ROGERIO CARLOS VIEIRA MACIEL

MELHORIA DA QUALIDADE DE SINAIS DE FALA DEGRADADOS POR RUÍDO ATRAVÉS DA UTILIZAÇÃO DE SINAIS SINTETIZADOS

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para a obtenção do título de Mestre em Engenharia.

São Paulo

2003

ROGERIO CARLOS VIEIRA MACIEL

MELHORIA DA QUALIDADE DE SINAIS DE FALA DEGRADADOS POR RUÍDO ATRAVÉS DA UTILIZAÇÃO DE SINAIS SINTETIZADOS

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para a obtenção do título de Mestre em Engenharia.

Área de Concentração:
Engenharia Elétrica

Orientador:
Prof. Dr. Phillip M. S. Burt

São Paulo

2003

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 11 de agosto de 2003.

Assinatura do autor

Assinatura do orientador

Maciel, Rogerio Carlos Vieira
Melhoria da qualidade de sinais de fala degradados por ruído através da utilização de sinais sintetizados. São Paulo, 2003.

Edição Revisada. 92p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Telecomunicações e Controle.

1. Processamento de fala 2. Redução de Ruído 3. Processamento digital de sinais I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Telecomunicações e Controle II. t

Agradecimentos

Ao professor Phillip Burt pela valorosa dedicação e participação constante no desenvolvimento deste trabalho.

Aos professores Fabio Violaro e Miguel Arjona, pela presença na banca examinadora e pelas importantes sugestões dadas no exame de qualificação.

Aos meus pais, pelo incentivo e estímulo.

À minha namorada, Cecília, por todo o seu carinho e compreensão.

A todos que, direta ou indiretamente, colaboraram na execução deste trabalho.

Esta dissertação recebeu apoio financeiro da Ericsson Telecomunicações, no âmbito do convênio de pesquisa intitulado "Algoritmos de redução de ruído com melhor desempenho para ruído não-estacionário e baixas relações sinal/ruído", firmado com a Fundação para o Desenvolvimento Tecnológico da Engenharia (FDTE) e o PTC/EPUSP.

Resumo

Este trabalho discute um novo método para a melhoria da qualidade de sinais de fala degradados por ruído aditivo branco de elevada intensidade (SNR segmentada variando de 10 a 3 dB). A técnica apresentada baseia-se na soma ponderada entre um sinal obtido por subtração espectral e um sinal sintetizado, produzido de acordo com o modelo digital de produção de fala (análise e síntese LPC). Para a estimação dos coeficientes LPC e período de pitch, foi implementado um pré-processador baseado na técnica de subtração espectral (modificada especificamente para este propósito), o que melhora significativamente a qualidade geral do sinal de fala sintetizado. A soma ponderada entre o sinal obtido por subtração espectral e o sinal sintetizado permite a reconstrução de regiões espectrais perdidas devido aos efeitos da aplicação da subtração espectral, bem como a eliminação do ruído musical. Os testes realizados com frases foneticamente balanceadas lidas por diversos locutores permitem observar que o método proposto oferece melhores resultados do que a subtração espectral. O sinal de fala obtido apresenta também maior clareza e naturalidade, sem o inconveniente do ruído musical.

Abstract

The present work discusses a new method to enhance speech signals degraded by white additive noise in high-noise environments (segmented SNR varying from 10 to 3 dB). The approach presented here is based upon a weighted sum involving a speech signal obtained from spectral subtraction and a synthetized speech signal, which is produced according to the concepts of the digital model of speech production (LPC analysis and synthesis). A spectral subtraction-based pre-processor was specifically implemented for LP coeficients and pitch period estimation, which significantly improves the overall quality of the synthetized speech signal. The weighted combination of these two speech signals allows the reconstruction of spectral regions lost due to the effects of spectral subtraction, as well as the elimination of musical noise. Tests conducted with phonetic-ballanced sentences from several speakers show that the proposed method offers better results than spectral subtraction, producing a more natural and clear enhanced speech signal, without the common artifact of musical noise.

Conteúdo

Lista de Figuras

Lista de Símbolos

Símbolos utilizados no capítulo 2

$s(n)$	Sinal de fala original (sem ruído)
$\eta(n)$	Ruído aditivo
$x(n)$	Sinal ruidoso
$H(e^{jw})$	Filtro de ponderação espectral
$\hat{S}(e^{jw})$	Estimador do sinal limpo da subtração espectral
$\hat{s}(n)$	Sinal gerado pela subtração espectral
$\mu(e^{jw})$	Valor médio da amplitude do espectro do ruído
$\theta_x(e^{jw})$	Fase do sinal de fala
a_k	Vetor de coeficientes LPC obtidos do sinal de fala original
b_k	Vetor de coeficientes LPC obtidos do sinal de fala processado
d	Distância de Itakura entre vetores de coeficientes LPC
D	Fator de dissemelhança
$w(n)$	Janela de sinal
N_{eff}	Duração efetiva da janela de sinal
SNR	Relação sinal-ruído
SNR_{seg}	Relação sinal-ruído segmentada
σ^2	Potência do ruído
α	Fator de sobre-estimação de ruído da subtração espectral

λ	Coeficiente de suavização da subtração espectral
$H_{\min}$	Limite espectral mínimo

Símbolos utilizados nos capítulos 3 e 4

$s(n)$	Amostras do sinal original
$\tilde{s}_{pred}(n)$	Amostras preditas através da Predição Linear
α_k	Coeficientes preditores
p	Ordem da predição linear
g	Ganho da predição linear
$e(n)$	Erro de predição
E	Erro quadrático de predição de curto-prazo
$H_o(e^{jw})$	Filtro ótimo do algoritmo LMAP
$P_s(e^{jw})$	Densidade espectral de potência do sinal limpo (sem ruído)
σ_d^2	Potência do ruído
α	Fator de sobre-estimação de ruído da subtração espectral
λ	Coeficiente de suavização da subtração espectral

Símbolos utilizados nos capítulos 5 e 6

$\phi(k)$	Auto-correlação
$R_n(k)$	Auto-correlação de curto-prazo
$w(n)$	Janela de sinal
$s_w(n)$	Sinal janelado
$D_{AMDF}(k)$	Função Diferença de Magnitude Média
$D_{CAMDF}(k)$	Função Diferença de Magnitude Média Circular
$P_{HPS}(k)$	Espectro de Produtos Harmônicos
$c(n)$	Cepstrum real
Z_n	Taxa de cruzamentos por zero de curto-prazo
$d(n)$	Sinal sintetizado
g	Ganho (LPC)
P	Período de pitch
α_k	Coeficientes LPC
$u(n)$	Vetor de excitação

Símbolos utilizados no capítulo 7

$\hat{s}(n)$	Sinal obtido por subtração espectral
$d(n)$	Sinal sintetizado
$v(n)$	Sinal sintetizado multiplicado pelos fatores de ponderação
$\tilde{s}(n)$	Sinal reconstruído
a	Fator de peso aplicado ao sinal $v(n)$
b	Fator de peso aplicado ao sinal $\hat{s}(n)$

Capítulo 1

Introdução

1.1 Melhoria da qualidade de sinais de fala - Speech Enhancement

A melhoria da qualidade dos sinais de fala presentes nos sistemas de telecomunicações (Speech Enhancement) tem sido foco de intensos estudos nas últimas décadas [1, 2, 3, 4]. Em praticamente todas as aplicações de transmissão de voz a qualidade da comunicação pode ser comprometida pela presença de elementos que degradam o sinal, como ruído ambiente, reverberação, perdas devidas à codificação em enlaces digitais e concorrência de outras conversações ou de outras fontes de sinal. Tais elementos podem afetar o sinal de diversas formas, reduzindo sua inteligibilidade, aumentando o cansaço do ouvinte, tornando a conversação pouco natural, ou, ainda, afetando a eficiência de outros sistemas que se utilizarão desses sinais posteriormente, como reconhecedores ou codificadores de voz. Os métodos de melhoria da qualidade dos sinais de fala buscam, portanto, identificar e extrair os elementos que degradam a qualidade do sinal, realçando a informação de fala, possibilitando assim uma melhor comunicação entre as partes envolvidas.

1.2 Principais métodos

Este trabalho considera a melhoria da qualidade de sinais de voz degradados por ruído aditivo branco, em ambientes em que se dispõe de apenas um microfone (que capta, portanto, o sinal somado ao ruído). Dentre as várias abordagens já propostas para a solução deste problema, é interessante destacar os principais métodos que, além da importância histórica, servem também como ponto de partida para desenvolvimentos mais sofisticados. Em geral, costuma-se dividir os algoritmos em grupos, de acordo com a maneira pela qual a fala é modelada e de acordo com a técnica utilizada para o processamento do sinal.

Dos métodos baseados em modelos estocásticos para o sinal de fala, por exemplo, destaca-se a subtração espectral, que busca suprimir o ruído através da subtração de uma estimativa do espectro do ruído presente no sinal de fala, estimativa que pode ser obtida durante períodos de silêncio. Essa operação em geral é realizada trabalhando-se com a transformada de Fourier ou auto-correlação do sinal.

Para os métodos baseados na estimação de parâmetros do modelo de produção de fala, destacam-se as técnicas de síntese do sinal a partir de parâmetros obtidos do sinal ruidoso (coeficientes LPC, por exemplo). Tais métodos envolvem o conhecimento do modelo digital de produção de fala, bem como dos aspectos perceptuais relacionados. A estimativa dos parâmetros, por ser o ponto principal dos métodos baseados em síntese, é uma das áreas que oferece as maiores possibilidades de estudo.

Outra classe importante de algoritmos envolve a utilização da informação da freqüência fundamental do sinal de fala e de seus harmônicos, isto é, informação relativa à periodicidade do sinal de fala. Tais algoritmos fazem uso de filtros de seleção de harmônicos e rastreamento da freqüência fundamental.

Em casos mais específicos, quando se dispõe de mais de um microfone, é possível também a aplicação de algoritmos adaptativos para o cancelamento do ruído. Nessas situações, um sinal de referência é usado para extração de in-

formações sobre o ruído presente na comunicação.

Para este trabalho, entretanto, serão estudadas apenas as características da subtração espectral e da síntese de fala a partir de parâmetros do modelo de produção da voz. A proposta é a obtenção de um sinal de melhor qualidade através da combinação dos resultados desses dois métodos.

1.3 Proposta de trabalho

Neste trabalho será proposto um novo método para a melhoria da qualidade de sinais de fala degradados por ruído aditivo branco, baseado na soma ponderada entre um sinal obtido pelo processo de subtração espectral e um sinal sintetizado através do modelo digital de produção de fala (análise LPC). O método apresentado considera a existência de uma única fonte de sinal (isto é, não envolve a utilização de redes de microfones para captação do sinal de fala), e será aplicado a sinais degradados por ruído de intensidade moderada a alta (SNR segmentada variando de 10 a 3 dB). Os objetivos principais do método serão a reconstrução de regiões espectrais perdidas pela aplicação da subtração espectral (cujo efeito é a sensação de abafamento no sinal) e a eliminação do ruído musical (efeito indesejável comum aos métodos de redução de ruído). Como resultado do processamento espera-se uma melhoria da qualidade geral do sinal de fala, aliada a um baixo ruído musical.

O princípio da proposta apresentada é a combinação dos efeitos positivos da subtração espectral e de um sinal sintetizado. Nas freqüências em que o sinal é pouco atenuado pela subtração espectral, a contribuição do sinal sintetizado é pequena; nas freqüências em que o sinal é muito afetado pela subtração espectral, entretanto, a contribuição do sinal sintetizado é grande. Tal procedimento tem por objetivo a obtenção de uma maior qualidade subjetiva no sinal de saída, maior do que a do sinal processado por subtração espectral e também maior do que se o sinal de saída fosse simplesmente o sinal sintetizado.

De acordo com o que foi proposto, o trabalho exige o estudo criterioso dos

seguintes assuntos: subtração espectral; características do aparelho fonador humano; modelo digital de produção de fala; técnicas para estimação de coeficientes LPC e período de pitch; funcionamento do sintetizador de fala; estimação de parâmetros em sinais ruidosos. O estudo destes tópicos será realizado em capítulos separados, sendo que, ao final de cada capítulo, haverá uma seção de análise de implementações e resultados, comentando os desenvolvimentos práticos realizados com relação a cada assunto estudado. Após o estudo dos componentes do sistema, serão discutidos os detalhes da implementação da estrutura completa do método (que envolve a implementação da soma ponderada entre o sinal obtido por subtração espectral e o sinal obtido através da síntese baseada no modelo digital de produção de fala), bem como os resultados e as possibilidades de futuros desenvolvimentos.

1.4 Organização do texto

Este documento está organizado da seguinte forma: No capítulo 2, a técnica de subtração espectral é estudada e são apresentados os resultados das implementações realizadas. No capítulo 3 é discutido o modelo digital de produção de fala, bem como os primeiros resultados práticos obtidos com o sintetizador. O capítulo 4 trata da formulação da análise LPC e das técnicas para estimação em ambientes ruidosos; também neste capítulo são apresentados os resultados dos testes realizados com sinais ruidosos. No capítulo 5 são apresentados os principais métodos para estimação do período de pitch, bem como dos resultados de implementações e comparações entre os algoritmos. O capítulo 6 trata dos detalhes da construção do sintetizador de fala e os refinamentos utilizados. O capítulo 7 apresenta a estrutura completa do método proposto no trabalho. Finalmente, no capítulo 8, são apresentadas as conclusões e sugestões de futuros estudos baseados neste trabalho.

Segue, junto ao texto, um CD com exemplos reais de aplicação do algoritmo apresentado. São arquivos de áudio contendo frases usadas no decorrer do desen-

volvimento do método, ilustrando os principais itens discutidos neste trabalho, como a subtração espectral, a síntese de fala baseada em parâmetros LPC e a combinação dos sinais produzidos. Através dos exemplos é possível ter uma percepção melhor das características do método.

Capítulo 2

Subtração espectral

2.1 Formulação

Baseada nas premissas de que o ruído é aditivo e que seu espectro de potência é conhecido, a técnica de subtração espectral busca subtrair, do sinal ruidoso, a informação referente ao espectro do ruído [2]. Consideremos que o sinal $s(n)$ foi somado ao sinal de ruído $\eta(n)$, resultando no sinal $x(n)$. Desta forma, temos:

$$x(n) = s(n) + \eta(n). \tag{2.1}$$

O ruído será considerado estacionário e ergódico. Desta forma, tomando a transformada de Fourier, temos:

$$X(e^{jw}) = S(e^{jw}) + \aleph(e^{jw}), \tag{2.2}$$

onde

$$x(n) \longleftrightarrow X(e^{jw}) \tag{2.3}$$

$$X(e^{jw}) = \sum_{n=-\infty}^{\infty} x(n)e^{-jwn} \tag{2.4}$$

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{jw}) e^{jwn} dw. \tag{2.5}$$

O propósito da subtração espectral, portanto, é a obtenção de uma estimativa do sinal limpo - que será chamada de $\hat{s}(n)$ - a partir do sinal ruidoso $x(n)$ e de um conhecimento prévio da estatística do ruído adicionado ao sinal, $\eta(n)$. Para facilidade de notação e compreensão, costuma-se introduzir o conceito do filtro de ponderação espectral $H(e^{jw})$, de fase nula, que representa a atenuação a ser aplicada ao espectro do sinal ruidoso de forma a remover suas componentes espectrais de ruído. Os coeficientes de $H(e^{jw})$ são calculados, quadro a quadro, a partir de informações referentes ao espectro do sinal ruidoso e à estatística do ruído, estimada em trechos de silêncio. O uso do filtro de ponderação leva ao estimador do sinal limpo, $\hat{S}(e^{jw})$:

$$\hat{S}(e^{jw}) = H(e^{jw}) X(e^{jw}). \tag{2.6}$$

No caso da subtração espectral, temos:

$$H(e^{jw}) = 1 - \frac{\mu(e^{jw})}{|X(e^{jw})|} \tag{2.7}$$

onde $\mu(e^{jw})$ representa o valor médio da amplitude do espectro do ruído, que será estimado em trechos de silêncio.

$$\mu(e^{jw}) = E\left\{ \left| \aleph\left(e^{jw}\right) \right| \right\}. \tag{2.8}$$

Desta forma, temos:

$$\hat{S}(e^{jw}) = \left[\left| X(e^{jw}) \right| - \mu(e^{jw}) \right] e^{j\theta_x(e^{jw})}, \tag{2.9}$$

que representa a formulação básica da subtração espectral. A informação de fase, $\theta_x(e^{jw})$, é obtida diretamente do sinal ruidoso, $x(n)$, e aplicada, sem alterações, ao sinal processado $\hat{s}(n)$.

2.2 Processamento em quadros

O sinal de fala pode ser considerado estacionário em curto-prazo, para quadros de duração típica de 20 a 30 ms [5]. Essa característica, aliada à necessidade de processamento em tempo real (com baixo atraso), conduz à técnica de processamento em quadros, com superposição e janelamento, prática comum em algoritmos de processamento de fala.

O sinal é dividido em quadros de tamanho N (no caso das simulações realizadas, foram utilizados sinais amostrados a 8 kHz e quadros de 256 amostras, ou 32 ms) e multiplicado por uma janela (janela de Hamming, por exemplo). Sobre o sinal janelado é calculada a transformada de Fourier, de modo a permitir a aplicação do método desejado ao espectro do sinal ruidoso, $X(k)$. Após a multiplicação pelos coeficientes $H(k)$ (filtro de ponderação), é calculada a transformada inversa de Fourier de $X(k)$. O sinal obtido, de volta ao domínio do tempo, produz um quadro de sinal de saída. Entretanto, a janela de análise não é deslocada de N amostras, mas de uma quantidade menor, M, de modo a produzir uma superposição entre os quadros (é comum utilizar-se uma superposição de $M = N/2$ amostras, ou mesmo valores maiores). Para garantir a correta superposição entre os quadros, o sinal que é obtido a partir da transformada inversa de fourier após a aplicação do filtro de ponderação é somado ao trecho com as M amostras produzidas no processamento do quadro anterior.

A Figura 2.1 ilustra o diagrama em blocos do processo de redução de ruído por subtração espectral. Os coeficientes $H(k)$ representam o filtro de ponderação. O espectro do sinal ruidoso, $X(k)$, é multiplicado pelo filtro $H(k)$, e o resultado é submetido à transformada inversa de Fourier, para obtenção do sinal $\hat{s}(n)$, após janelamento e soma.

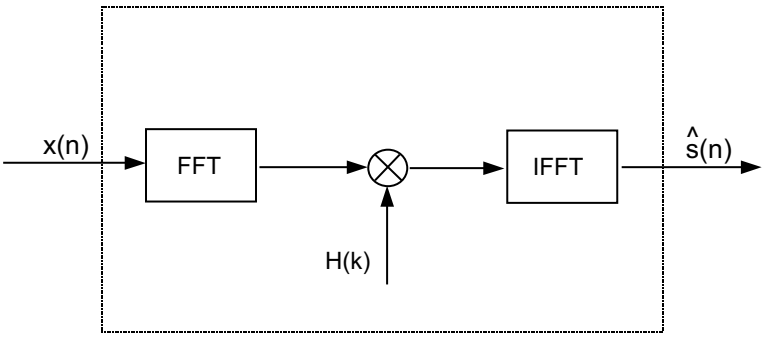

Figura 2.1: Diagrama em blocos do processo de Subtração Espectral, ilustrando a presença do filtro de ponderação espectral $H(k)$.

2.3 Ruído residual

Apesar de proporcionar a redução do nível de ruído, a subtração espectral traz consigo um sério inconveniente. Devido à característica aleatória do ruído, é possível que, num dado quadro de silêncio, o valor da amplitude do espectro de ruído, em algumas freqüências, esteja ligeiramente acima do valor médio da amplitude do ruído, $\mu(e^{jw})$. Neste caso, o filtro de ponderação $H(e^{jw})$ permitirá a manutenção de algumas raias isoladas oriundas do sinal ruidoso. Ao longo do processamento, tal fenômeno será responsável pelo surgimento de senóides aleatoriamente distribuídas no tempo e na freqüência, que permanecerão no sinal resultante e serão percebidas pelos ouvintes. Este efeito indesejado é comumente chamado de ruído musical. Por outro lado, é possível também que o valor da amplitude do espectro de ruído esteja abaixo do valor médio da amplitude do ruído, $\mu(e^{jw})$, em algumas freqüências. Neste caso, para evitar o aparecimento de valores negativos na magnitude do espectro, tais amplitudes têm seus valores igualados a zero.

2.4 Modificações

Os efeitos indesejados decorrentes da aplicação da subtração espectral motivaram o desenvolvimento de variações no método, com o objetivo de reduzir o ruído musical e aumentar a inteligibilidade do sinal processado. As principais variações serão descritas a seguir.

2.4.1 Suavização das amplitudes espectrais

A suavização dos coeficientes do filtro de ponderação $H(e^{jw})$ contribui para a diminuição da intensidade do ruído musical. Em geral, a suavização é obtida com filtros de média móvel aplicados a cada um dos coeficientes de $H(e^{jw})$. A aplicação da suavização, porém, deve levar em conta a característica de estacionariedade de curto-prazo dos sinais de fala. Durante a produção da fala, os trechos estacionários são curtos, variando tipicamente de 20 a 30 ms. A utilização de janelas extensas de suavização, portanto, pode prejudicar a inteligibilidade de trechos de sinal que apresentam variações rápidas, como inícios de fonemas e consoantes plosivas (p, t, k, por exemplo). A escolha dos filtros de suavização deve levar em conta o compromisso entre a redução do ruído musical e a perda das informações associadas às variações rápidas do sinal de fala.

2.4.2 Retificação e Limite Espectral Mínimo

A magnitude dos filtros de ponderação $H(e^{jw})$ não pode ser menor do que zero. Entretanto, há situações em que, devido à característica aleatória do ruído, a magnitude do espectro do sinal ruidoso, $|X(e^{jw})|$, pode ser menor do que a magnitude média do espectro do ruído, $\mu(e^{jw})$. Nestes casos, é comum aplicar-se o conceito de retificação de meia-onda, isto é, se a magnitude do espectro do sinal ruidoso for menor do que a magnitude média do espectro do ruído, $H(e^{jw})$ será nulo em tais freqüências. A aplicação isolada deste método, porém, contribui para o aumento do ruído musical. Uma alternativa é a utilização de um valor mínimo

para $H(e^{jw})$, também conhecido como Minimum Spectral Floor, de maneira a mascarar o efeito do ruído musical com um ruído residual de baixa intensidade. O valor mínimo de $H(e^{jw})$ pode ser pré-fixado ou obtido a partir de informações do espectro de ruído de quadros adjacentes.

2.4.3 Atenuação adicional durante trechos de silêncio

O conteúdo de energia de $\hat{S}(e^{jw})$, quando comparado à potência média do ruído, $\mu(e^{jw})$, oferece um bom indicador da presença de sinal de fala dentro de um certo quadro sob análise. Se o trecho for silencioso, $\hat{S}(e^{jw})$ consistirá apenas do resíduo de ruído que permanece após o processamento. Algumas possibilidades de tratamento dessa situação são: manter o sinal inalterado, atenuar sua amplitude, reduzir a saída a zero. Os testes subjetivos indicam que atenuar a amplitude do sinal nesses quadros é a melhor alternativa [2], pois mantém o resíduo de ruído que garante a naturalidade do sinal resultante, porém atenua sua intensidade.

2.4.4 Sobre-estimação do ruído

Até o momento, a discussão sobre subtração espectral envolveu a estimação do sinal limpo a partir da aplicação de uma regra de ponderação obtida da estimativa da magnitude média do espectro de ruído. Apesar de efetiva na redução do ruído, o efeito negativo do ruído musical é o principal inconveniente desta técnica. Uma alternativa para redução do ruído musical é a sobre-estimação do ruído, que consiste em multiplicar a magnitude média do espectro do ruído por um fator de sobre-estimação, com o propósito de reduzir a possibilidade de que raias espúrias de ruído permaneçam no sinal resultante e produzam ruído musical. A sobre-estimação de ruído, entretanto, pode eliminar informação do sinal de fala, principalmente os formantes de ordem superior, cujas amplitudes são baixas. O efeito da eliminação dos formantes mais fracos é uma sensação de abafamento no sinal, ou uma sensação de não-naturalidade no sinal processado. A taxa de sobre-estimação de ruído, todavia, pode ser escolhida de modo a buscar o melhor

compromisso entre distorção espectral e ruído musical.

2.4.5 Modificação no cálculo do filtro de ponderação espectral

Outra possibilidade de alteração é a utilização de expoentes aplicados ao termo referente à magnitude do espectro do sinal ruidoso e ao termo referente ao valor médio da amplitude do espectro do ruído, no cálculo de $H(e^{jw})$. Com isso, temos:

$$\mu(e^{jw}) = E\left\{\left|\aleph\left(e^{jw}\right)\right|^{\gamma}\right\}, \tag{2.10}$$

e a expressão para obtenção dos coeficientes $H(e^{jw})$ passa a ser:

$$H(e^{jw}) = 1 - \frac{\mu(e^{jw})}{|X(e^{jw})|^{\gamma}} \tag{2.11}$$

onde γ, o expoente aplicado à magnitude do espectro do sinal ruidoso, pode ser ajustado de maneira a produzir melhores resultados gerais. Neste trabalho, foi utilizado $\gamma = 2$.

2.5 Avaliações de desempenho

Uma das etapas importantes no processo de desenvolvimento de algoritmos de melhoria da qualidade do sinal de fala é a avaliação de seu desempenho. Em primeira análise, a melhoria na relação sinal-ruído pode ser vista como fator de mérito de um sistema de redução de ruído. Entretanto, a relação sinal-ruído não exprime toda a gama de detalhes envolvidos no processo. Os aspectos perceptuais, tais como inteligibilidade, efeitos indesejados, cansaço do ouvinte e qualidade do sinal resultante devem ser levados em conta na apreciação da qualidade de um método. Existem basicamente duas abordagens para a avaliação dos algoritmos de redução de ruído: a medição objetiva, que utiliza o cálculo de medidas de distância entre o sinal original e o sinal processado, e a avaliação subjetiva, que

leva em conta a reação dos ouvintes diante do sinal processado, em comparação com o sinal ruidoso e com o sinal original (sem ruído).

Medidas objetivas

As medidas objetivas buscam caracterizar, numericamente, as diferenças entre dois sinais de fala. A informação oferecida por uma medida objetiva deve estar relacionada a aspectos perceptuais, isto é, uma distância elevada calculada entre dois sinais deve indicar que os ouvintes têm percepções distintas ao escutá-los, enquanto que uma distância pequena calculada entre outros dois sinais deve indicar que, perceptualmente, tais sinais são muito semelhantes. A soma dos quadrados das diferenças das amostras, por exemplo, é uma medida de distância com pouco valor perceptual, pois leva em conta apenas os valores das amostras dos sinais. É possível que dois sinais com características perceptuais muito semelhantes apresentem distância muito elevada segundo esse critério, e vice-versa (isso se deve basicamente à relativa insensibilidade da audição humana à distorção de fase, o que torna as medições baseadas em diferenças amostra-a-amostra pouco eficazes). Por outro lado, a soma dos quadrados das diferenças das amplitudes espectrais pode apresentar uma melhor representação perceptual, por levar em conta o espectro dos sinais. Várias medidas de distância têm sido utilizadas nos sistemas de análise e síntese de fala, com o intuito de refletir numericamente a informação perceptual dos sinais, como, por exemplo, a distância cepstral, distância RMS logarítmica, medida do cosseno hiperbólico, dentre outras [6]. Neste trabalho, utilizamos como referência para os testes objetivos a distância de Itakura [7], por suas importantes características perceptuais. A distância de Itakura está baseada na diferença entre os modelos de pólos obtidos para o sinal de referência e para o sinal sob análise, e é computada entre conjuntos de coeficientes LPC estimados em quadros sincronizados (de duração variando tipicamente de 25 a 35 ms). Por ser baseada na análise LPC, a distância de Itakura é útil na avaliação das técnicas de estimação de coeficientes LPC discutidas ao longo deste trabalho. Sua prin-

cipal característica é oferecer uma medida da semelhança entre dois vetores de coeficientes LPC ou, em outras palavras, uma medida da probabilidade de que o vetor de coeficientes b_k, obtido de um sinal degradado ou processado, tenha sido computado a partir do segmento real de fala com os coeficientes LPC verdadeiros a_k.

O desenvolvimento do equacionamento da distância de Itakura foge ao escopo deste trabalho. Todavia, é importante levar em conta a análise quantitativa do fator de dissemelhança D [8]. O fator de dissemelhança D é obtido a partir da distância de Itakura e da duração efetiva do segmento de sinal que está sendo analisado, da seguinte forma:

$$D = d.N_{eff}, \tag{2.12}$$

onde d representa a distância de Itakura e N_{eff} representa a duração efetiva do segmento de sinal analisado. Dada uma janela de sinal $w(n)$, de tamanho N, a duração efetiva do segmento, N_{eff}, pode ser obtida de acordo com:

$$N_{eff} = \frac{1}{N} \sum_{n=0}^{N-1} w^2(n). \tag{2.13}$$

Para uma janela de Hamming, por exemplo, $N_{eff} = 0,3975N$.

Estudos realizados [8] indicam que, para quadros de duração de 25 a 35 ms ($N = 256$, por exemplo), o valor de $D = 24$ representa o limiar de significância estatística, isto é, quadros de fala que geram um fator de dissemelhança D maior que 24 podem ser considerados estatisticamente diferentes. Indo além, o valor de $D = 72$ é definido como sendo o limiar de significância perceptual, isto é, quadros de fala que produzem um fator de dissemelhança D maior que 72 podem ser considerados perceptualmente distintos. A Figura 2.2 ilustra os valores do fator de dissemelhança D obtidos para três pares de sinais de fala. Os gráficos mostram a envoltória espectral dos sinais (obtidos pela análise LPC) e seus respectivos fatores de dissemelhança.

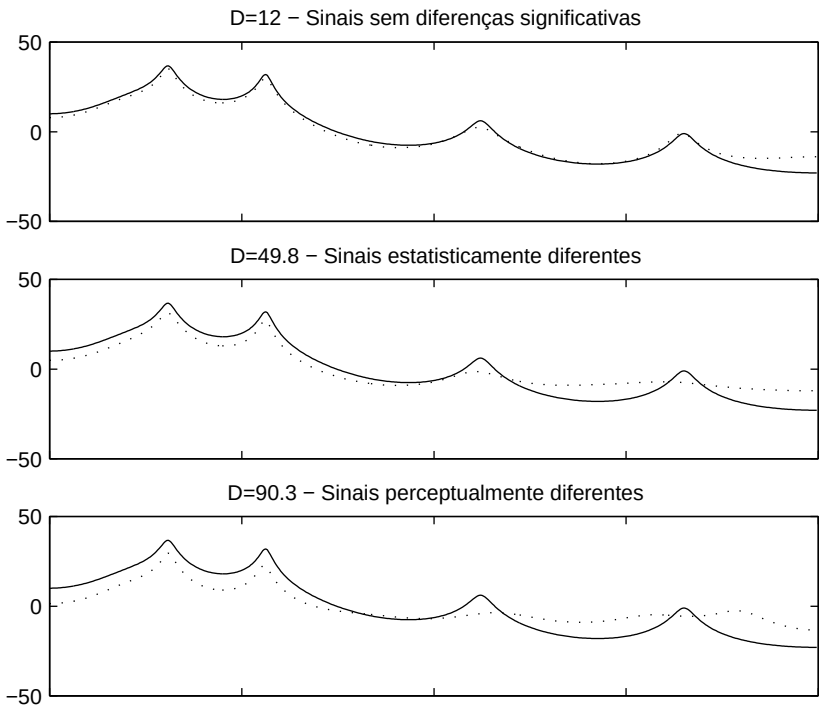

Figura 2.2: Ilustração da interpretação do fator de dissemelhança D. No gráfico superior são mostrados os envelopes espectrais de dois sinais que não apresentam diferenças significativas. No gráfico intermediário, são apresentados dois sinais cujas diferenças são estatisticamente significativas. No gráfico inferior, são apresentados sinais cujas diferenças são perceptualmente significativas.

Medidas subjetivas

Apesar de eficientes no que tange à característica perceptual dos sinais, as medidas objetivas nem sempre são capazes de identificar fatores importantes que diferenciam sinais de fala. Para tanto, é importante a análise subjetiva dos sinais processados. Neste trabalho, procurou-se avaliar a qualidade do sinal processado em termos de clareza, inteligilibilidade, cansaço do ouvinte, presença de ruído musical e outros efeitos indesejados, melhoria em comparação com o sinal obtido por subtração espectral e eficiência na remoção do ruído.

2.6 Implementação e análise de resultados

A primeira tarefa realizada foi o estudo e implementação da subtração espectral. Os objetivos desta fase foram a verificação das características, vantagens e desvantagens do método, bem como os efeitos das modificações sugeridas no resultado final do algoritmo. O estudo da técnica de subtração espectral é importante pois o sinal produzido através deste método será combinado ao sinal sintetizado para obtenção de um sinal de maior qualidade subjetiva. Desta forma, é necessário obter-se o melhor resultado possível para esta técnica, de maneira a contribuir positivamente no sinal final.

2.6.1 Sinais utilizados

Como o trabalho envolve a redução de ruído em sistemas de comunicação (em especial a telefonia), foram utilizados sinais de fala amostrados a 8 kHz, 16 bits. Para facilidade de comparação e padronização da análise, os mesmos sinais foram utilizados em todas as demais etapas do trabalho.

O material de fala utilizado nas simulações foi extraído da base de dados TIMIT, que é composta de um conjunto de frases foneticamente balanceadas, lidas por diversos locutores e criadas especificamente para o desenvolvimento e teste de aplicações de processamento de fala. Apesar de faladas em língua inglesa, a utilização deste conjunto de frases é interessante por permitir a comparação com outros estudos realizados na área, tendo em vista sua larga utilização em avaliações de algoritmos de processamento de fala. Durante o desenvolvimento da tese, foram utilizadas frases de três locutores, totalizando 15 sentenças, com duração média de 4 segundos cada.

Relação sinal-ruído

Para construir o sinal ruidoso, o sinal original foi somado a ruído branco. Para a determinação da potência do ruído, foi inicialmente calculada a potência do

sinal original e considerada a taxa de relação sinal-ruído desejada, conforme a formulação básica para o cálculo de SNR:

$$SNR = 10 \log_{10} \frac{\sum_{n} s^2(n)}{\sum_{n} \eta^2 (n)} \tag{2.14}$$

onde $s(n)$ representa o sinal original e $\eta(n)$ o ruído branco adicionado. Tal formulação, porém, é considerada uma estimativa pobre da qualidade do sinal de fala degradado por distorções e ruído [9, 10], por não estar relacionada a nenhum aspecto subjetivo da qualidade do sinal de fala. Uma medida de qualidade mais efetiva pode ser obtida se a relação sinal-ruído for avaliada através da média dos valores calculados para segmentos curtos de sinal. Como a energia dos quadros do sinal de fala é variante no tempo, tal estratégia contribui para equilibrar os pesos atribuídos aos trechos de maior e menor intensidade do sinal de fala. A SNR segmentada [11] é definida como:

$$SNR_{seg} = \frac{1}{M} \sum_{j=0}^{M-1} 10 \log_{10} \left[\frac{\sum_{n=m_j-N+1}^{m_j} s^2(n)}{\sum_{n=m_j-N+1}^{m_j} \eta^2 (n)} \right], \tag{2.15}$$

onde m_j representa os limites de cada um dos M quadros, cada um de tamanho N. Para cada quadro (com 15 a 25 ms de duração), é computada uma SNR local e o resultado final é obtido tomando-se a média dos valores individuais. Costumam ser utilizados limites inferiores e superiores para a SNR local, tipicamente 0 dB e 35 dB, isto é, se a SNR local for maior do que 35 dB, será feita igual a 35 dB e, se for menor do que 0 dB, será feita igual a 0 dB.

Considerando-se a medida de relação sinal-ruído tradicional (não-segmentada), o sinal ruidoso, utilizado para os testes do sistema completo de melhoria da qualidade do sinal de fala, apresenta SNR=10 dB. No entanto, utilizando-se a SNR segmentada, o sinal de teste apresenta SNR_{seg}=3,4 dB. Para os testes específicos dos blocos do sistema e avaliação dos algoritmos, foram também utilizados sinais com SNR_{seg} de 5 e 10 dB.

2.6.2 Análise da subtração espectral

A implementação da técnica de subtração espectral foi realizada a partir da Equação 2.11 (com $\gamma = 2$) e das modificações discutidas ao longo do texto. Para ruído branco, temos:

$$E\left\{\left|\aleph\left(e^{jw}\right)\right|^2\right\} = \sigma^2 \sum_{n=0}^{M-1} w^2(n) = \overline{M}\sigma^2, \tag{2.16}$$

onde M representa o número de amostras por quadro, $w(n)$ representa a janela de sinal utilizada e σ^2 representa a potência do ruído (estimada em quadros de silêncio). Para facilidade de notação, pode-se utilizar a grandeza $\overline{M}$ (número efetivo de amostras por quadro), que leva em conta o formato da janela de sinal aplicada (para janela retangular, $\overline{M} = M$; para janela de Hamming, $\overline{M} = 0,3975M$). Com essas alterações, a obtenção dos coeficientes $H(k)$ foi realizada utilizando-se a seguinte expressão:

$$H(k) = 1 - \frac{\alpha\overline{M}\sigma^2}{|X(k)|^2}, \tag{2.17}$$

onde $\overline{M}$ representa o número efetivo de amostras por quadro; $X(k)$ descreve o espectro do sinal ruidoso, obtido a partir da transformada de Fourier do sinal $x(n)$, multiplicado pela janela de sinal; e α descreve o fator de sobre-estimação do ruído.

A suavização dos coeficientes $H(k)$ foi obtida de acordo com:

$$H(k)_n = \lambda H(k)_{n-1} + (1-\lambda)\left[1 - \frac{\alpha\overline{M}\sigma^2}{|X(k)|^2}\right], \tag{2.18}$$

onde λ representa o coeficiente de suavização (variando de 0 a 1), e $H(k)_{n-1}$ representa o valor do coeficiente no quadro $n-1$.

Além da suavização, o conceito do Limite Espectral Mínimo também foi utilizado. Todo valor de $H(k)$ inferior a $H_{\min}$ é substituído por $H_{\min}$.

Coeficientes e fatores de ajuste

Durante os testes, foi possível observar o compromisso entre os diversos fatores que podem ser ajustados na subtração espectral, evidenciando as seguintes características:

1. a não utilização dos modificadores (isto é, considerando o fator de sobre-estimação $\alpha = 1$, o fator de suavização $\lambda = 0$ e o limite $H_{\text{min}} = 0$) produz como resultado um sinal com uma considerável taxa de ruído musical, que o torna incômodo aos ouvintes.

2. a aplicação do fator de sobre-estimação α reduz o ruído musical, porém introduz uma distorção espectral no sinal, produzindo uma sensação de abafamento na voz.

3. a utilização do fator de suavização λ também reduz o ruído musical, mas valores elevados (acima de 0,8) passam a introduzir uma certa reverberação no sinal.

4. a utilização do limite H_{min} contribui para a substituição do ruído musical por um ruído branco de nível reduzido, o que torna o sinal resultante mais natural.

A partir dessas observações, foi realizado um processo de ajuste conjunto desses fatores, de maneira a se obter o melhor resultado possível, com ruído musical reduzido e baixa distorção espectral. Os valores $H_{\text{min}} = 0,05$, $\alpha = 8$ e $\lambda = 0,6$ foram os que ofereceram o melhor resultado.

Capítulo 3

Modelo digital do sinal de fala

O processo de melhoria da qualidade do sinal de fala discutido neste trabalho envolve a soma ponderada de um sinal obtido por subtração espectral e um sinal sintetizado. Desta forma, é importante analisar-se o modelo digital do sinal de fala utilizado no processo de síntese. O estudo das características fisiológicas do aparelho fonador humano e suas implicações no processo de produção da fala permitem uma modelagem muito mais eficiente do sistema, bem como uma maior capacidade de reproduzir os detalhes envolvidos na geração da voz.

3.1 Mecanismos de produção da fala

O aparelho fonador humano é capaz de produzir uma variada gama de sons, com diferentes nuances, entonações e detalhes. Tal capacidade advém da grande flexibilidade do trato vocal (conjunto de órgãos responsáveis pela produção da voz) que, de acordo com seu posicionamento e movimentação, permite a geração de diversos tipos de sons. De modo simplificado, o trato vocal assemelha-se a um tubo, que se inicia na glote (abertura entre as cordas vocais), passa pela faringe (conexão entre o esôfago e a boca), pela cavidade oral, e termina nos lábios, com um comprimento médio de 17cm em um homem adulto. Cada uma das regiões apresenta uma diferente secção transversal, que pode variar de zero (totalmente

fechada) até cerca de 20cm^2. O fluxo de ar, vindo dos pulmões, pode ser interrompido por uma vibração constante da glote (que produz os fonemas sonoros) ou passar pelo trato vocal, gerando turbulência (que produz os fonemas surdos, ou não-sonoros)[12]. Diante dessas características, já é possível identificar alguns elementos do modelo de produção de fala: o trato vocal, que age como um tubo ressonante (semelhante ao de instrumentos de sopro) e a excitação (fluxo de ar vindo dos pulmões). Numa primeira análise, os sons produzidos pelo aparelho fonador humano são divididos em apenas duas classes principais: sonoros (produzidos pela passagem de ar através da glote, com a tensão das cordas vocais ajustada de maneira a vibrar numa certa freqüência, gerando pulsos quase periódicos de ar que excitam o trato vocal) e não-sonoros, ou surdos (gerados através da produção de uma região de constrição - usualmente na boca - através da qual o ar é forçado a passar numa velocidade suficiente para produzir turbulência). As vogais, os sons nasais e os ditongos são exemplos de fonemas sonoros, enquanto que algumas consoantes (s, x, p, t) são exemplos de fonemas não-sonoros. Algumas consoantes mesclam características dos dois tipos (z e g, por exemplo).

O trato vocal, por assemelhar-se a um conjunto de tubos, apresenta ressonâncias em freqüências determinadas pelas relações entre as dimensões das diversas seções transversais tomadas ao longo de sua extensão. Essas freqüências de ressonância são chamadas de formantes, e permitem diferenciar os diferentes sons produzidos, de acordo com o posicionamento da língua e da boca (Figuras 3.1 e 3.2).

A excitação, no caso de fonemas sonoros, tem a forma de uma sequência quase periódica de pulsos, espaçados de um intervalo denominado de período de pitch. O espectro da excitação, portanto, é composto de raias espaçadas de um intervalo que é o inverso do período de pitch. No caso dos trechos não-sonoros, a excitação toma a forma de um fluxo de ar turbulento, que apresenta características de sinal ruidoso. Por isso, considera-se que a excitação de trechos não-sonoros é composta de ruído branco, de envoltória espectral também praticamente plana. Ao atravessar o trato vocal, o espectro da excitação é moldado de acordo com as

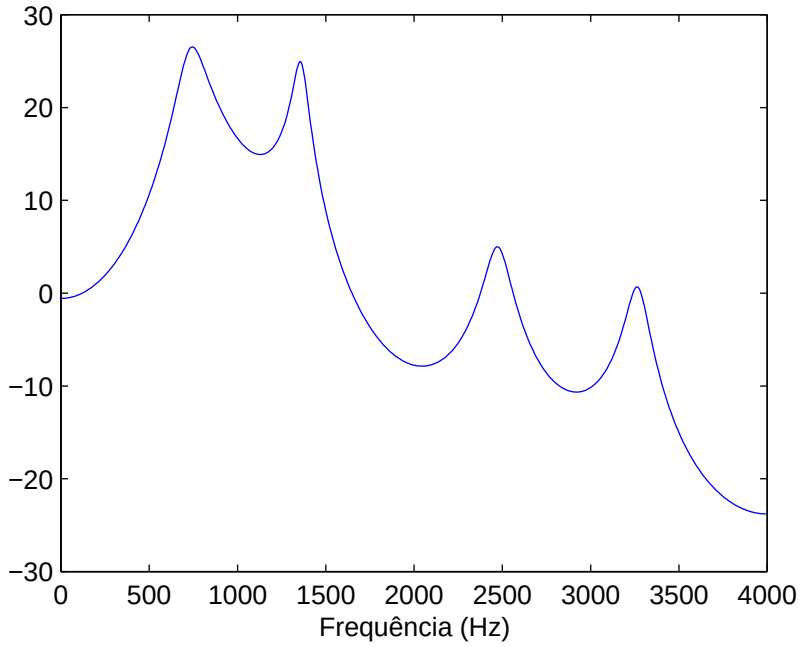

Figura 3.1: Ressonâncias do trato vocal para a vogal /a/ da língua portuguesa.

ressonâncias (freqüências formantes), resultando no sinal de fala que conhecemos.

3.2 Analogia com filtros digitais

A análise da propagação da excitação ao longo dos tubos do trato vocal está além do escopo deste trabalho. Todavia, as características dessa propagação permitem a associação do comportamento do trato vocal a modelos auto-regressivos, que nos conduzem a filtros IIR variantes no tempo, compostos apenas por pólos. O efeito combinado dos pólos desses filtros modela as ressonâncias (freqüências formantes) do trato vocal. Apesar de haver algumas classes de sons que teoricamente exigiriam zeros para os modelos (sons nasais, por exemplo), o efeito de um zero na função de transferência pode ser obtido através do acréscimo de mais pólos [5]. Em geral, os sistemas de análise e síntese de sinal realizam a obtenção dos coeficientes do filtro através de métodos de predição linear (LPC), que também

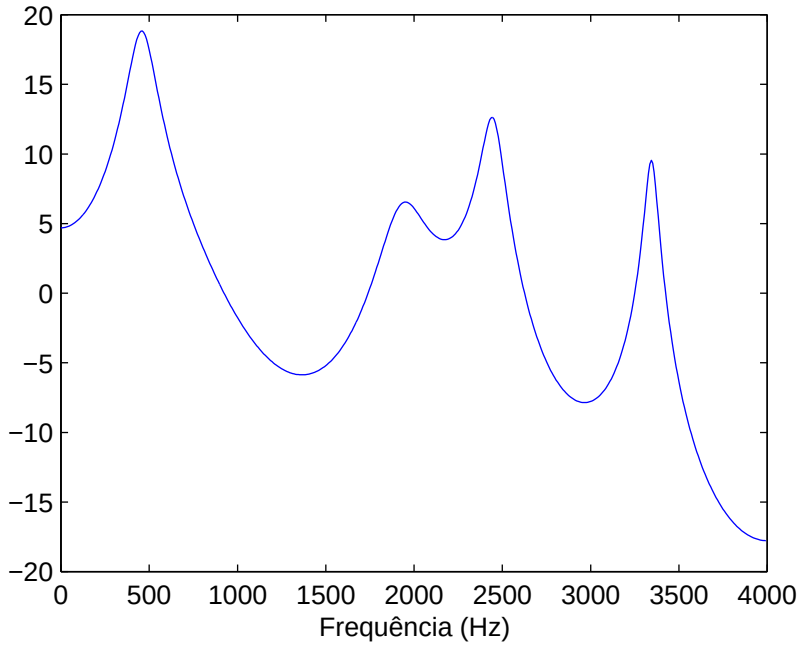

Figura 3.2: Ressonâncias do trato vocal para a vogal /e/ da língua portuguesa.

serão utilizados neste trabalho. O modelo digital para a produção de fala, portanto, leva em conta a existência da excitação (trem de pulsos ou ruído branco, dependendo da classe de som), ganho e filtro IIR (coeficientes LPC), de acordo com o esquema ilustrado na Figura 3.3

3.3 Estimação de parâmetros

Diante do exposto, surge a necessidade da correta obtenção dos parâmetros que serão utilizados pelo sintetizador para a produção da fala. Em aplicações de síntese a partir de texto (Text-to-Speech), os parâmetros podem ser criados tomando-se por base informações linguísticas e fonéticas (em sistemas de síntese por regra), ou obtidos em condições praticamente ideais (sem ruído) do sinal de fala do locutor que oferece sua voz ao sistema (síntese concatenativa). Em aplicações de codificação de voz e transmissão digital, porém, tais parâmetros

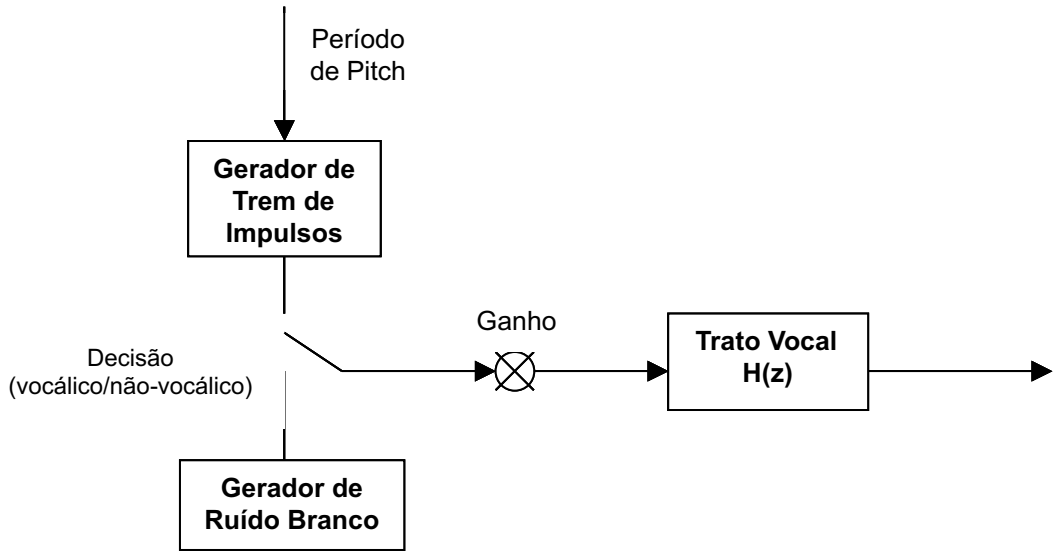

Figura 3.3: Diagrama em blocos do modelo digital de produção de fala.

são obtidos, em geral, de sinais de fala sujeitos a ruído, que devem ser corretamente reproduzidos após a transmissão. Na atividade proposta, o principal objetivo é a estimação dos parâmetros do sintetizador (ganho, período de pitch, coeficientes LPC e tipo de excitação) a partir do sinal ruidoso. Por isso, especial atenção deve ser dada às características das técnicas de estimação com sinais ruidosos.

3.4 Implementação inicial e análise de resultados

Durante as etapas iniciais do desenvolvimento deste trabalho, foram realizados testes para verificar as características, vantagens e desvantagens do modelo de produção de fala considerado. Foram utilizados alguns sinais limpos, isto é, sem adição de ruído, e processos clássicos para estimação dos principais parâmetros (ganho, período de pitch e coeficientes LPC).

Para análise e síntese do sinal, foram utilizados, inicialmente, quadros de 256 amostras, sem superposição, para uma primeira observação de resultados. Ao longo do trabalho, foi acrescentada a superposição de quadros, com janelamento, o que traz benefícios ao processo de análise e síntese. Para a estimação dos coeficientes LPC, em geral são utilizados polinômios de ordem 12, 14 ou 16. Nas simulações foram usados polinômios de ordem 16, devido a uma melhor qualidade subjetiva observada nos testes, principalmente em baixas relações sinal-ruído. Além disso, foi observado também que, em muitos casos, mais de um par de pólos contribui para a modelagem de um único formante, melhorando sua estimativa. Para a estimação de pitch, foi implementada a versão simplificada do algoritmo SIFT (a ser discutido em detalhes no capítulo relativo à estimação de pitch).

Para comparação e análise da qualidade dos resultados, foram tomadas algumas vogais isoladas, trechos de consoantes e palavras simples, que permitiram verificar o comportamento do sintetizador. A utilização de palavras e frases completas exige um tratamento específico da variação dos parâmetros (discutido no capítulo referente aos aperfeiçoamentos do processo de síntese de sinal de fala). Um dos primeiros problemas que surgem na utilização do modelo é o tratamento das variações dos parâmetros. Como discutido anteriormente, todos os parâmetros participantes do sintetizador são variantes no tempo, o que torna importante a análise do compromisso entre a qualidade da estimativa e a suavidade das transições. Quando quadros maiores são utilizados, é possível obter, por exemplo, uma melhor estimativa dos coeficientes LPC para uma vogal sustentada, ou mesmo uma melhor estimativa do pitch para uma região cuja freqüência fundamental se mantém constante. Entretanto, as rápidas variações da fala são facilmente perdidas quando se utilizam quadros grandes (512 amostras, por exemplo). Em contrapartida, a utilização de quadros pequenos (64 ou 128 amostras, por exemplo), provoca erros maiores nas estimativas. Durante os testes, a melhor relação de compromisso foi obtida com quadros de 256 amostras (32 milissegundos, para uma freqüência de amostragem de 8 kHz).

Para produzir a excitação, foi utilizado ruído branco para quadros não-sonoros

e pulsos igualmente espaçados para quadros sonoros. (No caso da excitação impulsiva, foi utilizado um algoritmo simples para manter a continuidade do espaçamento entre os pulsos, pois as descontinuidades no vetor de excitação provocam efeitos desagradáveis no sinal sintetizado).

Os testes iniciais demonstraram que o sintetizador apresenta um bom desempenho quando os parâmetros são obtidos diretamente do sinal limpo (sem ruído). Entretanto, ficaram evidentes alguns pontos sensíveis (que foram tratados com mais cuidado ao longo do desenvolvimento do trabalho), como por exemplo: a necessidade de suavizar as variações do ganho entre quadros; a necessidade de equalizar os ganhos da excitação ruidosa e impulsiva; o problema da falha na decisão entre quadros sonoros e não-sonoros, realizada pelo estimador de pitch; a necessidade de suavizar o vetor de excitação de modo a acomodar variações grandes de período de pitch entre quadros; o tratamento dos estados internos do filtro LPC de síntese, que é variante no tempo. Tais aperfeiçoamentos serão discutidos no capítulo referente à sintese do sinal de fala.

Capítulo 4

Estimação dos coeficientes LPC

4.1 Histórico e características

O conhecimento das características do sinal a ser analisado traz importantes benefícios à estimação de seu conteúdo espectral. De acordo com o tipo do sinal, modelos específicos podem ser aplicados, com o objetivo de extrair com a máxima precisão as informações de interesse, inclusive na presença de ruído. No caso dos sinais de fala, o modelo auto-regressivo, com parâmetros obtidos através da técnica de predição linear [5], tem sido uma das principais ferramentas de estimação do espectro, com significativas vantagens quando comparado a outras abordagens, sendo utilizado inclusive como padrão em sistemas de comunicação digital [13]. Pelas próprias características do método, a estimação LPC tende a ser mais robusta na presença do ruído do que métodos não-paramétricos, por ter sua formulação baseada nas características específicas dos sinais de fala. Entretanto, a estimação LPC tende a impor ao ruído características de fala, isto é, pode considerar que o ruído também contém informação de voz, produzindo efeitos indesejados. Por isso, diversos métodos alternativos têm sido propostos, visando aprimorar a estimativa LPC em sinais contaminados por ruído, levando-se em conta que os efeitos de alargamento de banda dos formantes principais e perda de magnitude nos formantes de ordem superior dificultam a utilização direta da

análise LPC em sinais ruidosos [14].

Neste trabalho, os coeficientes LPC são utilizados no bloco responsável pela síntese de um sinal que será somado ao sinal $\hat{s}(n)$, obtido por subtração espectral. A estimação, porém, será feita a partir de um sinal contaminado por ruído intenso. Sendo assim, vários métodos foram analisados, com o objetivo de buscar o melhor procedimento de estimação para aplicação no sintetizador.

4.2 Métodos de cálculo

O conceito principal envolvido na Predição Linear é a possibilidade de que uma amostra de fala possa ser aproximada por uma combinação linear de amostras passadas. Através da minimização do erro quadrático entre as amostras reais e as amostras preditas linearmente, é possível obter-se um conjunto de coeficientes preditores, também chamados de coeficientes LPC.

$$\tilde{s}_{pred}(n) = \sum_{k=1}^{p} \alpha_k s(n-k), \tag{4.1}$$

onde $s(n-k)$ representa as amostras passadas, α_k representa os coeficientes LPC (de ordem p) e $\tilde{s}_{pred}(n)$ representa a amostra predita.

A técnica de predição linear está intimamente ligada ao modelo de produção de fala, tratado anteriormente. O filtro $H_{LPC}(z)$, cuja resposta em freqüência está associada às ressonâncias do trato vocal, está diretamente relacionado aos coeficientes preditores α_k, de acordo com a expressão:

$$H_{LPC}(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}. \tag{4.2}$$

O cálculo dos coeficientes α_k envolve a minimização do erro quadrático de predição. O erro de predição, $e(n)$, é definido como:

$$e(n) = s(n) - \tilde{s}_{pred}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k). \tag{4.3}$$

Dado um quadro de sinal de fala, o objetivo será, portanto, encontrar os coeficientes preditores α_k que minimizem o erro quadrático de predição de curto-prazo, dado por:

$$E = \sum_n e^2(n). \tag{4.4}$$

O desenvolvimento completo do equacionamento da predição linear está além do escopo deste trabalho. O que se propõe, neste momento, é uma visão geral do problema, bem como uma análise inicial de dois métodos muito utilizados para cálculo dos coeficientes: o Método da Auto-correlação e o Método da Covariância.

Opcionalmente, pode-se definir o erro quadrático médio de predição de curto prazo, dividindo-se E pelo tamanho do segmento sob análise. Entretanto, tal constante é irrelevante para o equacionamento, de modo que será utilizada a forma definida em 4.4.

Substituindo-se 4.3 em 4.4, e desenvolvendo as equações, chegamos a:

$$E = \sum_n \left\{ s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right\}^2, \tag{4.5}$$

ou

$$E = \sum_n s^2(n) - 2\sum_{k=1}^{p} \alpha_k \sum_n s(n)s(n-k) + \sum_n \left\{ \sum_{k=1}^{p} \alpha_k s(n-k) \right\}^2. \tag{4.6}$$

Para minimizar o erro quadrático de predição E, pode-se fazer $\frac{\partial E}{\partial \alpha_i} = 0$, para $1 \leq i \leq p$:

$$\frac{\partial E}{\partial \alpha_i} = 0 = -2\sum_n s(n)s(n-i) + 2\sum_n \left\{ \sum_{k=1}^{p} \alpha_k s(n-k) \right\} s(n-i). \tag{4.7}$$

Rearranjando os termos de 4.7, temos:

$$\sum_n s(n)s(n-i) = \sum_{k=1}^{p} \alpha_k \left(\sum_n s(n-k)s(n-i) \right), \tag{4.8}$$

ou

$$c(i,0) = \sum_{k=1}^{p} \alpha_k c(k,i). \tag{4.9}$$

Esta equação é conhecida como equação de predição linear, ou equação de Yule-Walker. Numerando as equações internas para cada valor de i, ela pode ser expressa na forma matricial:

$$\overline{c} = \underline{C}\overline{\alpha}, \tag{4.10}$$

onde:

$$\overline{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_p \end{bmatrix} \quad \underline{C} = \begin{bmatrix} c(1,1) & c(1,2) & ... & c(1,p) \\ c(2,1) & c(2,2) & ... & c(2,p) \\ ... & ... & ... & ... \\ c(p,1) & c(p,2) & ... & c(p,p) \end{bmatrix} \quad \overline{c} = \begin{bmatrix} c(1,0) \\ c(2,0) \\ ... \\ c(p,0) \end{bmatrix}$$

A solução do conjunto de equações envolve, portanto, uma inversão de matrizes, de modo a obtermos:

$$\overline{\alpha} = \underline{C}^{-1}\overline{c}, \tag{4.11}$$

que é conhecido como o Método da Covariância. Como a matriz de covariância é simétrica, há maneiras simplificadas de fazer sua inversão. Uma possibilidade é a utilização da decomposição de Cholesky, em que a matriz de covariância é fatorada em matrizes triangulares superiores e inferiores. Tal desenvolvimento, entretanto, não será descrito neste trabalho.

A partir de uma outra interpretação dos limites da minimização do erro quadrático de predição (isto é, forçando a existência de dados apenas dentro do quadro de sinal a ser utilizado), é possível calcular a solução da equação de predição linear através do Método da Auto-correlação:

$$\overline{\alpha} = \underline{R}^{-1}\overline{r}, \tag{4.12}$$

onde:

$$\overline{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_p \end{bmatrix} \quad \underline{R} = \begin{bmatrix} r(0) & r(1) & ... & r(p-1) \\ r(1) & r(0) & ... & r(p-2) \\ ... & ... & ... & ... \\ r(p-1) & r(p-2) & ... & r(0) \end{bmatrix} \quad \overline{r} = \begin{bmatrix} r(1) \\ r(2) \\ ... \\ r(p) \end{bmatrix}$$

Conforme discutido no início do tópico, o desenvolvimento dessas equações não está no escopo deste trabalho. Entretanto, é importante observar que $\underline{R}$ é simétrica, e todos os elementos ao longo da diagonal principal são iguais. Desta forma, sua matriz inversa sempre existe, e as raízes estão localizadas no interior do círculo unitário do plano Z, o que garante que o filtro resultante será sempre estável. Além disso, a forma Toeplitz da matriz (elementos constantes ao longo das diagonais) permite uma forma de solução muito eficiente, o algoritmo de Levinson-Durbin.

4.3 Estimação LPC em sinais ruidosos

A primeira abordagem para o problema é a estimação direta dos parâmetros LPC a partir do sinal ruidoso. Conforme discutido na literatura, o método extrai uma boa informação sobre os picos espectrais, todavia, a estimativa fica bastante comprometida pela presença do ruído ([5, 14, 15]). Em geral, podem ser observados desvios na posição dos formantes, perda de formantes de menor intensidade e uma baixa definição dos formantes, isto é, a amplitude dos formantes é menor do que no sinal original, conforme ilustra a Figura 4.1.

4.3.1 LMAP - Processo iterativo

Proposto por Lim e Oppenheim [16], o algoritmo LMAP utiliza-se de um processo iterativo para realizar a estimativa LPC de um sinal de fala degradado por ruído.

O método iterativo procura resolver a estimativa MAP (Maximum A-Posteriori) dos parâmetros envolvidos na produção da fala. Quando se tem à disposição apenas a observação do sinal degradado, as equações para a resolução do estimador

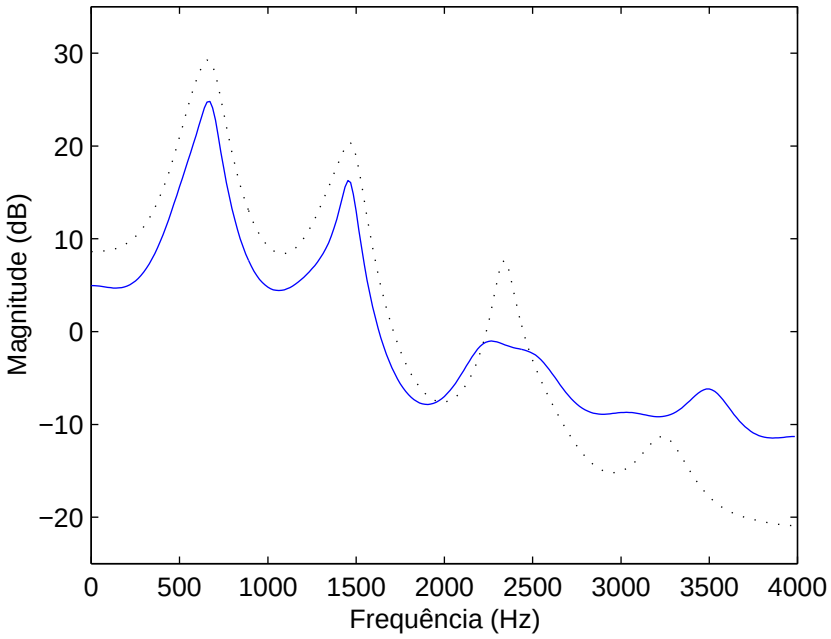

Figura 4.1: Ilustração do erro cometido na estimativa LPC de um sinal ruidoso. (Sinal sem ruído em linha pontilhada e sinal ruidoso em linha cheia). SNR segmentada = 5 dB.

MAP tornam-se não-lineares e de difícil resolução. Para contornar o problema, propõe-se a solução através da aplicação de um procedimento iterativo, que requer apenas a solução de um conjunto de equações lineares.

O algoritmo iterativo, denominado MAP Linearizado (LMAP), inicia-se com uma estimativa inicial dos parâmetros LPC, a_k. Após a estimação inicial (que pode ser realizada analisando-se o sinal ruidoso ou utilizando os resultados do quadro anterior), o sinal observado é filtrado por um filtro ótimo, $H_o(e^{jw})$, obtido a partir dos parâmetros LPC estimados inicialmente (coeficientes a_k e ganho g). A filtragem é realizada através da aplicação dos coeficientes do filtro $H_o(e^{jw})$ à FFT do quadro sob análise. O filtro $H_o(e^{jw})$ é obtido de acordo com a expressão:

$$H_o(e^{jw}) = \frac{P_s(e^{jw})}{P_s(e^{jw}) + \sigma_d^2}, \tag{4.13}$$

onde σ_d^2 representa a potência do ruído. $P_s(e^{jw})$ representa a densidade espectral de potência do sinal limpo, $s(n)$, e é estimada, a cada iteração, por:

$$P_s(e^{jw}) = \frac{g^2}{\left|1 - \sum_{k=1}^{p} a_k e^{-jkw}\right|^2}, \tag{4.14}$$

onde g representa o ganho do quadro, a_k representa os coeficientes LPC e p representa a ordem do filtro LPC utilizado.

O sinal observado, filtrado pelo filtro ótimo, $H_o(e^{jw})$, é o ponto de partida para a próxima iteração. Os autores sugerem a realização de 3 ou 4 iterações, no máximo. Um número maior de iterações provoca um estreitamento da banda dos formantes, prejudicando o resultado final do algoritmo.

Nos testes realizados, foi possível observar que o algoritmo apresenta um bom desempenho para quadros não-sonoros (cuja excitação é composta por ruído branco) e um desempenho inferior para quadros sonoros (cuja excitação é composta por trens de pulsos).

4.3.2 Melhorias no processo iterativo - uso de restrições

Apesar das características positivas do método iterativo, duas desvantagens motivaram o desenvolvimento de melhorias no procedimento: a redução da banda dos formantes e seu deslocamento em freqüência, quando iterações adicionais são utilizadas e a variação na posição dos pólos, de quadro a quadro, que contribuem para uma sensação de não-naturalidade do sinal processado [17]. A abordagem sugerida trabalha com restrições aplicadas internamente aos quadros (entre uma iteração e outra) e restrições aplicadas entre quadros adjacentes, de maneira a:

a) garantir uma trajetória de formantes mais próxima do real (os pólos não se apoximam demais da circunferência unitária, produzindo bandas passantes anormalmente estreitas);

b) reduzir a variação na posição dos pólos, entre um quadro e outro, isto é, evitar variações bruscas nas características relacionadas ao trato vocal;

c) garantir que os filtros correspondentes aos coeficientes LPC sejam sempre estáveis.

As restrições aplicadas internamente aos quadros envolvem, basicamente, o controle da movimentação dos pólos (para que não se aproximem demais da circunferência unitária) e o controle da movimentação dos coeficientes de autocorrelação, para cada iteração realizada. Entre quadros adjacentes, é realizada a suavização dos parâmetros espectrais. Como não é adequada a aplicação da suavização diretamente sobre os coeficientes LPC, outras representações espectrais são utilizadas, dentre as quais destaca-se a representação através dos coeficientes LSP (Line Spectrum Pairs) [18]. Os coeficientes LSP apresentam características adequadas para a interpolação, inclusive com a garantia de que o filtro LPC obtido a partir da transformação inversa será estável.

Coeficientes LSP

Os coeficientes LSP são obtidos a partir da decomposição do polinômio $A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$ em outros dois polinômios, $P(z)$ e $Q(z)$, de forma que:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{4.15}$$

e

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}). \tag{4.16}$$

Os dois polinômios obtidos, de ordem $p + 1$, possuem suas raízes na circunferência unitária. Os ângulos das raízes são chamados de Line Spectrum Pairs. Se as raízes de $P(z)$ e $Q(z)$ estiverem intercaladas, as raízes de $A(z)$ estarão dentro do círculo unitário. Os ângulos das raízes de $P(z)$ correspondem, aproximadamente, aos ângulos das raízes de $A(z)$ (freqüências formantes), e a separação entre uma raiz específica de $P(z)$ e a raiz mais próxima de $Q(z)$ fornece uma idéia da largura da banda dessa ressonância. A partir dessas características, os coeficientes LSP podem ser utilizados na análise da movimentação dos pólos, bandas passantes dos formantes e estabilidade do filtro resultante.

4.3.3 Subtração da componente de ruído na auto-correlação

Outra possibilidade de tratamento do problema é a compensação do ruído durante a fase de cálculo dos coeficientes, isto é, a subtração da componente de ruído da função de auto-correlação utilizada no algoritmo de obtenção dos coeficientes [19]. A função de auto-correlação de um sinal contaminado por ruído branco tem a seguinte característica:

$$R(i) = \begin{cases} R_s(i) + \sigma^2, & \text{para } i = 0 \\ R_s(i), & \text{para } i \neq 0 \end{cases} \tag{4.17}$$

Desta forma, os parâmetros LPC podem ser estimados subtraindo-se a potência do ruído σ_d^2 do primeiro termo da função de auto-correlação. Apesar da aparente eficácia, a simples subtração sugerida pode gerar um filtro instável, além de produzir variações quadro-a-quadro nos coeficientes gerados. Para contornar o problema, a subtração do ruído pode ser feita passo a passo, num processo iterativo, de maneira a garantir a estabilidade dos coeficientes resultantes. A quantidade de ruído a ser subtraído aumenta a cada iteração. Se o filtro resultante deixar de ser estável, o processo é interrompido.

A subtração da componente de ruído na função de auto-correlação guarda semelhanças com a técnica de subtração espectral usada como pré-processamento para estimação LPC (a ser discutida a seguir). Em ambos os casos, busca-se subtrair do sinal a componente devida ao ruído. Na subtração espectral, tal componente reflete-se num valor constante a ser subtraído da magnitude do espectro do sinal ruidoso. Na técnica baseada na função de auto-correlação, subtrai-se a potência do ruído de seu primeiro termo. Entretanto, a técnica da utilização da subtração espectral como pré-processamento não oferece o risco de produzir filtros LPC instáveis, pois a estimativa LPC é sempre realizada sobre um sinal real, obtido após o janelamento e retorno do espectro processado ao domínio do tempo.

4.3.4 Estimação após pré-processamento

Conforme descrito por Ahmed [3] e Kang e Fransen [20], o pré-processamento utilizando subtração espectral é ferramenta útil na estimação dos parâmetros LPC de sinais ruidosos. Avaliações objetivas e subjetivas demonstraram o benefício da técnica, com ruídos de diferentes tipos e intensidades.

Usualmente, a técnica de subtração espectral é utilizada como recurso único para a melhoria da qualidade de um sinal de fala degradado por ruído. O esforço de ajuste e calibração do método (envolvendo as modificações discutidas no capítulo referente à subtração espectral) volta-se totalmente para a qualidade do sinal resultante. Neste cenário, entretanto, é importante que seja observado o compromisso entre a quantidade de ruído subtraído do sinal, o ruído musical produzido e a qualidade do sinal processado. A utilização do valor correto da potência do ruído não elimina o ruído musical, devido à presença de senóides aleatórias no sinal reconstruído [2]. Tal efeito, além de trazer uma sensação de artificialidade ao sinal processado, também contribui para o cansaço do ouvinte. O aumento da quantidade de ruído subtraído do sinal (sobre-estimação do ruído) reduz o ruído musical, porém aumenta a degradação do sinal processado, que passa a ter uma característica mais "abafada", devido à atenuação ou mesmo eliminação dos formantes de ordem superior. Por esse comportamento, a subtração espectral aplicada isoladamente tornou-se pouco atrativa para aplicações de redução de ruído, e alternativas têm sido buscadas.

No caso da estimação de coeficientes LPC para posterior síntese de sinal, entretanto, o cenário é ligeiramente diferente. As senóides aleatórias do sinal processado, responsáveis pelo ruído musical, podem não possuir efeito tão pronunciado sobre os coeficientes LPC estimados, desde que tais variações não sejam de grande amplitude. Os testes realizados neste trabalho demonstraram que uma taxa reduzida de ruído musical - que já seria intolerável como sub-produto de um algoritmo isolado de redução de ruído - permite uma melhor estimativa de coeficientes LPC em comparação com a estimativa direta realizada sobre o sinal

ruidoso. O ponto ótimo de posicionamento da taxa de sobre-estimação espectral localiza-se, portanto, numa região que produziria algum ruído musical (se fosse considerado apenas o resultado isolado do método), mas que garante uma maior qualidade na estimativa do espectro. Desta forma, a subtração espectral com menor sobre-estimação de ruído surge como uma importante ferramenta de pré-processamento para estimação dos coeficientes LPC usados na síntese do sinal que será combinado ao sinal obtido por subtração espectral.

4.4 Implementação e análise de resultados

Para a avaliação das técnicas para estimação dos coeficientes LPC, foram realizados testes isolados (avaliados através das medidas de distância e da observação das respostas em freqüência dos filtros obtidos) e em conjunto com o sintetizador (cuja avaliação foi feita escutando-se os sinais sintetizados). Maior importância foi dada à avaliação subjetiva, principalmente no que diz respeito à percepção de melhoria na qualidade do sinal sintetizado.

A primeira característica observada na análise dos sinais sintetizados com parâmetros obtidos diretamente do sinal ruidoso foi a diferença perceptual provocada pela falta de definição dos formantes no espectro estimado. Conforme ilustrado na Figura 4.1, algumas regiões do espectro tendem a ficar planas quando a estimação é feita sobre um sinal ruidoso. Com isso, o espectro plano da excitação deixa de ser corretamente filtrado, causando uma sensação de zumbido no sinal sintetizado. Este é um dos principais problemas enfrentados na estimação LPC em sinais ruidosos.

Como alternativas para melhorar a estimativa, foram analisados os algoritmos LMAP e a técnica de pré-processamento para estimação LPC. O algoritmo LMAP foi implementado conforme discutido anteriormente, com restrições aplicadas através da suavização dos coeficientes (utilizando parâmetros LSP). A Figura 4.2 ilustra o efeito da aplicação do algoritmo sobre um quadro de sinal ruidoso.

Para o pré-processamento utilizando subtração espectral, foram analisados

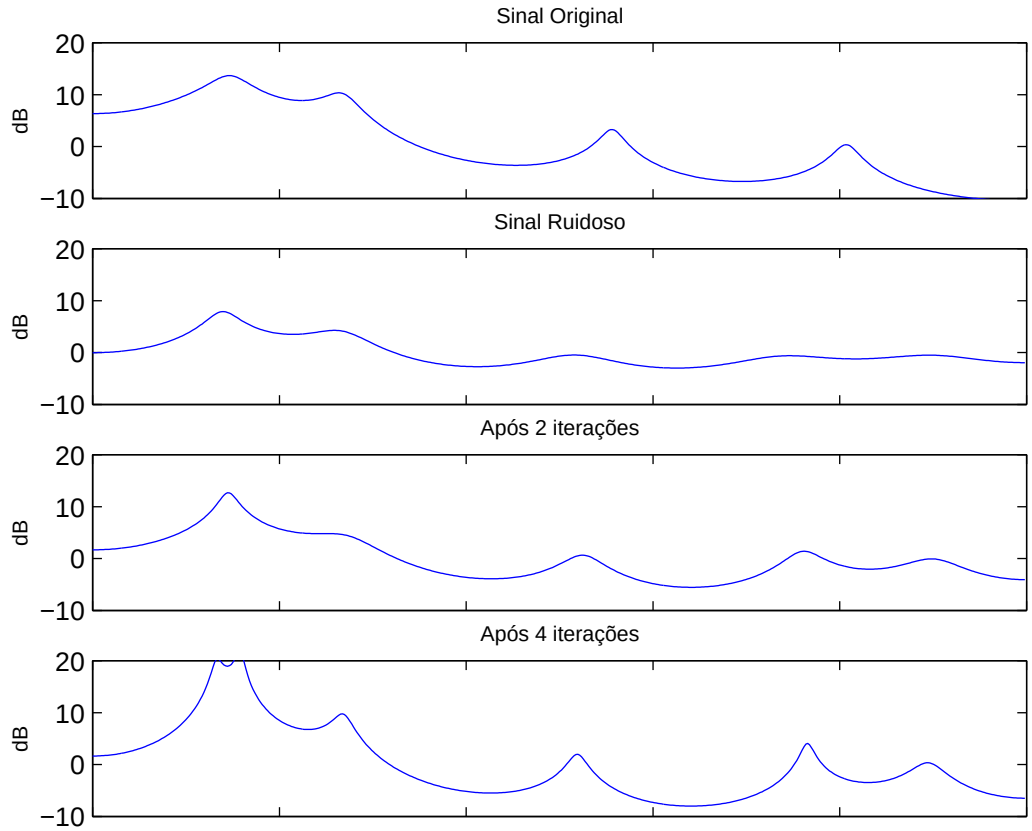

Figura 4.2: Variação na Resposta em Freqüência dos coeficientes LPC utilizando algoritmo LMAP. O gráfico mostra as Respostas em Freqüência dos coeficientes obtidos a partir do sinal original, do sinal ruidoso e após a aplicação do algoritmo LMAP (2 e 4 iterações).

os resultados objetivos e subjetivos com o intuito de ajustar os parâmetros do algoritmo. A análise objetiva, utilizando a distância de Itakura, indicou uma melhor proximidade espectral com o sinal original no caso de utilização de um fator de sobre-estimação de ruído α próximo de 2, e um fator de suavização de coeficientes $\lambda = 0,6$. Nos testes subjetivos (realizados através da avaliação de sinais sintetizados), esses valores dos parâmetros também foram considerados os mais adequados. A Figura 4.3 ilustra a variação do fator de dissemelhança

(obtido a partir da distância Itakura) médio em função da variação do fator de sobre-estimação de ruído α, para um sinal de fala com relação sinal-ruído segmentada de 5 dB. O resultado indica que realizar o pré-processamento com subtração espectral é sempre mais vantajoso do que não processar o sinal (caso em que $\alpha = 0$), e que o ponto ótimo é obtido para um valor de α menor do que o valor que se utilizaria para eliminar totalmente o ruído musical em aplicações nas quais a subtração espectral é o único recurso para redução do ruído ($\alpha = 8$, por exemplo).

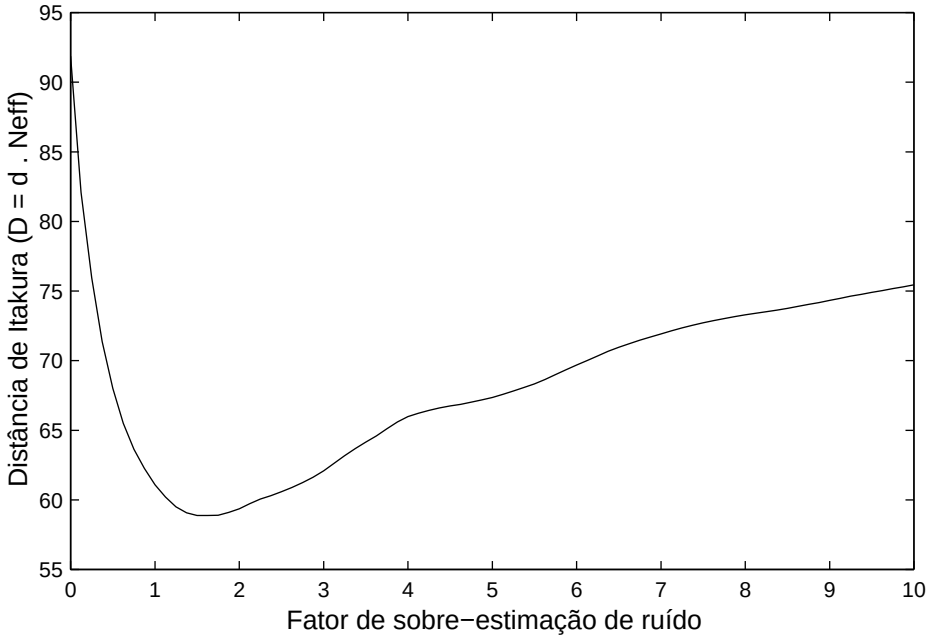

Figura 4.3: Ilustração da variação do fator de dissemelhança D (obtido a partir da distância de Itakura) em função do fator de sobre-estimação de ruído utilizado no pré-processamento por subtração espectral.

A técnica de estimação após pré-processamento apresentou melhores resultados subjetivos, por isso foi utilizada na estrutura completa do método.

Capítulo 5

Estimação do período de pitch

O período de pitch, como visto anteriormente, também possui fundamental importância dentro do modelo de produção de fala. No aparelho fonador humano, o período de pitch representa o intervalo de tempo entre os movimentos de fechamento da glote, que define a característica da excitação a ser aplicada ao trato vocal. Em termos perceptuais, a freqüência de pitch relaciona-se à sensação de vozes agudas e graves. No modelo digital de produção de fala, o período de pitch determina o intervalo entre os pulsos de excitação que serão aplicados ao filtro LPC de síntese. Da correta estimação do período de pitch depende a correta síntese da freqüência fundamental do sinal de fala.

5.1 Características dos algoritmos

Em geral, o algoritmo de estimação de pitch realiza duas funções: a análise e diferenciação de quadros sonoros e não-sonoros, e a estimação do período de pitch propriamente dito para os quadros classificados como sonoros. Para maior facilidade de utilização dessas informações dentro das rotinas de processamento de fala, costuma-se atribuir ao período de pitch o valor zero nos quadros não-sonoros. Desta forma, os demais blocos de síntese, cujo funcionamento depende da classificação do tipo de excitação dos quadros, podem facilmente extrair essa in-

formação (quadro sonoro ou não-sonoro) diretamente do valor estimado do pitch.

Da mesma forma que os demais parâmetros, o período de pitch também possui faixas de valores válidos, bem como restrições quanto à continuidade e taxa de variação entre quadros. Tais informações são utilizadas pelos algoritmos para validar e suavizar os valores estimados. Em alguns algoritmos, as restrições sobre os valores válidos de pitch servem inclusive como redutores de complexidade computacional, pois evitam cálculos desnecessários.

5.2 Algoritmos e técnicas de estimação

Nas últimas décadas foram propostas diversas técnicas para estimação de pitch, e várias análises sobre os resultados de implementação dessas técnicas foram publicadas, levando em conta a complexidade computacional, os erros de decisão sobre quadros sonoros/não-sonoros, os desvios na estimativa e a possibilidade de estimar-se a metade ou o dobro do valor correto para o pitch (um erro muito comum no processo de estimação) [21, 22, 13, 12]. De forma geral, os algoritmos buscam identificar as variações periódicas no sinal de fala e, ao mesmo tempo, calcular o nível de certeza que se tem de que o trecho de sinal realmente é periódico (que caracteriza um quadro sonoro).

Neste trabalho, foram feitas avaliações de várias técnicas de estimação de pitch, com o objetivo de verificar seu comportamento tanto em condições ideais (sinal limpo), quanto nas condições de utilização do algoritmo de melhoria do sinal de fala (sinal ruidoso). Além da qualidade da estimativa, foi avaliada também a carga computacional de cada algoritmo.

5.2.1 Estimação direta a partir da auto-correlação do sinal

Uma abordagem inicial para o problema consiste em analisar a periodicidade do sinal calculando sua função de auto-correlação. Para um sinal determinístico de tempo discreto, de energia finita, a auto-correlação é definida como:

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k). \tag{5.1}$$

Para um sinal periódico, ou aleatório, utilizamos:

$$\phi(k) = \lim_{N\to\infty} \frac{1}{(2N+1)} \sum_{m=-N}^{N} x(m)x(m+k). \tag{5.2}$$

De qualquer forma, fica evidente que, se o sinal for periódico, com período P, a função de auto-correlação obedecerá à seguinte propriedade:

$$\phi(k) = \phi(k+P). \tag{5.3}$$

Além disso, sabemos que a função de auto-correlação terá picos nas amostras $0, \pm P, \pm 2P$,.... Desta forma, é possível estimar-se a periodicidade de um sinal avaliando o segundo máximo de sua função de auto-correlação (o primeiro máximo da função de auto-correlação situa-se na origem). Para análise de sinais de fala, é importante considerar a função de auto-correlação de curto prazo, que leva em conta o janelamento do sinal. Aplicando-se o janelamento ao sinal, chegamos à seguinte expressão:

$$R_n(k) = \sum_{m=0}^{N-1-k} [x(n+m)w(m)][x(n+m+k)w(k+m)], \tag{5.4}$$

onde w representa a janela de sinal de N amostras a ser aplicada.

Para realizar a estimação do pitch, basta calcular-se a função de auto-correlação do quadro de sinal, e buscar o segundo máximo. Para reduzir a complexidade computacional, é possível realizar o cálculo apenas na faixa de interesse para valores de pitch. A decisão entre quadro sonoro e não-sonoro é realizada comparando-se o segundo máximo da auto-correlação com um limiar pré-definido. Se o valor estiver abaixo de tal limiar, o quadro é não-sonoro. O algoritmo é de fácil implementação e tem uma carga computacional relativamente baixa. No entanto, apresenta pouca robustez com o aumento do nível de ruído.

5.2.2 Algoritmo AMDF

O algoritmo AMDF (Average Magnitude Difference Function, ou Função Diferença de Magnitude Média) foi proposto por Ross et al. em 1974 [23], e permite a estimativa do período de pitch com baixa complexidade computacional em processadores de ponto fixo, por ser implementado apenas com operações de subtração, adição e valor absoluto, em contraste com as operações de adição e multiplicação da função de auto-correlação. A AMDF é definida como:

$$D_{AMDF}(k) = \sum_{n=0}^{N-k-1} |(s_w(n+k) - s_w(n)|, \tag{5.5}$$

onde $s_w(n)$ representa o sinal janelado.

Para obtenção do período de pitch, calcula-se $D_{AMDF}(k)$ para os valores válidos de pitch, e busca-se o primeiro valor de mínimo do resultado obtido. Para a decisão sobre o tipo de quadro (sonoro ou não-sonoro) pode-se analisar o valor do pico obtido em comparação com um limiar, ou a relação entre o valor de pico obtido e a média dos demais valores calculados em $D_{AMDF}(k)$. O algoritmo é de fácil implementação e baixa complexidade computacional, mas apresenta uma quantidade elevada de erros de estimação, que levam à metade ou o dobro do valor correto. Isso se deve principalmente à variação na quantidade de itens a serem somados na função $D_{AMDF}(k)$, conforme a variação de k, o que dificulta a identificação do verdadeiro pico na curva resultante.

5.2.3 Algoritmo CAMDF

Como alternativa ao método anterior, o algoritmo CAMDF (Circular Average Magnitude Difference Function, ou Função Diferença de Magnitude Média Circular), [24], busca corrigir alguns problemas que surgem da utilização do AMDF. A principal vantagem do CAMDF é a não ocorrência da variação da quantidade de valores participantes da soma em função das diferentes distâncias consideradas no cálculo, devido à característica circular da somatória adotada. No AMDF, a

curva decrescente dos picos, devido à variação na quantidade de valores presentes na somatória, dificulta a análise dos valores, como visto anteriormente. Isto resulta, em muitos casos, na estimação do dobro ou da metade do valor correto do período de pitch. A CAMDF pode ser definida como:

$$D_{CAMDF}(k) = \sum_{n=0}^{N-1} |s_w(mod(n+k, N)) - s_w(n)|, \tag{5.6}$$

onde $s_w(n)$ representa o sinal janelado.

Da mesma maneira que ocorre com a AMDF, não é necessário calcularem-se todos os coeficientes da CAMDF, mas apenas aqueles que representam os valores válidos ou esperados para o período de pitch.

5.2.4 Estimação a partir do Espectro de Produtos Harmônicos

Como alternativa na estimativa do pitch utilizando-se métodos não-paramétricos, o Espectro de Produtos Harmônicos (Harmonic Product Spectrum) busca identificar a freqüência fundamental do sinal através da multiplicação de versões comprimidas do espectro do sinal considerado [12]. O algoritmo consiste da obtenção da DFT do sinal ruidoso, $X(k)$, e realização de um produtório (com $R = 5$ ou 6) dos espectros comprimidos pelo fator r. Com este procedimento, raias múltiplas da freqüência fundamental são multiplicadas, evidenciando a freqüência de pitch. Após a multiplicação dos espectros, é possível identificar a freqüência fundamental analisando-se o pico resultante. A decisão sobre quadros sonoros e não-sonoros pode ser feita comparando-se o pico obtido com um limiar. Este algoritmo também é de fácil implementação e sua carga computacional é menor do que o algoritmo SIFT (discutido na seqüência). Entretanto, os resultados da estimativa não são muito bons na presença de ruído intenso.

$$P_{HPS}(k) = \prod_{r=1}^{R} |X(rk)|. \tag{5.7}$$

5.2.5 Estimação a partir do Cepstrum Real

O modelo de produção de voz descreve o sinal de fala como produto da filtragem de uma excitação (na forma de trem de pulsos ou ruído) pelo filtro definido pelas ressonâncias características do trato vocal. O sinal de fala, portanto, é dado pela convolução entre a excitação e a resposta impulsiva do filtro que representa o trato vocal. A análise cepstral (realizada tomando-se a transformada inversa de Fourier do logaritmo da transformada direta de Fourier do sinal) é uma maneira de separar a informação de pitch da informação do trato vocal. Em outras palavras, é uma forma de realizar a de-convolução de um sinal. No domínio da freqüência, a convolução torna-se uma multiplicação. Desta forma, utilizando a propriedade dos logaritmos ($logAB = logA + logB$), a multiplicação pode ser transformada numa adição. A definição de cepstrum real é dada por:

$$c\left(n\right) = F_{TFTD}^{-1}\left\{\log\left|F_{TFTD}\left\{s\left(n\right)\right\}\right|\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left|S_c\left(w\right)\right|e^{jwn}dw, \tag{5.8}$$

onde

$$S_c(w) = \sum_{n=-\infty}^{\infty} s\left(n\right)e^{-jwn}. \tag{5.9}$$

A unidade n, do eixo do cepstrum, é chamada de quefrência, em analogia a freqüência.

Da definição do cepstrum, é possível demonstrar que as características referentes à excitação ficam expostas em altas quefrências, enquanto que as características do trato vocal aparecem em baixas quefrências. Sendo assim, é possível realizar-se uma estimativa do pitch através do pico do cepstrum em altas quefrências.

A estimação de pitch baseada no cepstrum real é mais robusta do que a estimativa feita diretamente através da auto-correlação. Entretanto, sua carga computacional é bem mais elevada.

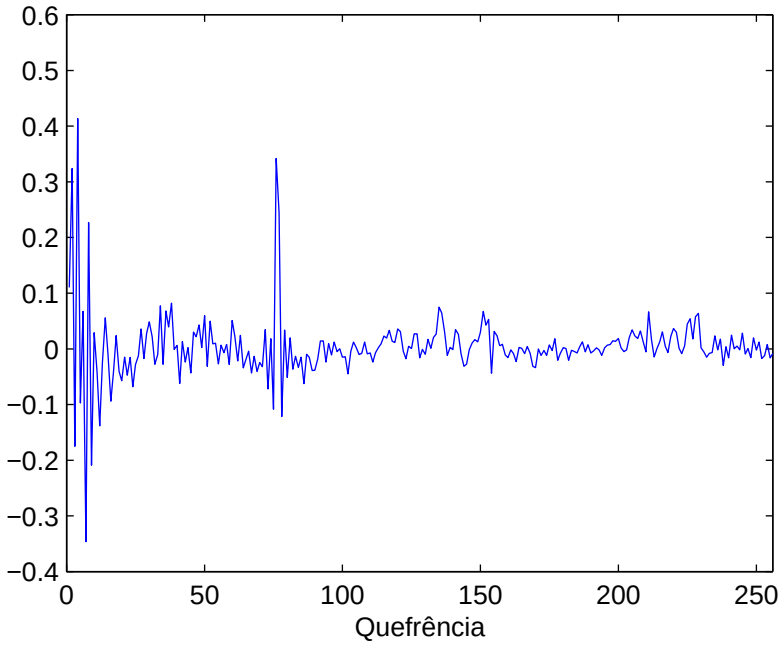

Figura 5.1: Cepstrum de um quadro de sinal de fala, ilustrando o pico que representa o período de pitch.

5.2.6 Algoritmo SIFT

O modelo auto-regressivo de produção de fala (predição linear) mostra que o sinal de fala, num dado instante, pode ser predito a partir das amostras anteriores, ponderadas pelos coeficientes do filtro correspondente ao trato vocal. Uma vez estimados os parâmetros do filtro, podemos realizar a filtragem inversa do sinal e obter o sinal de excitação que o gerou, também denominado erro de predição. É possível demonstrar-se que o erro de predição, no caso de um fonema sonoro, é máximo nos instantes em que ocorrem os pulsos de excitação (fechamento da glote)[12]. A partir dessa informação, pode-se avaliar o período de pitch observando os intervalos em que ocorrem os máximos no sinal de erro de predição. Esta foi a principal motivação para o desenvolvimento do algoritmo SIFT (Simple Inverse Filtering Tracking), proposto por Markel [25]. Por levar em conta

o modelo de produção de fala, o método tende a apresentar maior estabilidade e precisão do que a simples análise de periodicidade do sinal. Nesse algoritmo, o sinal de entrada é inicialmente submetido a uma filtragem passa-baixas, com freqüência de corte de cerca de 900 Hz, de modo a aproveitar a região do espectro com maior informação sobre a excitação, e também para possibilitar a realização de uma dizimação (redução da taxa de amostragem) em 4 ou 5 vezes, de maneira a reduzir a complexidade computacional do algoritmo. A taxa de amostragem pode ser reduzida de 8 para 2 kHz, por exemplo. O sinal dizimado é, então, analisado pelo método da auto-correlação, para obtenção de um filtro LPC de ordem baixa ($p = 4$, tipicamente, devido à existência de 1 ou 2 formantes nesta faixa de freqüências). Aplica-se, então, a filtragem inversa a esse sinal, para obtenção do erro de predição, cuja envoltória espectral é apoximadamente plana. Sobre o sinal de erro de predição é calculada a função de auto-correlação, e seu maior pico (dentro da faixa considerada válida) corresponde ao período de pitch (Figura 5.2). A classificação de quadro sonoro ou não-sonoro pode ser feita comparando-se o valor máximo da auto-correlação com um limiar pré-definido. De todos os algoritmos estudados, é o que apresenta maior complexidade computacional.

5.3 Análise da taxa de cruzamentos por zero

A diferenciação entre quadros sonoros e não-sonoros é realizada, em princípio, pelo mesmo algoritmo que estima o período de pitch. Em geral, o valor de energia (obtido do pico da função de auto-correlação ou do cepstrum, por exemplo) é comparado a um limiar de energia calculado a partir de informações sobre o sinal, indicando a existência de um quadro sonoro. Surgem daí alguns problemas, principalmente no que diz respeito à determinação do limiar de detecção. O limiar deve ser posicionado de acordo com a potência do ruído e potência média do sinal de fala. Há situações, porém, nas quais a obtenção e utilização do limiar não é suficiente para a correta detecção de um quadro sonoro. Pode ocorrer a classificação incorreta dos quadros não-sonoros como sendo sonoros (limiar abaixo do

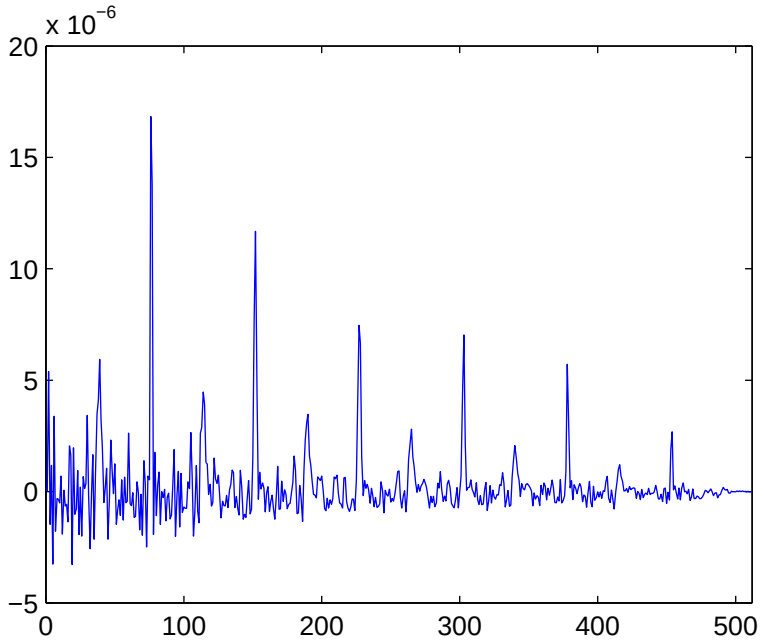

Figura 5.2: Auto-correlação do erro de predição de um quadro de sinal de fala, ilustrando o pico que representa o período de pitch.

necessário) ou a perda de quadros sonoros, isto é, quadros erroneamente classificados como não-sonoros (limiar acima do permitido). No primeiro caso, os efeitos são mais desagradáveis, pois nessa situação o sintetizador produz erroneamente um quadro sonoro. No segundo caso, o efeito também é ruim, porém, a sensação é de uma vogal sussurrada. Um procedimento de testes deve ser realizado para calibrar o limiar de modo a fornecer a maior precisão possível ao algoritmo.

Uma possibilidade adicional de identificação de quadros sonoros em sinais de fala é a utilização da taxa de cruzamentos por zero. Um cruzamento por zero ocore quando amostras sucessivas do sinal possuem diferentes sinais algébricos. A taxa de ocorrência de tais cruzamentos é uma medida simplificada do conteúdo de freqüências do sinal. Uma taxa de cruzamento por zeros elevada indica que existe uma componente ruidosa no sinal sob análise - portanto, é provável que o quadro seja não-sonoro. Uma taxa de cruzamentos por zero baixa, aliada a uma energia

relativamente alta, indica que o quadro é sonoro. Essa informação foi utilizada, neste trabalho, de maneira a contribuir para a distinção entre quadros sonoros e não-sonoros. O limiar de energia é, propositadamente, colocado ligeiramente acima do valor obtido nos testes. Desta forma, alguns quadros sonoros passam a ser classificados como não-sonoros. No entanto, o algoritmo leva em conta a diferença entre o valor de energia e o limiar. Se esse valor for baixo, significa que ainda existe a possibilidade de que o quadro seja sonoro. É feita, nesse caso, a análise da taxa de cruzamentos por zero. Se a taxa estiver dentro do previsto para quadros sonoros, a classificação é alterada.

A função para cálculo da taxa de cruzamentos por zero de curto prazo pode ser definida como:

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|w(n-m), \tag{5.10}$$

onde

$$sgn[x(n)] = \begin{cases} 1, & \text{para } x(n) \geqslant 0 \\ -1, & \text{para } x(n) > 0 \end{cases} \tag{5.11}$$

e

$$w(n) = \begin{cases} \frac{1}{2N}, & \text{dentro da janela de sinal} \\ 0, & \text{fora da janela de sinal} \end{cases} \tag{5.12}$$

A análise da taxa de cruzamentos por zero oferece uma indicação do conteúdo em freqüência do sinal, mas as distribuições estatísticas das taxas de cruzamentos, quando se comparam quadros sonoros e não-sonoros, apresentam regiões de intersecção, de maneira que uma decisão precisa e inequívoca não pode ser realizada apenas tomando-se por base a taxa de cruzamentos por zero [12]. O conceito pode, entretanto, ser utilizado como etapa adicional na classificação, como foi realizado neste trabalho. Apesar de simples e de baixa complexidade computacional, tal procedimento contribui para uma sensível melhoria na classificação

dos quadros, reduzindo a incidência de erros devidos ao mau posicionamento do limiar de energia.

5.4 Implementação e análise de resultados

5.4.1 Análise dos algoritmos

Todos os algoritmos mencionados foram testados com o objetivo de verificar sua eficácia, robustez a ruído e complexidade computacional. Inicialmente, foram realizados testes com sinais sem ruído. Mesmo em condições ideais, já foi possível observar a baixa eficácia dos algoritmos baseados na estimação direta a partir da auto-correlação. Por essa razão, tais algoritmos não foram utilizados nos demais testes. Com o aumento da intensidade do ruído, os algoritmos que apresentaram melhores resultados foram: SIFT, CAMDF e Cepstrum. A Figura 5.3 ilustra os valores estimados para o período de pitch (em quantidade de amostras) para o sinal de teste, ainda sem adição de ruído. O algoritmo SIFT, apesar de sua elevada complexidade computacional, foi o que apresentou melhor robustez na presença de ruído. A Figura 5.4 ilustra os valores de período de pitch estimados ao longo de um sinal de fala com SNR segmentada de 10dB. É possível observar a perda de eficácia do algoritmo baseado no Cepstrum real, e uma semelhança nos valores estimados através dos algoritmos SIFT e CAMDF. Na Figura 5.5, entretanto, o mesmo sinal é submetido a uma SNR Segmentada de 5dB. Neste caso, é possível observar a superioridade do algoritmo SIFT. Foram realizados testes com diversos sinais de fala, e confirmada a maior qualidade do algoritmo SIFT em comparação com os demais.

5.4.2 Suavização dos valores

Os valores estimados para o período de pitch devem ser suavizados de maneira adequada para que não introduzam descontinuidades significativas na etapa de síntese. Entretanto, uma simples suavização linear, por um filtro de média móvel,

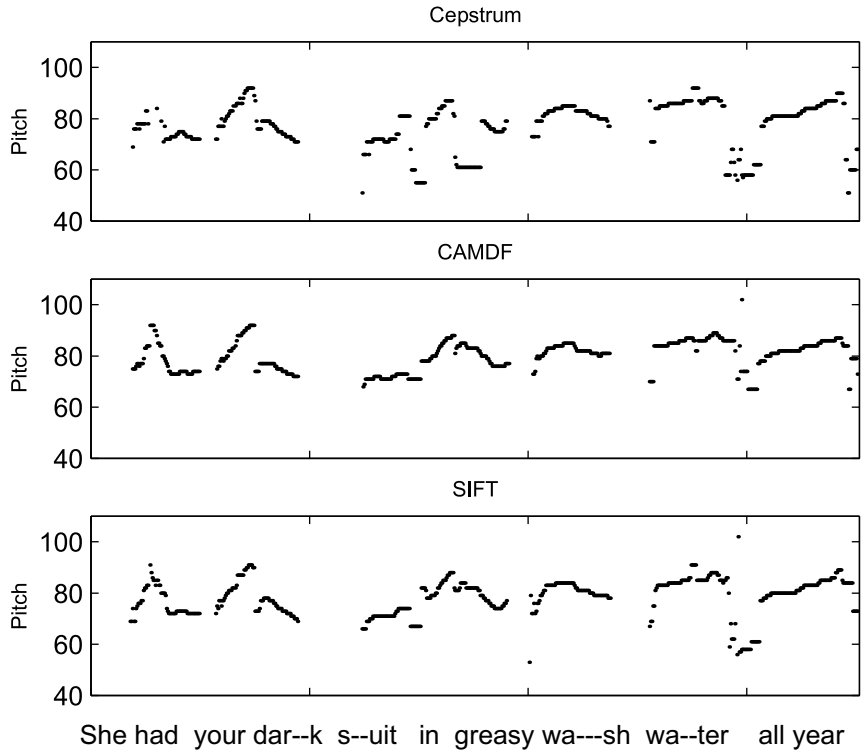

Figura 5.3: Comparação entre as estimativas de pitch realizadas ao longo de um sinal de fala sem adição de ruído.

pode produzir resultados indesejados nas estimativas. Por exemplo: no caso de um valor estimado erroneamente como sendo o dobro do valor correto (situação relativamente comum na estimação de pitch), uma suavização linear introduziria um erro grosseiro na freqüência fundamental do quadro seguinte, comprometendo o processo de síntese. Nesse caso, um filtro de mediana móvel seria muito mais adequado [12]. Desta forma, erros dessa natureza são mais bem administrados. Além do filtro de mediana, é importante também uma verificação adicional nos quadros que são candidatos a transição sonoro/não-sonoro. Isto é, após uma sequência de quadros sonoros, representando uma vogal, por exemplo, é necessária uma checagem especial no primeiro quadro que se apresenta como não-sonoro, para confirmar sua condição. Um dos problemas comuns no sintetizador é o surgi-

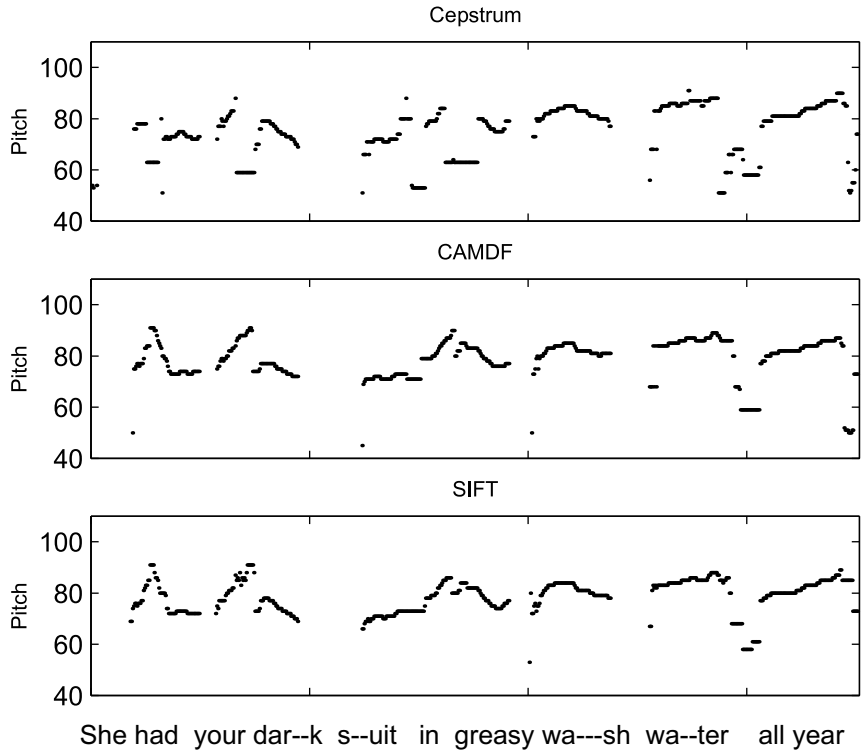

Figura 5.4: Comparação entre as estimativas de pitch realizadas ao longo de um sinal de fala com SNR segmentada de 10dB.

mento de descontinuidades curtas, produzidas por um único quadro não-sonoro cercado por quadros sonoros. Essa verificação pode ser feita através de uma redefinição momentânea do limiar de decisão, ou mesmo através da comparação da taxa de cruzamentos por zero do quadro em questão com os valores obtidos anteriormente.

Além das suavizações realizadas durante o processo de estimação, a etapa de síntese também trabalha no sentido de garantir transições suaves no espaçamento dos pulsos, conforme será discutido no capítulo correspondente.

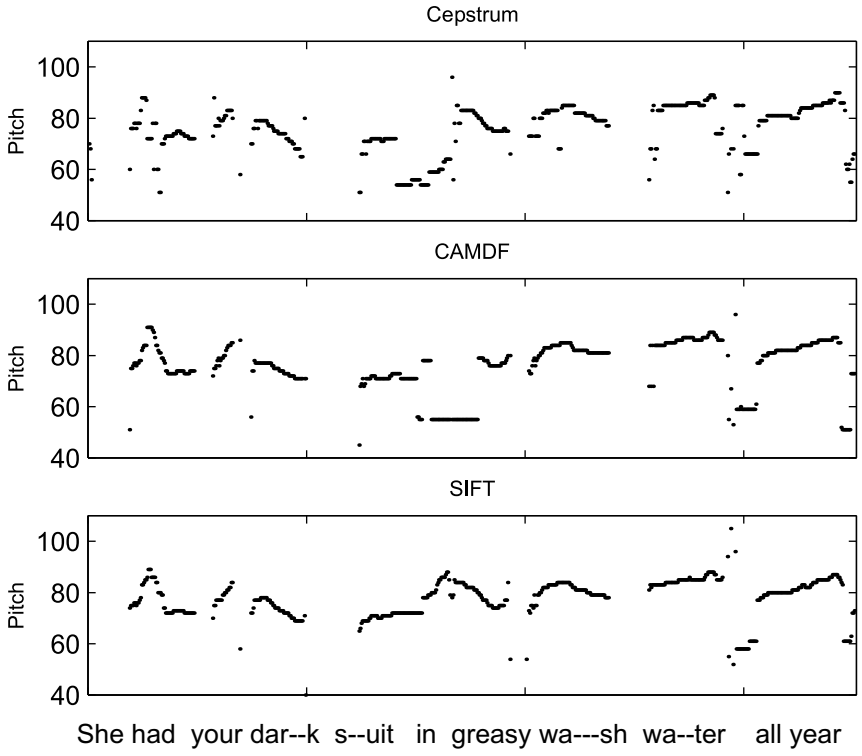

Figura 5.5: Comparação entre as estimativas de pitch realizadas ao longo de um sinal de fala com SNR segmentada de 5dB.

Capítulo 6

Síntese do sinal de fala

6.1 Características gerais

O sinal sintetizado é obtido de acordo com o modelo digital de produção de fala discutido anteriormente. Os parâmetros de controle utilizados pelo sintetizador são o ganho g, os coeficientes LPC α_k, o período de pitch P e o indicador de quadro sonoro ou não-sonoro. As informações de período de pitch e tipo de quadro podem ser agrupadas no vetor de excitação, $u(n)$, que conterá pulsos no caso de quadros sonoros ou ruído branco no caso de quadros não-sonoros. Os parâmetros são processados pelo sintetizador de maneira a gerar um quadro de sinal de fala, $d(n)$, de acordo com a expressão:

$$d(n) = \sum_{k=1}^{p} \alpha_k d(n-k) + g.u(n). \tag{6.1}$$

6.2 Implementação e análise de resultados

6.2.1 Geração da excitação

A excitação a ser aplicada ao sintetizador, $u(n)$, pode ser impulsiva, em caso de quadros sonoros, ou ruidosa, em caso de quadros não-sonoros. O algoritmo de geração da excitação deve trabalhar, portanto, com essas duas possibilidades.

Excitação ruidosa

A excitação ruidosa é composta basicamente de ruído branco, multiplicado pelo valor do ganho, g. Apesar de simples, a implementação mostrou a necessidade de algumas correções, principalmente com relação a descontinuidades no ganho. Como a análise é feita quadro a quadro, é comum a ocorrência de variações bruscas nas energias dos quadros sintetizados. O reflexo de tais variações é o aparecimento de estalos durante as frases, principalmente nas regiões de quadros não-sonoros. Para solucionar o problema, foi implementada uma suavização interna do ganho dos quadros não-sonoros, de maneira a produzir uma variação suave, evitando descontinuidades. A suavização produz uma transição progressiva do ganho dentro do quadro sob análise, levando em conta o valor de ganho do quadro anterior.

Excitação impulsiva

A geração da excitação impulsiva envolve considerações especiais. Nos testes iniciais, simplesmente foi utilizado um trem de pulsos igualmente espaçados, com o espaçamento dado pelo período de pitch. Apesar de funcional, tal procedimento produz uma aspereza no sinal sintetizado, sempre que há trechos em que o período de pitch varia de maneira rápida entre quadros. Para contornar esse problema, foi utilizada uma forma de suavização interna para os pulsos de excitação. Para a geração da excitação do quadro n, é levado em conta o período de pitch do quadro anterior, $n-1$. Desta forma, os primeiros pulsos do quadro n terão espaçamento intermediário entre o valor de pitch do quadro $n-1$ e do quadro n. A utilização desse procedimento e da suavização do ganho (da mesma forma que foi utilizada para a excitação ruidosa) reduziram significativamente as asperezas observadas no sinal sintetizado.

6.2.2 Aperfeiçoamentos no sintetizador

Sincronismo entre os sinais para a soma ponderada

Conforme mencionado na proposta de trabalho, a reconstrução do espectro do sinal (que foi degradado pela aplicação da subtração espectral) é realizada através da soma ponderada com o sinal sintetizado. Esta soma, entretanto, pode ser prejudicada pela falta de sincronismo entre o sinal $\hat{s}(n)$, obtido por subtração espectral, e o sinal $d(n)$, sintetizado a partir de parâmetros estimados do sinal ruidoso pré-processado.

Assim como foi feito nos outros tópicos deste trabalho, é interessante iniciar a análise partindo-se de uma condição ideal - neste caso, a possibilidade de se obter um sinal sintetizado a partir de parâmetros espectrais idênticos aos do sinal original (não degradado por ruído). Nesta situação, é simples observar-se que a soma do sinal sintetizado $d(n)$ com o sinal $\hat{s}(n)$ seria severamente prejudicada pela falta de sincronismo entre os sinais.

Para verificar o problema, foram realizados alguns testes com o intuito de avaliar o efeito isolado do sincronismo no resultado final do algoritmo. Inicialmente, foi analisado o efeito de soma de um sinal de fala (não ruidoso) com um sinal sintetizado a partir desse mesmo sinal, com e sem sincronismo, isto é, considerando duas condições distintas para a montagem do vetor de excitação do sinal sintetizado:

a) criação de um vetor de pulsos igualmente espaçados, obedecendo aos critérios discutidos no capítulo referente à síntese (continuidade do vetor, suavidade nas variações entre quadros, etc). Neste caso, as restrições à posição dos pulsos devem-se apenas aos critérios de continuidade e suavização.

b) obtenção da posição dos pulsos a partir da análise do sinal de fala. O algoritmo de análise calcula o erro de predição do sinal e, a partir da estimativa do período de pitch, busca as prováveis posições da excitação dentro do quadro de sinal. O período de pitch é utilizado para a criação de janelas de busca em torno das regiões mais prováveis para a presença de um pulso. O sinal de erro

de predição é, então, analisado e, havendo pulsos dentro das janelas de busca, serão utilizados no vetor de excitação. Desta forma, as restrições à posição dos pulsos obedecem às características do sinal de fala original. A Figura 6.1 ilustra o algoritmo de busca, mostrando o sinal de fala, o erro de predição e as posições dos pulsos obtidos após a análise do sinal de erro de predição.

A análise dos resultados permite observar que surge uma importante degradação quando não há sincronismo entre os sinais somados, em condições ideais. Quando a análise é feita em condições reais, isto é, com sinal ruidoso e soma com sinal sintetizado, o efeito é bem menos pronunciado, entretanto, ainda importante.

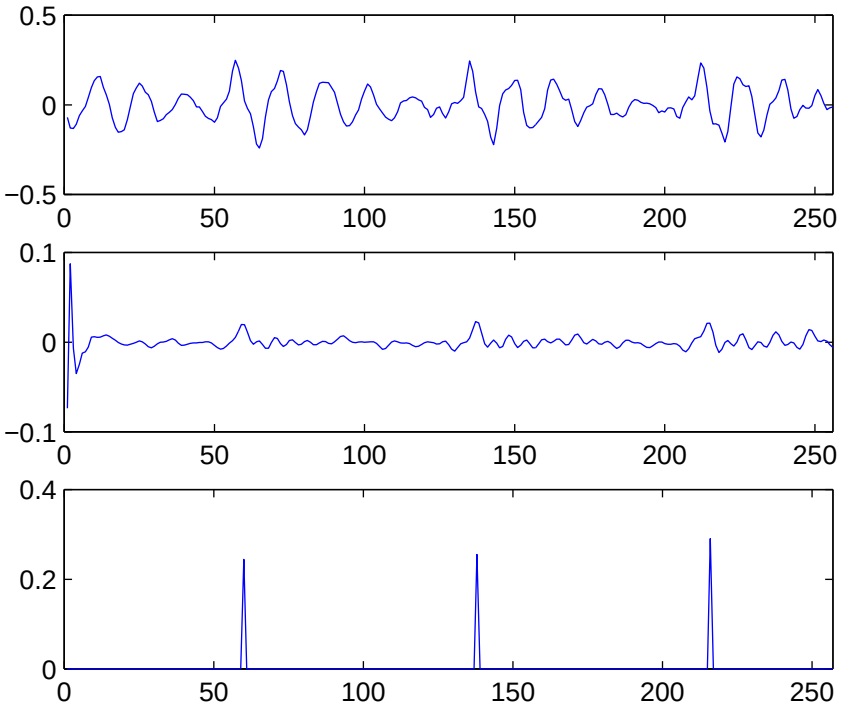

Figura 6.1: Ilustração do algoritmo de captura da posição dos impulsos de excitação. Quadro superior: sinal de fala. Quadro intermediário: erro de predição. Quadro inferior: posição dos impulsos de excitação.

Suavização dos coeficientes LPC

Como tratado no início do capítulo, o bloco de síntese produz quadros de fala a partir das informações referentes à excitação e à resposta em freqüência do trato vocal (através dos coeficientes LPC). As amostras sintetizadas são obtidas de acordo com:

$$d(n) = \sum_{k=1}^{p} \alpha_k d(n-k) + g.u(n), \tag{6.2}$$

onde α_k representa o vetor de coeficientes LPC, g representa o ganho do quadro, e $u(n)$ representa a excitação. No caso de quadros sonoros, a excitação é composta de pulsos adequadamente espaçados, de maneira que a expressão 6.2 pode ser considerada a resposta livre do sistema para os instantes nos quais a excitação é nula. Desta forma, a atualização dos coeficientes LPC deve ser realizada de maneira sincronizada com os pulsos de excitação. Como os quadros têm duração fixa, e geralmente englobam de 3 a 11 períodos de pitch em seu interior (para freqüências típicas de pitch variando entre 100 e 350 Hz), surge a necessidade de interpolação dos coeficientes, internamente a um quadro, de maneira a garantir uma transição suave dos coeficientes entre os quadros [5].

A suavização dos coeficientes LPC, entretanto, exige considerações especiais. O filtro de síntese deve ser, obrigatoriamente, estável. A simples interpolação dos coeficientes LPC, todavia, não atende a este requisito, pois não há garantia de estabilidade do filtro resultante. Para evitar esse problema, é necessário trabalhar-se com representações alternativas, cujos parâmetros possam ser interpolados com garantia de estabilidade do filtro resultante. Exemplos de tais parâmetros são os Coeficientes de Reflexão (RC's), Razão de Áreas (Log Area Ratios) e os parâmetros LSP (Line Spectrum Pairs). Neste trabalho, foram utilizados os parâmetros LSP para interpolação (conforme discutido anteriormente), com bons resultados. A Figura 6.2 ilustra a diferença nos espectrogramas de dois sinais, um sintetizado sem utilizar interpolação e outro com interpolação. Em termos perceptuais, a ausência da interpolação gera uma certa aspereza no sinal

sintetizado, causada pela variação brusca dos parâmetros LPC entre os quadros.

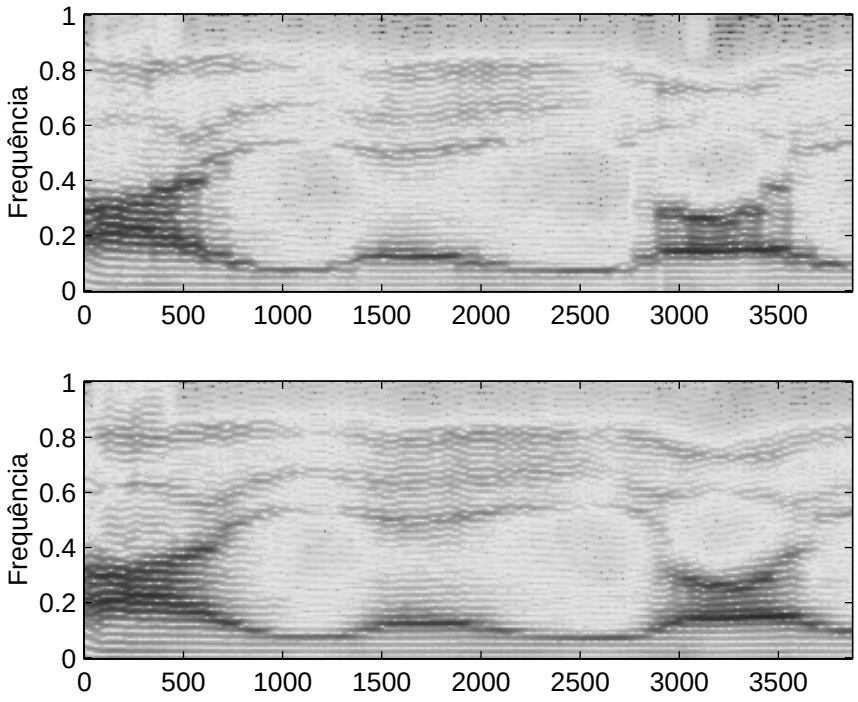

Figura 6.2: Ilustração do efeito da interpolação LPC usando coeficientes LSP. No quadro superior é mostrado o espectrograma de um sinal sintetizado sem interpolação. No quadro inferior, é mostrado o mesmo sinal, porém com interpolação de coeficientes.

Capítulo 7

Estrutura completa do método proposto

Conforme discutido nos capítulos iniciais, a proposta deste trabalho é o estudo de um novo método de melhoria da qualidade de sinais de fala degradados por ruído aditivo branco, que tem por objetivos principais aproveitar as características positivas da subtração espectral e acrescentar informação espectral através da síntese de um sinal de fala cujos parâmetros são obtidos do sinal ruidoso. A reconstrução do espectro é obtida através da soma ponderada do sinal obtido por subtração espectral e do sinal sintetizado. O método visa a obtenção de um sinal resultante cujas características perceptuais sejam superiores ao método de subtração espectral, porém sem produção de ruído musical.

A Figura 7.1 ilustra o diagrama em blocos do método. O sinal ruidoso, $x(n)$, é tratado por dois blocos simultaneamente. No primeiro deles são obtidos os parâmetros do modelo digital de produção de fala e é gerado o sinal sintetizado, $d(n)$. No outro bloco, é realizado o processo de subtração espectral sobre o sinal ruidoso, produzindo o sinal $\hat{s}(n)$. A partir dos coeficientes do filtro de ponderação da subtração espectral, $H(k)$, são obtidos os fatores de ponderação que, multiplicando o sinal $d(n)$, produzem o sinal $v(n)$. Os fatores de ponderação (cujo cálculo será discutido a seguir) têm a função de acrescentar a informação

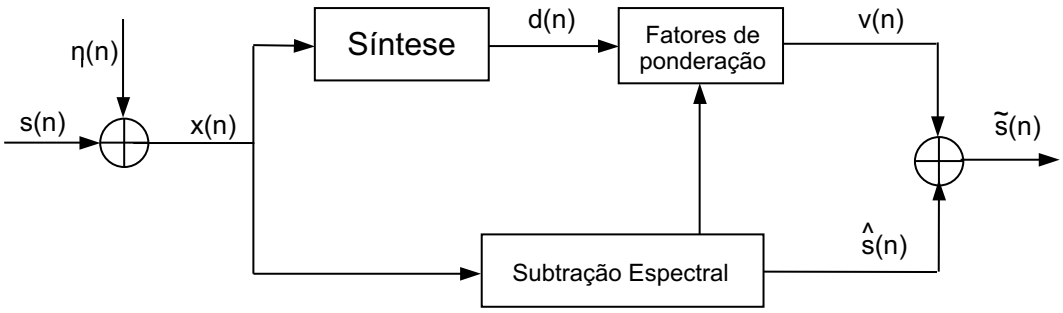

Figura 7.1: Diagrama em blocos do método proposto

espectral do sinal sintetizado ao sinal obtido por subtração espectral, isto é, nas freqüências em que o sinal é pouco atenuado pela subtração espectral, a contribuiç ão do sinal sintetizado deve ser pequena, porém, nas freqüências em que o sinal é muito afetado, a contribuição do sinal sintetizado deve ser maior. O sinal reconstruído, $\tilde{s}(n)$, que é o produto final do método, é obtido da soma entre o sinal $\hat{s}(n)$ (resultado da subtração espectral) e o sinal $v(n)$ (resultado da aplicação dos fatores de ponderação ao sinal sintetizado).

7.1 Algoritmos e técnicas utilizados

Para a avaliação geral do método proposto, foram utilizados os algoritmos e técnicas que apresentaram melhor desempenho nos testes. Para esta avaliação, não foi levada em conta a complexidade computacional dos algoritmos utilizados. O que se buscou foi o melhor desempenho na presença de ruído intenso (relações sinal-ruído segmentadas de 5dB e inferiores) para todas as técnicas aplicadas. Ao final do capítulo será apresentado um resumo dos parâmetros utilizados em cada

bloco.

7.1.1 Subtração espectral

Foi utilizada a formulação usual da subtração espectral, porém com os modificadores discutidos anteriormente (suavização de coeficientes, limite espectral mínimo e sobre-estimação de ruído) para a obtenção do sinal $\hat{s}(n)$. Neste caso, como se espera um sinal sem ruído musical, o fator de sobre-estimação α deve ser elevado, de modo a eliminar o ruído musical. Conforme tratado nos capítulos anteriores, o método de subtração espectral apresenta um compromisso importante entre a quantidade de ruído removido, a perda de informação espectral (formantes mais tênues) e o ruído musical. À medida que se diminui o ruído musical, através da suavização dos coeficientes e da sobre-estimação do ruído, mais formantes são eliminados do sinal. Isto provoca uma sensação de abafamento no sinal resultante, pois apenas os formantes mais intensos permanecem. Apesar do problema de abafamento, o sinal obtido através de tal procedimento (sobre-estimação de ruído e suavização de coeficientes) possui muito pouco ou nenhum ruído musical, o que é vantajoso para o resultado final do processo.

7.1.2 Estimação de pitch

Foi utilizado o algoritmo SIFT (Simple Inverse Filtering Tracking), em conjunto com a análise da taxa de cruzamentos por zero para auxiliar na decisão sobre quadros sonoros e não-sonoros. Para suavização dos valores, foi aplicado um filtro de mediana móvel de 7 amostras.

7.1.3 Estimação de coeficientes LPC

Foi aplicado o pré-processamento baseado na subtração espectral. A subtração espectral usada neste ponto, entretanto, é diferente daquela que é aplicada para a geração do sinal $\hat{s}(n)$. Para a estimação dos coeficientes LPC, o fator de sobre-estimação de ruído, α, é menor do que aquele aplicado na obtenção do sinal

$\hat{s}(n)$, como discutido anteriormente. Após o pré-processamento, os coeficientes são obtidos através de uma análise LPC de ordem 16.

7.1.4 Sintetizador

O sintetizador foi construído com base no modelo digital do sinal de fala, conforme tratado no capítulo referente à síntese de fala. A geração da excitação impulsiva, entretanto, foi realizada de maneira a manter o sincronismo do sinal sintetizado com o sinal $\hat{s}(n)$.

Diante do que foi exposto, é possível construir o diagrama em blocos do sintetizador, como ilustrado na Figura 7.2. O sinal ruidoso, $x(n)$, é pré-processado para a obtenção dos sinais que serão utilizados na estimativa do período de pitch e dos coeficientes LPC. Dois blocos de pré-processamento são utilizados de modo a permitir variações independentes nos parâmetros da subtração espectral de cada um deles. O estimador de pitch obtém a informação do tipo de quadro (sonoro ou não-sonoro) e, no caso de quadro sonoro, do seu período de pitch. Neste último caso, o bloco de obtenção de sincronismo analisa o sinal com o objetivo de encontrar as posições corretas para os pulsos de excitação. Com essas informações, o bloco de geração de excitação produz o trem de pulsos ou o ruído branco que servirão como excitação para o filtro de síntese. Em paralelo, o estimador LPC obtém os coeficientes e o ganho do quadro a partir do sinal pré-processado. De posse de todos esses parâmetros, o sintetizador produz um quadro de sinal.

7.2 Fatores de ponderação

Os fatores de ponderação da soma dependem da informação obtida do processo de subtração espectral, isto é, dos coeficientes de atenuação das amplitudes do espectro do sinal ruidoso. Para cada quadro de fala, o algoritmo de subtração espectral produz os coeficientes do filtro de ponderação espectral $H(k)$, de acordo com (2.17). O espectro do sinal ruidoso é, então, multiplicado por tais coeficientes.

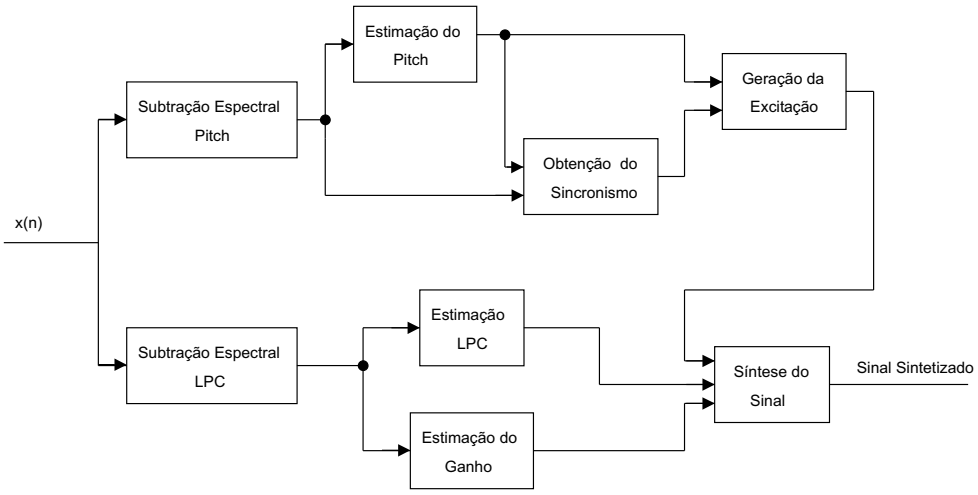

Figura 7.2: Diagrama em blocos do sintetizador

$$\hat{S}(k) = H(k)X(k), \tag{7.1}$$

onde a magnitude de $H(k)$ varia entre 0 e 1, e $X(k)$ representa o espectro de tempo discreto do sinal ruidoso.

Além de eliminar informação espectral referente ao ruído, tal procedimento também elimina os formantes mais fracos, como analisado anteriormente, pois está sendo utilizado um fator de sobre-estimação de ruído durante o processo de subtração espectral. Para recuperar essa informação, propõe-se somar o sinal $\hat{S}(k)$ ao sinal $V(k)$, que é obtido por:

$$V(k) = D(k)[1 - H(k)], \tag{7.2}$$

onde $D(k)$ representa o espectro do sinal sintetizado, $d(n)$. $D(k)$ é obtido da DFT do sinal sintetizado $d(n)$, que foi produzido com base nas informações de pitch, ganho e coeficientes LPC extraídas do sinal ruidoso. O fator $[1-H(k)]$ tem

a função de trazer, a partir do sinal sintetizado, informações espectrais perdidas devido ao processo de subtração espectral.

Desta forma, o sinal resultante, $\tilde{S}(k)$, conterá a informação espectral corretamente obtida da subtração espectral, e também a informação recuperada através do processo de síntese. Pode-se ainda adicionar pesos aos sinais $V(k)$ e $\hat{S}(k)$, para a composição do sinal reconstruído, $\tilde{S}(k)$, resultando em

$$\tilde{S}(k) = aV(k) + b\hat{S}(k). \tag{7.3}$$

A soma dos sinais $V(k)$ e $\hat{S}(k)$ é realizada, portanto, no domínio da freqüência. O sinal de fala reconstruído, $\widetilde{s}(n)$, é obtido através da transformada inversa de $\hat{S}(k)$, com janelamento e superposição de quadros, conforme descrito no capítulo referente à técnica de subtração espectral.

7.3 Resumo dos parâmetros utilizados no programa

7.3.1 Ambiente usado nas simulações

As simulações e testes dos algoritmos foram realizados em um PC com processador AMD Athlon XP 1.4 GHz, rodando sistema operacional Windows 2000. Todo o software foi escrito em MatLab versão 5.3. Como o objetivo do trabalho sempre esteve relacionado à análise dos resultados, e não propriamente à velocidade de execução, poucas otimizações de código foram aplicadas. Algumas rotinas poderiam ser reduzidas e tratadas de maneira a se tornarem mais eficientes, mas isso reduziria a flexibilidade e a possibilidade de alterações rápidas para verificação das possibilidades de variação nos algoritmos. Levando em conta esse ambiente, o tempo médio para o processamento de uma frase de 3 segundos de duração foi de 12 segundos.

O programa foi estruturado de maneira a permitir rápidas modificações em

parâmetros e testes de novas funcionalidades. Todas as principais variáveis do algoritmo (tais como janelamento, fatores da subtração espectral, coeficientes e filtros) estão distribuídas em grupos de parâmetros, de fácil alteração. Desta forma, é possível testar o efeito isolado da mudança da quantidade de coeficientes LPC, da taxa de superposição de quadros ou da ponderação da soma, por exemplo, sem precisar modificar substancialmente o próprio programa.

Os parâmetros mais importantes utilizados nas simulações são descritos a seguir, de acordo com os principais blocos do programa.

Tamanho dos quadros e superposição

O tamanho dos quadros pode ser configurado no programa, bem como a superposição entre eles. Os melhores resultados nas simulações foram obtidos com quadros de 256 amostras. Para uma taxa de amostragem de 8 kHz, isso representa quadros de 32 ms. O uso de quadros de 256 amostras também facilita a execução das transformadas de Fourier necessárias aos blocos de subtração espectral. Foi utilizada uma superposição entre quadros de 224 amostras. Apesar de tornar o programa mais lento, a superposição de uma maior quantidade de amostras mostrou-se mais eficaz nos testes subjetivos. Para uma aplicação real, provavelmente seria necessária alguma modificação no sentido de evitar o excesso de carga computacional proporcionado por esse valor de superposição. Para o janelamento do sinal, foi utilizada a janela de Hanning.

Subtração espectral

O bloco de subtração espectral é responsável por produzir o sinal de fala que será posteriormente somado ao sinal sintetizado. O objetivo desta subtração espectral é proporcionar o mínimo de ruído musical que for possível, mesmo que isso implique na perda de formantes mais fracos, pois estes poderão ser recuperados a partir do sinal sintetizado. Os principais parâmetros usados foram:

- Limite espectral mínimo (Minimum Spectral Floor): $H_{\min} = 0,05$

- Fator de sobre-estimação de ruído: $\alpha = 8$ (evita o ruído musical, porém provoca a perda de formantes mais fracos)

- Fator de suavização dos coeficientes do filtro de ponderação espectral $H(e^{jw})$: $\lambda = 0,6$.

Pré-processamento para estimação dos coeficientes LPC usando subtração espectral

Conforme discutido ao longo do texto, o pré-processamento baseado em subtração espectral se mostrou muito eficiente na estimativa dos coeficientes LPC. A principal diferença, quando comparado ao bloco anterior, é a menor taxa de sobre-estimação de ruído. Apesar de permitir a produção de ruído musical, uma menor taxa de sobre-estimação de ruído contribui para uma melhor estimativa dos coeficientes. Os parâmetros usados foram:

- Limite espectral mínimo (Minimum Spectral Floor): $H_{\min} = 0,05$

- Fator de sobre-estimação de ruído: $\alpha = 2$

- Fator de suavização dos coeficientes do filtro de ponderação espectral $H(e^{jw})$: $\lambda = 0,9$.

Estimação do período de pitch

A estimação do período de pitch foi feita com o algoritmo SIFT. De acordo com a proposta do algoritmo, o sinal de fala é inicialmente filtrado (por um filtro FIR passa-baixas de 16 coeficientes e freqüência de corte de 900 Hz). A análise LPC interna (aplicada após a dizimação do sinal, que reduz sua taxa de amostragem para 2 kHz) é realizada com um polinômio de ordem 4. Após o cálculo da autocorrelação do sinal de erro de predição, o máximo obtido é comparado a um limiar, para decisão sobre o tipo de quadro (sonoro ou não-sonoro). Este limiar foi definido como $8.10^{-5}/N$, onde N representa o tamanho dos quadros utilizados.

O limiar foi obtido a partir de testes subjetivos e observações do perfil de valores de pitch estimados para diversos sinais.

A taxa de cruzamentos por zero também foi utilizada para uma verificação adicional. Se o quadro é considerado não-sonoro (conforme descrito acima), mas o valor máximo da função de auto-correlação do erro de predição é maior do que 90% do limiar, é calculada a taxa de cruzamentos por zero, para uma verificação mais precisa a respeito do tipo de quadro que está sendo analisado. Se a taxa de cruzamentos por zero, nesse caso, for inferior a um limiar, o quadro é considerado sonoro. Esse limiar foi definido como 0,6.

Além desses procedimentos, foi utilizado também um filtro de mediana móvel de 7 amostras para os valores de pitch de quadros sonoros. Tal filtro tem o objetivo de eliminar valores aberrantes, estimados erroneamente pelo algoritmo.

Obtenção do sincronismo entre o sinal sintetizado e o sinal obtido por subtração espectral

Para encontrar a posição dos pulsos de excitação do sinal ruidoso, o algoritmo utilizado analisa o erro de predição do sinal ruidoso, após uma análise LPC de ordem 8, buscando as posições de ocorrência de picos, que correspondem às posições dos pulsos de excitação.

Uma vez encontrado um pico no sinal de erro de predição, a rotina procura os próximos picos levando em conta o período de pitch estimado e uma janela de busca, isto é, a rotina analisa a região em volta da posição provável do próximo impulso (4 amostras antes e 4 amostras depois da posição provável). Após encontradas as posições, o vetor de excitação a ser aplicado ao sintetizador é construído.

Combinação dos sinais produzidos

Os sinais a serem combinados (sintetizado e obtido da subtração espectral) são multiplicados por um fator de ponderação antes de serem somados (parâmetros a e b da Equação 7.3). Os melhores resultados foram obtidos para $a = 0,3$

(multiplicador do sinal sintetizado) e $b = 0,7$ (multiplicador do sinal obtido por subtração espectral).

Capítulo 8

Conclusões

8.1 Resultados obtidos

Durante todo o trabalho foram realizados testes de avaliação das técnicas estudadas, de maneira individual, com o objetivo de verificar suas características, vantagens e desvantagens para a aplicação no método proposto. Além dos testes individuais, porém, foram realizadas avaliações da qualidade geral do método, utilizando sinais de fala com frases completas, com diversos locutores e diferentes relações sinal-ruído.

Como resultado geral, é possível observar que a qualidade do sinal obtido é superior à que se consegue apenas com a subtração espectral, e superior também à do sinal sintetizado escutado isoladamente. Para os quadros não-sonoros, nota-se uma maior naturalidade e, para os quadros sonoros, uma maior clareza, o que também contribui para uma melhor sensação auditiva. Pelas próprias características do método, tais como a suavização de parâmetros e estimação LPC após pré-processamento, não há descontinuidades no sinal devido a variações bruscas de parâmetros ou filtros IIR instáveis. Exemplos de arquivos de áudio, ilustrando as diversas fases do desenvolvimento do método proposto, estão disponíveis no CD que acompanha este documento (arquivo "Exemplos.htm").

Alguns pontos, porém, poderiam ser melhorados, como, por exemplo, o pro-

blema do zumbido que aparece no sinal sintetizado. Esse zumbido surge, principalmente, em regiões sonoras de baixa relação sinal-ruído local (tipicamente inícios de fonemas e trechos de menor energia), e soa de maneira semelhante a um trem de pulsos não-filtrado. Conforme discutido no capítulo referente à estimação dos coeficientes LPC, um dos problemas da estimação realizada a partir de sinais ruidosos é a perda da definição dos formantes, isto é, como as ressonâncias devidas aos formantes são menos pronunciadas do que deveriam, grande parte do espectro da excitação deixa de ser adequadamente eliminada do sinal, produzindo a sensação de zumbido. Apesar de restrito a algumas regiões (principalmente inícios de fonemas), o zumbido é percebido no conjunto do sinal, o que diminui a qualidade geral do método.

8.2 Propostas de extensão do trabalho

Durante a execução dos testes e das implementações discutidas neste documento, foi possível observar algumas possibilidades de extensão do estudo que, apesar de fugirem ao escopo deste trabalho, poderiam ser explorados em outros trabalhos mais avançados, com o objetivo de melhorar a qualidade do sinal sintetizado que é somado ao sinal obtido por subtração espectral.

8.2.1 Métodos alternativos para obtenção do sinal sintetizado

Como tratado nos capítulos anteriores, o objetivo do método proposto é a melhoria da qualidade do sinal de fala degradado por ruído através da combinação entre um sinal obtido por subtração espectral e um sinal sintetizado obtido a partir do sinal ruidoso. A obtenção deste sinal sintetizado envolve a estimação dos parâmetros do modelo de produção de fala: período de pitch, ganho, tipo de excitação e coeficientes LPC. A melhoria na estimativa de cada um desses parâmetros pode trazer também sensíveis melhorias ao método como um todo.

Desta forma, o método proposto pode se beneficiar do avanço no estudo de cada um dos temas relacionados ao modelo de produção de fala. A estimação dos coeficientes LPC, por exemplo, tem recebido especial atenção nas últimas décadas, e continua sendo foco de importantes estudos. Neste trabalho, foram avaliados principalmente dois processos: estimação iterativa (algoritmo LMAP), e a estimação a partir de um sinal pré-processado utilizando subtração espectral. Todavia, poderiam ser aplicados outros métodos de estimação, incluindo até abordagens estatísticas (uso de misturas gaussianas e modelos de Markov, por exemplo), com o intuito de trazer maior precisão à estimativa dos coeficientes.

8.2.2 Melhoria do processo de síntese do sinal através do acréscimo de informações de excitação

O algoritmo construído para a síntese do sinal de fala utiliza, como excitação, um trem de pulsos no caso de quadros sonoros e ruído branco no caso de quadros não-sonoros. Entretanto, há métodos de codificação de voz (CELP e RELP, por exemplo) que aplicam técnicas específicas para a obtenção do sinal de excitação a ser usado na síntese. Apesar de tais métodos terem por objetivo a reconstrução completa do sinal de entrada (o que seria inconveniente para o caso de redução de ruído), o conceito poderia ser estudado no sentido de permitir a construção de uma excitação mais adequada ao processo de melhoria de qualidade do sinal de fala. Desta forma, o algoritmo teria como objetivo a extração de informações adicionais sobre a excitação a ser aplicada ao sintetizador, de modo a produzir um sinal sintetizado com qualidade superior à que se pode obter usando apenas trem de pulsos e ruído branco como excitação.

8.2.3 Utilização de informações vocais do locutor para melhoria na estimação dos parâmetros

Outra linha para possível melhoria do método é a utilização de informações vocais do locutor para estimação dos parâmetros LPC. Apesar de restringir a aplicação do método a situações nas quais o usuário do sistema é previamente conhecido, tal abordagem poderia contribuir para o refinamento da estimativa dos coeficientes LPC na presença de ruído intenso. Um algoritmo de análise de máxima verossimilhança poderia buscar, dentro de um conjunto de amostras representativas do padrão vocal do locutor, aquelas que melhor completam uma estimativa precária obtida em situação de ruído intenso. O padrão vocal do locutor seria obtido em momentos de ausência de ruído, de maneira a construir uma base de dados para direcionar o funcionamento do algoritmo na presença de ruído.

Bibliografia

[1] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Transactions on Audio and Eletroacoustics*, vol. AU-16, pp. 165–168, Junho 1968.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, Abril 1979.

[3] M. S. Ahmed, "Comparison of noisy speech enhancement algorithms in terms of LPC perturbation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 121–125, Janeiro 1989.

[4] B. L. Pellom e J. H. L. Hansen, "An improved (Auto:I, LSP:T) constrained iterative speech enhancement for colored noise environments," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 573–579, Novembro 1998.

[5] B. S. Atal e S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, pp. 637–655, Abril 1971.

[6] A. H. Gray e J. D. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 380–391, Outubro 1976.

[7] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, pp. 67–72, Fevereiro 1975.

[8] M. R. Sambur e N. S. Jayant, "Lpc analysis/synthesis from speech inputs containing quantizing noise or additive white noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 488–494, Dezembro 1976.

[9] B. J. McDermott, "Multidimensional analysis of circuit quality judgments," *Journal of the Acoustical Society of America*, vol. 45, pp. 774–781, mar 1969.

[10] B. J. Dermott, C. Scagliola, e D. J. Goodman, "Perceptual and objective evaluation of speech processed by adaptive differential PCM," em *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 581–585, 1978.

[11] J. John R. Deller, J. H. L. Hansen, e J. G. Proakis, *Discrete-time Processing of Speech Signals*. IEEE Press, 2000.

[12] L. R. Rabiner e R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[13] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technology*, pp. 40–49, Abril 1982.

[14] J. Tierney, "A study of LPC analysis of speech in additive noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 389–397, Agosto 1980.

[15] S. M. Kay, "Noise compensation for autoregressive spectral estimates," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 292–303, Junho 1980.

[16] J. S. Lim e A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 197–210, Junho 1978.

[17] J. H. L. Hansen e M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, pp. 795–805, Abril 1991.

[18] F. Itakura, "Line spectral representation of linear predictive coeficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, p. S35, 1975.

[19] J. S. Qifang Zhao, Tetsuya Shimamura, "Linear predictive analysis of noisy speech," em *Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing*, pp. 585–588, 1997.

[20] G. S. Kang e L. J. Fransen, "Quality improvement of LPC-processed noisy speech by using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 939–942, Junho 1989.

[21] C. A. McGonegal, L. R. Rabiner, e A. E. Rosemberg, "A semiautomatic pitch detector (SAPD)," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, pp. 570–574, Dezembro 1975.

[22] L. R. Rabiner, M. J. Cheng, A. E. Rosemberg, e C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 399–418, Outubro 1976.

[23] M. J. Ross, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, pp. 353–362, 1974.

[24] W. Zhang, G. Xu, e Y. Wang, "Pitch estimation based on circular AMDF," em *Proc. IEEE ICASSP*, vol. 1, pp. 341–344, 2002.

[25] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-20, pp. 367–377, Dezembro 1972.