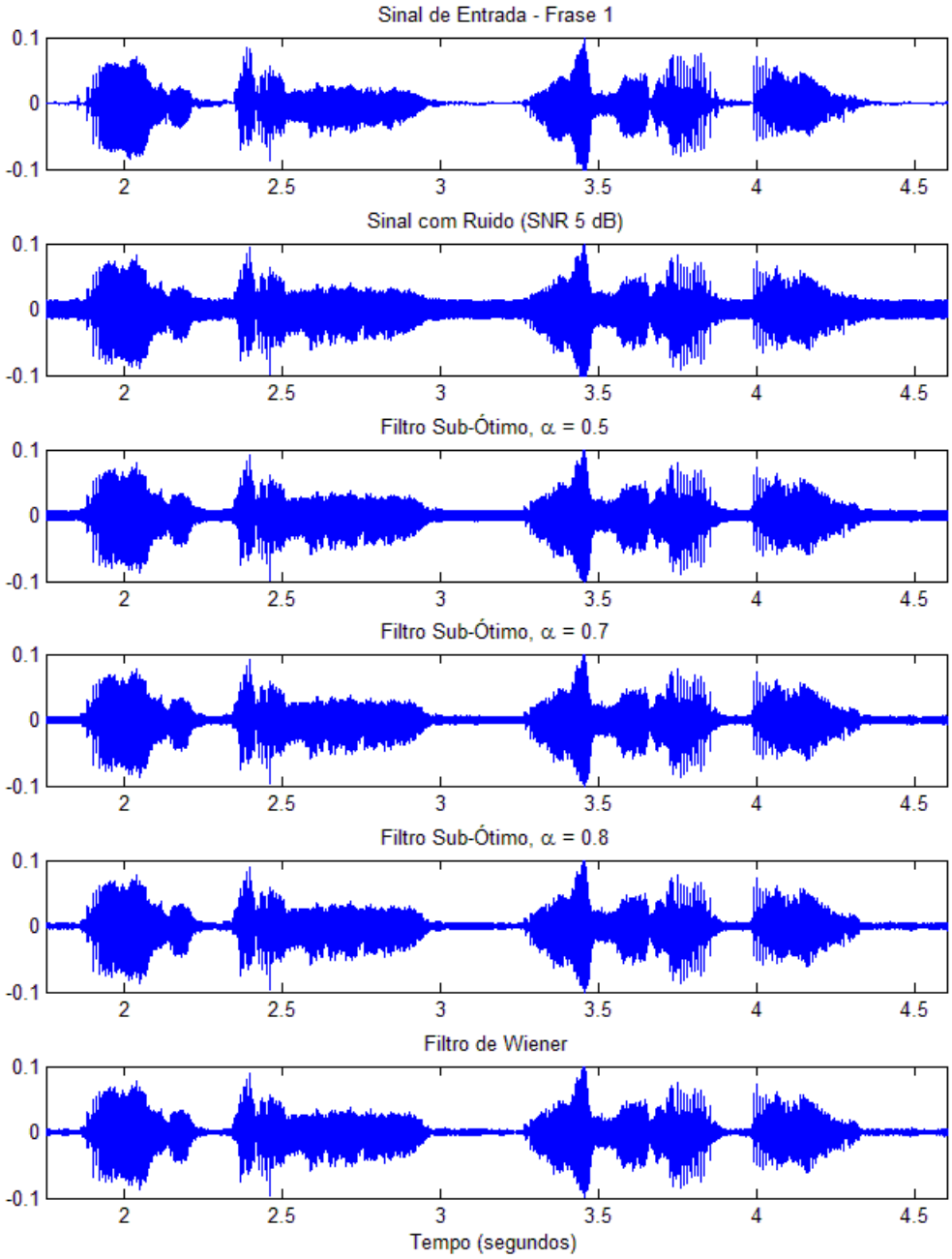

Figura C.10 Sinais de saída dos filtros obtidos para Frase 1 e SNR 5 dB para $L = 10$.

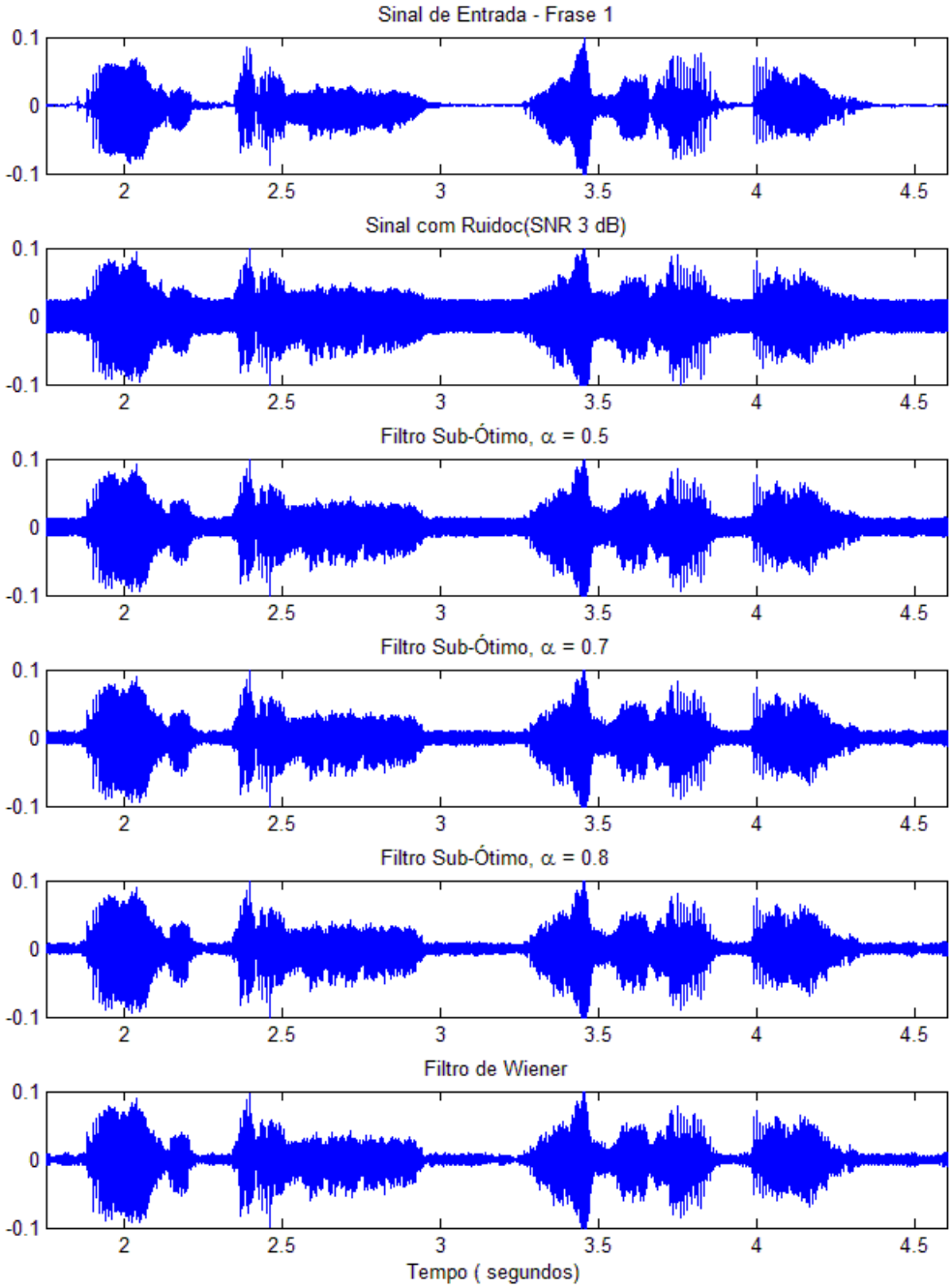

Figura C.11 Sinais de saída dos filtros obtidos para Frase 1 e SNR 3 dB para $L = 10$.

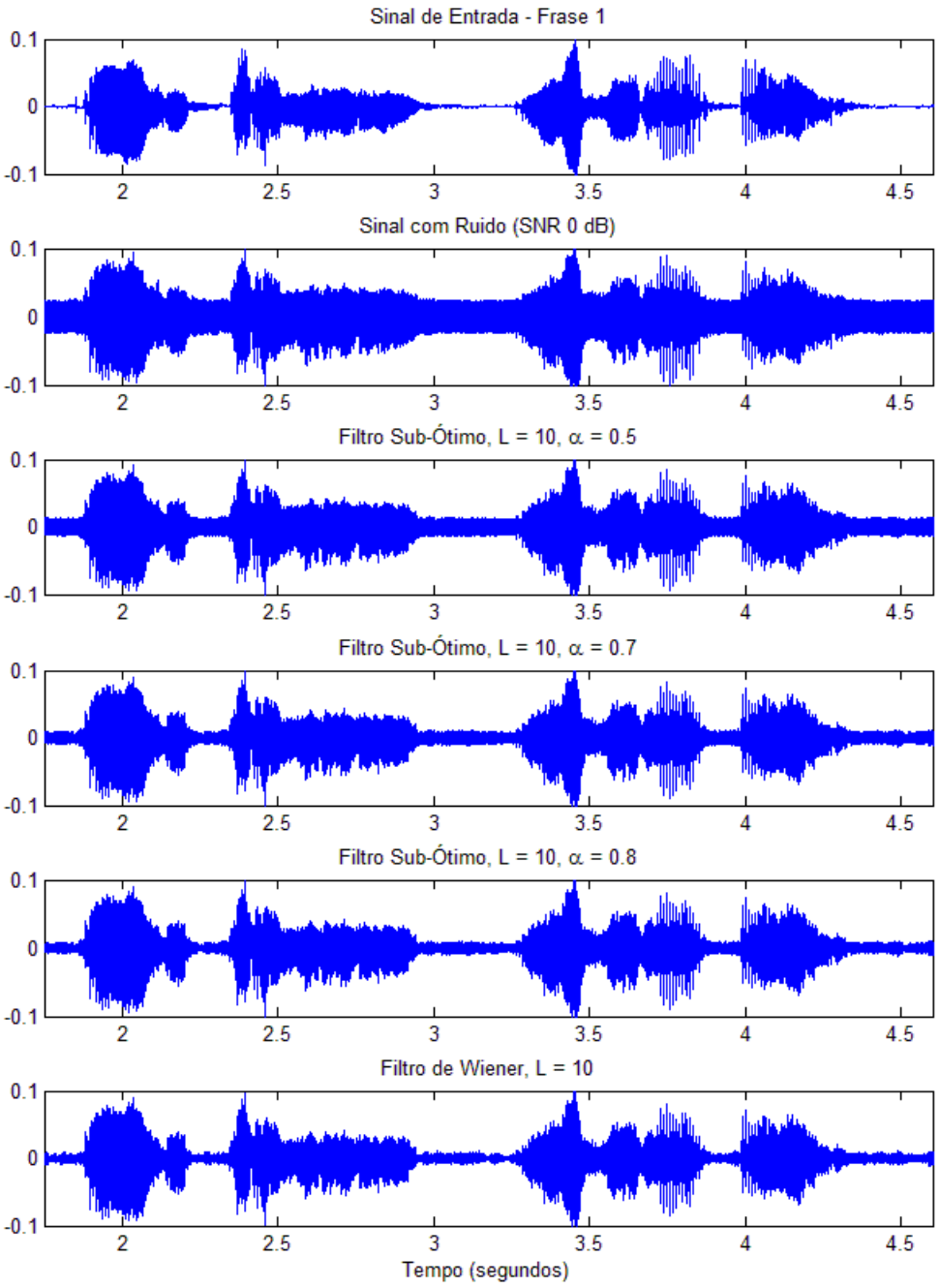

Figura C.12 Sinais de saída dos filtros obtidos para Frase 1 e SNR 0 dB para $L = 10$.

Apêndice D
Resultados do Reconhecimento

				Sub-Ótimo		
Frase	Número de Palavras	Sem Filtragem	Wiener	$\alpha = 0.8$	$\alpha = 0.7$	$\alpha = 0.5$
Frase 1	6	6	4	5	5	6
Frase 2	9	5	5	5	5	5
Frase 3	10	2	1	2	2	2
Frase 4	10	10	7	7	7	8
Frase 5	7	6	5	5	5	6
Frase 6	6	3	2	3	3	3
Frase 7	7	4	4	4	4	4
Frase 8	6	3	1	2	3	3
Frase 9	11	10	9	9	9	9
Frase 10	8	6	5	6	6	6
Frase 11	7	4	4	4	4	4
Frase 12	7	7	7	7	7	7
Frase 13	7	6	6	6	6	6
Frase 14	5	1	1	1	1	1
Frase 15	8	5	5	5	5	5
Frase 16	6	6	6	6	6	6
Frase 17	5	4	3	4	4	4
Frase 18	7	3	3	3	3	3
Frase 19	8	7	7	7	7	7
Frase 20	6	6	6	6	6	6
Total	146	104	91	97	98	101
Percentual de acerto		71.2%	62.3%	66.4%	67.1%	69.2%

Tabela D.1 Palavras reconhecidas corretamente para entrada sem ruído inserido.

				Sub-Ótimo		
Frase	Número de Palavras	Sem Filtragem	Wiener	$\alpha = 0.8$	$\alpha = 0.7$	$\alpha = 0.5$
Frase 1	6	4	5	5	5	4
Frase 2	9	5	5	5	5	5
Frase 3	10	2	1	1	2	2
Frase 4	10	6	6	6	6	6
Frase 5	7	2	2	2	2	2
Frase 6	6	0	1	1	1	1
Frase 7	7	3	2	2	4	3
Frase 8	6	1	1	1	3	2
Frase 9	11	5	6	6	7	6
Frase 10	8	5	5	5	5	5
Frase 11	7	3	3	3	3	3
Frase 12	7	2	6	6	6	5
Frase 13	7	6	5	5	5	6
Frase 14	5	0	1	1	1	1
Frase 15	8	4	4	4	4	4
Frase 16	6	3	4	4	5	5
Frase 17	5	4	4	4	4	4
Frase 18	7	2	3	3	3	3
Frase 19	8	2	3	3	5	4
Frase 20	6	1	3	3	3	2
Total	146	60	70	72	79	73
Percentual de acerto		41.1%	47.9%	49.3%	54.1%	50.0%

Tabela D.2 Palavras reconhecidas corretamente para entrada com SNR 20 dB.

				Sub-Ótimo		
Frase	Número de Palavras	Sem Filtragem	Wiener	$\alpha = 0.8$	$\alpha = 0.7$	$\alpha = 0.5$
Frase 1	6	4	5	4	5	4
Frase 2	9	4	5	5	6	5
Frase 3	10	2	2	2	2	2
Frase 4	10	3	5	6	3	3
Frase 5	7	1	1	2	2	1
Frase 6	6	0	0	0	0	0
Frase 7	7	2	1	2	3	3
Frase 8	6	1	2	1	1	1
Frase 9	11	4	5	5	6	5
Frase 10	8	5	5	5	5	4
Frase 11	7	1	3	3	2	3
Frase 12	7	2	2	2	2	2
Frase 13	7	4	5	5	6	5
Frase 14	5	0	1	0	1	1
Frase 15	8	2	3	2	2	2
Frase 16	6	2	2	1	4	1
Frase 17	5	2	1	2	3	2
Frase 18	7	3	3	3	3	3
Frase 19	8	3	3	3	5	4
Frase 20	6	1	1	2	3	2
Total	146	46	55	55	64	53
Percentual de acerto		$31,5\%$	$37,7\%$	$27,7\%$	$43,8\%$	$36,3\%$

Tabela D.3 Palavras reconhecidas corretamente para entrada com SNR 15 dB.

				Sub-Ótimo		
Frase	Número de Palavras	Sem Filtragem	Wiener	$\alpha = 0.8$	$\alpha = 0.7$	$\alpha = 0.5$
Frase 1	6	4	5	4	5	4
Frase 2	9	3	2	3	4	3
Frase 3	10	2	2	2	2	2
Frase 4	10	0	0	2	2	1
Frase 5	7	0	0	1	1	1
Frase 6	6	0	0	0	0	0
Frase 7	7	2	2	2	3	3
Frase 8	6	0	0	1	2	1
Frase 9	11	2	3	4	5	2
Frase 10	8	2	2	2	2	2
Frase 11	7	1	2	1	1	1
Frase 12	7	2	2	2	2	2
Frase 13	7	2	3	2	2	2
Frase 14	5	0	1	1	1	1
Frase 15	8	2	2	2	2	2
Frase 16	6	1	1	1	2	1
Frase 17	5	1	1	3	3	0
Frase 18	7	2	2	2	4	2
Frase 19	8	3	3	3	5	4
Frase 20	6	1	1	2	3	2
Total	146	30	34	40	51	36
Percentual de acerto		$20,5\%$	$23,3\%$	$27,4\%$	$34,9\%$	$24,7\%$

Tabela D.4 Palavras reconhecidas corretamente para entrada com SNR 10 dB.

				Sub-Ótimo		
Frase	Número de Palavras	Sem Filtragem	Wiener	$\alpha = 0.8$	$\alpha = 0.7$	$\alpha = 0.5$
Frase 1	6	2	2	2	3	3
Frase 2	9	3	2	2	2	2
Frase 3	10	2	2	2	2	2
Frase 4	10	0	0	0	1	1
Frase 5	7	1	0	1	1	1
Frase 6	6	0	0	0	0	0
Frase 7	7	2	2	2	2	3
Frase 8	6	0	0	1	2	1
Frase 9	11	2	2	2	2	2
Frase 10	8	0	1	1	2	1
Frase 11	7	1	1	1	2	1
Frase 12	7	1	2	2	2	2
Frase 13	7	1	2	2	2	2
Frase 14	5	0	1	0	0	0
Frase 15	8	1	1	1	1	1
Frase 16	6	1	0	1	2	1
Frase 17	5	0	1	2	2	1
Frase 18	7	1	2	2	2	2
Frase 19	8	2	2	2	2	2
Frase 20	6	0	0	1	1	1
Total	146	20	23	27	33	29
Percentual de acerto		$13,7\%$	$15,8\%$	$18,5\%$	$22,6\%$	$19,9\%$

Tabela D.5 Palavras reconhecidas corretamente para entrada com SNR 5 dB.

				Sub-Ótimo		
Frase	Número de Palavras	Sem Filtragem	Wiener	$\alpha = 0.8$	$\alpha = 0.7$	$\alpha = 0.5$
Frase 1	6	0	1	2	2	2
Frase 2	9	0	1	1	1	0
Frase 3	10	1	2	2	2	1
Frase 4	10	0	0	0	1	0
Frase 5	7	1	0	1	1	2
Frase 6	6	0	0	0	0	0
Frase 7	7	1	1	2	2	1
Frase 8	6	0	0	1	1	2
Frase 9	11	2	2	2	2	2
Frase 10	8	0	1	1	2	1
Frase 11	7	0	1	2	2	1
Frase 12	7	1	2	2	2	2
Frase 13	7	1	2	2	2	2
Frase 14	5	0	0	0	0	0
Frase 15	8	1	1	1	1	1
Frase 16	6	0	0	0	2	1
Frase 17	5	0	1	2	2	0
Frase 18	7	2	2	2	2	2
Frase 19	8	1	1	1	1	1
Frase 20	6	0	0	1	1	1
Total	146	11	18	25	29	22
Percentual de acerto		$7,5\%$	$12,3\%$	$17,1\%$	$19,9\%$	$15,1\%$

Tabela D.6 Palavras reconhecidas corretamente para entrada com SNR 3 dB.

				Sub-Ótimo		
Frase	Número de Palavras	Sem Filtragem	Wiener	$\alpha = 0.8$	$\alpha = 0.7$	$\alpha = 0.5$
Frase 1	6	0	1	2	2	1
Frase 2	9	0	1	0	1	0
Frase 3	10	0	2	1	2	0
Frase 4	10	1	0	1	0	0
Frase 5	7	0	0	1	1	0
Frase 6	6	0	0	0	0	0
Frase 7	7	1	0	1	2	1
Frase 8	6	0	0	1	1	1
Frase 9	11	2	2	2	2	2
Frase 10	8	0	1	1	2	1
Frase 11	7	0	1	1	1	1
Frase 12	7	1	2	2	2	2
Frase 13	7	1	1	2	2	2
Frase 14	5	0	0	0	0	0
Frase 15	8	1	1	1	1	1
Frase 16	6	0	0	1	2	0
Frase 17	5	0	0	0	1	1
Frase 18	7	1	1	1	1	1
Frase 19	8	1	1	1	1	1
Frase 20	6	0	0	0	0	0
Total	146	9	14	19	24	15
Percentual de acerto		$6,2\%$	$9,6\%$	$13,0\%$	$16,4\%$	$10,3\%$

Tabela D.7 Palavras reconhecidas corretamente para entrada com SNR 0 dB.