Silvana Luciene do Nascimento Cunha

Sistema Multicanal Adaptativo para Supressão de Ruído Usando Arranjo de Microfones

Dissertação submetida ao corpo docente da Coordenação dos Cursos de Pós-Gradução em Engenharia Elétrica da Universidade Federal da Paraíba - Campus II como parte dos requisitos necessários para obtenção do grau de Mestre em Engenharia Elétrica.

Benedito G. Aguiar Neto - Dr.-Ing.
Orientador

Campina Grande, Paraíba, Brasil

C972s Cunha, Silvana Luciene do Nascimento.
Sistema multicanal adaptativo para supressão de ruído usando arranjo de microfones / Silvana Luciene do Nascimento Cunha. - Campina Grande, 1994.
91 f.

Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal da Paraíba, Centro de Ciências eTecnologia, 1994.
Referências.
"Orientação : Prof. Dr. Benedito Guimarães Aguiar Neto".

1. Processamento de Sinais. 2. Ruído. 3. Microfones - Arranjo. 4. Dissertação - Engenharia Elétrica. I. Aguiar Neto, Benedito Guimarães. II. Universidade Federal da Paraíba - Campina Grande (PB). III. Título

CDU 621.391(043)

SISTEMA MULTICANAL ADAPTATIVO PARA SUPRESSÃO DE RUÍDO USANDO ARRANJO DE MICROFONES

SILVANA LUCIENE DO NASCIMENTO CUNHA

Dissertação Aprovada em 26.09.1994

BENEDITO GUIMARÃES AGUIAR NETO, Dr.-Ing., UFPB
Orientador

MARCELO SAMPAIO DE ALENCAR, Ph.D., UFPB
Componente da Banca

JOÃO CESAR MOURA MOTA, Dr., UFC
Componente da Banca

CARLOS FELIPE DE BRITO JACCOUD, Mestre, EMBRATEL
Componente da Banca

CAMPINA GRANDE - PB
Setembro - 1994

Dedico este trabalho ao meu esposo Washington e

a minha filha Isabella.

Mensagem

"Direi do Senhor: Ele é o meu Deus., o meu refúgio, a minha fortaleza, e nele confiarei." (Salmos:91;2)

Agradecimentos

Gostaria de expressar meus sinceros votos de gratidão as seguintes pessoas e instituições:

Ao meu esposo Washington, por todo apoio e dedicação durante a realização deste trabalho, sem os quais este trabalho não seria possível.

Aos meus pais, Euclides (*in memorian*) e Olindina, bases do meu caráter e da minha determinação.

À todos os meus irmãos, em especial a minha irmã Nilza e ao meu irmão Sandro, que me deram apoio e me incentivaram em todos os momentos.

Ao professor Benedito G. Aguiar Neto pela orientação.

Aos professores Marcos Antônio G. Brasileiro, Marcelo Sampaio de Alencar, Rômulo R. Maranhão do Vale, Fátima Vieira Turnell pelos ensinamentos e pelo estímulo.

Ao professor Antônio Marcos N. de Lima, pela atenção dispensada em alguns momentos difíceis deste trabalho.

Ao amigo Carlos Allan, sem a sua presteza e amizade, este trabalho não teria sido concluído.

Ao amigo Paulo César Cortez pela amizade e apoio constantes.

Ao casal amigo Fátima e Medeiros e a minha amiga Isabel Lausanne, pelo apoio nos momentos mais difíceis.

A todos que contribuiram de alguma forma para a realização deste trabalho: Rosângela, Silvana Porto, Wallington, Glauco, Joseana, Rinaldo, Adriano Fábio e demais companheiros do LAPS.

A Ângela, por sua amizade, eficiência e dedicação junto a secretaria executiva da COPELE.

Ao CNPq, órgão financiador deste trabalho.

A Universidade Federal da Paraíba-Campus II, pela oportunidade oferecida.

Resumo

Esta dissertação consiste em um estudo, uma avaliação e uma implementação em *software* de um sistema adaptativo de supressão de ruídos aplicado a sinais de voz degradados. Na entrada do sistema é utlizado um arranjo de quatro microfones para a recepção dos sinais. Após os sinais serem captados pelo arranjo, são alinhados em fase e enviados para processamento. O alinhamento é feito baseado no coeficiente de correlação entre os canais, proporcionado um ganho de 2 a 4 dB na relação sinal-ruído. A supressão do ruído é realizada, ainda, através da filtragem dos sinais por um filtro Wiener-Kolmogoroff adaptativo. Um pós-processamento das medidas de correlação cruzada aumenta o desempenho do sistema e um ganho adicional de até 6 dB pode ser obtido na saída do filtro. O sinal melhorado obtido na saída do filtro é adequado para ser usado como entrada para sistemas de transmissão e sistemas de reconhecimento de voz. Avaliações subjetivas através de testes de escuta informais indicam um melhoramento substancial na qualidade e na inteligibilidade do sinal.

Abstract

This dissertation consists of a study, evaluation and software implementation of an adaptive noise suppression system, applied to degraded speech signals. An array of four microphones is used, at the system input, to receive the signals. After the signal reception, the signals are phase aligned and then processed. The alignment is made based on the cross-correlation coefficient between the channels, giving a gain of 2 to 4 dB in the signal-to-noise ratio. In addition, the noise suppression is achieved by filtering the signals using an adaptive Wiener-filter. The post-processing of cross-correlation measurements increases the system performance and an additional gain of about 6 dB at the filter output can be obtained. The enhanced signal, obtained at the filter output, can be used as an input to transmission systems and to speech recognition systems. Informal subjective listening tests indicates an improving in the signal quality and intelligibility.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Símbolos

$s(n)$ - Sinal de voz

$x(n)$ - Sinal de voz degradado

$r(n)$ - Ruído aditivo

$\hat{S}_{yy}(w,k)$ - Densidade Espectral de Potência a curto intervalo de tempo do sinal de voz estimado por subtração espectral

$\hat{S}_{xx}(w,k)$ - Densidade Espectral de Potência a curto intervalo de tempo do sinal de voz degradado

$\hat{S}_{rr}(w,k)$ - Densidade Espectral de Potência a curto intervalo de tempo do ruído

$N(w,k)$ - Espectro a curto intervalo de tempo do ruído

$Y(w,k)$ - Espectro a curto intervalo de tempo do sinal de voz estimado por subtração espectral

$\theta(w,k)$ - Fase do espectro de $\mathrm{Y}(w,k)$

$X(w,k)$ - Espectro a curto intervalo de tempo do sinal de voz degradado

$x(n,k)$ - k-ésimo segmento do sinal de voz degradado

$s(n,k)$ - k-ésimo segmento do sinal de voz original

$\hat{s}(n,k)$ - Estimação de $s(n,k)$

$R_{xx}(\tau)$ - Autocorrelação de $x(n)$

$R_{xs}(i)$ - Correlação cruzada entre $x(n)$ e $s(n)$

$R_{ss}(i)$ - Autocorrelação de $s(n)$

$H\text{opt}(\Omega)$ - Função de transferência do filtro Wiener-Kolmogoroff

$S_{rr}(\Omega)$ - Densidade espectral de potência de $r(n)$

$S_{xx}(\Omega)$ - Densidade espectral de potência de $x(n)$

$Q(\Omega, k)$ - Fator relativo de degradação

L - Número de amostras num dado segmento

$d(n)$ - Ruído de fundo

r(n) - Entrada de referência do cancelador adaptativo de ruído

$\hat{s}$(n) - Estimação do sinal $s(n)$

$y_i(t)$ - Sinal no i-ésimo microfone

R_{k0} - Valor máximo de correlação do sinal no k-ésimo microfone

$\tau_{\max}$ - Valor de τ para o máximo valor de correlação

$\tau'(m, n)$ - Atraso do sinal no microfone na coluna m e na linha n do arranjo bidimensional

$h(n)$ - Resposta ao impulso do filtro WK

σ_e^2 - Variância do erro de estimação $r(n)$

$h_{j,opt}$ - Coeficientes ótimos do filtro WK

$\mathbf{R}_{xs}$ - Matriz de correlação entre $x(n)$ e $s(n)$

$\mathbf{R}_{xx}$ - Matriz de autocorrelação de $x(n)$

$\mathbf{h}_{opt}$ - Matriz dos coeficientes ótimos do filtro WK

$F\{\cdot\}$ - Operador transformada de Fourier

d - Distância entre dois microfones do arranjo

$m_i(n)$ - Sinal captado pelo i-ésimo microfone

M_i - i-ésimo microfone

$x_i(n)$ - Sinal degradado captado pelo i-ésimo microfone após o alinhamento de fase

$x_s(n)$ - Sinal médio obtido entre os sinais alinhados $x_i(n)$

N - Número adicional de canais (microfones) do arranjo

τ_v - Atraso de tempo variável

τ_f - Atraso de tempo fixo

τ_0 - Atraso de tempo para o valor máximo de correlação entre dois canais do arranjo

$\tilde{\tau}_0$ - Estimação de τ_0

l_1 - Distância do microfone 1 à fonte de som

l_i - Distância do i-ésimo microfone à fonte de som

v - Velocidade do som = $3x10^8$ m/s

$\rho_{k1}(\tau)$ - Coeficiente de correlação cruzada entre o sinal captado pelo microfone 1 e o sinal captado pelo k-ésimo microfone

τ_k - Atraso de tempo no k-ésimo microfone

$\tilde{\tau}_k$ - Estimação para τ_k

$X_i(l)$ - Transformada Discreta de Fourier de $x_i(n)$

$A(l)$ - Auto-densidade espectral dos sinais $x_i(n)$

$C(l)$ - Densidade espectral cruzada dos sinais $x_i(n)$

$S(l)$ - Densidade espectral do sinal de voz original

N_{ij} - Erro de estimação em $C(l)$

$V(l)$ - Variância do ruído

$Cm(l)$ - Estimação modificada de $C(l)$

$P(l)$ - Estimação pós-processada de $S(l)$

α(l) - Fator de redução dependente da freqüência

$Vm(l)$ - Estimação modificada de $V(l)$

Lista de Abreviaturas

SNR - Relação sinal-ruído (*Signal-to-Noise Ratio*)

DFT - Transformada discreta de Fourier (*Discrete Fourier Transform*)

FFT - Transformada rápida de Fourier (*Fast Fourier Transform*)

DEP - Densidade Espectral de Potência

SP1 - Sinal de Prova 1

SP2 - Sinal de Prova 2

SegSNR-G - Ganho na relação sinal-ruído segmental

SNRseg - Relação sinal-ruído segmental (*Segmental Signal-to-Noise Ratio*)

LPC - Codificação por predição linear (*Linear Prediction Coding*)

A/D - Analógico/Digital

VLSI - Integração em altíssima densidade (*Very Large Scale Integration*)

Filtro WK - Filtro Wiener-Kolmogoroff

HMM's - Modelos de Markov escondidos (*Hidden Markov Models*)

Capítulo 1

Introdução

A degradação de voz por ruído acústico ambiental produz uma perda considerável na qualidade perceptiva e na inteligibilidade de uma conversação, em sistemas de comunicação, e causa uma redução significativa da taxa de reconhecimento em sistemas de reconhecimento de voz [1, 2, 3].

O interesse em processamento de sinais para melhoramento de voz vem desde a Segunda Guerra Mundial, em virtude da dificuldade de comunicação de pilotos de aeronaves com o controle da terra devido ao alto nível de ruído presente nas suas cabines (cabine de helicóptero, por exemplo). Nesses sistemas se faz necessário um processamento do sinal de voz, antes da transmissão, de forma a reduzir o efeito do ruído aditivo que seria transmitido juntamente com o sinal de voz.

Atualmente, com o surgimento da telefonia móvel celular, aumenta cada vez mais o número de usuários de aparelhos telefônicos em automóveis e o ruído do próprio automóvel bem como o ruído de trânsito podem tornar-se inconvenientes na conversação entre usuários. A qualidade e a inteligibilidade do sinal de voz podem ser seriamente prejudicadas pela presença do ruído. Assim, aumenta cada vez mais a necessidade de pesquisa em algoritmos para processamento de sinais a serem aplicados também nesta área.

Os rápidos avanços que têm ocorrido na tecnologia de *hardware* para processamento de sinais em tempo real vem contribuindo para aumentar o interesse de engenheiros

e pesquisadores em melhoramento de voz degradada. Entre estes, estão engenheiros trabalhando nos problemas de comunicação de voz, tais como no desenvolvimento de codificadores de voz, e audiologistas ajudando pessoas com problemas auditivos [4].

Os sistemas de redução de ruído para melhoramento de sinais de voz são divididos, em geral, em sistemas de supressão de ruído e sistemas de cancelamento de ruído [5]. Estes sistemas podem ainda ser classificados como monocanal e multicanal dependendo do número de entradas de sinal disponíveis [6]. Sistemas de cancelamento são casos típicos de sistemas multicanal, já que têm pelo menos uma entrada de referência para o ruído.

O sucesso do cancelamento adaptativo de ruído é dependente da obtenção de uma entrada de referência externa para o ruído que seja descorrelacionada com o sinal e altamente correlacionada com o ruído aditivo. Na maioria das aplicações, um sinal de ruído de referência externo pode não estar disponível em um segundo sensor. Conseqüentemente, para formar um sinal de ruído de referência, se poderia assumir que o ruído é estacionário e que o sinal médio determinado durante períodos classificados como "silêncio" seja representativo do ruído. No entanto, como o ruído é raramente estacionário o método não é bom, pois uma amostra finita pode ser insuficiente para estimar o ruído e, além disso, a decisão de silêncio não é livre de erros. Embora seja difícil formar uma entrada de referência para o ruído, é muito fácil formar uma entrada de referência para o sinal de voz original. Além disso, técnicas de canal único têm a desvantagem de serem limitadas a ruídos semi-estacionários além de introduzirem distorção no sinal [7].

Assim, surgem sistemas com mais de uma entrada de referência para o sinal de voz [8, 9, 10]. Já que voz é quase-periódica, uma seção de voz atrasada por uma pequena quantidade (um ou dois períodos de *pitch*) será altamente correlacionada com o sinal de voz e altamente descorrelacionada com o ruído aditivo [7].

Arranjos unidimensionais e bidimensionais de transdutores tem sido aplicados em radar e sonar [11, 12] por mais de quarenta anos. Recentemente, a necessidade de um microfone unidirecional, seguido de processamento do sinal, tem surgido nas áreas de teleconferência e entrada para reconhecimento de voz. Isto tem produzido cada vez

mais trabalhos de pesquisa no uso de técnicas de arranjo para estas aplicações [9, 10, 3].

Esta dissertação estuda e avalia um sistema de supressão de ruído que reduz o ruído recebido usando um arranjo de microfones bidimensional com quatro microfones. Com uma fonte única numa distância moderada dos microfones, os sinais gravados são réplicas atrasadas uma das outras e escalonadas por um fator de ganho pequeno [3]. O ganho de diretividade do arranjo é usado para redução do ruído. Além disso, os sinais são alinhados automaticamente e um sinal médio obtido dos sinais alinhados é filtrado por um filtro Wiener adaptativo. Um algoritmo de pós-processamento proporciona um melhoramento adicional pelo processamento das medidas de correlação cruzada, diminuindo o efeito do ruído. O ruído é estimado das características espectrais do sinal degradado e um processamento é feito no domínio da freqüência usando a transformada rápida de Fourier (FFT- *Fast Fourier Transform*). Um sistema similar foi proposto em [13], sendo que o mesmo não realiza alinhamento automático dos sinais como realizado pelo sistema aqui proposto e implementado. Em [13] os atrasos foram simplesmente colocados como valores fixos correspondendo à direção do locutor desejado. O sinal melhorado obtido na saída do sistema implementado pode ser usado como entrada para sistemas de transmissão e sistemas de reconhecimento de voz.

No capítulo 2 desta dissertação é feita uma revisão bibliográfica de algumas das várias técnicas utilizadas no processamento digital para melhoramento de sinais de voz degradados. São vistas técnicas monocanal e multicanal, além do conceito de cancelamento adaptativo e algumas aplicações dos sistemas mostrados. Estes sistemas, entre outros, serviram de base para o trabalho em questão e poderá servir de apoio bibliográfico para trabalhos futuros.

O Capítulo 3 apresenta uma revisão teórica sobre o filtro Wiener-Kolmogoroff, por ser o filtro utilizado pelo sistema apresentado para redução do ruído, sendo pois de grande importância no contexto deste trabalho.

O Capítulo 4 descreve o sistema multicanal adaptativo para supressão de ruído implementado. Cada unidade do sistema é descrita e analisada e sua importância destacada dentro do sistema.

No Capítulo 5 estão contidos, além das condições experimentais realizadas no sistema proposto, os resultados obtidos. É feita, então, uma avaliação do desempenho do sistema através da relação sinal-ruído e avaliações subjetivas por testes de escuta informais.

O Sexto e último Capítulo apresenta as conclusões a partir dos resultados obtidos, além de sugestões para futuros trabalhos a serem realizados neste campo de pesquisa.

Capítulo 2

Técnicas utilizadas para melhoramento de sinais de voz degradados

2.1 Introdução

Considere a situação onde uma pessoa está operando um sistema de comunicação que transmite voz contaminada por qualquer ruído acústico captados por um microfone (Fig. 2.1). Assume-se que a contribuição de ruído do canal de transmissão à mensagem

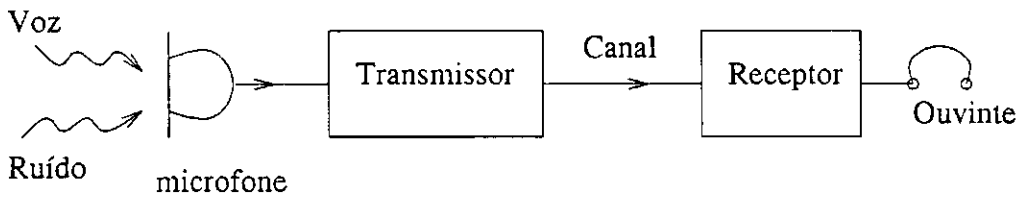


Figura 2.1: Sistema de comunicação sem processamento de voz.

é desprezível e o ouvinte, que está localizado num ambiente calmo (sem a presença

do ruído interferente), tem a responsabilidade de interpretar a mensagem [14]. Esta situação é típica em sistemas de comunicação de aeronaves à terra, onde o ruído é adicionado no terminal de transmissão mais do que no canal de transmissão.

O problema do processamento de voz antes de sua degradação pelo ruído aditivo cresce nesse tipo de situação. Este problema recebeu considerável atenção durante e após a Segunda Guerra Mundial. Uma força de motivação maior por trás deste interesse foi a dificuldade que pilotos experimentaram na recepção de mensagens verbais do controle da terra devido ao alto nível de ruído das cabines de helicópteros. O interesse foi renovado após os anos setenta devido parcialmente aos avanços em processamento de sinais que proporcionaram avanços nas pesquisas em algoritmos sofisticados [4].

A questão a ser resolvida é o projeto de algum sistema de processamento localizado no receptor ou transmissor tal que a mensagem transmitida seja mais inteligível ao ouvinte do que se nenhum processador estivesse presente (Fig. 2.2). Está claro que o processador deve processar voz mais ruído já que os dois não podem ser separados. O caso do ruído acústico ambiental está em contraste à situação clássica na qual o ruído é elétrico e adicionado no canal. No último caso, a voz pode ser processada antes que

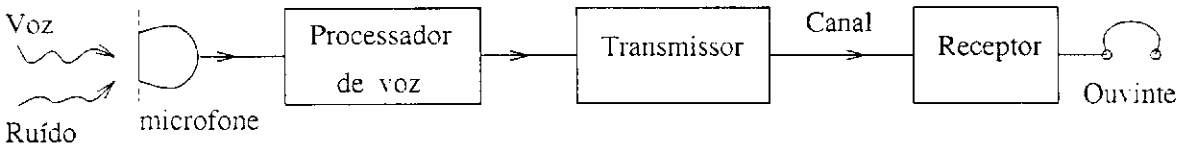


Figura 2.2: Sistema de comunicação com processamento de voz.

o ruído seja adicionado e assim o processador pode estar localizado no transmissor; além disso, outro processador pode estar localizado no receptor se quaisquer transformações inversas forem necessárias (Ex: pré–ênfase e de–ênfase) [15].

Investigações para melhorar transmissão e recepção em um ambiente de alto ruído ambiental tem sido geralmente limitadas ao estudo de dispositivos protetores de ouvidos [16] ou ao uso de microfones especiais [14].

Os sistemas de redução de ruído em sinais de voz dividem-se, de uma forma geral, em sistemas de supressão de ruído e sistemas de cancelamento de ruído [5]. Esses sistemas

podem, ainda, ser classificados como sistemas monocanais ou sistemas multicanais. Esta classificação é feita em função do número de entradas disponíveis de sinal [6]. Os sistemas de cancelamento de ruído são casos típicos de sistemas multicanal, pois dispõem de pelo menos uma entrada de referência para o ruído.

Uma característica dos ruídos acústicos ambientais, presentes no sinal a ser transmitido, diz respeito a degradação do sinal de forma contínua, ou seja, todas as amostras do sinal são degradadas. Conseqüentemente, técnicas não-seletivas devem ser usadas por permitirem uma filtragem contínua do sinal degradado a fim de se conseguir a remoção do ruído. Estas técnicas incluem os chamados *Métodos de Supressão de Ruído* [1, 2], os quais são baseados em Filtros Ótimos e/ou na *Teoria de Estimação Espectral* [17] a curtos intervalos de tempo.

Neste capítulo são apresentadas algumas técnicas de supressão de ruído, dentre as quais, técnicas de canal único bem como técnicas de supressão de ruído em sistemas multicanais. É visto, ainda, o conceito de *cancelamento* adaptativo de ruído e é feita uma breve comparação com os sistemas de *supressão* de ruído. Algumas características dos sistemas vistos a seguir são utilizadas como base para a implementação do método proposto neste trabalho.

2.2 Supressão de ruído acústico em voz usando subtração espectral

Ruídos de fundo acusticamente adicionados à voz podem degradar o desempenho de processadores digitais de voz usados para aplicações tais como compressão, reconhecimento e verificação [18]. Sistemas digitais de voz deverão ser usados numa variedade de ambientes e seu desempenho deve ser mantido em um nível próximo daquele medido para um sistema onde o sinal de voz de entrada é livre de ruído. Para garantir uma credibilidade contínua, os efeitos do ruído de fundo podem ser reduzidos pelo uso de técnicas de redução de ruído [17].

Microfones de cancelamento de ruído, embora essenciais para ambientes de ruído

extremamente elevado, tais como cabine de helicóptero, oferecem pouca ou nenhuma redução acima de 1 kHz [17]. Assim, cresce o esforço na obtenção de técnicas adequadas para eliminar ou reduzir os efeitos da contaminação de voz pelo ruído.

A tecnica de supressão de ruído por subtração espectral pertence a uma classe de sistemas de melhoramento de voz que explora a noção de que estaá na magnitude espectral, ao invés da fase, a principal característica de informação para inteligibilidade e qualidade da voz. Nesta classe de sistemas, a voz degradada é inicialmente segmentada e a cada segmento são determinadas estimativas espectrais a curto intervalo de tempo da voz degradada e do ruído que, subtraídas entre si, representam a estimação espectral do sinal de voz.

A subtração espectral pode ser realizada a partir da subtração de estimações da densidade espectral de potência do sinal degradado e do ruído, tomadas ao longo de k-ésimos intervalos curtos de tempo

$$\hat{S}_{yy}(w,k) = \hat{S}_{xx}(w,k) - \hat{S}_{rr}(w,k) \tag{2.1}$$

As estimações espectrais na Eq. (2.1) para o sinal degradado e o ruído, são obtidas através do quadrado das respectivas amplitudes espectrais $\mid X(w,k) \mid$ e $\mid N(w,k) \mid$ determinadas para o k-ésimo segmento considerado:

$$\hat{S}_{xx}(w,k) \simeq \mid X(w,k) \mid^2 \tag{2.2}$$

e

$$\hat{S}_{rr}(w,k) = \hat{E}[\mid N(w,k) \mid^2] \tag{2.3}$$

$\hat{E}[\mid N(w,k) \mid^2]$ representa a estimação do ruído através da média das estimações obtidas ao longo dos k-ésimos segmentos considerados.

De forma semelhante, a estimação da densidade espectral de potência a curto intervalo de tempo $\hat{S}_{yy}(w,k)$ do sinal de voz é dada por:

$$\hat{S}_{yy}(w,k) \cong \mid Y(w,k) \mid^2 \tag{2.4}$$

O algoritmo de subtração espectral será dado por:

$$\begin{cases} \hat{S}_{yy}(w,k) = \mid X(w,k) \mid^2 - E[\mid N(w,k) \mid^2] & \text{para } \mid X(w,k) \mid^2 > E[\mid N(w,k) \mid^2] \\ \hat{S}_{yy}(w,k) = 0 & \text{caso contrário} \end{cases} \tag{2.5}$$

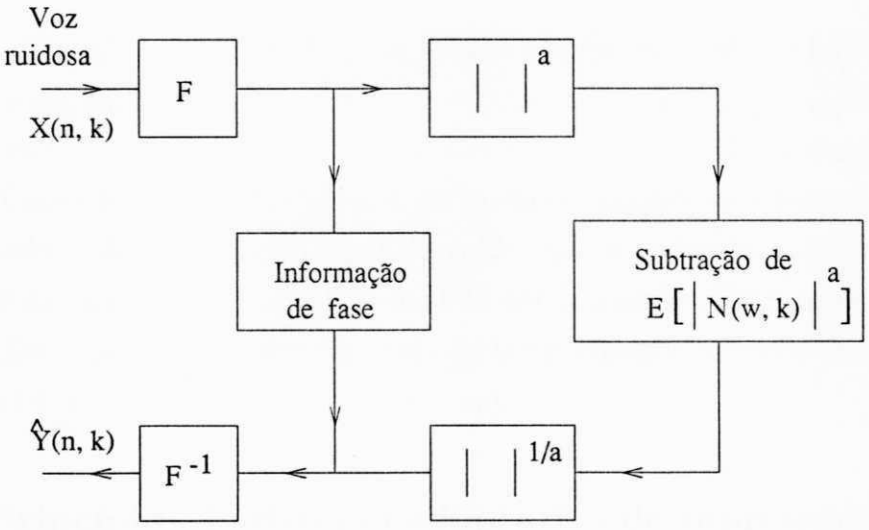
Teremos, portanto, que a amplitude espectral de curtos intervalos de tempo do sinal de voz será:

$$\mid Y(w,k) \mid = \{\mid X(w,k) \mid^2 - E[\mid N(w,k) \mid^2]\}^{1/2} \tag{2.6}$$

e o seu espectro de curto intervalo de tempo dado por:

$$Y(w,k) = \mid Y(w,k) \mid \exp(j\theta(w,k)) \tag{2.7}$$

onde $\theta(w,k)$ é a fase do espectro do próprio sinal degradado utilizada para reconstruir o sinal melhorado.



FONTE: Lim (1986) [4], p. 3136

Figura 2.3: Generalização do método de subtração espectral para melhoramento de voz.

Um sistema de melhoramento de voz baseado numa generalização da Equação (2.5) é mostrado na Figura 2.3. Se o resultado após a subtração de $E[\mid N(w,k)\mid^2]$ é menor do que zero, o mesmo é igualado a zero. Quando a constante "a" na Figura 2.3 é igual a 2, o sistema corresponde ao método de subtração do espectro de potência ou, simplesmente, subtração espectral.

O sistema na Figura 2.3, com a = 1, foi avaliado em [17] quando a degradação é devido ao ruído de helicóptero e em [5] para ruído de automóveis. Os resultados baseados no *Diagnostic Rhyme Test* indicam que na relação SNR na qual o escore de inteligibilidade do material de voz não-processada é cerca de 84%, o sistema não melhora a inteligibilidade, mas melhora a qualidade.

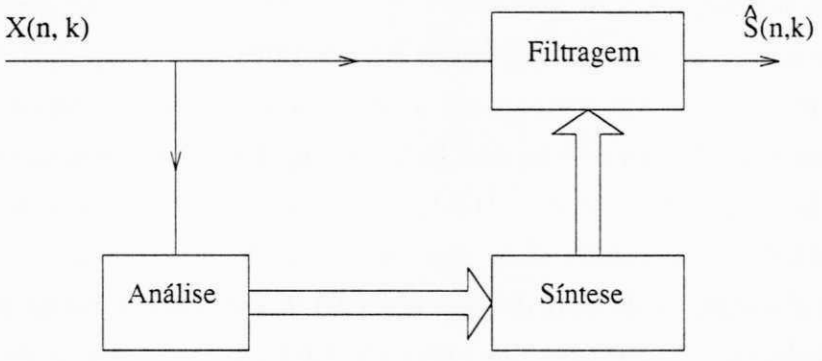
2.3 Método de supressão de ruído por balanceamento espectral adaptativo

Em [1], é estudado um método de supressão de ruído no qual a redução do ruído é efetuada por um balanceamento adaptativo da amplitude espectral do sinal degradado. O balanceamento espectral consiste em se determinar uma função de transferência de um filtro Wiener-Kolmogoroff, obtida a partir de estimações do espectro do sinal de voz degradado e do espectro do sinal de ruído, que é utilizado para obter-se uma modificação da amplitude espectral do sinal de voz degradado. Este método é baseado em estimações espectrais a curto intervalo de tempo usando a transformada discreta de Fourier (DFT - *Discrete Fourier Transform*).

2.3.1 Princípios do sistema adaptativo de supressão de ruído

A Figura 2.4 mostra a estrutura geral de um sistema adaptativo de supressão de ruído. Este tipo de sistema utiliza, em geral, algum tipo de filtro adaptativo cuja resposta ao impulso é determinada em função das propriedades estatísticas do sinal a ser melhorado e do ruído. Em geral, estas propriedades não são disponíveis nestes sistemas e o seu conhecimento prévio não é possível se os sinais são não-estacionários.

Assim, as informações estatísticas necessárias dos sinais são obtidas a partir do sinal degradado e a curtos intervalos de tempo, nos quais considera-se que os sinais sejam quase estacionários. Para sinais de voz, considera-se estacionariedade para intervalos de 16 ms a 32 ms. Os coeficientes do filtro são calculados e atualizados a cada novo segmento e considera-se que o sinal e o ruído sejam descorrelacionados.



FONTE: Aguiar Neto (1989) [2]

Figura 2.4: Estrutura geral de um sistema adaptativo de supressão de ruídos.

Para a supressão do ruído, parte-se da Equação de Wiener-Hopf [19, 20], a ser apresentada no capítulo 3, com a qual são calculados os coeficientes ótimos do filtro Wiener-Kolmogoroff. Para o cálculo dos coeficientes, são necessárias apenas a auto-correlação do sinal de voz degradado $x(n)$, $R_{xx}(i)$, e a correlação cruzada entre o sinal degradado $x(n)$ e o sinal original $s(n)$, $R_{xs}(i)$.

O algoritmo de supressão de ruído é dado por [1]:

$$Hopt(\Omega) = \begin{cases} 1 - \frac{S_{rr}(\Omega)}{S_{xx}(\Omega)} & \text{para } S_{rr}(\Omega) < S_{xx}(\Omega) \\ 0 & \text{para } S_{rr}(\Omega) \geq S_{xx}(\Omega) \end{cases} \tag{2.8}$$

onde $S_{rr}(\Omega)$ representa a densidade espectral do ruído, $S_{xx}(\Omega)$ a densidade espectral do sinal degradado e $Hopt(\Omega)$ a função de transferência do filtro de supressão de ruído.

Assim, da Equação (2.8), pode-se observar que a função de transferência $H\text{opt}(\Omega)$ permite a supressão do ruído sem a necessidade do conhecimento ou da estimação da estatística do sinal de voz.

2.3.2 Balanceamento espectral adaptativo

No sistema de supressão de ruído por balanceamento espectral adaptativo, de acordo com a Equação (2.8), a filtragem do sinal degradado é realizada através de modificações da amplitude espectral deste sinal. No entanto, devido à não estacionariedade dos sinais de voz e do ruído, a Equação (2.8) não pode ser usada diretamente. Assim, faz-se necessária a estimação das DEP's (DEP - Densidade Espectral de Potência) $S_{xx}(\Omega)$ e $S_{rr}(\Omega)$, a curtos intervalos de tempo, onde há maior possibildade dos sinais apresentarem estacionariedade. A DEP do sinal degradado é estimada nos intervalos de atividade de voz, enquanto a DEP do ruído é estimada nos intervalos de pausa [1].

Assim, substituindo-se as DEP's por suas estimações, obtém-se:

$$\text{Hopt}(\Omega) = \begin{cases} 1 - Q(\Omega, k) & \text{para } 0 < Q(\Omega, k) < 1 \\ 0 & \text{para } Q(\Omega, k) \geq 1 \end{cases} \tag{2.9}$$

com

$$Q(\Omega, k) = \frac{\hat{E}[\mid N(\Omega, k) \mid^2]}{\mid X(\Omega, k) \mid^2}$$

onde o numerador representa a estimação no k-ésimo intervalo de tempo da DEP do ruído e o denominador é a estimação da DEP do sinal de voz degradado no k-ésimo intervalo. Q(Ω,k) é independente do nível absoluto do sinal degradante e é denominado em [21] como "fator relativo de degradação".

A estimação do espectro de um segmento do sinal de voz original é obtida através de um balanceamento espectral do k-ésimo segmento do sinal degradado:

$$\hat{\text{S}}(\Omega, k) = \text{Hopt}(\Omega, k) \cdot \mid X(\Omega, k) \mid \exp(j\theta(\Omega, k)) \tag{2.10}$$

Considera-se que a fase do sinal de voz degradado representa uma aproximação razoável da fase do sinal de voz original, já que o ouvido humano é insensível a degradações de fase [1].

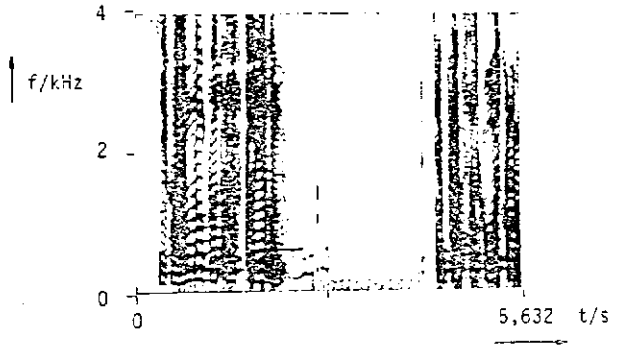
Em [2], o sinal de voz degradado é dividido em segmentos de L amostras e multiplicados por uma janela de Hamming com 50% de superposição.

A atualização da DEP do sinal degradante estimada é ativada durante os intervalos de pausas. Nos intervalos de atividade de voz são utilizados os valores da DEP do último intervalo de pausa. Para uma melhor adaptação às propriedades estatísticas variantes no tempo do sinal de voz, bem como para obter uma melhor resolução no espectro, o segmento de análise escolhido é de 16 a 32 ms, pois nesse intervalo o sinal de voz pode ser considerado como quase estacionário [1, 22].

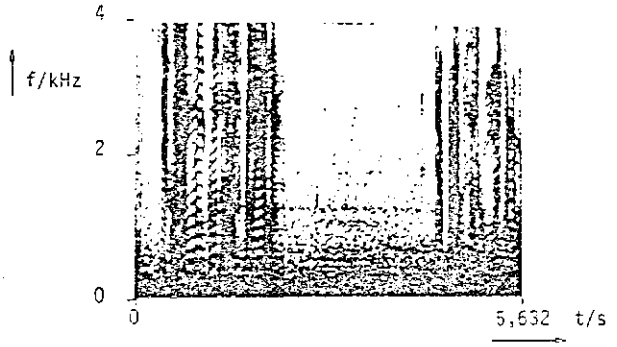
Estimações espectrais errôneas podem atenuar ou suprimir algumas componentes espectrais do sinal de voz, produzindo um ruído residual. Assim, ao invés da utilização da estimação da DEP do último segmento do intervalo de pausa considerado, pode-se usar a estimação até o penúltimo segmento. Para evitar isso, pode-se considerar como se os intervalos de atividade de voz tivessem uma maior duração. Além disso, para que a impressão subjetiva da perturbação nos intervalos de pausas e nos intervalos de voz corresponda à mesma, podem ser fixados valores mínimos a serem atingidos pelos valores espectrais nos intervalos de pausas iguais ao valores fixados nos intervalos de atividade de voz [2].

Os resultados obtidos para este método apresentam um ganho de até 5,18 dB para um sinal de voz falado por um interlocutor masculino e degradação por ruído de automóvel (SP1 - Sinal de Prova 1), e de 4, 93 dB para voz falada por dois interlocutores masculinos (SP2 - Sinal de Prova 2). Para a relação sinal ruído segmental foram obtidos ganhos de 8,54 e 15,99 dB, respectivamente [1].

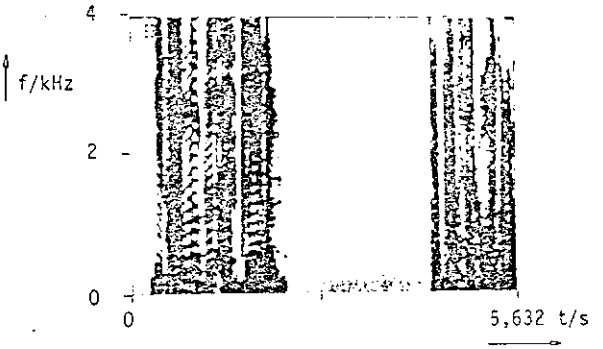
Na Figura 2.5, mostrada a seguir, pode-se observar a eficácia do método de supressão de ruídos, através do espectograma do sinal de voz após o tratamento. Pode-se perceber que o método é eficaz principalmente nos intervalos de pausas. Como o sinal SP2 possui um intervalo de pausa maior, o ganho obtido na SegSNR (Relação sinal-ruído segmental - *Segmental Signal-to-Noise Ratio*) foi maior do que para o sinal SP1. No entanto, percebe-se também, comparando-se a Figura 2.5(a) e a Figura (2.5)(b) que segmentos de baixa energia do sinal original são parcialmente suprimidos se estes



(a)



(b)



(c)

FONTE: Aguiar Neto (1987) [5], p. 56

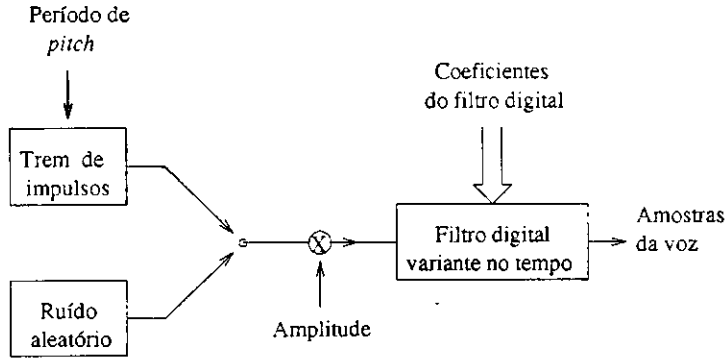
Figura 2.5: Espectograma para o sinal de prova SP2: a) Sinal de voz original, b) Sinal de voz degradado por ruído de automóvel e c) Sinal de voz tratado.

segmentos são interpretados como pausas.

Pelos resultados obtidos, pode-se concluir que o êxito de um método de supressão de ruído é bastante dependente de uma detecção de pausas livre de erros, já que a adaptação do algoritmo é realizada apenas nos intervalos de pausas, pela atualização dos valores espectrais estimados para o sinal degradante. O melhoramento da voz degradada depende ainda da estacionariedade do sinal. Se o sinal for fortemente não-estacionário, os parâmetros do filtro atualizados durante os intervalos de pausas, dificilmente terão o mesmo efeito nos intervalos de atividade de voz. Isto ocorre porque o espectro não pode ser atualizado nestes intervalos.

2.4 Sistema para melhoramento de voz baseado no modelo de produção da fala

Um outro método para melhoramento de voz tenta explorar o modelo para a produção de fala. Neste método, a voz é tipicamente modelada pela resposta de um sistema linear, representando o trato vocal, excitado por um trem de pulsos periódico para sons sonoros e por uma fonte de ruído aleatório faixa-larga para sons surdos, como mostra a Figura 2.6.



FONTE: Lim (1986) [4], pag. 3137

Figura 2.6: Um modelo de produção de voz

Como o trato vocal muda sua forma em função do tempo, o filtro digital na Figura 2.6 que representa o trato vocal é, em geral, variante no tempo. Entretanto, sobre um curto intervalo de tempo, o filtro digital pode ser aproximado como um sistema linear invariante no tempo [23].

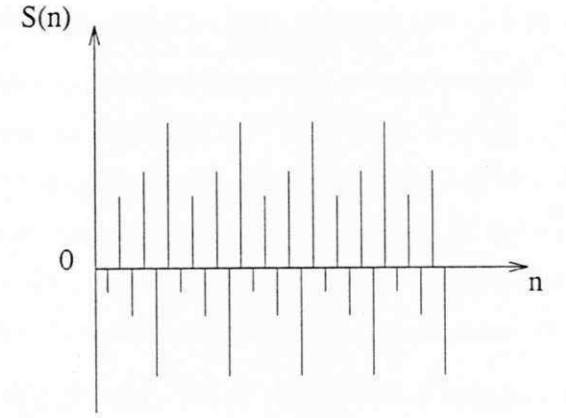
Numa técnica de melhoramento de voz que explora o modelo de produção de fala em questão, os parâmetros do modelo de voz são primeiro estimados e depois a voz é gerada por um sistema de síntese baseado no mesmo modelo de voz ou através do projeto de um filtro com os parâmetros do modelo estimados. É feita, então, a filtragem da voz ruidosa.

Vários sistemas de melhoramento de voz diferentes foram desenvolvidos usando este método, com o trato vocal modelado por um sistema de apenas pólos e os parâmetros do modelo de voz estimados pelo método da máxima verossimilhança para a detecção da presença do ruído. O desempenho destes sistemas não foi avaliado por um teste subjetivo. Audição informal, no entanto, indica que a qualidade da voz é melhorada enquanto o melhoramento na inteligibilidade não é muito claro [4].

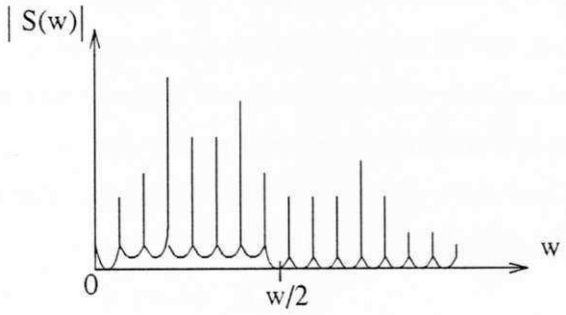
2.5 Supressão de ruído baseado na filtragem pente

Outro método para melhoramento de voz está baseado na exploração do fato de que os sons sonoros possuem formas de onda pseudo-periódicas. Especificamente, a periodicidade de uma forma de onda no tempo manifesta-se no domínio da freqüência como harmônicas múltiplas da freqüência fundamental.

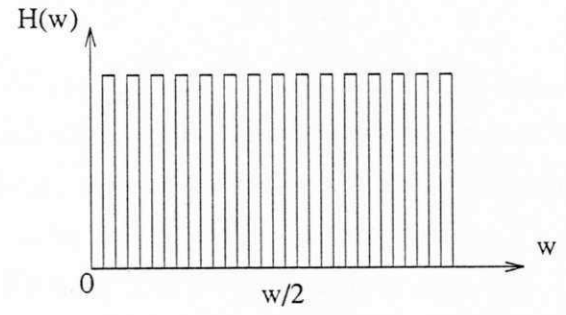
Na Figura 2.7(a) é mostrado a descrição temporal de um segmento de uma forma de onda periódica, e na Figura 2.7(b) é mostrado o espectro de magnitude associado. Como pode ser visto na Figura 2.7(b), a energia de um sinal periódico está concentrada em faixas de freqüências. Já que os sinais interferentes, em geral, têm energia sobre todas as faixas de freqüências, de forma que a informação exata da freqüência fundamental está disponível, um filtro pente como mostrado na Figura 2.7(c) pode reduzir o ruído enquanto preserva o sinal. Um filtro adaptativo que é baseado no conceito da filtragem pente e que parcialmente justifica o fato de que voz sonora é apenas aproxi-



(a)



(b)



(c)

FONTE: Lim (1986) [4], pag. 3137

Figura 2.7: Conceito da filtragem pente (a) Uma forma de onda periódica; (b) Magnitude espectral da forma de onda em (a); (c) Função de transferência de um filtro pente ideal.

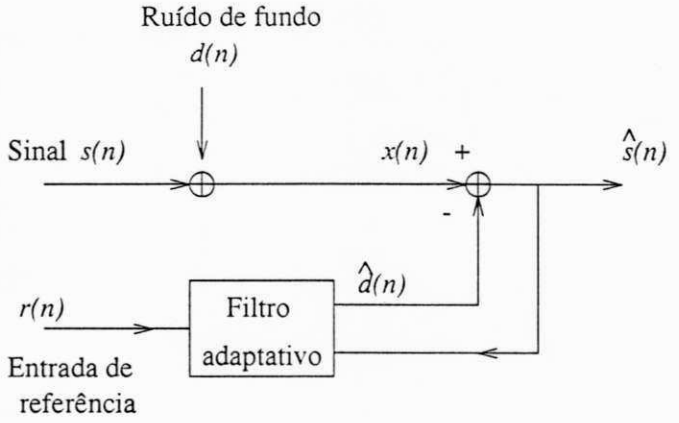
madamente periódica foi avaliado por Lim, Oppenheim e Braida [24] quando a degradação é devido ao ruído aleatório faixa-larga. A informação do período fundamental usada no processamento foi obtida da voz livre de ruído. Os resultados do teste mostram que mesmo com informação do período fundamental exata, as técnicas de filtragem adaptativa tendem a diminuir a inteligibilidade em várias valores de SNR. A despeito da diminuição da inteligibilidade, a voz processada por um filtro adaptativo soa "menos ruidosa" devido à capacidade do sistema de aumentar a relação sinal-ruído.

Os sistemas de melhoramento de voz discutidos acima são aplicáveis ao caso onde existe apenas uma entrada degradada. Quando mais de uma entrada está disponível para processamento, pode-se obter um melhor desempenho, aumentando-se assim a qualidade do sinal de voz. Cada uma das entradas individuais pode ser processada separadamente usando os sistemas de melhoramento de voz discutidos acima e depois combinadas adequadamente com o restante. Para o processamento de diferentes entradas, vários algoritmos de processamento de sinais tem sido desenvolvidos, nos quais a correlação de ruído nas várias entradas é explorada e um melhoramento significativo é possível em determinadas aplicações. Um exemplo de tais algoritmos é o algoritmo de cancelamento adaptativo de ruído [11].

2.6 Cancelamento adaptativo de ruído

Considere um ambiente no qual a entrada principal tem o sinal de voz $s(n)$ e o ruído $d(n)$ descorrelacionados; e a entrada de referência tem o ruído $r(n)$ descorrelacionado com $s(n)$ mas correlacionado de alguma forma desconhecida com o ruído $d(n)$. O cancelador de ruído adaptativo, conforme mostrado na Figura 2.8, filtra adaptativamente o ruído de referência $r(n)$ para estimar o ruído $d(n)$ o qual é então subtraído da entrada principal $x(n)$, representando a voz degradada, para obter uma estimação do sinal de voz.

O filtro de cancelamento adaptativo é tipicamente um filtro de resposta ao impulso finita cujos coeficientes são adaptados pela minimização da potência em ŝ(n). Pode ser mostrado que a minimização da potência de ŝ(n) de fato minimiza o erro



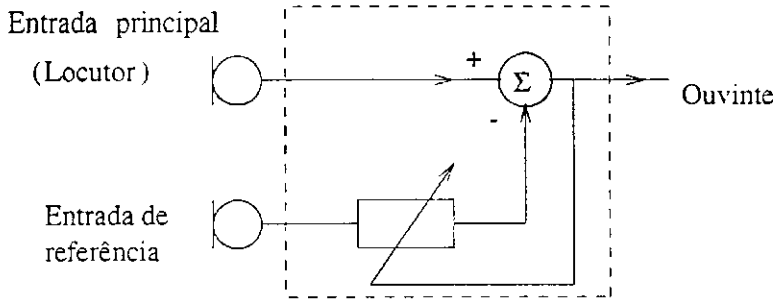
FONTE: Lim (1986) [4], p. 3138

Figura 2.8: Um algoritmo para cancelamento adaptativo de ruído.

médio quadrático entre s(n) e ŝ(n) e algoritmos tem sido desenvolvidos para estimar os coeficientes do filtro [11].

O algoritmo de cancelamento adaptativo de ruído foi aplicado num ambiente simulado no qual uma pessoa falou num microfone em uma sala onde forte interferência acústica estava presente [7]. O sinal neste microfone formou a entrada principal. Um segundo microfone, funcionando como entrada de referência, foi colocado na sala longe do locutor e perto da fonte de interferência acústica (Figura 2.9). O melhoramento na relação sinal–ruído obtida neste experimento usando a técnica de cancelamento adaptativo de ruído está em torno de 20 dB [11]. A despeito deste melhoramento significativo no desempenho e à capacidade do sistema de adaptar-se às mudanças estatísticas e ao movimento dos microfones, a técnica de cancelamento adaptativo de ruído é limitada na prática, já que a entrada de referência contém o sinal s(n) bem como o ruído. Neste caso, o cancelador de ruído tentará cancelar o sinal tanto quanto o ruído degradante.

Assim, o sucesso de sistemas adaptativos de cancelamento de ruído depende da obtenção de uma entrada de ruído de referência externa, que seja descorrelacionada com o sinal e altamente correlacionada com o ruído aditivo. Na maioria das aplicações,



Cancelador adaptativo de ruído

FONTE: Widrow (1975) [11], pag. 1704

Figura 2.9: Cancelamento de ruído em sinais de voz.

entretanto, um sinal de ruído de referência não pode ser captado diretamente por um segundo microfone. Para evitar este problema, pode-se obter um sinal de referência de ruído assumindo-se que o ruído é estacionário e que o sinal médio determinado durante períodos classificados como "silêncio" sejam representativos do ruído. No entanto, este método pode não apresentar bons resultados, pois o ruído é raramente estacionário, uma vez que uma dada sequência finita do ruído pode ser insuficiente para estimá-lo precisamente. Além disso, a decisão de silêncio não é livre de erros [7].

A dificuldade de se obter uma entrada de ruído de referência é evidente pelas razões citadas acima. No entanto, há uma maior facilidade na obtenção de uma entrada de referência para o sinal de voz [7]. Dessa forma, cresce o interesse em um sistema de supressão de ruído que, utilizando um maior número de entradas de referência para o sinal, apresente um desempenho melhor do que os sistemas existentes.

2.7 Sistema multicanal de supressão de ruído para uso em sistemas de reconhecimento de voz

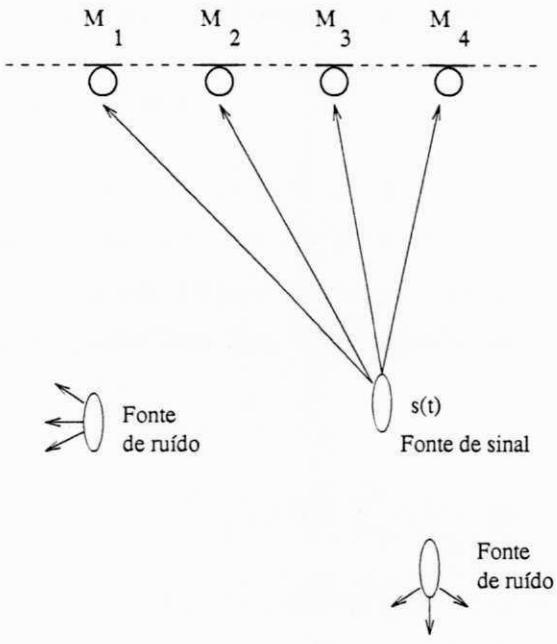
Sistemas de melhoramento de voz podem ser usados como um pré-processador de sinais em sistemas de comunicação ou sistemas de reconhecimento de voz com um

ou vários locutores. Através da correlação cruzada entre os sinais de entrada, tenta-se localizar e focalizar numa única fonte de interesse, mesmo quando muitas fontes competindo estão presentes.

Tipicamente, sistemas complexos de reconhecimento de voz tem um desempenho mais uniforme para relações sinal-ruído (SNR) entre +50 dB e +25 dB. Para situações ruidosas, seu desempenho não piora gradualmente, mas cai abruptamente. Em ambientes como escritórios, carros e fábricas, estas condições de SNR favoráveis podem somente ser garantidas com o uso de microfones labiais, que podem tornar-se muito inconvenientes. Em muitas aplicações práticas, uma SNR na faixa de 0 a +20 dB é desejável. Boa robustez de ruído nos sistemas existentes requerem o uso de supressão de ruído de canal único ou técnicas de cancelamento [25, 26], a maioria delas obtidas da subtração espectral [17]. Entretanto, estas técnicas estão limitadas ao ruído semi-estacionário. Outra desvantagem das técnicas de canal único é que a maioria introduz distorção no sinal. O uso de canais de gravação múltiplos e técnicas de alinhamento de sinais é o próximo passo lógico para uma maior extensão da faixa de operação de sistemas de reconhecimento de voz, com uma possibilidade adicional de usar o sistema simultaneamente como um sistema de melhoramento de voz para propósitos de comunicação. Um sistema para melhoramento de voz para ser usado como entrada de um sistema de reconhecimento de voz apresenta os seguintes critérios de projeto [3]:

- Apresentar distorção não-detectável no processo de reconhecimento ;
- Usar uma técnica eficiente de redução de ruídos e
- Garantir os objetivos acima, com um mínimo de informação sobre o locutor e sua localização.

O último critério de projeto do sistema requer alguma explicação adicional. Como são permitidas fontes múltiplas de mesma natureza presentes simultaneamente, deve ser fornecido também um mecanismo de inteligência que selecione uma delas. O critério usado é o das médias de energia a longo intervalo de tempo. A fonte que tiver o sinal de nível mais elevado com janelas de tempo de 5 segundos será o alvo, ou seja, o sinal a ser selecionado.



FONTE: Compernolle (1990) [3], p. 435

Figura 2.10: Recepção em 4 microfones de um sinal direto e interferência reverberante

A configuração exata e/ou o espaçamento dos microfones do arranjo não exercem influência nos algoritmos implementados para processamento do sinal de voz. O sinal é significativamente mais forte no caminho direto do que nos caminhos secundários da fonte de som no arranjo. As fontes de ruído interferentes contribuem de forma reverberante por supostamente estarem longe do arranjo (Figura 2.10).

Os sinais gravados são réplicas um do outro, com amplitudes diferentes. Diferenças existentes entre os microfones, podem aumentar o efeito do ângulo de chegada dos sinais. Este efeito pode ser minimizado pelo uso de bons microfones unidirecionais [3].

2.7.1 Métodos para alinhamento de sinais

2.7.1.1 Método Atraso e Soma

Flanagan [9] propôs uma unidade de alinhamento de sinais pelo método de atraso e soma para uso em grandes auditórios. Depois que a direção da fonte predominante e os atrasos, $\tilde{\tau}_k$, associados a cada um dos microfones existentes no arranjo são detectados, os sinais são alinhados em fase e somados. A solução de atraso e soma, $\hat{s}(t)$, é obtida como:

$$\hat{s_1}(t) = \frac{1}{N+1}[y_0(t) + \sum_{k=1}^{N} y_k(t - \tilde{\tau}_k)] \tag{2.11}$$

onde N representa o número adicional de canais (microfones) e $y_0(t)$ o sinal captado pelo canal predominante e $y_k(t)$; k=1, 2, ..., N é o sinal captado pelo k-ésimo canal[1].

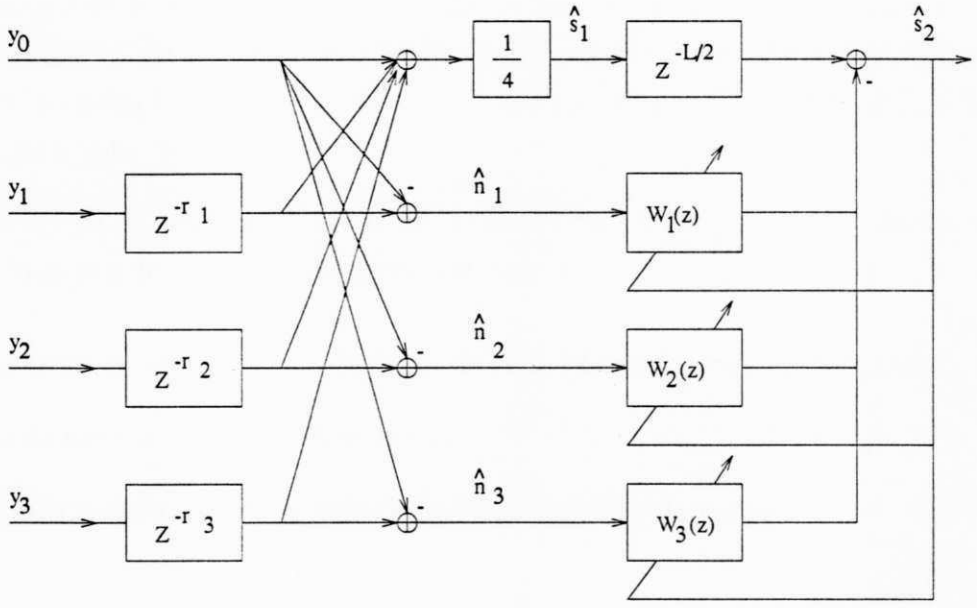
Esta é uma solução simples, de baixo custo em *hardware* para processamento digital de sinais. O método, entretanto, produz uma supressão de ruído quase uniforme em todas as direções e sem adaptação à situação de ruído específica. A atenuação máxima do ruído é de 10log (N+1) dB, que é obtida no caso de fontes de ruído descorrelacionadas. Conseqüentemente, Para obter-se um alto ganho na relação sinal-ruído, é necessário um número elevado de microfones. No entanto, grandes arranjos são impraticáveis em várias situações, sendo possível apenas para grandes auditórios [9].

2.7.1.2 Método de Griffiths-Jim

O alinhamento por atraso e soma é usado como o primeiro estágio de um sistema adaptativo. A soma e as diferenças são calculadas de todos os canais alinhados em fase como mostrado na Figura 2.11. O sinal $\hat{s}_1(t)$, resultante da soma entre os sinais recebidos dos canais y_1, y_2 e y_3 serve como uma referência melhorada do sinal. Ao mesmo tempo, as diferenças mútuas $\hat{n_1}$, $\hat{n_2}$ e $\hat{n_3}$ entre os canais alinhados em fase e o canal de referência y_0 servem como entradas de referência de ruído para um cancelador de ruído

[1]Não confundir com canal de transmissão. Aqui canal significa microfone.

multicanal adaptativo (Widrow and Stearns, 1985; Ferrara and Widrow, 1981) [11, 8] e produz o sinal $\hat{s}_2(t)$ na saída da unidade de alinhamento Griffiths-Jim.



FONTE: COMPERNOLLE (1990) [3], pag. 436

Figura 2.11: Alinhamento por método Griffths-Jim de 4 canais.

Os sinais resultantes obtidos pela média, $\hat{s}_1(t)$, e na saída do sistema, $\hat{s_2}(t)$, são dados por:

$$\hat{s}_1(t) = \frac{1}{N+1}[y_0(t) + \sum_{k=1}^{N} y_k(t - \tilde{\tau_k})] \tag{2.12}$$

$$\hat{n_k}(t) = y_k(t - \tilde{\tau_k}) - y_o(t) \tag{2.13}$$

$$\hat{s_2}(t) = \hat{s}_1(t) - \sum_{k=1}^{N} w_k(t) * \hat{n}_k(t) \tag{2.14}$$

A escolha do canal de referência, bem como a numeração dos canais são feitas de forma arbitrária. Em [3], na implementação original desta unidade de alinhamento, foi

forçada a priori uma direção para o canal predominante definida com atrasos τ_k fixos. Para aplicações de voz, este método pode ser usado somente se puderem ser impostas restrições acerca da posição do locutor. Isto pode ser possível no caso de ajudas de escuta, por exemplo, onde uma direção de captação fixa pode ser definida na frente do usuário da ajuda de escuta.

A estrutura de Griffiths–Jim é bem adequada para voz e pode ser usada em sistemas onde se deseja solucionar os seguintes problemas:

- determinação da direção da fonte de som ou canal predominante (alvo);
- melhoramento do sinal detectado e
- ajuste dos filtros adaptativos para cancelamento do ruído.

2.7.2 Alinhamento com chaveamento e captação de alvo adaptativa

2.7.2.1 Implementação da seção de Atraso e Soma

A seção de atraso e soma é implementada de tal forma que permita uma posição flexível do locutor além de calcular os atrasos de forma adaptativa. Os τ_k's são determinados em função do valor máximo das correlações cruzadas dos respectivos canais. O alinhamento de fase é determinado por:

$$R_{k0}(\tau) = E[y_k(t) \cdot y_0(t - \tilde{\tau}_k)] \tag{2.15}$$

onde $R_{k0}(\tau)$ é a correlação cruzada entre o canal 0 ou de referência e um dos k canais secundários e $\tilde{\tau}_k$ representa o instante em que é encontrado o valor máximo de $R_{k0}(\tau)$, ou seja, o atraso do sinal.

Os melhores resultados são obtidos se o valor esperado na Equação (2.15) é estritamente calculado nos períodos em que o alvo está presente. A confiabilidade dos valores estimados depende da mobilidade do locutor e da relação sinal–ruído. Para um locutor

fixo e uma SNR > 6 dB, um segmento de voz de 250 ms é exigido para estimativas confiáveis do atraso de tempo [3].

2.7.2.2 Griffiths-Jim Chaveado

A estrutura do Griffiths-Jim padrão descrita acima falha em aplicações de voz. Este método leva a uma boa solução quando as referências de ruído são "limpas", isto é, se elas não contém sinais de voz. Em aplicações de voz, estas sempre contém considerável fuga de sinal devido às contribuições multipercurso. O uso da estrutura Griffits–Jim é somente bem sucedida em situações de relação sinal-ruído fortemente negativas. Em outras situações, cancelamento de sinal e degradação muito audível podem ocorrer. Referências de ruído limpas podem ser obtidas se a função de transferência do locutor aos microfones é conhecida com grande precisão (Faucon e Le Bouquin, 1985) [27]. No entanto, isto só pode ser obtido "*off-line*" e nenhuma implementação adaptativa satisfatória tem sido demonstrada.

Para evitar problemas de cancelamento de sinal, é imperativo que a adaptação dos coeficientes do filtro, $w(k)$, seja interrompida quando o sinal alvo está presente. Conseqüentemente, a determinação da condição de "presença de sinal" é crucial à implementação da unidade de alinhamento usando Grifiths-Jim chaveado.

2.7.2.3 Condições multipercurso - dereverberação

Na determinação dos atrasos de fase, supõe-se que o caminho direto domine as reflexões anteriores, caso contrário a determinação dos atrasos de fase não funcionará. Entretanto isto permite reverberação muito forte. A única condição pedida ao locutor é falar alguma coisa na direção do arranjo de microfones, o que parece muito plausível em circunstâncias adversas.

Com a determinação adequada dos atrasos de fase, o alinhamento pelo método de atraso e soma melhorará na proporção do caminho direto versus as ondas refletidas. Conseqüentemente, além da supressão do ruído, esta seção também proporciona dereverberação. Para um sistema de reconhecimento de voz, especialmente para um sistema

usando processamento de sinal baseado em LPC, isto é uma vantagem adicional.

2.7.2.4 Resolução espacial e movimentação de alvos

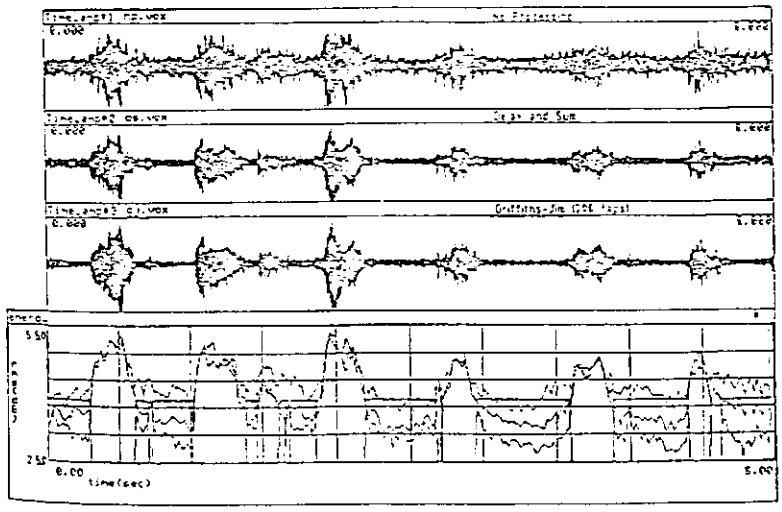
A resolução nos atrasos da unidade de alinhamento de fase é de um período de amostragem, e este determina a resolução do ângulo do alinhador principal. Conseqüentemente, é impossível seguir, lentamente, um alvo em movimento.

Uma freqüência de amostragem de 10 kHz e dois microfones numa distância de 20 cm implicam numa exatidão de ângulo de apenas 15 graus. Isto faz com que o sistema de captação de alvo (sinal desejado) dê verdadeiros saltos quando o locutor se move ao redor do mesmo. Por esta razão, deve-se escolher freqüências de amostragens superiores a 10 kHz.

2.7.3 Experimentos para melhoramento de voz

Compernolle [3] utilizou, em seus experimentos para melhoramento de voz, um arranjo de 4 microfones, com o locutor de 1 m a 2 m de distância do arranjo e uma taxa de amostragem de 20 kHz. Para a fonte de ruído, foi utilizado um rádio mal sintonizado com uma distância de 1 m a 2 m do arranjo de forma que o som ruidoso atuasse de forma refletida.

A Figura 2.12 mostra a entrada, a soma e o atraso processados, as formas de onda processadas no Griffiths-Jim e a energia para uma sentença com SNR de 12 dB. Os períodos para os quais o cancelador se adapta são indicados por barras fortes no gráfico da energia no nível de limiar. A solução de atraso e soma produz um melhoramento na relação sinal ruído de cerca de 5 dB e o cancelador de ruído adaptativo aumenta mais 5 dB.



FONTE: Compernolle (1990) [3], pag. 439

Figura 2.12: Formas de onda no tempo e energia a curto intervalo de tempo para os sinais de entrada e o sinal filtrado do alinhador adaptativo.

2.7.4 Experimentos para reconhecimento de voz

Um sistema de reconhecimento de voz usando modelo de Markov escondido foi utilizado para avaliação [3]. O sistema foi desenvolvido para aplicações telefônicas e usado para tarefas de pequeno vocabulário de palavras isoladas multilocutor e independente do locutor. Desempenhos típicos quando usado para o reconhecimento dos dígitos de Dutch são:

- multilocutor (6), escritório silencioso: > 99 %
- independente do locutor, qualidade telefônica: 95 %

Em salas reverberantes, em distância de aproximadamente 2 m do arranjo, o desempenho cai de 99 % para 80 % correto, enquanto num escritório não reverberante padrão, distâncias de 2 a 3 m podem ser facilmente toleradas.

Para o caso do alinhamento por atraso e soma foi obtido um ganho de +4 dB e para o Griffiths-Jim um ganho de +5 a +8 dB [3].

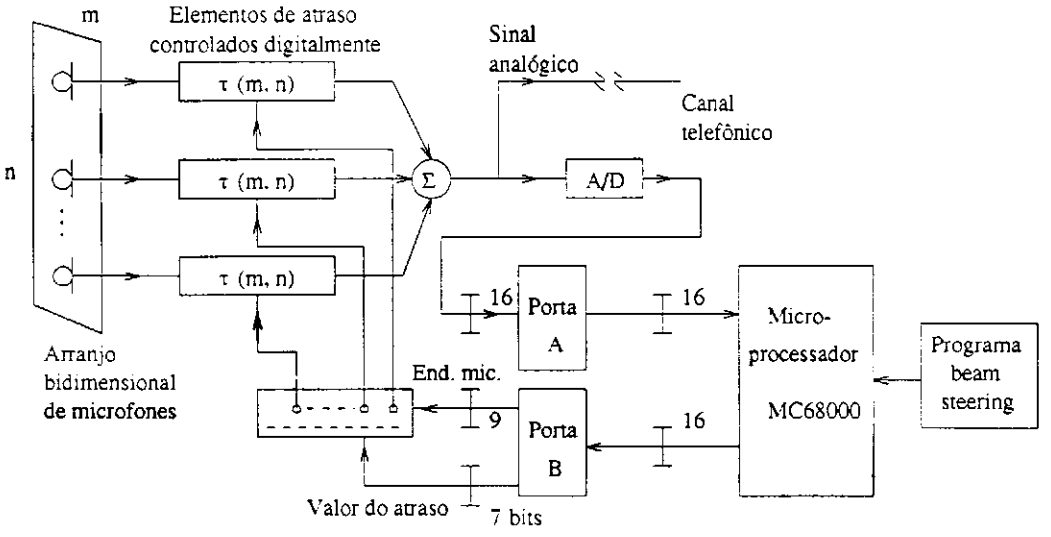
2.8 Sistema com arranjo de microfones conduzido por computador para transdução em grandes ambientes

A qualidade do som captado em grandes ambientes tais como auditórios, salas de conferências, ou salas de aula é prejudicada pela reverberação e pelas fontes de ruído interferentes. Estas degradações podem ser minimizadas por um sistema transdutor que discrimina do som que chega de todas as direções o som da fonte desejada. Um arranjo de microfones bidimensional pode ser eletronicamente conduzido para acompanhar esta diretividade. Flanagan (1985) [9], mostra a teoria, o projeto e a implementação de um sistema de microprocessador que utiliza um arranjo bidimensional, localiza automaticamente e dirige-se ao locutor dominante ativo em salas moderadamente grandes. O sistema é projetado, principalmente, para comunicação de voz e para habilitar teleconferência de grandes grupos interativos. A maximização da relação do som direto com o som reverberante e com o ruído implica na colocação do sistema de microfones tão próximo quanto for possível da fonte desejada. Mas em grandes salas e com grandes grupos - tais como ocorrem para discussões interativas entre salas de conferências e salas de aula - esta proximidade é impraticável. É inconveniente passar microfones perto de cada um dos conferencistas ou distribuir microfones para cada um dos participantes. Além disso, uns poucos microfones unidirecionais isolados, localizados em distâncias substanciais das fontes, não podem atuar satisfatoriamente.

Sistemas com microfones altamente diretivos, que podem ser apontados para a fonte desejada do momento - preferível automaticamente - são, portanto, de maior interesse.

2.8.1 Sistema de *hardware* experimental

Um sistema controlado a microprocessador é mostrado na Figura 2.13. Para cada saída de microfone do arranjo é atribuído um atraso controlado digitalmente. As saídas de todos os elementos do arranjo são somadas para produzir a saída do arranjo. Esta

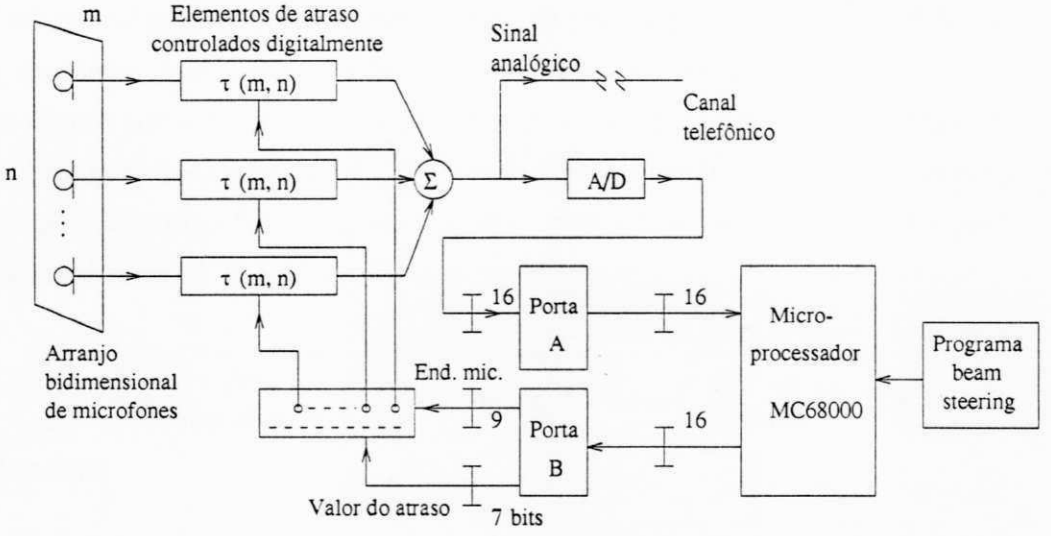


FONTE: Flanagan (1985) [9] pag. 1515

Figura 2.13: Sistema de microprocessador para controlar o arranjo bidimensional.

saída é o sinal recebido pela unidade de alinhamento quando é conduzido à direção especificada pelo ajuste de $\tau'(m, n)$. O microprocessador (MC68000) tem duas portas de entrada/saída para comunicação com o equipamento do arranjo. A porta de saída (porta B) fornece palavras de 16 bits lidas de tabelas armazenadas no computador. Nove bits de cada palavra representam o endereço de um microfone e sete bits representam o valor do atraso a ser aplicado naquele canal microfônico. A porta de entrada (porta A), também de 16 bits, recebe uma versão digitalizada do sinal de saída do arranjo de um conversor analógico-digital (A/D). Este sinal pode ser processado por um programa de alinhamento de sinais no microprocessador. O programa pode ser desenvolvido num computador central de propósito geral e carregado um microprocessador através de uma conexão telefônica.

Na realidade, dois conjuntos de *hardware* de atraso e soma são implementados para formar e processar duas unidades de alinhamento simultaneamente. Isto é feito para dar ao equipamento do arranjo uma capacidade de varrer-enquanto-explora (*track-while-scan*) automático para captação de sons em ambientes espaçosos.



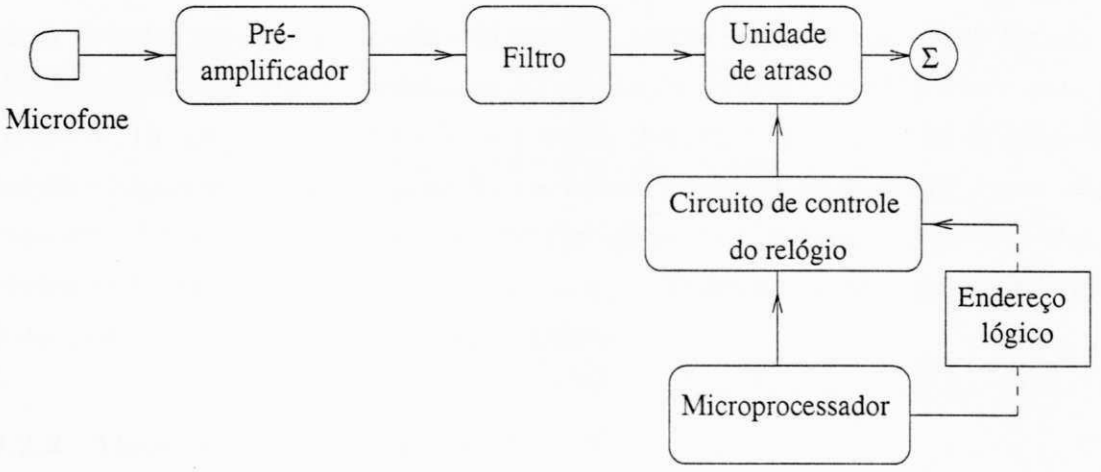
FONTE: Flanagan (1985) [9] pag. 1515

Figura 2.13: Sistema de microprocessador para controlar o arranjo bidimensional.

saída é o sinal recebido pela unidade de alinhamento quando é conduzido à direção especificada pelo ajuste de $\tau'(m, n)$. O microprocessador (MC68000) tem duas portas de entrada/saída para comunicação com o equipamento do arranjo. A porta de saída (porta B) fornece palavras de 16 bits lidas de tabelas armazenadas no computador. Nove bits de cada palavra representam o endereço de um microfone e sete bits representam o valor do atraso a ser aplicado naquele canal microfônico. A porta de entrada (porta A), também de 16 bits, recebe uma versão digitalizada do sinal de saída do arranjo de um conversor analógico-digital (A/D). Este sinal pode ser processado por um programa de alinhamento de sinais no microprocessador. O programa pode ser desenvolvido num computador central de propósito geral e carregado um microprocessador através de uma conexão telefônica.

Na realidade, dois conjuntos de *hardware* de atraso e soma são implementados para formar e processar duas unidades de alinhamento simultaneamente. Isto é feito para dar ao equipamento do arranjo uma capacidade de varrer-enquanto-explora (*track-while-scan*) automático para captação de sons em ambientes espaçosos.

Cada canal microfônico também inclui um pré–amplificador de 50 dB de ganho, um filtro analógico rejeita-faixa de frequência de corte de 4 kHz, o circuito de atraso, uma lógica digital para decodificar as palavras de endereço de 9 bits e de atraso de 7 bits, como mostrado na Figura 2.14. Todo o circuito do canal é implementado numa placa de circuito impresso feita sob encomenda. Cada placa contém circuito para quatro canais microfônicos.



FONTE: Flanagan (1985) [9], pag. 1515

Figura 2.14: Diagrama de blocos para um canal microfônico do arranjo.

O posicionamento dos captadores de fontes às direções desejadas é executado com a maior velocidade através de valores pré–calculados de valores de atraso τ'(m,n). Uma tabela completa pode ser lida e emitida pelo circuito de direção em 600 μs. Esta velocidade para posicionamento do elemento de captação é muito mais rápida do que as medições nos sinais de voz para decidir para onde apontar o arranjo. Esta velocidade é, também, suficientemente rápida para permitir a formação de dois elementos de captação, com dois conjuntos separados de circuito de atraso. Isto permite que um elemento aja com o objetivo de pesquisa continua, enquanto o segundo ou principal, é posicionado no locutor ativo dominante.

2.8.2 Programa para controle de captação automático

2.8.2.1 Características de voz

Um problema fundamental em controle automático de captação é a separação das fontes de voz das fontes de não–voz. Um algoritmo de deteção simples pode ser formado pela exploração de um conhecimento a priori de características específicas de voz. Sons de voz sonoros são produzidos pela ação vibratória das cordas vocais. A maioria da energia acústica irradiada em voz está contida nestes sons sonoros. Suas formas de onda tipicamente exibem periodicidade na excitação do trato vocal (período nominal de cerca de 10 ms para homens e 5 ms para mulheres), e um alto fator de pico. Seu espectro é caracteristicamente passa-baixa, caindo em cerca de 8–10 dB/oitava acima de aproximadamente 500 Hz. A voz também apresenta surtos de energia seguidos de períodos de baixa energia ou de pausas no tempo. Todas estas características ajudam a distinguir entre voz e ruído de fundo contínuo.

2.8.2.2 Algoritmo de deteção de voz

Pequenos, econômicos, os microprocessadores disponíveis atualmente são limitados em espaço e memória. Conseqüentemente, para processamento em tempo real, Flanagan et al [9] utilizaram estratégias rudimentares. Para efetuar a pesquisa do alvo, foi feita uma quantização da área da audiência em direções angulares sobrepostas e os setores varridos seqüencialmente. Os sinais recebidos de cada direção em questão podem ser então comparados um com o outro.

Utilizando-se dois elementos de captação para processamento (usando os componentes mostrados na Figura 2.13), o elemento principal é sempre posicionado no locutor dominante do momento, enquanto o elemento de pesquisa está varrendo a área da audiência para uma nova fonte de voz. Se duas pessoas estão falando ao mesmo tempo, a saída do elemento de pesquisa pode ser opcionalmente usado para transmissão deste sinal. Mais elementos podem, é claro, ser formados, dependendo do desejo de investir em *hardware* para captação e alinhamento de sinais e da velocidade e capacidade do microprocessador para controlar os elementos de captação.

Para processamento da saída do arranjo recebida em qualquer localização explorada, há duas medidas simples que podem ser rapidamente calculadas (porque não requerem multiplicações). São os valores da magnitude de pico do sinal dentro de um intervalo de tempo, e a magnitude média sobre o mesmo intervalo. O comprimento da função janela do tempo é escolhido tal que tipicamente no mínimo um período de fundamental esteja contido dentro da janela. Isto implica um intervalo de tempo de análise da ordem de 10 ms. O número de amostras tomadas dentro desta janela deverá ser suficientemente grande tal que os valores das magnitudes de pico e média sejam confiavelmente estimados.

Em [9], foi utilizado um algoritmo para a deteção de um sinal de voz baseado na relação da magnitude média a curto intervalo de tempo e a longo intervalo de tempo de cada direção do elemento de captação. Uma decisão de que o sinal está presente é feita se esta relação excede um limiar prescrito. Se mais de uma direção do elemento de captação produz uma relação que ultrapasse este limiar, a direção do elemento de captação que tem a energia a curto intervalo de tempo maior é escolhida como a direção na qual o elemento de captação principal deve apontar.

O sistema implementado por Flanagan et al [9], parece altamente vantajoso para teleconferência de grandes grupos e o desenvolvimento de microeletrônica para processamento sofisticado de sinais pode suportar algumas aplicações. Extensões para implementações puramente digitais virão a medida que as economias de VLSI proliferarem. Crucial à operação adequada do transdutor de "busca de voz" (*speech-seeking*) é que tenha o suporte computacional necessário (dedicado ao transdutor) para fazer julgamentos razoáveis acerca das características da fonte de som. As diferenças físicas entre os sinais de voz, música, e ruído interferente, tanto contínuo quanto transitório, representam conhecimento programado no microprocessador de busca-de-sinal (*signal-seeking*).

As várias técnicas vistas acima de processamento de sinais de voz e suas respectivas aplicações servem como referências básicas para o sistema de supressão de ruído proposto neste trabalho. A seguir será vista a teoria básica sobre o filtro Wiener-Kolmogoroff utilizado pelo sistema na supressão do ruído.

Capítulo 3

O Filtro Wiener-Kolmogoroff

3.1 Introdução

O método usual de estimação de um sinal degradado pelo ruído aditivo é passá-lo por um filtro que tende a suprimir o ruído enquanto o sinal permanece relativamente inalterado. O projeto de tais filtros é domínio da filtragem ótima, que originou do trabalho pioneiro de Wiener e foi estendido e melhorado pelo trabalho de Kalman, Bucy e outros [28, 29, 30].

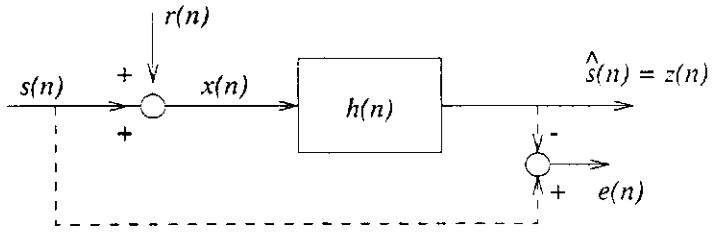
Filtros usados para o propósito acima podem ser fixos ou adaptativos. O projeto de filtros fixos é baseado no conhecimento a priori tanto do sinal quanto do ruído. Por outro lado, filtros adaptativos têm a capacidade de ajustar seus próprios parâmetros automaticamente e seu projeto requer pouco ou nenhum conhecimento a priori das características do sinal ou do ruído. Filtros fixos são em sua maioria inaplicáveis em supressão de ruído, porque as funções de correlação e correlação cruzada das entradas principal e secundária(s) são geralmente desconhecidas e freqüentemente variam com o tempo. Filtros adaptativos requerem o "aprendizado" das estatísticas inicialmente e seguem-nas se elas variam lentamente [11].

O filtro Wiener-Kolmogoroff é bastante utilizado em sistemas de supressão de ruído. Ele está incluido na categoria dos filtros ótimos e é utilizado em sistemas não-seletivos

que permitem uma filtragem contínua do sinal degradado. O filtro Wiener é aplicado na redução de ruído, redução de distorção e na predição linear [2].

3.2 Estrutura do filtro Wiener-Kolmogoroff (filtro WK)

A estrutura básica do filtro WK é mostrada na Figura 3.1. O objetivo do filtro WK é a estimação do sinal original $s(n)$ a partir do sinal degradado $x(n) = s(n) + r(n)$, onde $r(n)$ representa o ruído. A resposta ao impulso do filtro, $h(n)$, é determinada em função das propriedades estatísticas do sinal a ser melhorado e do ruído. Estas propri-



FONTE: Aguiar Neto (1987) [5], pag. 20

Figura 3.1: Estrutura geral de um filtro Wiener-Kolmogoroff

edades geralmente não são disponíveis e um conhecimento a priori das mesmas também não é possível se os sinais são não-estacionários. Logo, as informações estatísticas dos sinais devem ser obtidas a partir do sinal degradado e em curtos intervalos de tempo, de forma que o sinal seja suficientemente estacionário no intervalo de tempo considerado.

A resposta ao impulso $h(n)$ de um sistema linear, como o da Figura 3.1, deve ser tal que leve ao menor erro possível $\epsilon(n)$ entre o sinal $s(n)$ e a sua respectiva estimação $\hat{s}(n) = h(n) * x(n)$ (onde $*$ é o operador convolucional), ou seja, a variância do erro seja mínima:

$$\sigma_e^2 = E[\{s(n) - \hat{s}(n)\}^2] \doteq \min \tag{3.1}$$

3.3 Determinação dos coeficientes ótimos do filtro Wiener-Kolmogoroff

Da Equação 3.1, sabe-se que o erro de estimação do sinal original pelo filtro deve ser mínimo. O sinal estimado $\hat{s}(n)$, na saída do filtro, é dado por:

$$\hat{s}(n) = x(n) * h(n) \tag{3.2}$$

ou ainda,

$$\hat{s}(n) = \sum_{j \in I} h(j) \cdot x(n-j) \tag{3.3}$$

deseja-se, então, encontrar coeficientes h(j), tais que o erro seja mínimo. Logo, substituindo-se a Equação 3.3 na Equação 3.1, obtém-se:

$$\sigma_e^2 = E[\{s(n) - \sum_{j \in I} h(j) \cdot x(n-j)\}^2] \doteq \min \tag{3.4}$$

Para que a Equação 3.4 seja satisfeita, toma-se a sua derivada em relação a h_i, ou seja,

$$\frac{\partial \sigma_e^2}{\partial h_i} \doteq \min, \quad i \in I \tag{3.5}$$

Assim,

$$\frac{\partial \sigma_e^2}{\partial h_i} = 2 \cdot E[\{s(n) - \sum_{j \in I} h_j \cdot x(n-j)\} \cdot \{-x(n-i)\}] \tag{3.6}$$

Do princípio da ortogonalidade [20, 19], segue que os coeficientes do filtro, $h(j)$, devem ser escolhidos de forma que a diferença entre $s(n)$ e a sua estimação seja ortogonal aos dados, isto é,

$$E[\{s(n) - \hat{s}(n)\}\{-x(n-i)\}] = 0$$

logo,

$$E[e(n) \cdot x(n-i)] = 0 \tag{3.7}$$

ou seja, o erro $e(n)$ e o sinal degradado $x(n)$ são ortogonais.

Das Equações 3.6 e 3.7, temos que:

$$\frac{\partial \sigma_e^2}{\partial h_i} = 0 = -2 \cdot E[s(n) \cdot x(n-i)] + 2 \cdot E[\sum_{j \in I} h_j \cdot x(n-j) \cdot x(n-i)]$$

logo,

$$E[s(n) \cdot x(n-i)] = E[\sum_{j \in I} h_j \cdot x(n-j) \cdot x(n-i)] \tag{3.8}$$

ou seja,

$$R_{xs}(i) = \sum_{j \in I} h_{j,\text{opt}} \cdot R_{xx}(i-j), \quad i \in I \tag{3.9}$$

A Equação 3.9 é a Equação de Wiener-Hopf [31], onde R_{xs}(i) representa a correlação cruzada entre $x(n)$ e $s(n)$, R_{xx}(i) a autocorrelação de $x(n)$ e $h_{j,\text{opt}}$ representa os coeficientes ótimos do filtro Wiener-Kolmogoroff.

Na forma matricial, temos a seguinte notação:

$$\mathbf{r}_{xs} = \mathbf{R}_{xx} \cdot \mathbf{h}_{\text{opt}} \tag{3.10}$$

$$\mathbf{h}_{\text{opt}} = \mathbf{R}_{xx}^{-1} \cdot \mathbf{r}_{xs} \tag{3.11}$$

onde $\mathbf{R}_{xx}$ é a matriz de covariância de estrutura Toeplitz[1] [32].

[1]Quando os elementos de uma matriz em qualquer diagonal são iguais, a matriz é dita Matriz Toeplitz Simétrica

3.4 Classificação dos filtros Wiener-Kolmogoroff

A resposta ao impulso $h(n)$ está no intervalo I = (A, B), com $-\infty \leq A \leq \infty$ e $-\infty \leq B \leq \infty$, determina o tipo do filtro, que podem ser classificados como:

3.4.1 Filtro causal

O filtro Wiener-Kolmogoroff é dito causal quando I=(0,∞), ou seja, para a estimação de uma amostra num determinado instante de tempo, precisa-se de todas as amostras anteriores.

Os coeficientes do filtro são calculados pela Equação 3.9 da seguinte forma:

$$R_{xs}(i) = \sum_{j=0}^{\infty} h_{j,\text{opt}} \cdot R_{xx}(i-j) \tag{3.12}$$

onde i = 0, ..., ∞.

3.4.2 Filtro causal finito

O filtro Wiener-Kolmogoroff é classificado como causal finito para I = (0, N). O filtro causal finito (Figura 3.2) é a implementação prática possível do filtro causal.

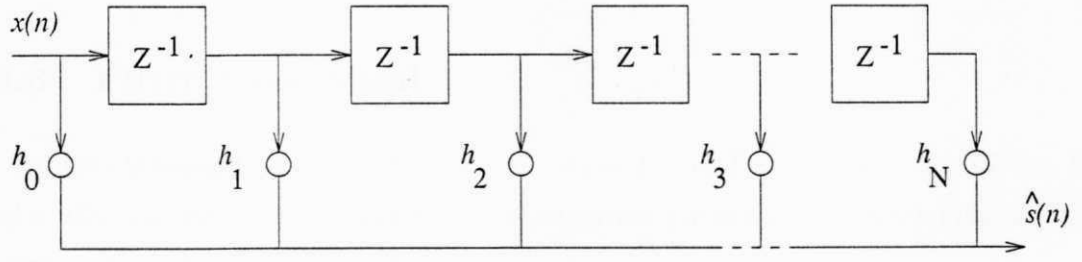


Figura 3.2: Estrutura do filtro WK causal-finito.

Os coeficientes do filtro WK causal finito são determinados a partir do sistema de equações:

$$R_{xs}(i) = \sum_{j=0}^{N} h_{j,\text{opt}} \cdot R_{xx}(i-j) \tag{3.13}$$

e

$$\hat{s}(n) = \sum_{j=0}^{N} h_j \cdot x(n-j) \tag{3.14}$$

onde i = 0, ..., N.

Em forma matricial:

$$\begin{bmatrix} R_{xx}(0) & R_{xx}(1) & R_{xx}(2) & \dots & R_{xx}(N) \\ R_{xx}(1) & R_{xx}(0) & R_{xx}(1) & \dots & R_{xx}(N-1) \\ R_{xx}(2) & R_{xx}(1) & R_{xx}(0) & \dots & R_{xx}(N-2) \\ \vdots & \vdots & \vdots & & \vdots \\ R_{xx}(N) & R_{xx}(N-1) & R_{xx}(N-2) & \dots & R_{xx}(0) \end{bmatrix} \cdot \begin{bmatrix} h_{0,\text{opt}} \\ h_{1,\text{opt}} \\ h_{2,\text{opt}} \\ \vdots \\ h_{N,\text{opt}} \end{bmatrix} = \begin{bmatrix} R_{xs}(0) \\ R_{xs}(1) \\ R_{xs}(2) \\ \vdots \\ R_{xs}(N) \end{bmatrix}$$

Para o filtro com um coeficiente, por exemplo, tem-se que:

$$R_{xs}(0) = h_{0,\text{opt}} \cdot R_{xx}(0) \tag{3.15}$$

Assim,

$$h_{0,\text{opt}} = \frac{R_{xs}(0)}{R_{xx}(0)} \tag{3.16}$$

3.4.3 Filtro não-causal

O filtro Wiener-Kolmogoroff é dito não-causal para I=$(-\infty, \infty)$. Este é um filtro ideal e não-realizável que requer todas as amostras passadas para a estimação de uma amostra em qualquer $-\infty \leq t \leq \infty$.

A obtenção dos coeficientes ótimos do filtro WK não-causal é feita através da seguinte equação:

$$R_{xs}(i) = \sum_{j=-\infty}^{\infty} h_{j,\text{opt}} \cdot R_{xx}(i-j), \quad i \in I \tag{3.17}$$

$$R_{xs}(i) = h_{\text{opt}}(i) * R_{xx}(i) \tag{3.18}$$

onde $h_{\text{opt}}(i)$, representa a resposta ótima ao impulso do filtro WK (Figura 3.3).

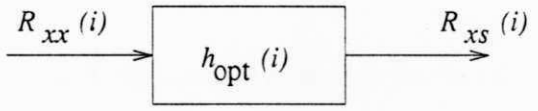


Figura 3.3: Filtro Wiener-Kolmogoroff com resposta ao impulso ótima.

3.4.3.1 Função de transferência do filtro WK não-causal

Sendo $F\{\cdot\}$ o operador transformada de Fourier e aplicando-o Equação 3.17, obtém-se:

$$F\{R_{xs}(i)\} = F\{h_{\text{opt}}(i) * R_{xx}\} \tag{3.19}$$

como

$$F\{R_{xs}(i)\} \rightarrow S_{xs}(\Omega),$$

$$F\{R_{xx}(i)\} \rightarrow S_{xx}(\Omega)$$

e

$$F\{h_{\text{opt}}(i)\} \rightarrow H_{\text{opt}}(\Omega)$$

e usando-se a propriedade da convolução da transformada de Fourier da Equação 3.19, obtém-se:

$$S_{xs}(\Omega) = H_{\text{opt}}(\Omega) \cdot S_{xx}(\Omega) \tag{3.20}$$

Assim, a função de transferência do filtro WK (Figura 3.4) será dada por [31]:

$$H_{\text{opt}}(\Omega) = \frac{S_{xs}(\Omega)}{S_{xx}(\Omega)} \tag{3.21}$$

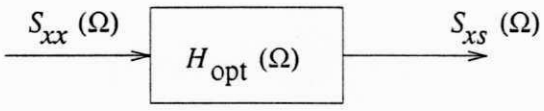


Figura 3.4: Filtro WK não-causal com função de transferência ótima.

Com o erro de estimação ortogonal ao sinal degradado, ou seja:

$$E[\{s(n) - \hat{s}(n)\} \cdot x(n-i)] = 0 \tag{3.22}$$

e desenvolvendo a Equação 3.22, obtém-se:

$$R_{sx}(i) = R_{\hat{s}x}(i) \tag{3.23}$$

Para o sinal degradado x(n) dado por x(n) = s(n) + r(n), tem-se que:

$$E[\{s(n) - \hat{s}(n)\} \cdot \{s(n) + r(n)\}] = E[(s(n) - \hat{s}(n)) \cdot s(n) + (s(n) - \hat{s}(n)) \cdot r(n)]$$

mas,

$$E[(s(n) - \hat{s}(n)) \cdot r(n)] = E[s(n) \cdot r(n)] - E[\hat{s}(n) \cdot r(n)]$$

Como $s(n)$ é ortogonal a $r(n)$, espera-se que a sua estimação $\hat{s}(n)$ também o seja. Assim, temos que:

$$E[s(n) \cdot r(n)] = 0$$

e

$$E[\hat{s}(n) \cdot r(n)] = 0$$

Assim,

$$E[(s(n) - \hat{s}(n))r(n)] = 0 \quad \text{e}$$

$$E[\{s(n) - \hat{s}(n)\}\{s(n) + r(n)\}] = E[\{s(n) - \hat{s}(n)\}s(n)]$$

$$\Rightarrow \quad E[\{s(n) - \hat{s}(n)\} \cdot s(n)] = 0 \tag{3.24}$$

ou seja, o erro de estimação é ortogonal a s(n). Logo,

$$\min\{\sigma_e^2\} = R_{ss}(0) - E[\sum_{j=-\infty}^{\infty} h_{\text{opt}}(j) \cdot x(n-j) \cdot s(n)]$$

$$= R_{ss}(0) - \sum_{j=-\infty}^{\infty} h_{\text{opt}}(j) \cdot R_{xs}(j)$$

$$\Rightarrow \min\{\sigma_e^2\} = \sigma_s^2 - \sum_{j=-\infty}^{\infty} h_{\text{opt}}(j) \cdot R_{xs}(j) \tag{3.25}$$

O sinal degradado $x(n)$ é dado por $x(n) = s(n) + r(n)$. Para encontrar-se a correlação entre $x(n)$ e $s(n)$ procede-se da seguinte forma:

$$E[x(n) \cdot s(n)] = E[(s(n) + r(n)) \cdot s(n)] \tag{3.26}$$

$$\rightarrow E[x(n) \cdot s(n)] = E[s(n) \cdot s(n)] + E[s(n) \cdot r(n)]$$

Para o caso de $s(n)$ e $r(n)$ descorrelacionados (o ruído descorrelacionado com o sinal) [5], temos que $E[s(n) \cdot r(n)] = 0$. Assim,

$$R_{xs}(i) = R_{ss}(i) \tag{3.27}$$

e

$$R_{xx}(i) = R_{ss}(i) + R_{rr}(i) \tag{3.28}$$

logo, no domínio da freqüência, tem-se que:

$$S_{xs}(\Omega) = S_{ss}(\Omega) \quad \text{e} \tag{3.29}$$

$$S_{xx}(\Omega) = S_{ss}(\Omega) + S_{rr}(\Omega) \tag{3.30}$$

Substituindo-se as Equações 3.29 e 3.30 na Equação 3.21, obtém-se [5]:

$$H_{\text{opt}}(\Omega) = \frac{S_{ss}(\Omega)}{S_{ss}(\Omega) + S_{rr}(\Omega)} \tag{3.31}$$

e

$$\min\{\sigma_e^2\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S_{ss}(\Omega)\cdot S_{rr}(\Omega)}{S_{ss}(\Omega) + S_{rr}(\Omega)} d\Omega \tag{3.32}$$

3.4.4 Filtro finito com atraso

O filtro Wiener-Kolmogoroff é dito ser finito, com atraso, para I = (-M, N-M), onde M e N podem assumir qualquer valor inteiro. Logo, um *buffer* deve ser usado para retenção das amostras para processamento. O cálculo dos coeficientes ótimos do filtro WK finito com atraso é calculado a partir da Eq. (3.17) por:

$$R_{xs}(i) = \sum_{j=-M}^{N-M} h_{j,\text{opt}} \cdot R_{xx}(i-j),\ i \in I;\quad I = \{-M, N-M\} \tag{3.33}$$

O vetor de correlação cruzada é dado por:

$$\mathbf{r}_{xs} = \{R_{xs}(-M), \ldots, R_{xs}(0), \ldots, R_{xs}(N-M)\}^T$$

A matriz de correlação é dada por:

$$\begin{bmatrix} R_{xs}(-M) \\ R_{xs}(-2) \\ R_{xs}(-1) \\ R_{xs}(0) \\ R_{xs}(1) \\ R_{xs}(2) \\ \vdots \\ R_{xs}(N) \end{bmatrix}$$

Exemplos:

- Para o intervalo I = {-2, 2} → filtro finito com atraso;

- Para o intervalo I = 0, N → filtro Wiener causal finito.

Neste trabalho é usado um filtro Wiener-Kolmogoroff causal, cujos coeficientes são calculados adaptativamente. Até a obtenção do sinal estimado na saída do sistema, o sinal de entrada passa por alguns processos como o alinhamento de fase e processamento na freqüência para o cálculo das funções de autocorrelação e correlação cruzada. O sinal estimado na saída é obtido pela convolução dos coeficientes do filtro Wiener e o sinal médio resultante da unidade de alinhamento do sistema. Maiores detalhes sobre cada unidade do sistema, bem como da utilização do filtro Wiener serão vistos no capítulo a seguir.

Capítulo 4

Sistema multicanal adaptativo para supressão de ruído

4.1 Introdução

Os sistemas de redução de ruído em sinais de voz dividem-se, de uma forma geral, em sistemas de supressão de ruído e em sistemas de cancelamento de ruído [5]. Esses sistemas podem, ainda, ser classificados como sistemas monocanal ou sistemas multicanal. Esta classificação é feita em função do número de entradas de sinal disponíveis [6]. Os sistemas de cancelamento de ruído são casos típicos de sistemas multicanal, pois dispõem de pelo menos uma entrada de referência para o ruído.

Nos sistemas de cancelamento de ruído pode ocorrer, além do ruído interferente, o cancelamento indesejável de componentes do sinal de voz. Dessa forma, busca-se um sistema para cancelar o ruído sem suprimir componentes do sinal de voz.

A despeito de décadas de esforços de pesquisadores, a tarefa para melhoramento de voz degradada por ruído está incompleta, já que poucos métodos tem resolvido os seguintes problemas [33]:

1. A dificuldade da deteção de intervalos de silêncio num sinal de voz ruidoso para estimar o nível de ruído ou para formar o sinal de referência em sistemas de supressão de ruído;

2. A falha no melhoramento dos segmentos surdos de voz;

3. O alto custo computacional;

4. A dependência do sistema em relação às condições de aprendizado.

Neste trabalho é apresentado um sistema de supressão de ruído multicanal, que procura eliminar ou pelo menos reduzir os problemas acima citados. Para suprimir o ruído é usado um arranjo de quatro microfones num sistema que explora as características de correlação do sinal de voz bem como as do ruído. O sinal melhorado obtido na saída do sistema é adequado para transmissão bem como para utilização como entrada em sistemas de reconhecimento de voz. A supressão do ruído é realizada através da filtragem do sinal degradado usando para este fim o filtro Wiener-Kolmogoroff cujos coeficientes são calculados adaptativamente. Os sinais captados por cada microfone são alinhados pela unidade de alinhamento de fase baseado no coeficiente de correlação cruzada entre os sinais captados por cada microfone. Além disso, um estágio adicional de pós-processamento das medidas de correlação cruzada é utilizado de forma a aumentar de forma significativa o desempenho do sistema.

No sistema proposto não há a necessidade da deteção de pausas para estimação do ruído, não havendo, portanto, a presença de erro de deteção nas pausas. O sistema apresenta um bom desempenho mesmo para uma pequena correlação entre os sinais de ruído dos diferentes canais. Os algoritmos implementados são simples e de baixo custo computacional. Além disso, o sistema independe de aprendizado para o seu funcionamento.

4.2 Sistema multicanal adaptativo para supressão de ruído usando arranjo de microfones

O sistema proposto utiliza um arranjo de quatro microfones para a obtenção dos sinais de referência. O arranjo de microfones usado é um quadrado de dimensões 0,6 m x 0,6 m, com uma distância de 0,6 m do locutor ao microfone principal (Figura 4.1). O microfone mais próximo ao locutor é considerado como referência (M_1).

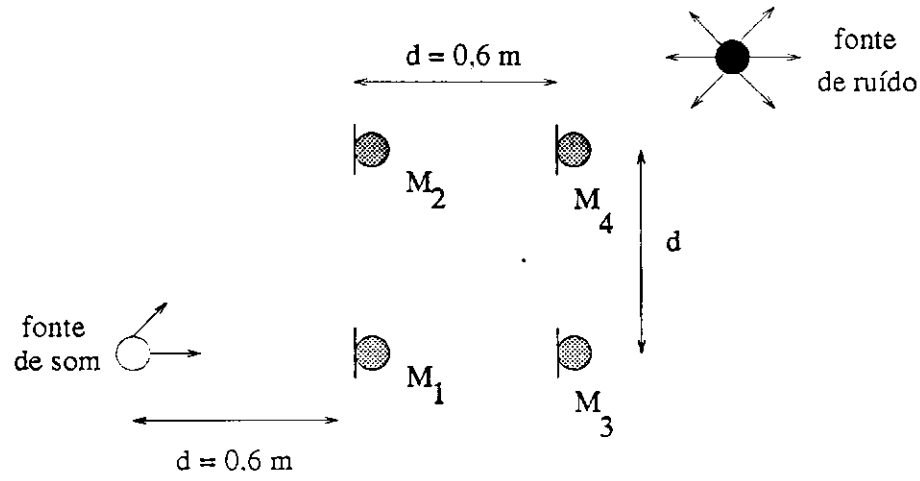


Figura 4.1: Arranjo bidimensional de microfones do sistema de supressão de ruídos.

O sistema multicanal adaptativo, ora proposto, utiliza o arranjo da Figura 4.1, onde a fonte de som (o locutor) está mais próximo ao microfone e o som atua de forma direta. A fonte de ruído, no entanto, fica mais distante do arranjo, de forma que haja um domínio dos sons refletidos. A diretividade do arranjo dos microfones proporciona um ganho que diminui o ruído ambiental. O uso de bons microfones unidirecionais reduz o efeito da degradação do sinal pela propagação multipercurso a um valor irrelevante [3].

A Figura 4.2, dada a seguir, mostra o sistema de supressão de ruído proposto [34].

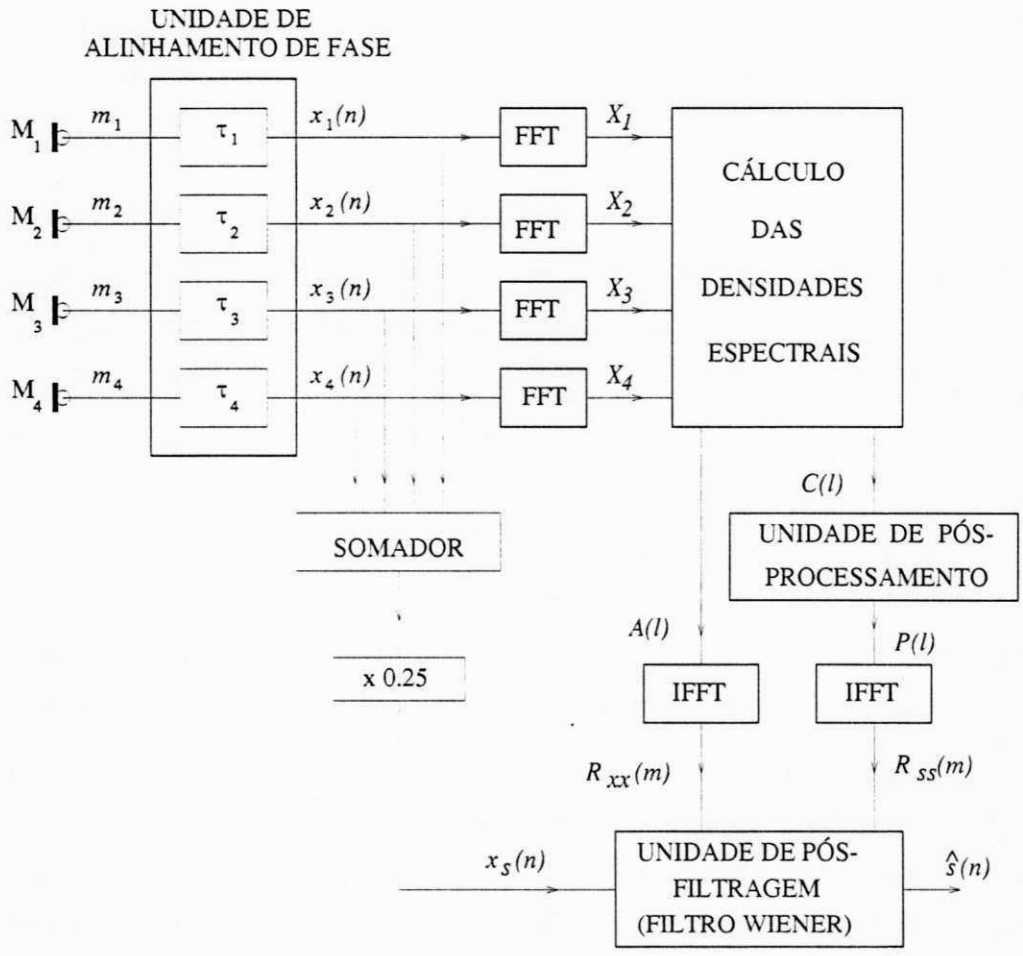


Figura 4.2: Diagrama de blocos do sistema de supressão de ruídos usando arranjo de quatro microfones.

Os sinais, $m_i(n)$, são captados pelos microfones M_1 ... M_4, tal que:

$$m_i(n) = s(n) + r_i(n); \quad i = 1, ..., 4 \tag{4.1}$$

onde $s(n)$ representa o sinal original e $r_i(n)$ representa as componentes do ruído interferente no i-ésimo canal.

Os sinais captados entram na unidade de alinhamento de fase, onde será estimado o atraso dos canais adicionais em relação ao canal de referência, aqui escolhido arbitrariamente como o microfone 1. Os sinais são ajustados de forma que os sinais sejam

alinhados em fase nos quatro canais, com relação ao sinal do microfone de referência, como se os sinais chegassem simultaneamente nos quatro microfones. Tem-se, portanto:

$$x_i(n) = m_i(n - \tau_i) \quad i = 1, \ldots, 4 \tag{4.2}$$

onde $m_i(n - \tau_i)$ é o sinal alinhado no i-ésimo microfone com atraso τ_i; i = 1, ..., 4, onde τ_1 é igual a zero ou a um determinado valor fixo.

Como um primeiro passo para a redução das componentes de ruído $r_i(n)$, é obtido um sinal médio entre os sinais já alinhados:

$$x_s(n) = \frac{1}{N+1} \sum_{i=1}^{N+1} x_i(n) \qquad i = 1, 2, \ldots, N \tag{4.3}$$

onde N representa o número de entradas adicionais, (N = 3, neste caso), e $x_i(n)$ o sinal degradado captado pela i-ésima entrada do sistema, já alinhado em fase.

A unidade de alinhamento de fase é vista a seguir de forma mais detalhada. Esta unidade funciona como um pré-processamento para os sinais captados pelos microfones. O sinal médio obtido a partir dos sinais alinhados já é uma versão melhorada dos sinais degradados na entrada do sistema. Este melhoramento se dá pelo fato de que, com o deslocamento dos sinais no tempo para a retirar-se o atraso, aumenta a correlação entre os sinais de voz, enquanto a correlação entre as componentes de ruído presentes nestes sinais torna-se menor. Assim, há um ganho na relação sinal-ruído destes sinais. O sinal médio é utilizado como entrada para a unidade de pós-filtragem.

4.3 Unidade de alinhamento de fase

O alinhamento correto dos sinais é de fundamental importância para o sistema de supressão de ruído como um todo. Isto porque um alinhamento incorreto dos sinais poderia introduzir distorção no sinal resultante, provocando uma queda no desempenho do sistema.

4.3.1 Alinhamento de fase dos sinais captados pelos microfones

Com uma fonte de sinal única, em uma distância moderada dos microfones, os sinais recebidos nos quatro microfones são réplicas atrasadas umas das outras e escalonadas por um fator de ganho pequeno. O uso de microfones unidirecionais de boa qualidade reduz o efeito espectral do ângulo de chegada a um valor mínimo, podendo este ser desprezado. Outros efeitos espectrais poderiam ser introduzidos através de diferenças entre os microfones individuais. Estes efeitos podem ser minimizados procurando-se usar microfones com curvas de respostas mais próximas possíveis umas das outras [3].

Todos os sinais captados pelos microfones são alinhados em fase e somados. Com N entradas adicionais, além da entrada de referência, o sinal resultante da média entre os sinais alinhados, $\hat{x}_s$, é dado por:

$$\hat{x}_s(n) = \frac{1}{N+1}[m_1(n) + \sum_{k=2}^{N+1} m_k(t - \tilde{\tau}_k)] \qquad k = 2, ..., N+1 \qquad (4.4)$$

onde, $m_1(n)$ é o sinal no microfone 1 e, m_k e $\tilde{\tau}_k$ representam, respectivamente, o sinal e seu atraso no k-ésimo microfone do arranjo.

4.3.2 Estimação do atraso entre os sinais captados pelo arranjo de microfones

A Figura 4.3, dada a seguir, apresenta a unidade de alinhamento de fase proposta, baseada no modelo de Kaneda e Tohyama [33].

No modelo proposto em [33], o arranjo consiste em apenas dois microfones. No entanto, o modelo pode ser usado para um número maior de canais já que, para os cálculos a serem efetuados para o alinhamento, serão tomados os sinais dois a dois. O mais próximo à fonte de som é tomado como referência, e os outros são denominados canais secundários.

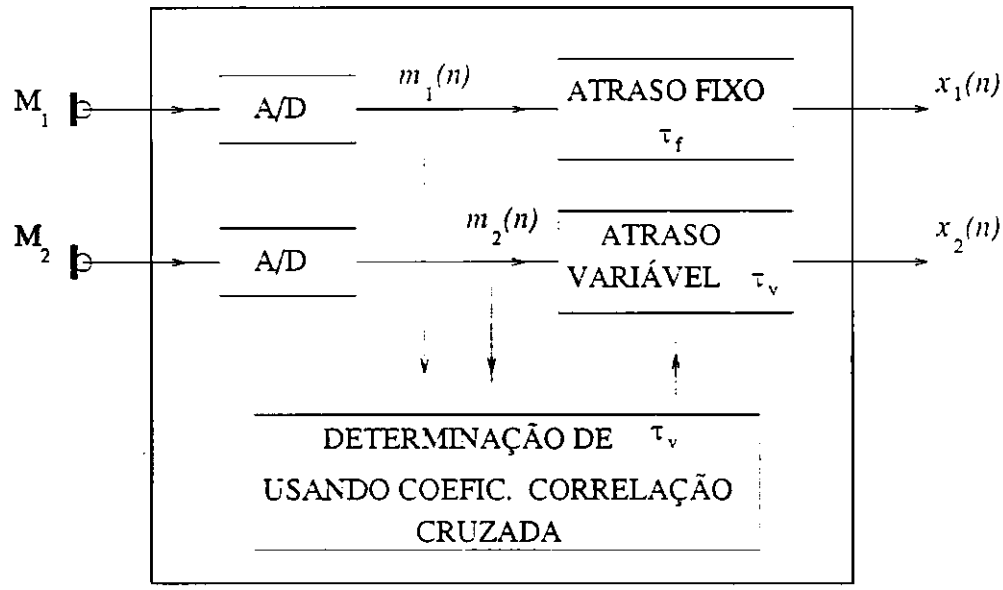


Figura 4.3: Alinhamento de fase entre os sinais de dois microfones M_1 e M_2.

Um atraso fixo, τ_f, é dado ao sinal do microfone 1 e um atraso variável, τ_v, no sinal do i-ésimo microfone, i = 2, ..., 4, é dado por:

$$\tau_v = \tau_f - \bar{\tau}_0 \tag{4.5}$$

onde $\bar{\tau}_0$ é uma estimação para τ_0 dado pela Equação (4.6), que representa o instante de tempo em que ocorre o valor máximo de correlação entre os sinais nos microfones.

$$\tau_0 = \frac{l_i - l_1}{v} \tag{4.6}$$

Os parâmetros l_1, l_i e v, são a distância da fonte do sinal ao microfone 1, a distância da fonte do sinal ao microfone M_i, i = 2, ..., 4 e v a velocidade do som, respectivamente.

Em [33, 3], para ajustar o atraso entre os sinais recebidos, é determinada a correlação cruzada entre os sinais e encontrado o instante de tempo em que ocorre o valor máximo de correlação. Este valor representa, então, o valor do defasamento entre os sinais.

No sistema ora proposto, o alinhamento de fase dos sinais é feito utilizando o coeficiente de correlação, ao invés da correlação cruzada como em [33, 3].

No alinhamento realizado através do cálculo do coeficiente de correlação, é estabelecido um limiar para o coeficiente em cada canal [34]. O valor do limiar é baseado no fato de que a distância entre os canais e a correlação entre os sinais nestes canais estão inversamente relacionados, ou seja, o aumento de um implica na diminuição do outro e vice-versa. Os valores dos limiares, para cada canal, são obtidos experimentalmente (ver Cap. 5). Quando o limiar é alcançado, é feito o alinhamento de fase tomando-se o ponto onde foi encontrado o limiar como o valor do atraso. Dessa forma, não há necessidade de processar todo o sinal para encontrar o atraso. O procedimento é efetuado nos passos descritos a seguir:

1. São tomados segmentos dos sinais recebidos no microfone de referência e no k-ésimo microfone onde $k = 2$, 3 e 4, e calculados os coeficientes de correlação ρ_{k1} de acordo com a Equação (4.7):

$$\rho_{k1}(\tau) = \frac{E[\mathrm{m}_k(n) \cdot \mathrm{m}_1(n - \tau_k)]}{E[\mathrm{m}_k^2(n)] \cdot E[\mathrm{m}_1^2(n)]} \tag{4.7}$$

   onde ρ_{k1} representa o coeficiente de correlação entre o sinal no k-ésimo microfone secundário e o sinal no microfone de referência m_1, tal que $\tilde{\tau}_k = \tau \ |_{\rho_{k_1} \geq \rho \text{ limiar}}$.

2. Os coeficientes calculados são comparados com limiares. Se o coeficiente encontrado for maior ou igual ao limiar, o valor de τ neste ponto é tomado como o ponto de início do sinal e assim é retirado o atraso. Caso contrário, o procedimento é repetido até que esta condição seja satisfeita.

Para o sinal do microfone 1, é dado um atraso fixo e o sinal captado pelo mesmo é tomado como referência para o cálculo dos atrasos correspondentes a cada um dos outros sinais captados pelos microfones secundários. Os limiares para os coeficientes de correlação cruzada foram encontrados experimentalmente. Para cada microfone foi atribuído uma estimativa inicial e a partir daí tentou-se chegar a um valor que desse um alinhamento de fase ótimo entre o sinal de um microfone M_k, $k = 2$, 3 e 4 e o microfone de referência M_1. A estimativa inicial, de forma intuitiva, foi dada baseado no fato de que a correlação entre sinais recebidos por dois microfones, espaçados a uma

determinada distância, diminui com o aumento da distância entre os mesmos e com o aumento da distância entre os mesmos e a fonte de som [33]. Dessa forma, o sinal no microfone M_2, terá um limiar mais alto, decrescendo para M_3 e M_4, respectivamente.

A Figura 4.4, dada a seguir, mostra a unidade de alinhamento para um arranjo de quatro microfones.

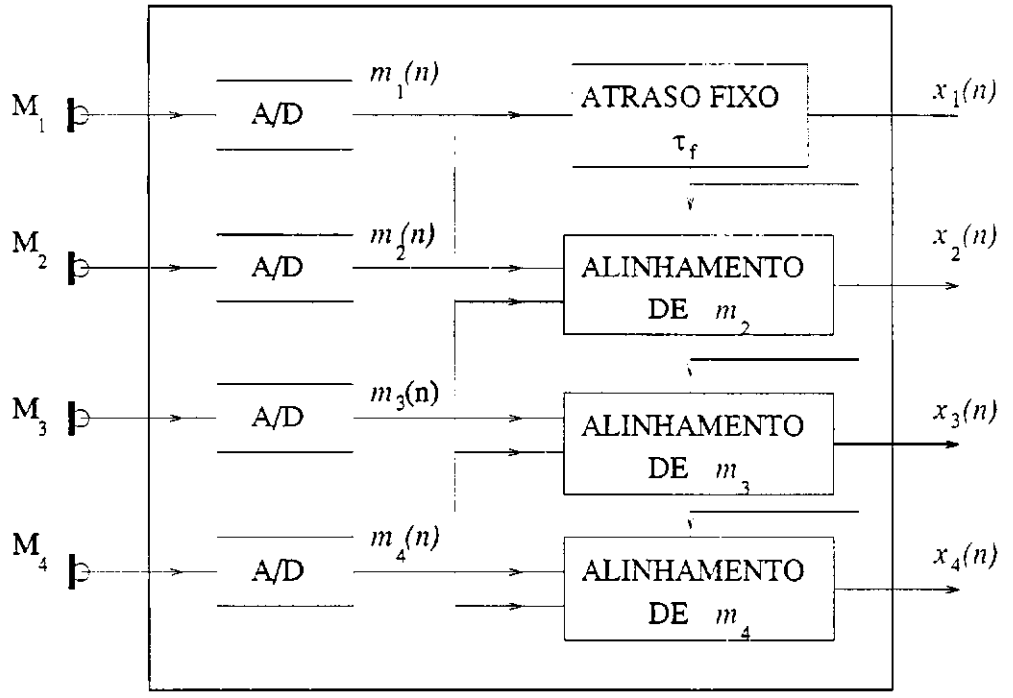


Figura 4.4: Unidade de alinhamento de fase.

Depois dos sinais estarem alinhados, é aplicada a Equação 4.3, obtendo-se, dessa forma, uma redução das componentes de ruído. O sinal médio obtido na saída da unidade de alinhamento, a partir da média entre os sinais já alinhados, é a entrada para o filtro Wiener e representa uma versão melhorada dos sinais degradados da entrada do sistema.

Alinhamento de fase com soma e ajuste do atraso entre os sinais, tem a principal vantagem de ser eficiente em simplicidade bem como no baixo custo em termos de requisitos de *hardware* para processamento digital de sinais. O método realiza uma supressão quase uniforme em todas as direções, sem adaptação à situação de ruído

específica e sem exigir estacionariedade do ruído. A atenuação de ruído máxima é 10 log (N + 1) dB, que é obtido quando a contribuição de ruído para todos os canais é igual e descorrrelacionada [3]. Assim, teoricamente, para um arranjo de quatro microfones (N = 4), teríamos uma atenuação máxima do ruído de cerca de 6 dB, aproximadamente.

Em [13], é apresentado um sistema de supressão de ruído que utiliza um arranjo de quatro microfones, sendo que o ajuste automático dos sinais não foi realizado. Os atrasos foram colocados como valores fixos correspondentes à direção do locutor desejado.

No sistema proposto, após os sinais estarem alinhados em fase, é encontrado o sinal médio resultante entre estes sinais (Eq. 4.3). O sinal resultante, como já foi dito, é um sinal melhorado, que é usado como entrada para o filtro Wiener-Kolmogoroff, na etapa de pós–filtragem, descrita a seguir.

4.4 Unidade de pós–filtragem adaptativa

Numa primeira etapa do sistema de supressão de ruído, o efeito das componentes de ruído r_i é reduzido pelo cálculo do sinal médio resultante entre os sinais alinhados em fase (Eq. 4.3).

No segundo passo, o ruído residual é diminuído adicionalmente pela pós-filtragem de x_s , produzindo o sinal de voz estimado $\hat{s}$. A adaptação do esquema de pós-filtragem é baseado no fato de que a correlação entre dois sinais recebidos por microfones numa sala reverberante diminui com o aumento da distância entre os microfones e a fonte de som, diminuindo também com o aumento da distância entre os microfones adjacentes. Conseqüentemente, para discriminar entre o sinal desejado $s(n)$ e as componentes de ruído interferentes $r_i(n)$, é necessário que o locutor desejado esteja relativamente próximo ao arranjo de microfones (o som direto domine), as fontes de ruído estejam mais distantes do arranjo (os sons refletidos dominem) e a distância entre os microfones adjacentes não seja muito pequena. Nestas suposições, as componentes de ruído recebidas podem ser consideradas como sendo mutuamente descorrelacionadas e o sistema completo pode

ser adaptado automaticamente. A adaptação é baseada nas estimativas a curto intervalo de tempo da autocorrelação e da correlação cruzada dos sinais dos microfones x_i, ..., x_4, que são avaliadas no domínio da freqüência, supondo-se que os sinais são descritos por um processo ergódico.

4.4.1 Determinação dos coeficientes do filtro Wiener-Kolmogoroff

O filtro Wiener adaptativo, implementado no domínio do tempo, é calculado a partir das medidas de correlação estimadas. Os coeficientes do filtro são obtidos a partir da Equação de Wiener–Hopf [19, 31].

Um método similar para redução de ruído foi investigado inicialmente em [33], usando um arranjo de dois microfones. Este trabalho apresenta uma variante do método proposto em [13], onde se utiliza um arranjo de quatro microfones para o cálculo das medidas de correlação cruzada e, adicionalmente, é levado a efeito um pós–processamento dessas medidas para reduzir o ruído residual. A unidade de alinhamento de fase proposta realiza o alinhamento automático dos sinais ao invés de colocar atrasos fixos de acordo com a direção do locutor como em [13].

Para determinação do filtro Wiener ótimo, consideremos a sequência no domínio do tempo:

$$x(n) = s(n) + r(n) \tag{4.8}$$

onde $s(n)$ é o sinal de voz, e $r(n)$ um sinal de ruído aditivo, descorrelacionado e estatisticamente independente de $s(n)$.

O filtro Wiener com coeficientes $h(j)$, definido na faixa do índice I := $0 \leq j \leq$ J, onde J é um numero inteiro positivo, produz o sinal estimado $\hat{s}(n)$, tal que

$$\hat{s}(n) = \sum_{j \in I} h(j) \cdot x(n-j) \tag{4.9}$$

A minimização do erro médio quadrático $E[(s(n) - \hat{s}(n))^2]$ leva à equação bem conhecida como Wiener–Hopf (ver Cap.3), que pode ser formulada aqui dos sinais $x(n)$ e $s(n)$, respectivamente como:

$$\sum_{j \in I} \mathrm{h}(j)\mathrm{R}_{xx}(i-j) = \mathrm{R}_{ss}(i), \quad i \in I \tag{4.10}$$

Para aplicação da Equação (4.10) no esquema de pós-filtragem da Figura 4.2, as funções $R_{xx}(\cdot)$ e $R_{ss}(\cdot)$ têm que ser estimadas dos sinais observados dos microfones $x_i(n); i = 1, \ldots, 4$. A função de autocorrelação do sinal degradado $x(n)$, $R_{xx}(\cdot)$, pode ser estimada diretamente de cada um dos sinais dos quatro microfones $x_i(n)$. A função de autocorrelação do sinal original, $R_{ss}(\cdot)$, pode ser estimada da correlação cruzada dos sinais dos dois microfones $x_i(n)$ e $x_j(n)$ se as componentes do ruído $r_1(n) \ldots r_4(n)$ forem mutuamente descorrelacionadas e independentes de $s(n)$. Assim,

$$\begin{aligned} E[x_i(n) \cdot x_j(n+m)] &= E[(s(n) + r_i(n)) \cdot (s(n+m) + r_i(n+m))] \\ &= R_{ss}(m) \quad \text{para i = j} \end{aligned} \tag{4.11}$$

4.4.2 Estimação das funções de autocorrelação

Os cálculos convolucionais para estimação de $R_{xx}(\cdot)$ e $R_{ss}(\cdot)$ são levados ao domínio da freqüência usando a transformada rápida de Fourier (FFT) com comprimento de bloco L. Para evitar "aliasing" no cálculo da autocorrelação [22], cada bloco de $L/2$ amostras consecutivas $\{x_i(n)\}$ é acrescida de $L/2$ amostras de valor zero e então transformada no domínio da freqüência para obter os coeficientes da FFT :

$$\{X_i(l)\}; \quad l = 0, \ldots, L-1 \quad \text{e} \quad i = 1, \ldots, 4. \tag{4.12}$$

De $X_1(l) \ldots X_4(l)$ é calculada $A(l)$, como uma estimação da auto-densidade espectral , tomando-se a média dos módulos ao quadrado dos coeficientes de Fourier $X_i(l)$, ou seja,

$$A(l) = \frac{1}{4} \sum_{i=1}^{4} \mid X_i(l) \mid^2 \quad l = 0, \ldots, L-1 \tag{4.13}$$

Tomando-se a FFT inversa de $A(l)$ obtém-se uma função no domínio do tempo que é uma estimação da função de autocorrelação $R_{xx}(\cdot)$.

A partir da média entre os produtos cruzados das FFT's dos sinais alinhados, obtém-se a estimação da densidade espectral cruzada entre os sinais nas k entradas:

$$C(l) = \frac{1}{6}\sum_{i=1}^{3}\sum_{j=i+1}^{4} X_i(l)\cdot X_j^*(l) \quad l = 0,\ldots,L-1 \tag{4.14}$$

onde $*$ denota o valor complexo conjugado de $X_j(l)$.

Os coeficientes $h(j)$ do filtro Wiener são, então, computados de acordo com a Equação 4.10, e o sinal $\hat{s}(n)$ é uma estimativa do sinal $x_s(n)$ através do filtro Wiener (Figura 4.5).

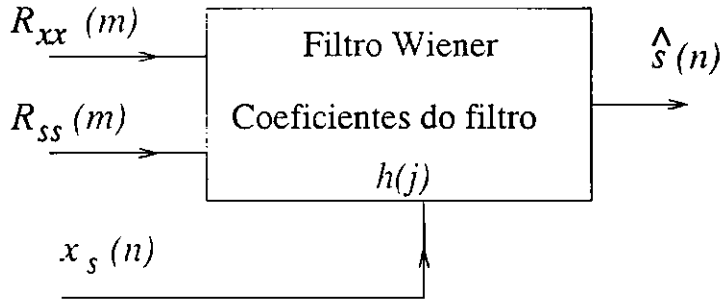


Figura 4.5: Diagrama de blocos do filtro Wiener-Kolmogoroff

A unidade de pós–filtragem adaptativa deve acompanhar as características estatísticas variantes no tempo do sinal de voz desejado. Dessa forma, o comprimento do bloco de amostras, utilizado para o processamento, é restringido a um valor no qual os sinais podem ser considerados estacionários. Assim, faz-se necessário a utilização de um segmento de comprimento L, tal que 16 ms $\leq$ L $\leq$ 32 ms, que mantenha a estacionariedade do sinal de voz [1]. Devido a esta restrição, a estimação da densidade espectral cruzada, $C(l)$, contém um erro de estimação [13] que provoca um ruído residual audível no sinal de saída $\hat{s}$. Este ruído residual pode ser reduzido através da implementação de um algoritmo de pós–processamento das medidas de correlação cruzada efetuadas pelo sistema, descrito no tópico a seguir.

A transformada inversa de $A(l)$ (valor real) leva a uma estimativa da função de autocorrelação do sinal de voz degradado $x(n)$, $R_{xx}(m)$. No caso de $C(l)$, como a mesma apresenta valor complexo para os seus coeficientes, a parte imaginária e os valores negativos serão considerados como contribuição do ruído residual. Assim, uma unidade de pós-processamento que explore as características estatísticas da Densidade Espectral de Potência [19] é adicionada ao sistema como uma forma de reduzir o erro de estimação em $C(l)$ ou, em outras palavras, o ruído residual.

Após a unidade de pós–processamento do sistema de supressão de ruído, é tomada a transformada inversa da estimação pós–processada de $C(l)$, obtendo-se assim uma estimação da função de autocorrelação do sinal $s(n)$.

As funções de autocorrelação do sinal degradado e do sinal de voz são estimadas de acordo com o procedimento acima, e servirão como parâmetros de entrada para o cálculo dos coeficientes do filtro Wiener-Kolmogoroff. O sinal na entrada do filtro é o sinal médio obtido na saída da unidade de alinhamento de fase.

4.5 Unidade de pós–processamento para as medidas de correlação cruzada

A unidade de pós–processamento tem como objetivo reduzir o ruído residual presente no sinal estimado, devido ao erro de estimação nas medidas de correlação cruzada. O pós–processamento é feito utilizando-se as propriedades das funções de autocorrelação, correlação cruzada e da densidade espectral de potência [19].

4.5.1 Estimação da Densidade Espectral de Potência - DEP do sinal de voz

No processo de estimação da DEP para o sinal de voz, é levado a efeito um pós-processamento para redução do ruído de estimação, baseado nas características estatísticas esperadas para a DEP [19]. Pelo uso da transformada inversa discreta

de Fourier da estimação da DEP é obtida, portanto, uma estimação da função de autocorrelação para o sinal de voz a ser utilizada para o cálculo dos coeficientes ótimos.

Podemos assumir que a estimação da densidade espectral cruzada $X_i(l) \cdot X_j^*(l)$ em $C(l)$, na Equação (4.14), pode ser modelada da seguinte forma:

$$X_i(l) \cdot X_j^*(l) = S(l) + N_{ij}(l) \tag{4.15}$$

onde $S(l)$ é a auto-densidade espectral do sinal de voz s(n) (de valor real) e $N_{ij}(l)$ é o erro de estimação de média zero (de valor complexo) com um ângulo de fase uniformemente distribuído sobre [0, 2π], que corresponderá à contribuição das componentes residuais do ruído. O erro de estimação $N_{ij}(l)$ é independente de $S(l)$.

As variâncias das partes real e imaginária de $N_{ij}(l)$ são definidas com

$$E\{Re^2[N_{ij}(l)]\} = E\{Im^2[N_{ij}(l)]\} = V(l) \tag{4.16}$$

Da Equação (4.15), o termo $N_{ij}(l)$, seria nulo se o ruído entre os canais fossem mutuamente descorrelacionados. Na ausência do sinal de voz s(m), o termo $N_{ij}(l)$ é idêntico a densidade espectral cruzada dos dois sinais de ruído $r_i(n)$ e $r_j(n)$. A variância $V(l)$, na Equação (4.16) representa, portanto, a potência do ruído recebida.

A suposição do modelo da Equação (4.15), é a ferramenta básica para o desenvolvimento do algoritmo de pós–processamento. A aplicabilidade deste algoritmo é confirmada pelo melhoramento no desempenho do sistema, realizado como se segue.

4.5.2 Redução do erro do ruído residual na estimação da DEP do sinal de voz

O efeito do erro de estimação na densidade espectral cruzada $C(l)$, $N_{ij}(l)$, é reduzido em duas etapas, a saber:

Etapa 1: Estimação modificada de $C(l)$

Para que $C(l)$ represente uma estimação de uma auto-densidade espectral, a mesma deve ser de valor real e a sua função de autocorrelação correspondente no domínio do tempo, $R_{ss}(m)$, tem que ser simétrica [19, 20]. Assim, podemos fazer uso de uma estimação modificada $Cm(l)$, onde:

$$Re[Cm(l)] = Re[C(l)] \quad e \quad Im[Cm(l)] = 0 \tag{4.17}$$

Como vimos da Equação (4.16), as variâncias das partes real e imaginária de $N_{ij}(l)$ são iguais. Assim, comparando-se com a estimação $C(l)$, a variância do erro de estimação em $Cm(l)$ é reduzida à metade, já que somente os valores reais de $N_{ij}(l)$ têm efeito sobre $Cm(l)$.

Etapa 2: Determinação de um fator de redução de ruído

Nesta etapa é determinado um fator de redução de ruído para obtenção de uma estimação pós–processada melhorada de $C(l)$.

Cada um dos coeficientes da densidade espectral $S(l), l = 0, \ldots, L-1$ podem ser estimados a partir de seis medidas básicas da densidade espectral cruzada que, de acordo com as Equações (4.15) e (4.17), podem ser escritas da seguinte forma:

$$Cm_{ij}(l) = Re\{X_i(l) \cdot X_j^*(l)\} = S(l) + Re\{N_{ij}(l)\} \tag{4.18}$$

onde ij é um elemento da quantidade de pares de índices IP := {12, 13, 14, 23, 24, 34}.

A estimação de $C(l)$ na Equação (4.14) é trocada pela estimação pós–processada $P(l)$, dada por:

$$P(l) = \alpha(l)\frac{1}{6}\sum_{ij \in IP} Cm_{ij}(l) \tag{4.19}$$

onde $\alpha(l)$ representa um fator de redução dependente da freqüência l

. $P(l)$ será o valor estimado para $S(l)$. O fator de redução $\alpha(l)$ é determinado pela

minimização do erro de estimação médio quadrático E[$(S(l)$ - $P(l))^2$]. Considerando-se, sem perda de generalidade, que as seis componentes $\{N_{ij}(l); ij \in IP\}$ são mutuamente descorrelacionadas e independentes de $S(l)$, o fator de redução dependente da freqüência, $\alpha(l)$, será dado por:

$$\alpha(l) = \frac{S^2(l)}{S^2(l) + \frac{1}{6}V(l)} \tag{4.20}$$

Pode-se observar na Equação (4.20), que quanto maior for a variância do erro de estimação $V(l)$, menor o fator de redução $\alpha(l)$.

Para que a Equação (4.20) possa ser avaliada, faz-se necessária a estimação de $S^2(l)$ e de $V(l)$. A estimação de ambos os termos é feita a partir dos seis valores medidos observados $\{Cm_{ij}(l); ij \in \text{IP}\}$, usando as propriedades da densidade espectral potência, como mostrado a seguir.

- Estimação de $S^2(l)$

  Como a densidade espectral $S(l)$ tem que ser não-negativa, os valores negativos de $Cm(l)$ são substituídos por zero. Os termos resultantes são elevados ao quadrado. Os valores resultantes são, então, tomados como estimação para $S^2(l)$ na Equação 4.20.

- Estimação de $V(l)$

Já que $S(l)$ é não-negativa, um valor negativo em $Cm_{ij}(l)$ deve ser causado pelo efeito do erro de estimação $N_{ij}(l)$ (Eq. 4.18). Conseqüentemente, os valores negativos observados em $\{Cm_{ij}(l)$; ij $\in$ IP$\}$ podem ser usados para estimação de $V(l)$.

Seja a quantidade de valores negativos observados denotada por M (M $\leq$ 6). É calculada a média:

$$Vm(l) = \frac{1}{M}\sum_{ij} Cm_{ij}^2(l) \tag{4.21}$$

para aqueles ij onde ij $\in$ IP e $Cm_{ij}(l) \leq 0$.

Assim, $Vm(l)$ é usada como a estimação de $V(l)$ na Equação (4.20).

A Equação (4.21) fornece uma sub-estimação da variância $V(l)$ se $S(l)$ não for igual a zero. No entanto, este comportamento não é crítico já que não há necessidade de reduzir a estimação da densidade espectral P(l) nas regiões de freqüência onde $S^2(l)$ é essencialmente maior que $V(l)$. Por outro lado, este procedimento de avaliar apenas os termos negativos de $Cm_{ij}(l)$ apresenta a vantagem de ser robusto contra pequenos desajustes da unidade de alinhamento. Isto é possível porque para pequenos desajustes na unidade de alinhamento, a componente de voz na densidade espectral cruzada é de valor complexo ao invés de valor real. No entanto, sua parte real será fortemente negativa. Dessa forma, uma sobre-estimação da variância $V(l)$ é evitada nestes casos [13].

Foi descrito acima o procedimento de pós-processamento das medidas de correlação cruzada para o sistema de supressão de ruído apresentado. Este procedimento reduz particularmente o ruído residual nas regiões de freqüência entre formantes. Assim, o sinal de voz na saída soa livre de ruído mesmo para altos níveis de ruído na entrada do sistema. $P(l)$ representa, então, uma estimação pós-processada de $C(l)$. Para obter-se a estimação da função de autocorrelação $R_{ss}(l)$, a ser usada no cálculo dos coeficientes do filtro Wiener, é tomada a transformada inversa de Fourier de $P(l)$. As FFT's inversas de $A(l)$ e $P(l)$, representando as estimações de $R_{xx}(l)$ e $R_{ss}(l)$ são então usadas como entrada no filtro Wiener-Kolmogoroff para obtenção dos coeficientes do filtro no domínio do tempo.

O pós-processamento permite, ainda, uma redução de ruído eficiente também na região de baixa freqüência. Quanto mais baixa a freqüência for, mais os dois sinais de ruído recebidos $r_i(n)$ e $r_j(n)$ parecem-se um com o outro, isto é, mais eles parecem estar correlacionados. Este efeito faz com que haja uma diminuição no desempenho de redução do ruído do filtro Wiener nas baixas freqüências. Entretanto, este efeito pode ser parcialmente compensado aumentando-se os valores de estimação da variância $Vm(l)$ na região de baixa freqüência. Assim, poderia ser usado um fator de acréscimo dependente da freqüência (que também dependa das dimensões do arranjo de microfones). Este fator seria então determinado das medidas estatísticas dos sinais

de ruído [13].

O sistema de redução de ruído é robusto contra alguma correlação residual entre os sinais de ruído captados também em outras regiões de freqüência. Esta propriedade é obtida através da estimação de V(l) de acordo com a Equação (4.21), já que é altamente provável que, mesmo para sinais de ruído (suavemente) correlacionados, no mínimo um dos seis termos de $\{Cm(l)\}$ seja negativo. A presença desse valor negativo indica um sinal de ruído naquela região de freqüência, que pode ser atenuado aplicando-se o algoritmo de pós-processamento.

Uma outra forma de calcular os coeficientes do filtro Wiener pode ser realizada através de cálculos efetuados no domínio da freqüência, o que pode diminuir o tempo de processamento do sistema. Esta forma será vista a seguir.

4.6 Cálculo dos coeficientes do filtro Wiener-Kolmogoroff no domínio da freqüência

No Capítulo 2, vimos a Equação de Wiener-Hopf para o cálculo dos coeficientes do filtro Wiener-Kolmogoroff (Eq. 3.17). A partir desta equação, chegamos a equação no domínio da freqüência para o cálculo dos coeficientes (Eq. 3.20), reescrita aqui da seguinte forma:

$$H\text{opt}(w) = \frac{S_{xs}(w)}{S_{xx}(w)} \tag{4.22}$$

onde $S_{xs}(w)$ é a Densidade Espectral de Potência (DEP) Cruzada entre $x(n)$ e $s(n)$, e $S_{xx}(w)$ é a DEP de $x(n)$ [20, 5].

Para um sinal degradado na forma $x(n) = s(n) + r(n)$ e considerando-se que $s(n)$ é descorrelacionado de $r(n)$, pode-se provar [20] que:

$$S_{xs}(w) = S_{ss}(w) \qquad \text{e} \tag{4.23}$$

$$S_{xx}(w) = S_{ss}(w) + S_{rr}(w) \tag{4.24}$$

Substituindo-se as Equações (4.23) e (4.24) na Equação (4.22), obtemos a seguinte expressão para os coeficientes do filtro Wiener no domínio da freqüência [5]:

$$H\text{opt}(w) = \frac{S_{ss}(w)}{S_{ss}(w) + S_{rr}(w)} \tag{4.25}$$

$$H\text{opt}(w) = \frac{S_{xx}(w) - S_{rr}(w)}{S_{xx}(w)} \tag{4.26}$$

Considerando-se os intervalos onde $S_{rr}(w) < S_{xx}(w)$ e $S_{nn}(w) > S_{xx}(w)$, obtém-se o algoritmo de supressão de ruídos dado por:

$$H\text{opt}(w) = \begin{cases} 1 - \frac{S_{rr}(w)}{S_{xx}(w)} & \text{para} \quad S_{rr}(w) < S_{xx}(w) \\ 0 & \text{para} \quad S_{rr}(w) \geq S_{xx}(w) \end{cases} \tag{4.27}$$

A função de transferência $H\text{opt}(w)$ dada pela Equação (4.27), mostra que pode-se obter a supressão de ruído, sem que seja conhecida ou estimada a estatística do sinal de voz, embora seja necessário o conhecimento da DEP do sinal degradado e do ruído.

Para a determinação da DEP do sinal degradado é usada a estimação da densidade espectral do sinal de voz, $A(l)$, dada pela Equação (4.13). Já para a estimação da DEP do ruído, é usada a variância $Vm(l)$ da Equação (4.21). Assim, tem-se que:

$$H\text{opt}(l) = \begin{cases} 1 - \frac{Vm(l)}{A(l)} & \text{para } Vm(l) < A(l) \\ 0 & \text{para } Vm(l) \geq A(l) \end{cases} \tag{4.28}$$

O desempenho do sistema para o cálculo dos coeficientes do filtro no domínio do tempo e no domínio da freqüência pode ser avaliado a partir da relação sinal-ruído obtida na saída do sistema. Além disso, uma avaliação subjetiva se faz necessária, através de testes de escuta para que o próprio usuário escolha o sistema que melhor se adeque às suas exigências. O melhoramento da relação sinal-ruído em voz está intrinsicamente relacionado ao melhoramento da inteligibilidade da voz. Entretanto, não há relação quantitativa entre os mesmos [7].

Os resultados obtidos pelo sistema de supressão de ruído proposto serão mostrados e avaliados no capítulo a seguir.

Capítulo 5

Avaliações e resultados experimentais

5.1 Introdução

O Sistema Multicanal Adaptativo de Supressão de ruído implementado foi avaliado de acordo com o ganho na relação sinal-ruído obtido na saída do sistema, em comparação a relação sinal-ruído dos sinais degradados na entrada do sistema. O Apêndice A mostra as formulações para os cálculos da SNR e SNR segmental, bem como do ganho obtido. Foram feitas, ainda, avaliações subjetivas através de testes de escutas informais. As condições e os resultados experimentais são comentados a seguir.

5.2 Condições de recepção de som e obtenção dos sinais

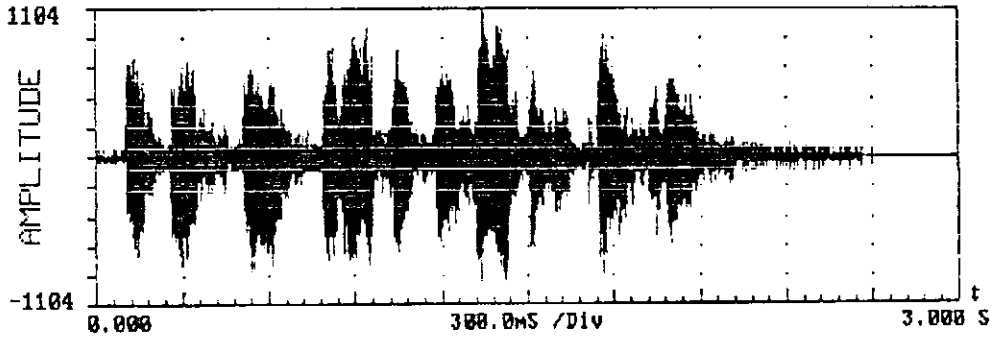
Para a recepção do som, foi fixado o arranjo de microfones num lugar determinado numa sala medindo aproximadamente $15m^2$ de área. O campo de ruído acústico é gerado por ruído de automóvel obtido a partir de uma gravação em fita k7 no interior de um automóvel com uma velocidade variando de 0 a 80 Km/h. Como sinal de teste, para o sinal de voz, foi utilizada uma frase falada por um interlocutor feminino:

"A questão será retomada no congresso". Esta frase foi escolhida de um grupo de frases foneticamente balanceado [35]. Esta escolha, em particular, deve-se à riqueza de fonemas diferentes. Na realidade, a frase escolhida não afeta o desempenho do sistema. No entanto, com esta frase, o algoritmo pode ser testado na presença de sons surdos onde há ocorrência de erros ou cancelamento do sinal [11].

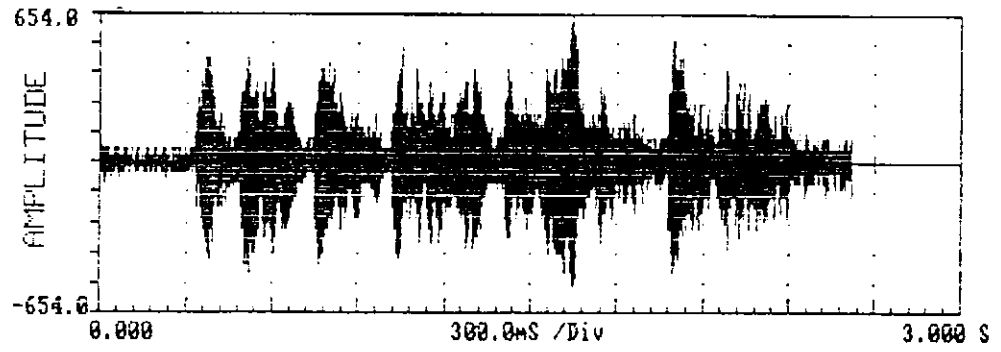
Foi montado o arranjo bidimensional, com quatro microfones, formando um quadrado medindo 0,6 m x 0,6 m. A fonte de som foi colocada distante 0,6 m do arranjo. A fonte de ruído foi colocada a cerca de 1,5 m do arranjo de forma a contribuir com um som refletido, enquanto o sinal de voz é recebido de forma direta pelo arranjo.

Assim, o microfone mais próximo da fonte de som.contendo o sinal de voz, é considerado como microfone ou canal de referência (microfone 1 ou canal 1). Os outros são considerados canais secundários.

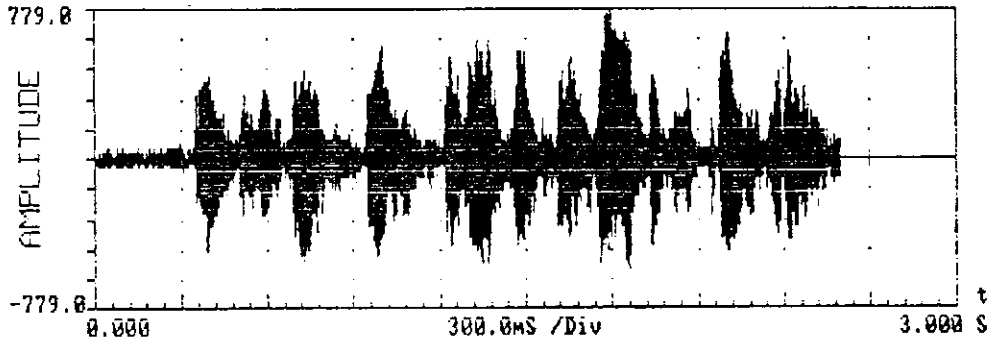
A Figura 5.1 mostra a forma de onda dos sinais captados pelos microfones. A Figura 5.1(a) mostra a forma de onda do sinal captado pelo canal 1, sendo este o sinal menos degradado pelo ruído por estar mais próximo à fonte de som do sinal de voz e mais distante da fonte de ruído. As Figuras 5.1(b), 5.1(c) e 5.1(d) dadas a seguir, mostram as formas de onda dos sinais captados pelos canais 2, 3 e 4, respectivamente. Sendo que o canal 4 apresenta o sinal mais degradado e mais atrasado, em relação ao canal de referência, captado pelo arranjo, por sua proximidade maior da fonte de ruído e maior distância da fonte de som.



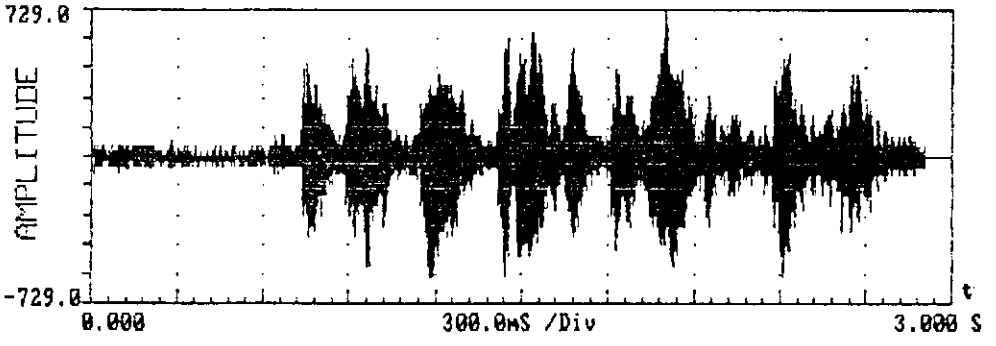
(a)



(b)



(c)



(d)

Figura 5.1: Sinais recebidos pelos microfones na entrada do sistema: (a) sinal captado pelo canal 1 ; (b) sinal captado pelo canal 2; (c) sinal captado pelo canal 3 e (d) sinal captado pelo canal 4.

Como era de se esperar, pode-se notar que o atraso de cada sinal é maior tanto quanto maior for a distância do microfone à fonte de som. Já a SNR diminui com o aumento da distância do microfone à fonte de som e com a proximidade maior da fonte de ruído.

O sinal no microfone 4 chega a ter uma SNR de aproximadamente -0,87 dB.

Os sinais foram gravados e digitalizados a uma freqüência de amostragem de 8 kHz, utilizando uma placa *Sound Blaster* de 8 bits, instalada num computador tipo IBM/PC.

Com a fonte de ruído presente, o sinal de voz é captado pelo arranjo, e cada microfone recebe o sinal degradado pelo ruído, com um fator de atraso e amplitudes diferentes. Os sinais entram na unidade de alinhamento de fase para ser retirado o atraso e só então são processados para a supressão do ruído.

A Figura 5.2 mostra o espectro do ruído de automóvel, utilizado como fonte de ruído para degradação dos sinais.

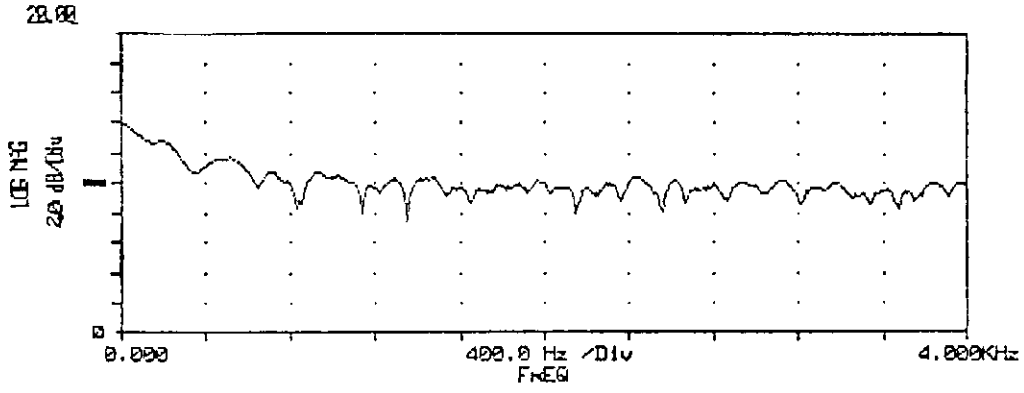


Figura 5.2: Espectro do ruído de automóvel.

5.3 Implementação da Unidade de Alinhamento de fase

O algoritmo para a unidade de alinhamento de fase do sistema foi implementado tendo como base as medidas de correlação cruzada entre os sinais captados pelos microfones do arranjo.

No sistema proposto, o alinhamento de fase dos sinais foi feito através da utilização do coeficiente de correlação entre os sinais ao invés do valor de correlação máxima como em [33, 3]. Esta decisão foi tomada através de observações experimentais, que demonstraram uma maior rapidez nos cálculos quando do uso do coeficiente de correlação.

Foi implementado, inicialmemte, um algoritmo que encontrasse o valor máximo de correlação cruzada entre os canais 1 e 2, 1 e 3, 1 e 4. O canal 1 é tomado como canal de referência por estar mais próximo da fonte de som do sinal de voz e, portanto, o sinal menos degradado pelo ruído. Considera-se a correlação entre os sinais de ruído presentes nos canais como desprezível. Ao ser encontrado o valor máximo de correlação, tomar-se-ia este valor como o ponto de início do sinal de interesse e estaria encontrado o valor do atraso. No entanto, para isto, seria necessário processar todo o sinal, o que tornaria inviável uma futura implementação deste sistema em tempo real.

Assim, foi implementado um algoritmo para efetuar o alinhamento dos sinais baseado no coeficiente de correlação cruzada ao invés, apenas, da correlação cruzada entre os sinais (Cap. 4).

Foram feitas várias medidas para a obtenção dos limiares para os coeficientes de correlação entre os canais. A curva mostrada na Figura 5.3, dada a seguir, mostra os coeficientes de correlação medidos para várias distâncias. Pode ser visto, na Figura 5.3, que a medida que a distância entre o locutor e os microfones é aumentada, o coeficiente de correlação diminui devido ao efeito dos sons refletidos na sala [33].

A curva cheia da Figura 5.3 foi obtida por Kaneda e Tohyama [33] para sinais de voz recebidos por dois microfones. Os pontos indicam os valores dos coeficientes de correlação obtidos para os sinais de voz captados pelos quatro microfones no sistema proposto.

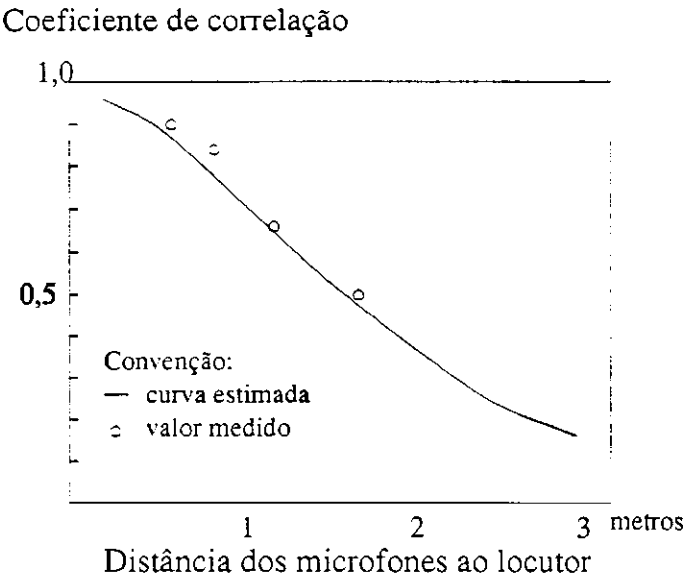


Figura 5.3: Coeficientes de correlação entre os sinais de voz recebidos pelo arranjo de quatro microfones.

Pode-se observar, a partir da Figura 5.3 que os valores obtidos como limiares para os coeficientes de correlação entre os canais 1 e 2, ρ_{12}, 1 e 3, ρ_{13}, e 1 e 4, ρ_{14} são 0,85, 0,65 e 0,5 respectivamente.

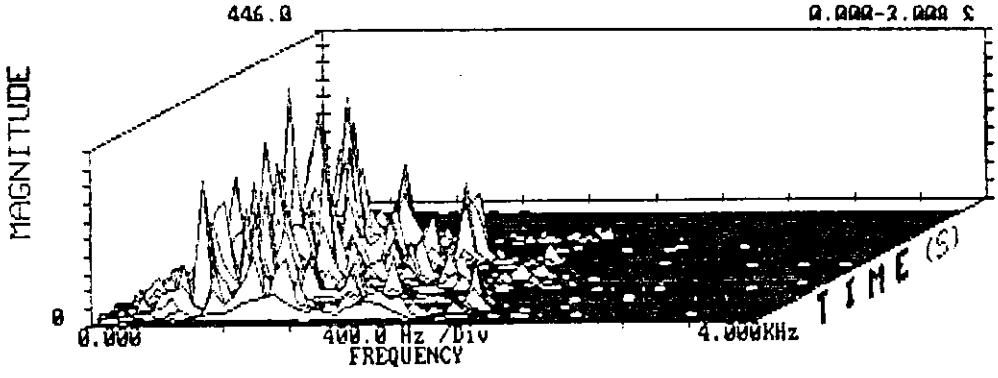
Além de buscar os valores ótimos para os limiares dos coeficientes de correlação entre os canais, é necessário encontrar qual o melhor comprimento de bloco de amostras a serem processadas para o alinhamento. No caso presente, foi utilizado um intervalo de tempo de análise da ordem de 8 ms ou 64 amostras a uma freqüência de amostragem de 8 kHz. Este comprimento contém um período de *pitch* da voz feminina (5 ms) [22, 9], utilizada como sentença de teste. Para um caso genérico pode-se alterar este comprimento para um valor maior que 10 ms a fim de conter pelo menos um período de *pitch* feminino e um masculino (10 ms) [9, 3]. O algoritmo implementado apresenta a facilidade de se poder mudar o comprimento do bloco de amostras sem a necessidade de nenhuma outra alteração.

A Tabela 5.1 dada a seguir, mostra os valores de ganho obtidos na relação sinal-ruído segmental (SegSNR-G) na saída da unidade de alinhamento.

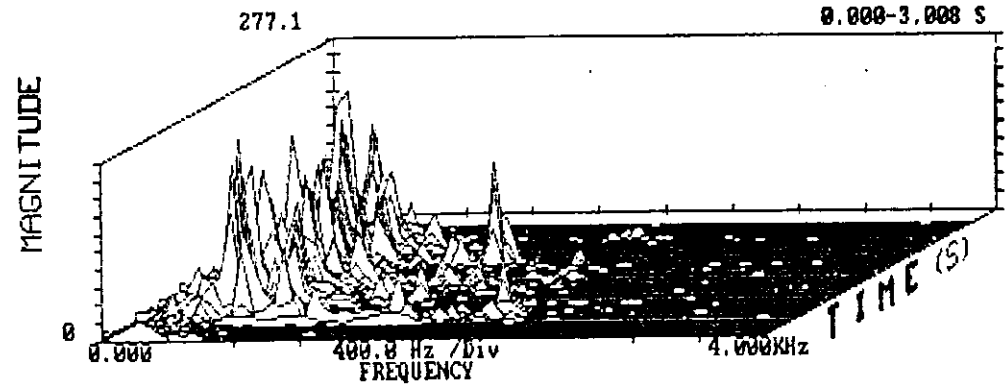
Sinal	SegSNR-G (dB)
m2	3,95
m3	2,86
m4	2,43
sinal médio	4.02

Tabela 5.1: Ganho na relação sinal-ruído segmental para os sinais na unidade de alinhamento de fase.

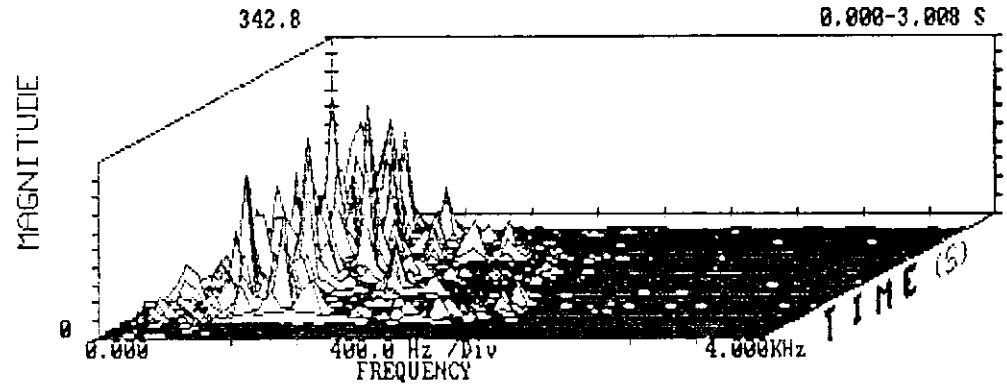
A partir do espectro dos sinais, pode-se ter uma visão das componentes de freqüência presentes em cada um dos sinais, de acordo com o grau de degradação. A Figura 5.4 mostra os espectros para cada um dos sinais de entrada captados pelo arranjo.



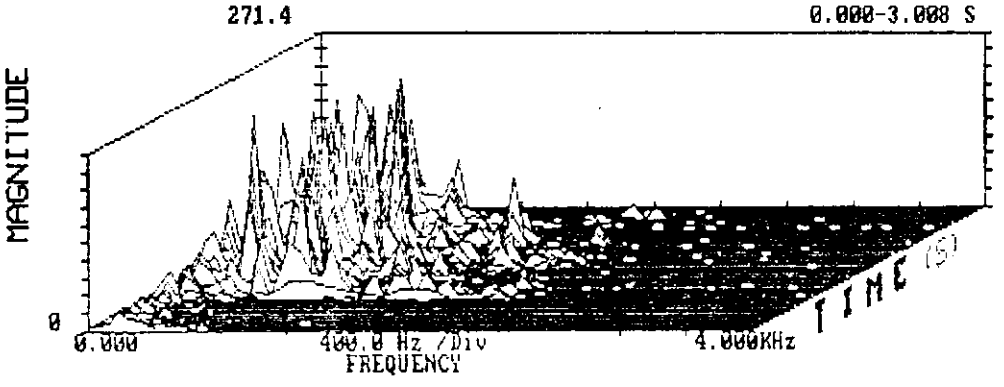
(a)



(b)



(c)

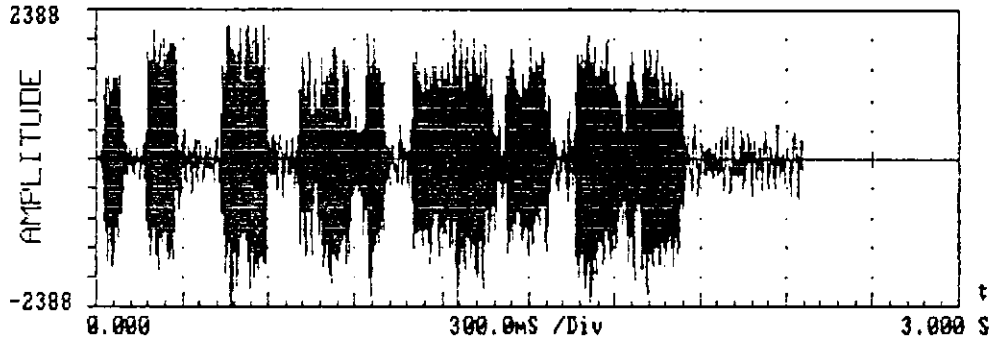


(d)

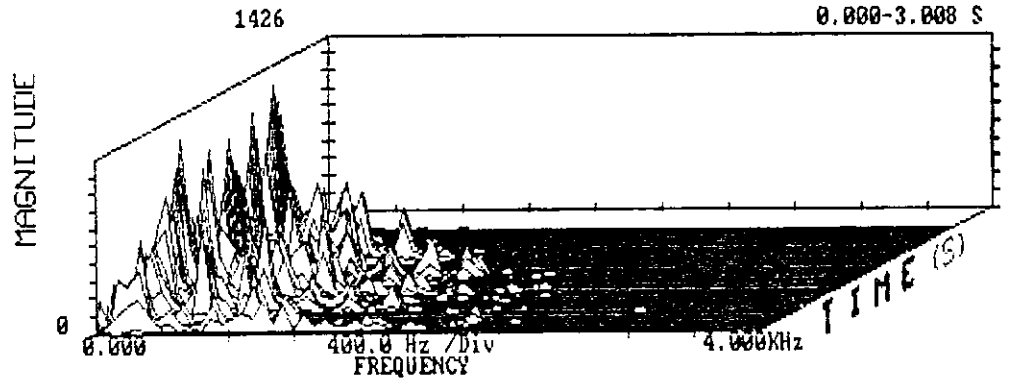
Figura 5.4: Espectro tridimensional para os sinais na entrada do sistema de supressão de ruído. (a) sinal captado pelo microfone 1; (b) sinal captado pelo microfone 2; (c) sinal captado pelo microfone 3 e (d) sinal captado pelo microfone 4.

Pode ser visto, também no domínio da freqüência, a diferença de fase e de amplitude entre os sinais captados por cada microfone, bem como a presença do ruído degradante.

Após o alinhamento dos sinais, é obtido um sinal resultante da média entre os sinais já alinhados, que representa uma versão melhorada dos sinais degradados da entrada do sistema. A Figura 5.5 mostra a forma de onda do sinal médio (Fig. 5.5(a)) e seu espectro de potência (Figura 5.5(b)).



(a)



(b)

Figura 5.5: Sinal médio obtido na saída da unidade de alinhamento de fase: (a) forma de onda ; (b) espectro de potência.

Após os sinais estarem alinhados, o sistema efetua um processamento desses sinais

no domínio da freqüência, como descrito no capítulo 4, seção 4.4.2. Daí, os sinais são submetidos a um pós-processamento e por fim, filtrados adaptativamente na unidade de pós-filtragem adaptativa.

5.4 Implementação da unidade de pós-filtragem adaptativa

A unidade de pós-filtragem adaptativa, consiste do filtro Wiener-Kolmogoroff adaptativo, com 16 coeficientes. A adaptação dos coeficientes do filtro é mostrada na Figura 5.6. O tamanho do segmento utilizado para processamento é de 16 ms ou 128 amostras ($L/2$ amostras). Entretanto, para evitar "*aliasing*" (Cap. 4) no cálculo da funções de autocorrelação são acrescentados mais $L/2$ amostras de valor zero a cada segmento. O segmento para processamento é, então, de $L = 256$ amostras.

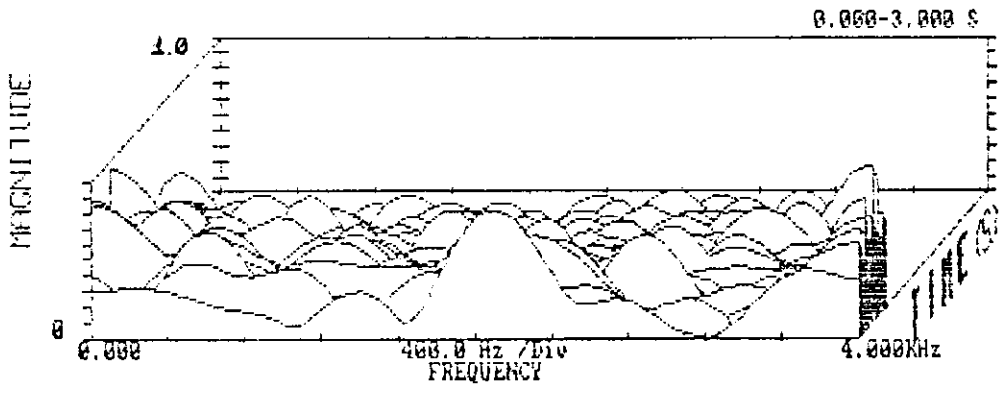


Figura 5.6: Adaptação dos coeficientes do filtro Wiener-Kolmogoroff adaptativo para cálculo no domínio do tempo.

Após a filtragem do sinal médio, foi proporcionado um ganho de aproximadamente 6 dB. A Tabela 5.2 mostra os valores de relação sinal-ruído para cada um dos sinais na

entrada do sistema e os valores obtidos para o sinal de saída. São mostrados, ainda, os valores de ganho na relação sinal ruído segmental obtidos.

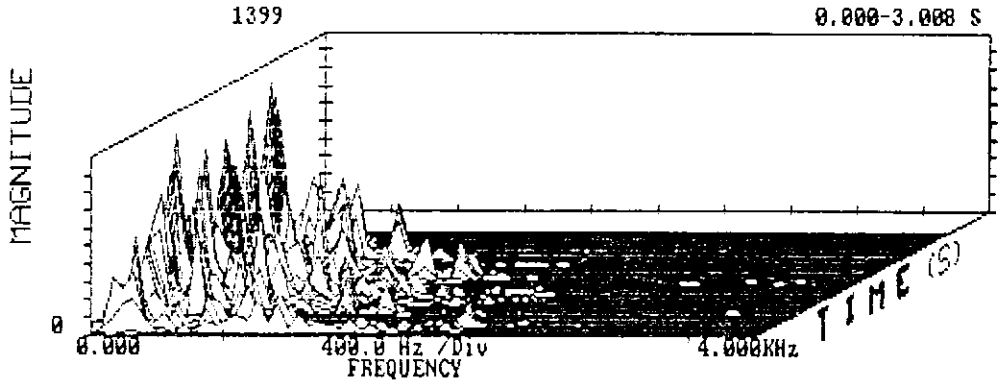
Sinal	SNR (dB)	SegSNR-G (dB)	SNR-Seg (dB)
m_1	4,51	4.51	5,72
m_2	2,91	3,34	3,41
m_3	0,48	2,86	1,98
m_4	-0.87	2.43	0.63
$\mathbf{x_s}$	6,85	4,02	7,33
$\hat{\mathbf{s}}$	13.97	6,12	15,49

Tabela 5.2: Valores de relação sinal-ruído e ganho obtidos pelo sistema de supressão usando arranjo de 4 microfones

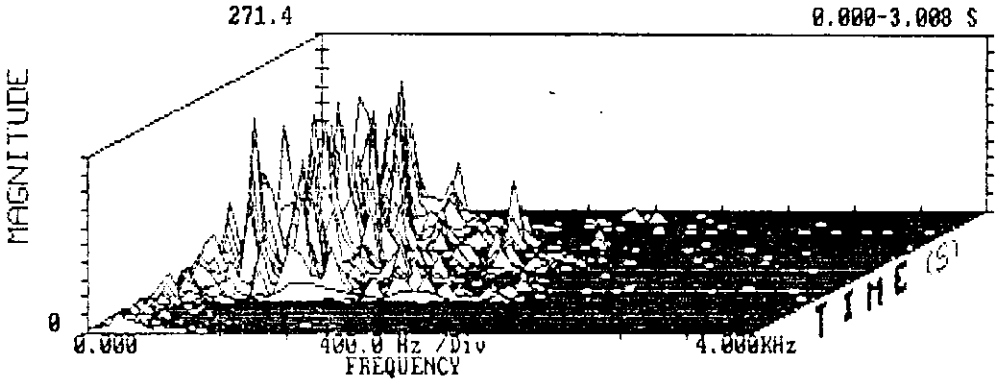
A primeira coluna da Tabela 5.2 relaciona os sinais de entrada e saída do sistema, onde m_1, m_2, m_3 e m_4, são os sinais captados pelo arranjo sem alinhamento, x_s e $\hat{s}$ representam o sinal médio obtido dos sinais alinhados e o sinal estimado na saída do filtro, respectivamente. A segunda coluna apresenta os valores de relação sinal-ruído destes sinais e a terceira e quarta colunas mostram os ganhos obtidos na SNR segmental (SegSNR-G) e na SNR (SNR-G). Todos os valores estão expressos em dB. O ganho de $\hat{s}$ é obtido em relação ao sinal na entrada do filtro, ou seja, x_s. O valor de SNR na saída do filtro é de 15,49 dB.

Para uma melhor avaliação dos sistema, podem ser vistos na Figura 5.7 os espectros do sinal original, do sinal mais degradado pelo ruído (sinal captado pelo microfone 4) e o sinal obtido na saída do filtro.

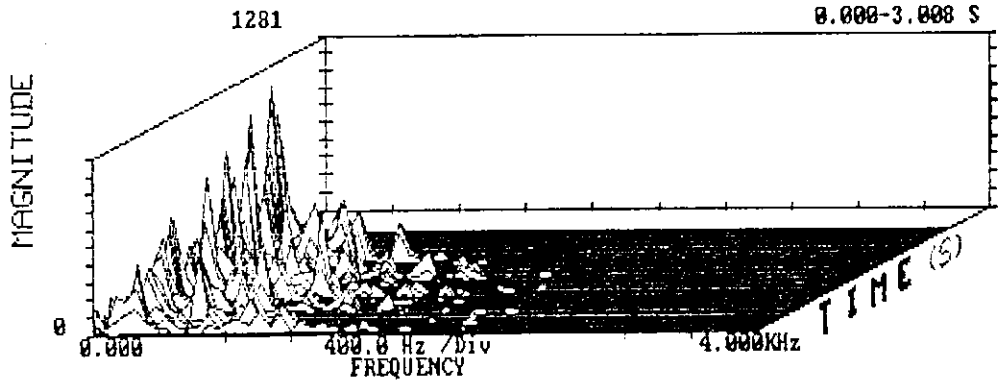
Pode-se observar que não há, neste caso, a forte presença do ruído na entrada do sistema (Figura 5.7(b)) e a sua supressão de forma significativa no sinal de saída (Figura 5.7(c)), comparados ao sinal original não-ruidoso.



(a)



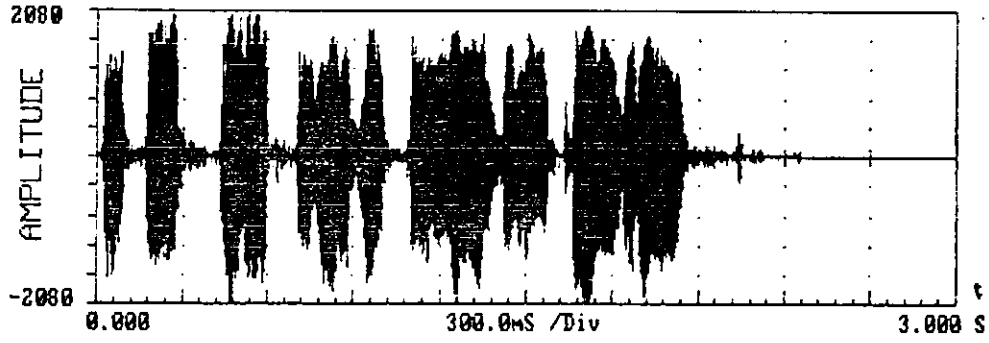
(b)



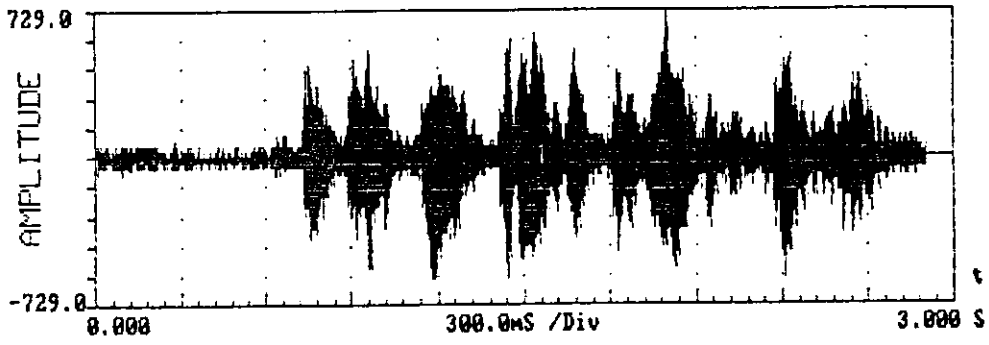
(c)

Figura 5.7: Descrição espectográfica do desempenho do sistema (a) sinal original; (b) sinal degradado na entrada do sistema e (c) sinal melhorado na saída do sistema.

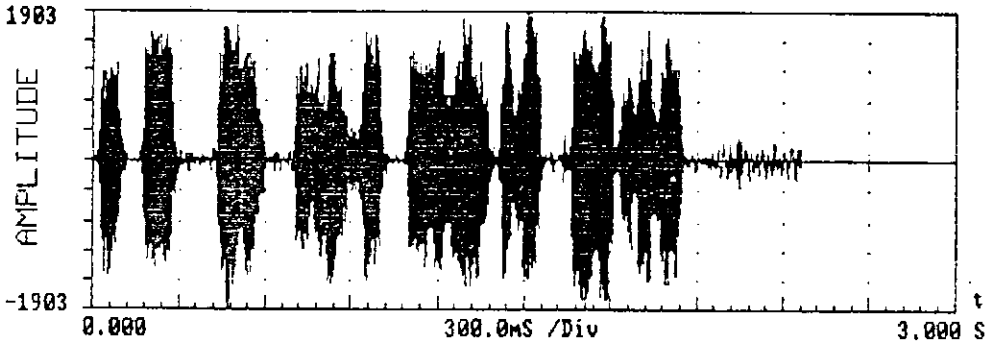
A Figura 5.8 mostra as formas de onda dos sinais da Figura 5.7. A Figura 5.8(a) mostra a forma de onda do sinal original, a Fig. 5.8(b) a forma de onda do sinal captado pelo microfone 4 e a Figura 5.8(c) a forma de onda do sinal obtido na saída do filtro. Assim, pode-se ter uma idéia do comportamento do sistema, observando-se os sinais no domínio do tempo. A Figura 5.8(b) é uma réplica da Fig. 5.1(d), repetida aqui para que se tenha uma melhor comparação do sinal estimado em relação ao sinal degradado.



(a)



(b)



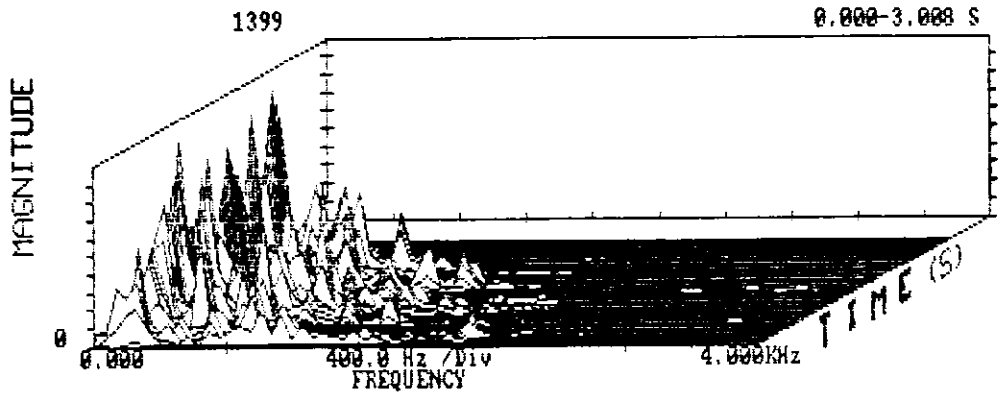
(c)

Figura 5.8: Descrição temporal dos resultados obtidos pelo sistema multicanal com arranjo de 4 microfones: (a) sinal original; (b) sinal degradado na entrada do sistema ($m_4(n)$ e (c) sinal melhorado na saída do sistema.

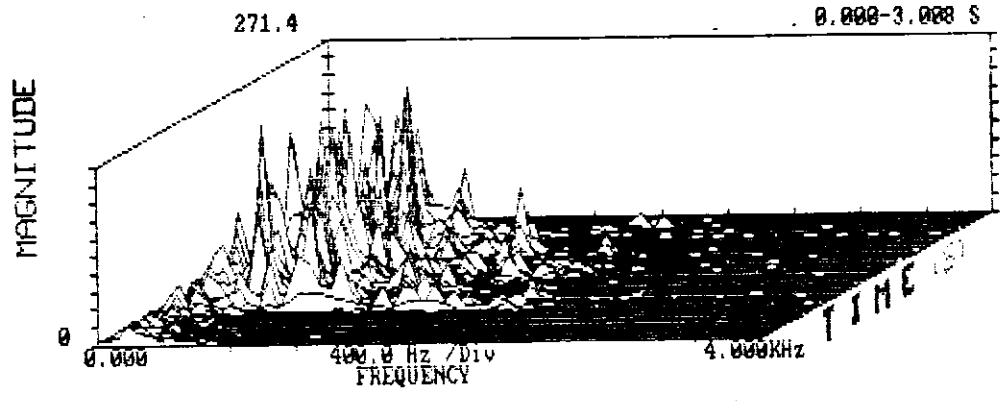
Como visto no capítulo 4, os coeficientes do filtro Wiener-Kolmogoroff podem ser calculados tanto por uma função implementada no domínio do tempo como no domínio da freqüência.

Através do cálculo dos coeficientes no domínio da freqüência elimina-se a inversão de matriz para obtenção dos mesmos, dando lugar a simples somas, divisões, subtrações e multiplicações. O algoritmo implementado no domínio da freqüência é cerca de 25% mais rápido do que aquele no domínio do tempo (observações experimentais). A Figura 5.9 dada a seguir, mostra os espectros do sinal original (Fig. 5.9(a)), do sinal mais ruidoso (m_4) na entrada do sistema (Fig. 5.9(b)) e do sinal estimado (Fig. 5.9(c)). Pode-se notar, que algumas componentes de freqüências mais altas do sinal não aparecem.

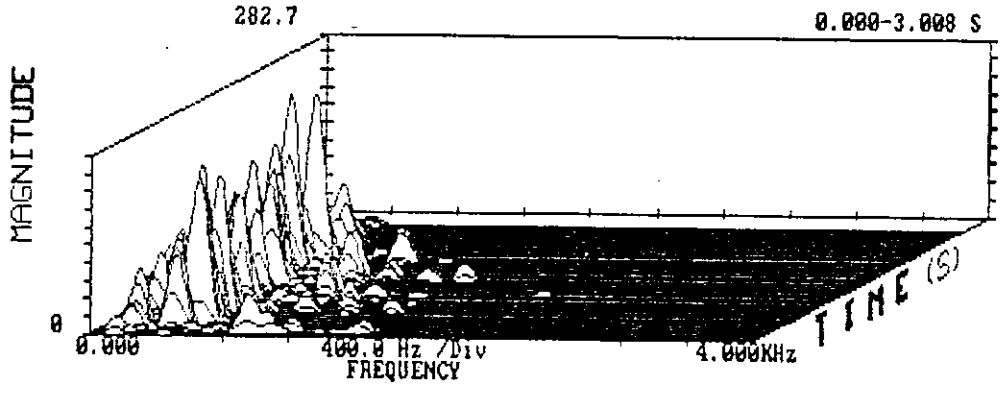
A ausência de algumas componentes de freqüências mais altas do sinal pode ser explicado pelo fato do algoritmo considerar essas freqüências de baixa amplitude como ruído. Isto pode ocorrer devido à degradação dessas freqüências por um nível de ruído que mascare este sinal.



(a)



(b)



(c)

Figura 5.9: Descrição espectográfica dos resultados obtidos pelo sistema com coeficientes no domínio da freqüência: (a) sinal original; (b) sinal de entrada de maior degradação pelo ruído e (c) sinal estimado na saída do filtro.

A Figura 5.10 mostra a adaptação dos coeficientes do filtro no domínio da freqüência. O ganho obtido na relação sinal-ruído para o sinal de saída foi de 5,68 dB, um pouco abaixo do ganho obtido no domínio do tempo (6,12). A Figura 5.10 mostra a adaptação dos coeficientes do filtro no domínio da freqüência, após a transformação inversa de Fourier. Pode-se notar, ainda, uma certa similaridade no comportamento comparando-se com o gráfico obtido para o cálculo dos coeficientes no domínio do tempo.

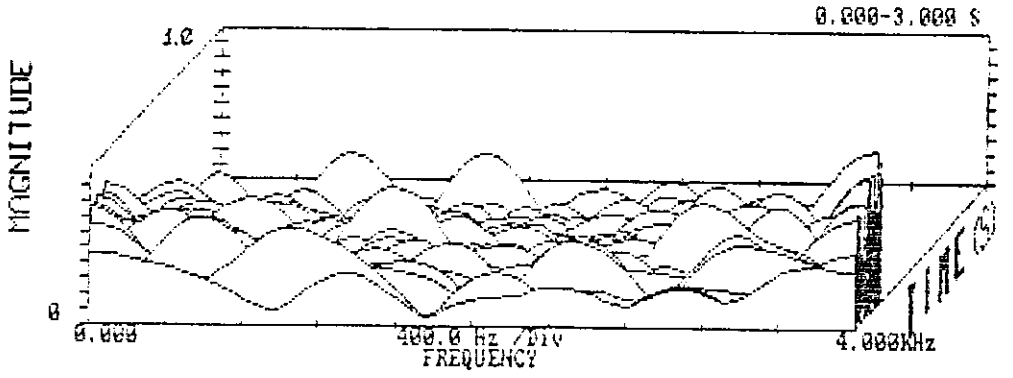


Figura 5.10: Adaptação dos coeficientes do filtro Wiener-Kolmogoroff no domínio da freqüência.

O estudo do algoritmo para cálculo dos coeficientes do filtro Wiener no domínio da freqüência será uma das etapas de continuidade deste trabalho. Pretende-se avaliar o seu comportamento variando-se alguns parâmetros que possam melhorar o desempenho do sistema, inclusive uma modificação no algoritmo através de ponderações na função Hopt(w), através de um balanceamento espectral como em [2].

5.5 Resultados obtidos em reconhecimento de voz

Foram feitos alguns testes preliminares do sistema de supressão multicanal ora proposto utilizando-o como um pré-processador para um sistema de reconhecimento de fala baseado em modelos de Markov escondidos implementado por Costa [36]. Inicialmente, um conjunto de dígitos (0 a 9) foi degradado para determinados valores de SNR. Foi feita, então, a análise de desempenho do referido reconhecedor considerando-se cada um desses valores.

Como sinais de entrada para o arranjo de microfones, foram colocados sinais degradados com SNRs de 8, 6, 4 e 0 dB para os microfones M_1, M_2, M_3 e M_4, respectivamente. Estes dígitos assim degradados apresentam uma taxa de erro de 80 %, sendo reconhecidos apenas os dígitos 5 e 9. Para os sinais resultantes após a filtragem, a taxa de erro caiu para 40 %, sendo reconhecidos os dígitos 0, 2, 3, 4, 5 e 9. Para uma degradação nos dígitos de 16, 14, 12 e 8 dB, na entrada do sistema, obteve-se após o processamento para supressão de ruído uma taxa de erro de erro de apenas 30%. Foram reconhecidos os dígitos 0, 2, 3, 4, 5, 8 e 9.

Este resultado pode ser considerado bom, em termos dos resultados obtidos pelo sistema de reconhecimento de voz, que destaca como um fator limitante em seu desempenho a não uniformidade do detetor de início e fim de palavras diante da presença do ruído [36]. Há, então, a necessidade da obtenção de um ganho maior na SNR do sistema de supressão para utilizá-lo como entrada em um reconhecedor de voz. Isto pode ser conseguido a partir da utilização de um número maior de microfones no arranjo, ou de outras variações no sistema a serem estudadas no futuro.

5.6 Comparação dos resultados obtidos em relação a outros sistemas

A implementação de um sistema em particular, envolve, além de suas próprias características, condições experimentais diversificadas tanto pelo ambiente do sistema

quanto pelos recursos tecnológicos envolvidos. Assim, ao comparar sistemas, estes aspectos devem ser levados em consideração diante da expectativa da obtenção de equivalência de resultados. Apesar disso, pode ser feita uma avaliação geral do desempenho do sistema implementado em relação a alguns dos sistemas encontrados na literatura.

O sistema implementado em [3] obteve, na saída da seção de atraso e soma equivalente à unidade de alinhamento do sistema aqui proposto, um valor de SNR de cerca de 4 dB. O resultado obtido, mostrado na Tabela 5.1 mostra um valor de SNR de até 4,02 dB. Para a estrutura de Griffts-Jim, em [3], foi obtido um ganho de +5 a + 8 dB, para alinhamento dos sinais. O ganho final obtido por Compernolle [3] para uma SNR de entrada entre +10 dB e + 20 dB foi de 9.0 dB para voz degradada por ruido aditivo. No caso presente, obteve-se um ganho de até 10 dB para sinais de entrada com SNR entre - 0,87 dB a + 4,51 dB.

Alguns dos sistemas estudados na literatura, não apresentam os valores de ganho obtidos através de alinhamento dos sinais de entrada [13, 33, 9]. O sistema que apresenta maior semelhança em relação a unidade de alinhamento é o sistema implementado por compernolle [3], com seção de atraso e soma.

Os resultados quantitativos, em termos de SNR podem ser considerados satisfatórios. Além disso, as avaliações subjetivas através de testes de escutas informais realizados demonstram que tanto a qualidade e quanto a inteligibilidade do sinal estimado na saída do sistema podem ser consideradas boas.

Em termos da taxa de reconhecimento de voz, os testes preliminares realizados mostram que o sistema reduz o ruído dos sinais degradados com SNRs de entrada variando de 0 a 16 dB, proporcionando uma taxa de reconhecimento do sistema baseado em HMM's (Modelos de Markov escondidos - *Hidden Markov Models*) implementado em [36] de até 70%. O sistema de Compernolle, também baseado em HMM's, com a seção de atraso e soma, apresenta uma taxa de reconhecimento de 80%.

Assim, os resultados obtidos podem ser considerados satisfatórios, diante das condições experimentais e aspectos tecnológicos envolvidos em relação aos demais sistemas.

Capítulo 6

Conclusão

Este trabalho teve como proposta a implementação de um sistema de supressão de ruído para melhoramento de sinais de voz degradados, baseado em características de sistemas existentes na literatura.

Assim, foi avaliado e proposto um sistema multicanal adaptativo para supressão de ruído cuja configuração dispunha de um arranjo bidimensional de microfones na entrada para receber o sinal degradado pelo ruído aditivo. O modelo apresentado baseou-se no modelo proposto por Zelinski em [13], bem como nas características dos modelos propostos em [33, 3], para a unidade de alinhamento de fase dos sinais de entrada utilizando como fonte de degradação o ruído de automóvel.

A principal contribuição deste trabalho foi, sem dúvida, a implementação da unidade de alinhamento de fase automática e o uso da filtragem adaptativa no domínio da freqüência além da comparação com a filtragem no domínio do tempo, o que facilitará, no futuro, uma implementação desse sistema em tempo real, bem como a de outros sistemas afins.

Ao final do trabalho, pode-se constatar que o sistema proposto foi implementado com sucesso. Neste sistema, a supressão do ruído foi levada a efeito através do alinhamento dos sinais de entrada, de um algoritmo de pós-processamento e finalmente pela filtragem realizada por meio de um filtro Wiener-Kolmogoroff adaptativo.

O sistema foi avaliado em termos quantitativos através dos valores de relação sinal-ruído e do ganho obtidos pelo sistema. Foi obtido um ganho de cerca de 10 dB na saída do filtro, o que demonstra um bom desempenho do sistema. Além disso, os testes subjetivos de escuta informais realizados indicam uma boa inteligibilidade do sinal. Uma avaliação geral do sistema leva às seguintes conclusões:

- Não é necessário um conhecimento a priori acerca das estatísticas do sinal de voz, bem como do ruído;
- As versões do ruído nos canais secundários não precisam ser correlacionadas com o ruído original - quanto mais descorrelacionadas, melhor o resultado;
- O sistema trabalha satisfatoriamente mesmo quando o próprio ruído consiste de sinais de voz;
- Não há limite no número de fontes de ruído que podem ser toleradas pelo sistema;
- Uma pequena correlação residual nos sinais de ruído captados não diminui o desempenho do sistema de forma perceptível;
- A unidade de alinhamento de fase é de extrema importância para o desempenho do sistema. Um alinhamento incorreto dos sinais pode gerar um sinal totalmente diferente do sinal original, sem possibilidade de recuperação da mensagem original;
- Pode-se obter um sistema de melhor desempenho, quanto maior o número de microfones do arranjo, mantendo-se praticamente a mesma base teórica para implementação do novo sistema;
- O sistema pode ser usado como entrada para sistemas de reconhecimento de fala desde que sejam feitos alguns melhoramentos adicionais para que sejam obtidos valores mais altos de SNR;

As principais dificuldades encontradas na implementação do sistema foram: a construção do arranjo de microfones, a obtenção dos sinais de prova devido a falta de

condições adequadas de gravação e, por fim, a falta de acesso a uma bibliografia mais recente que trate de sistemas de melhoramento de voz pela supressão de ruído de automóvel.

Algumas sugestões podem ser feitas para a continuidade deste trabalho ou de trabalhos afins na área de redução de ruído em sinais de voz:

- A implementação de um sistema com um número maior de microfones, para permitir um ganho maior na relação sinal ruído do sistema;
- Um estudo mais profundo para a obtenção de novos algoritmos para a unidade de pós-processamento do sistema;
- A implementação do sistema em tempo real, necessitando para isso da montagem de um protótipo com arranjos bidimensionais e da montagem de uma placa em hardware que sirva de unidade de pré-processamento de sinais a serem usados em sistemas de reconhecimento de fala e de verificação de locutor, por exemplo;
- A aplicação do arranjo de microfones em sistemas de supressão de ruído para uso em sistemas de teleconferência em grandes auditórios;
- Pode ser feita, ainda, um avaliação do sistema de supressão de ruído de automóvel em termos de compatibilidade presente e futura em relação aos sistemas de telefonia móvel atualmente utilizados.

Apêndice A

Cálculo da relação sinal-ruído

Uma medida de avaliação de desempenho de vários sistemas encontrados na literatura é a relação sinal-ruído. No entanto, cada sistema em particular pode utilizar uma forma diferente de calcular esta medida [32]. Alguns calculam-na em função da variância [1], outros em função do valor máximo [8], e assim por diante. Dessa forma, torna-se necessário destacar de que forma é feito o cálculo da SNR para que, ao comparar o desempenho do sistema com outros, leve-se em conta estas diferenças individuais.

A relação sinal-ruído (SNR - Signal-to-Noise Ratio) para o sistema implementado neste trabalho é calculada por [5]:

$$\text{SNR} = 10log\{\sigma_s^2/\sigma_e^2\} \qquad \text{[dB]} \tag{A.1}$$

onde σ_s^2 representa a variância do sinal de voz e σ_e^2 representa a variância do erro de estimação, dada por:

$$\sigma_e^2 = E[\text{e}^2(n)] \qquad \text{[dB]} \tag{A.2}$$

onde $e(n) = s(n)$ - $\hat{s}(n)$, e $\hat{s}(n)$ representa o sinal estimado.

Como o processamento do sinal é feito segmentalmente, é calculada a relação sinal-ruído no segmento (SNR_k):

$$\mathrm{SNR}_k = 10log\{\sigma^2_{s,k}/\sigma^2_{e,k}\} \qquad [\mathrm{dB}] \tag{A.3}$$

onde $\sigma^2_{s,k}$ e $\sigma^2_{e,k}$ representam as variâncias do sinal e do ruído no k-ésimo segmento.

A relação sinal-ruído segmental (SegSNR) é dada por:

$$\mathrm{SegSNR} = \frac{1}{\mathrm{NSEG}} \sum_{k=1}^{\mathrm{NSEG}} \mathrm{SNR}_k \qquad [\mathrm{dB}] \tag{A.4}$$

O ganho na relação sinal-ruído pode ser calculado por:

$$\mathrm{SNR} - \mathrm{G} = \mathrm{SNR} - \mathrm{SNR}_k \qquad [\mathrm{dB}] \tag{A.5}$$

E o ganho na relação sinal-ruído segmental por:

$$\mathrm{SegSNR} - \mathrm{G} = \mathrm{SegSNR} - \mathrm{SNR_k} \qquad [\mathrm{dB}] \tag{A.6}$$

Referências

[1] B. G. Aguiar Neto. Melhoramento de sinais de voz através de método de supressão de ruídos por balanceamento espectral adaptativo. *Anais do VI SBT*, pag. 239-243, setembro 1988.

[2] B. G. Aguiar Neto. Melhoramento de voz degradada por método baseado em subtração espectral adaptativa. *Anais do VII SBT*, pag. 54-58, setembro 1989.

[3] D. V. Compernolle et al. Speech recognition in noisy environments with the aid of microphone arrays. *Speech Communication*, 9:432–442, August 1990.

[4] J. S. Lim. Speech enhancement. *Proc. ICASSP*, pages 3135–3142, 1986.

[5] B. G. Aguiar Neto. *Signalaufbereitung in digitalen Sprachübertragunsgssysten.* Doktor-ingenieur genehmigte dissertation, Technishen Universität Berlin, Berlin, 1987.

[6] D. Becker. Einzelworterkennung in störgeräuscherfüllter ungebung. *Tagunsgband der Konferenz Elektronishe Sprachsignalverarbeitung - Berlin*, pages 111–117, 1990.

[7] M. R. Sambur. Adaptive noise cancelling for speech signals. *IEEE Trans. on ASSP*, 26(5), October 1978.

[8] E. R. Ferrara Jr. and Bernard Widrow. Multichannel adaptive filtering for signal enhancement. *IEEE Trans. on ASSP*, 29(3):766–770, June 1981.

[9] J. L. Flanagan et al. Computer steered microphone arrays for sound transduction in large rooms. *J. Acoust. Soc. Am.*, 78(5):1508–1518, November 1985.

[10] Y. Kaneda and J. Ohga. Adaptive microphone array system for noise reduction. *IEEE Trans. on ASSP*, 34(6):1391–1400, December 1986.

[11] B. Widrow et al. Adaptive noise cancelling: Principles and applications. *Proc. of the IEEE*, 63(12):1692–1716, December 1975.

[12] W. F. Gabriel. Adaptive arrays - an introduction. *Proc. of IEEE*, 64(2):239–279, February 1976.

[13] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. *Proc. of the Intern. Conf. on ASSP - New York*, pages 2578–2581, 1988.

[14] H. Drucker. Speech processing in a high ambient noise environment. *IEEE Trans. on Audio and Eletroacoustics*, 13(2):165–168, June 1968.

[15] M. Schwartz. *Transmissão de Informação, Modulação e Ruído.* Mcgraw-Hill - New York, 1979.

[16] K. C. Kuyter. Effects of ear protective devices on the intelligibility of speech in noise. *J. Acoust. Soc. Am.*, 18:413–417, 1946.

[17] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *Trans. on ASSP*, 27(2), April 79.

[18] M. R. Sambur and N. S. Jayant. LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise. *IEEE Trans. on ASSP*, 24:488–494, December 1976.

[19] B. P. Lathi. *An introduction to random signals and communication theory.* Intertext Books, London, 1970.

[20] A. Papoulis. *Signal Analysis.* McGraw-Hill,, New York, 1985.

[21] P. Vari. Verfahrun zur digitalen verbesserung gestorter sprache. *Tekade Tech. Mitteilungun*, pages 70–76, 1983.

[22] L. R. Rabiner and R. W. Shafer. *Digital processing of speech signals.* Prentice-Hall, Inc. - New Jersey, 1978.

[23] M. M. Sondhi, C. E. Schmidt, and L. R. Rabiner. Improving the quality of a noise speech signal. *The Bell System Technical Journal*, 60(8), October 1981.

[24] J. S. Lim, A. V. Oppenheim, and L. D. Braida. Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-26:354–358, August 1978.

[25] J. E. Porter and S. F. Boll. Optimal estimators for espectral restoration of noisy speech. *Proc. IEEE ICASSP*, 2:18A21–4, March 1984.

[26] D. Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer speech and language*, 3(2):151–168, 1989.

[27] G. Faucon and R. Le Bouquin. Traitement d'antenne pour la réduction de bruit sur la parole. *Proc. 12th Gretsi Symposium - Juan Les Pins*, pages 517–520, June 1989.

[28] N. Wiener. *Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engeneering Applications.* Wiley, New York, 1949.

[29] R. Kalman and R. Bucy. New results in linear filtering and prediction theory. *Trans. ASME, ser. D, J. Basic Eng.*, 83:95–107, December 1961.

[30] T. Kailath. A view of three decades of linear filtering theory. *IEEE Trans. Inform. Theory*, IT-20:145–181, March 1974.

[31] P. Noll. Statistische nachrichtentheorie. *Skript Institut für Fernmeldetechnik der TU Berlin*, 86-87.

[32] N. S. Jayant and P. Noll. *Digital Coding of Waveforms - Principles and Applications to Speech and Video.* Prentice-Hall, Inc. - New Jersey, 1984.

[33] Y. Kaneda and M. Tohyama. Noise suppression signal processing using 2-point received signals. *Electronics and Communications in Japan*, 67-A(12):19–28, 1984.

[34] Silvana L. N. C. Costa and Benedito G. Aguiar Neto. Adaptive multichannel system for speech enhancement using microphone array. *International Telecommunications Symposium of IEEE*, pages 152–155, August 94.

[35] A. Alcaim, J. A. Solewicz, and J. A. Moraes. Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no Português falado no Rio de Janeiro. *Relatório Técnico, CETUC - PUC - Rio de Janeiro.*

[36] Washington C. de A. Costa. Reconhecimento de fala utilizando Modelos de Markov Escondidos (HMMs) de densidades contínuas. *Dissertação de Mestrado*, Universidade Federal da Paraíba, junho 1994.