Estudo e Aperfeiçoamento de um Vocoder de Transformada Senoidal

Fábio Augusto Ribeiro do Nascimento

Dissertação apresentada ao Instituto Nacional de Telecomunicações, como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Francisco José Fraga da Silva

Santa Rita do Sapucaí
2004

Dissertação defendida e aprovada em 30/07/2004 pela comissão julgadora:

(Prof. Dr. Francisco José Fraga da Silva / INATEL)

(Prof. Dr. Fábio Violaro / UNICAMP)

(Prof. Dr. Carlos Alberto Ynoguti / INATEL)

Coordenador do Curso de Mestrado

Prof. Dr. Adonias Costa da Silveira

Dedicatória

À minha família.

Agradecimentos

- À minha família
- Ao Prof. Dr. Francisco José Fraga
- Ao Prof. Dr. Carlos Alberto Ynoguti
- Ao Prof. Carlos Nazareth Mota
- Aos amigos do mestrado

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

CELP	*Coded Excited Linear Prediction* - Predição Linear de Excitação por Código
FFT	*Fast Fourier Transform* - Transformada Rápida de Fourier
ITU	*International Telecommunications Union* - União de Telecomunicações Internacionais
LPC	*Linear Prediction Coding* - Codificação por Predição Linear
LSF	*Line Spectral Frequency* - Freqüência Espectral Discreta
MELP	*Mixed Excitation Linear Prediction* - Predição Linear de Excitação Mista
MOS	*Mean Opinion Score* - Classificação de Opinião Média
MSE	*Mean-Squared-Error* - Erro Médio Quadrático
PESQ	*Perceptual Evaluation of Speech Quality* - Avaliação Perceptiva de Qualidade de Fala
SEEVOC	*Spectral Envelope Estimation Vocoder* - Vocoder de Estimativa de Envoltória Espectral
STC	*Sinusoidal Transform Coding* - Codificação por Transformada Senoidal
STFT	*Short-Time Fourier Transform* - Transformada de Fourier de Tempo Curto
TIMIT	*Texas Instrument and the Massachusetts Institute of Technology*

Lista de Símbolos

ω_l	Freqüência de cada componente senoidal l
ϕ_l	Fase de cada componente senoidal l
Φ_s	Fase de sistema
Φ_a	Fase final da função de sistema
φ	Fase da função de sistema proposta por Chang e Wang
Ψ	Fase de sistema do filtro *all-pole*
$\hat{\theta}$	Fase sintética do sinal de fala *all-pole*
ω	Freqüência, em radianos
ω_0	Freqüência fundamental, em radianos
ω_c	Freqüência de corte dependente de sonoridade, em radianos
ϵ	Erro médio quadrático
ρ	Função de Verossimilhança do valor do Pitch (Pitch Likelihood Function)
$\hat{\sigma}$	Parâmetro ganho do modelo *all-pole*
s	Sinal de fala discreto no domínio do tempo
S	Sinal de fala discreto no domínio da freqüência
n	Índice de uma amostra no domínio do tempo
k	Índice de uma amostra no domínio da freqüência
q	Índice de um quadro
$\hat{e}$	Sinal de excitação estimado representado no domínio do tempo
$\hat{E}$	Sinal de excitação estimado representado no domínio da freqüência
n_0	Intervalo de ajuste, em amostras, necessário ao sincronismo dos pulsos de pitch de quadros vizinhos (Onset time)
$\hat{s}$	Sinal de fala sintético

N	Número de amostras
L	Número de componentes senoidais
A_l	Amplitude de cada componente senoidal l
T_0	Período fundamental do sinal de fala sonoro, em amostras
P_0	Período de pitch estimado do sinal de fala, em amostras
$ssinc$	Função *sinc* modificada
w	Função Janela de Hamming
w_s	Função Janela de Superposição de Síntese
T	Diferença, entre o índice da amostra temporal inicial do quadro q e a do quadro $q+1$ (incremento entre quadros)
$\bar{A}$	Vetor Envelope do Espectro
D	Função de aproximação de um pico do espectro
F	Intervalo equivalente ao valor da freqüência fundamental em radianos
P_v	Probabilidade de sonoridade
P_s	Potência do sinal de voz
SNR	Relação Sinal Ruído
H	Função de Transferência do Filtro Composto
H_s	Função de Transferência
H_a	Função de Transferência do Filtro *all-pole*
R	Função de autocorrelação do sinal de fala
$\mathbf{a}$	Vetor parâmetros coeficientes do filtro *all-pole*
p	Ordem do filtro *all-pole*
$\tilde{g}$	Parâmetro ganho decodificado

Resumo

Um modelo básico de representação senoidal da fala é apresentado e técnicas recentes de codificação dos parâmetros do modelo são estudadas. Como complemento, apresentam-se novos métodos para a melhoria da qualidade da fala a 2,4 e 4,8 kbps utilizando a representação senoidal. Um dos métodos apresentados é o refinamento na estimativa do pitch, uma vez que a precisão dessa estimativa é essencial para um modelo harmônico senoidal da fala. Outra proposta é um procedimento mais eficiente para a estimativa dos parâmetros do envelope do espectro, o qual explora a característica psico-acústica conhecida como *Subjective Loudness* do sistema de audição humano. Esta técnica mostrou-se capaz de reduzir a quantidade de parâmetros necessários para representar o envelope do espectro sem interferir na qualidade final da fala. Propõe-se também uma abordagem alternativa para a etapa de composição da fase sintética, de tal modo que um melhor custo de processamento é obtido sem degradação de qualidade. Os resultados experimentais indicam que o uso de todos esses métodos em combinação com o sistema básico de codificação senoidal aumenta a eficiência do Vocoder de Transformada Senoidal operando a 2,4 e 4,8 kbps, melhorando-lhe a qualidade final da fala e o seu custo computacional.

Palavras-chave: Codificação senoidal, Excitação senoidal, Modelo harmônico, Modelo senoidal.

Abstract

The basic sinusoidal analysis/synthesis system is presented and recent parameters codification techniques are studied. In addition, new methods for improving the speech quality of Sinusoidal Transform Vocoders at 2,4 and 4,8 Kbps are presented. One of them is a refinement in the pitch estimation, since pitch accuracy is essential for the efficiency of the harmonic sine-wave speech model. Another proposal is a more efficient procedure for estimation of the spectral envelope parameters, which exploits the subjective loudness psychoacoustic characteristic of the human hearing system. This technique reduces the amount of envelope parameters with no reduction in the speech quality. An alternative approach for the synthetic phase composition is also proposed, so that a better computational cost is obtained. Experimental results indicate that the use of all these methods in combination with the basic sinusoidal analysis/synthesis system enhances the performance of a Sinusoidal Transform Vocoder operating at 2,4 and 4,8 kbps, improving its quality and computational cost.

Keywords: Sinusoidal coding, Sinusoidal excitation, Harmonic model, Sine-wave model.

Capítulo 1

Introdução

1.1 Motivação

Uma das abordagens para a execução da tarefa de codificar eficientemente os sinais de fala é a utilização de um modelo de produção da mesma. O modelo mais difundido e utilizado pelos codificadores paramétricos é composto por um sinal de excitação, simulando o fluxo de ar proveniente dos pulmões e atravessando a laringe, aplicado à entrada de um filtro linear variante no tempo que simula as características ressonantes do trato vocal. Assim, o sinal de fala pode ser considerado como o resultado da passagem do sinal de excitação através desse filtro. Deste modo, o desafio proposto é promover uma eficiente modelagem, tanto do trato vocal pelo filtro, quanto do fluxo de ar (pulsado ou turbulento) pelo sinal de excitação. Em algumas aplicações, nas quais uma perda considerável na qualidade é aceitável, é suficiente assumir que a excitação pode ser modelada por apenas dois sinais: um trem de impulsos para representar a excitação sonora e ruído branco para representar a excitação surda. Tal modelo ficou conhecido como modelo binário de excitação [RABI78]. Para as aplicações que demandam uma qualidade de fala superior foram desenvolvidas, ao longo das últimas décadas, alterações para o modelo binário de excitação. Uma das alterações propostas para modelar o sinal de excitação é considerá-lo como sendo uma combinação de componentes senoidais com suas amplitudes, freqüências e fases peculiares [MCAU86]. A motivação para essa representação senoidal é que a excitação, no caso de sinais sonoros, quando perfeitamente periódica (o que é uma aproximação do caso real), pode ser representada por uma decomposição em Série de Fourier, ou seja, com uma composição linear de ondas senoidais harmônicas poderíamos representar perfeitamente uma excitação periódica. Generalizando, as ondas senoidais do modelo seriam não harmônicas se o sinal de excitação não fosse periódico, como no caso de sinais de fala surdos. Assim, ao definir um critério que extraia cor-

retamente as informações de amplitude, freqüência e fase dessas ondas senoidais, estaríamos, ao mesmo tempo, definindo um novo conjunto de parâmetros que poderiam ser devidamente codificados, com o objetivo de obter um vocoder de boa qualidade, baixa taxa de bits e um custo de processamento razoável.

1.2 Objetivos

O objetivo principal deste trabalho é demonstrar como um conjunto simplificado de parâmetros pode ser obtido através da representação senoidal, e como ele pode ser eficientemente codificado, de maneira a possibilitar reprodução da fala com boa qualidade, custo de processamento razoável e baixa taxa de bits, especificamente 2400 e 4800 bps. Este tipo de codificação é conhecido na literatura como codificação de fala por transformação senoidal (STC, do inglês: "Sinusoidal Transform Coding") [MCAU95]. Outro objetivo é a avaliação de seu desempenho e comparação de seus resultados com os codificadores da mesma categoria.

1.3 Desenvolvimento e Contribuições da Dissertação

A primeira etapa do trabalho foi o estudo dos conceitos básicos referentes à representação senoidal da fala. O aprofundamento teórico deste modelo evidenciou seu amplo potencial de aplicação na área de codificação de voz. A etapa seguinte foi o estudo detalhado do conceito de representação senoidal da fala, partindo-se do pressuposto da excitação senoidal do trato vocal e chegando-se, através da exploração de propriedades particulares aos sinais de voz, à determinação segura dos parâmetros necessários para a sua representação. A seguir, a etapa de codificação de cada parâmetro da representação senoidal foi pesquisada. As técnicas utilizadas nesse trabalho foram:

- um modelo harmônico para as freqüências das ondas senoidais (no caso de sons surdos as componentes não estão relacionadas harmonicamente, porém há técnicas que viabilizam a utilização do modelo harmônico ainda nestes casos);

- um modelo híbrido para a fase, com a combinação de um modelo de fase mínima só de pólos (conhecido do inglês como *all-pole* e lembrando que fase mínima significa todos os pólos e zeros no interior do círculo de raio unitário no plano z) e de um modelo de fase não mínima(pólos e zeros, sendo que os zeros não estão necessariamente todos em $|z| < 1$) para as fases das ondas senoidais e

- um modelo *all-pole* para a codificação das amplitudes [MCAU95].

Isto significa que não foi codificada uma freqüência específica para cada componente senoidal, mas um conjunto harmônico destas, montado a partir da freqüência fundamental do trecho do sinal de fala em análise (isto é uma generalização para o caso de quadros sonoros). As fases originais das componentes senoidais do sinal não são transmitidas e um modelo para elas é desenvolvido de forma a simulá-la integralmente no decodificador. As amplitudes das senóides também não são codificadas uma a uma, em seu lugar é codificado um conjunto de coeficientes capaz de representar a envoltória descrita pelas amplitudes no espectro de tempo curto. Cada uma dessas três técnicas é complexa e exige, por isso, um estudo dedicado, cujo detalhamento será exposto ao longo dessa dissertação. As técnicas acima foram estudadas e implementadas. Como fruto deste estudo e dos resultados práticos observados após a implementação das técnicas existentes, foram propostas diversas modificações, visando a obtenção de melhor qualidade da fala codificada. Tais modificações e aperfeiçoamentos propostos aos algoritmos de codificação encontrados na literatura constituem as contribuições científicas desse trabalho. Dentre elas, podemos destacar: um método para uma estimativa mais precisa da freqüência fundamental da fala no trecho de voz em análise (se o trecho for sonoro); contribuição para a determinação, no decodificador, da fase introduzida pelo trato vocal e principalmente a exploração de conceitos psicoacústicos referentes à audição humana, resultando em uma codificação mais eficiente da envoltória do espectro da fala. Estas contribuições foram também organizadas em forma de um artigo que foi oportunamente publicado no ICME2004 (IEEE International Conference on Multimedia and Expo) [FABI04]. O artigo está apresentado no Anexo B.

1.4 Organização da Dissertação

No capítulo seguinte, faz-se uma breve revisão histórica da literatura a respeito dos principais métodos envolvendo codificação senoidal da fala. Adiante, no capítulo 3, apresenta-se o modelo básico de representação da fala através de componentes senoidais. No capítulo 4 descreve-se o algoritmo extrator de pitch, ou freqüência fundamental, desenvolvido no presente trabalho. A informação de pitch é utilizada na substituição das freqüências das componentes senoidais por um conjunto de componentes senoidais harmonicamente relacionadas. O capítulo 5 descreve o modelo de reconstrução da fase das componentes senoidais que leva em consideração a fase imposta pelo trato vocal e a fase do sinal de excitação, seja ele surdo, sonoro ou uma combinação dos dois. No capítulo 6, apresenta-se o método utilizado para a representação paramétrica da envoltória do espectro

de tempo curto do sinal de voz.A descrição do sistema completo é consolidada no capítulo 7, onde descrevem-se os processos de quantização dos parâmetros do modelo a fim de obter a taxa de 2400 ou 4800 bps. Finalmente, no capítulo 8 são mostrados os resultados em termos de qualidade da fala reconstruída, bem como uma discussão e comparação com relação ao desempenho de outros codificadores padronizados da mesma categoria. As conclusões e sugestões para trabalhos futuros são apresentadas no capítulo 9.

Capítulo 2

Revisão da literatura

Como vimos, na representação senoidal a abordagem que se faz do modelo de excitação é, ao invés de usar seqüências de impulsos como no sistema multipulso [ATAL82] ou seqüências arbitrárias (ou palavras-código) como no sistema CELP [SCHR85], assumi-la como sendo o resultado da combinação de componentes senoidais de amplitudes, freqüências e fases particulares [MCAU86].

Há ainda uma variedade de outras abordagens para a representação de sinais de fala que são baseadas em modelos senoidais. O vocoder de fase [FLAN66] foi, talvez, a primeira tentativa de representar a forma de onda do sinal de fala por um conjunto de funções de faixa estreita de freqüências: um conjunto fixo de filtros passa-faixa é utilizado, onde define-se que cada um deixe passar somente uma componente de onda senoidal. O desvio de freqüência da onda senoidal em relação à freqüência central de cada filtro é estimado baseando-se na derivada da fase do sinal de saída de cada filtro. Este desvio de freqüência é quantizado e utilizado na síntese do vocoder.

PORTNOFF [PORT81] refinou o vocoder de fase representando cada componente de onda senoidal como resultado das contribuições da excitação e do trato vocal. As freqüências das ondas senoidais neste modelo foram forçadas a se relacionarem harmonicamente.

Outra evolução do vocoder de fase foi proposta por MALAH [MALA79], que assumiu que as freqüências das ondas senoidais eram harmônicas e assim definiu um banco de filtros adaptativo ao pitch, assegurando então, a grosso modo, uma onda senoidal por filtro.

O método de análise nesses sistemas não modela e estima explicitamente as componentes senoidais, mas, ao invés, as vê como saídas de um banco de filtros passa-faixa uniformemente espaçados. A forma de onda sintética pode ser vista como sendo a soma das saídas modificadas desse banco de filtros.

Embora fala de qualidade razoável tenha sido sintetizada utilizando essas

técnicas, o fato de se codificar somente a variação da fase significa que a informação de fase absoluta é perdida, e isso implica em degradação. Uma abordagem diferente foi tomada por HEDELIN [HEDE81], que propôs um modelo senoidal independente de pitch e a fase de cada onda senoidal como sendo a integral da freqüência instantânea associada, mas que ainda perde informação de fase absoluta.

Outro sistema de codificação baseado em componentes senoidais foi desenvolvido por ALMEIDA & SILVA [ALME84]. Em contraste com a abordagem de HEDELIN, o sistema deles usa uma estimativa do pitch durante a fala sonora para estabelecer um conjunto harmônico de ondas senoidais. As fases das ondas senoidais são computadas nas freqüências harmônicas do sinal e identificadas no espectro (Transformada de Fourier de Tempo Curto). Para compensar qualquer erro que possa ser introduzido como resultado da representação harmônica das ondas senoidais, uma forma de onda residual é codificada juntamente com os parâmetros das ondas senoidais em questão. Para representar a fala surda, o modelo usa um conjunto de funções base de faixa estreita [MARQ88].

Outra abordagem para representar fala surda no contexto do modelo senoidal é a geração explícita de ruído via filtragem linear de ruído branco, sempre que componentes de fala surda sejam detectados nas diferentes bandas. Esta abordagem, desenvolvida por GRIFFIN & LIM [GRIF88], conhecida como vocoder de Excitação Multibanda (MBE - Multi-Band Excitation), utiliza reconstrução baseada em superposição e adição na sintetização de fala em bandas caracterizadas como surdas. Para bandas caracterizadas como sonoras o sistema usa um "banco de osciladores", o qual é simplesmente outro termo para a análise e síntese senoidal a ser estudada como base desse trabalho.

Mais recentemente, KLEIJN & HAAGEN [KLEI95] usaram uma versão do sistema senoidal para rastrear a variação dos parâmetros das componentes senoidais e assim estabeleceram uma base para codificação da fala a 2400 b/s. Finalmente, MCAULAY & QUATIERI [MCAU95] propuseram, além da técnica de análise e síntese caracterizada pelas amplitudes, freqüências e fases das componentes senoidais, um modelo harmônico para as freqüências das ondas senoidais (harmônico não é um termo correto, pois serve tanto para quadros sonoros quanto para surdos, porém esta nomenclatura está sendo utilizada em respeito aos autores McAulay e Quatieri), um modelo de fase mínima para as fases das ondas senoidais e um modelo só de pólos (*all-pole model*) para a codificação das amplitudes das ondas senoidais. O sistema básico de representação senoidal da fala é apresentado a seguir.

Capítulo 3

O Modelo Senoidal de Análise e Síntese da Fala

No modelo de produção da fala descrito por Rabiner [RABI78], a sua forma de onda é assumida como sendo a saída de um sistema linear variante no tempo que representa as características do trato vocal e cuja entrada é uma forma de onda de excitação proveniente do pulmão, passando pelas cordas vocais. A função de excitação é geralmente representada por um trem de pulsos periódicos durante a fala sonora, onde o intervalo entre cada pulso corresponde ao "pitch" do locutor, e representada como um sinal semelhante a ruído durante fala surda. De maneira alternativa, tal excitação binária (sonora/surda) pode ser substituída por uma soma de ondas senoidais [MCAU86]. A justificativa para essa representação senoidal é que a excitação sonora, quando perfeitamente periódica, pode ser representada por uma decomposição em série de Fourier. As ondas senoidais no modelo não serão harmônicas se o sinal de excitação não for periódico (excitação surda ou mista). Passando essa representação senoidal da excitação através do trato vocal (sistema linear), obtém-se a representação senoidal da forma de onda da fala. Em um dado intervalo de análise, suficientemente pequeno de modo a permitir que seja desprezada qualquer modificação do trato vocal, a fala pode ser descrita pelo sinal de tempo discreto

$$s(n) = \sum_{l=1}^{L} A_l cos(\omega_l n + \phi_l) \tag{3.1}$$

onde A_l, ω_l e ϕ_l representam a amplitude, freqüência e fase de cada componente senoidal e L é o número de ondas senoidais utilizadas na representação. A viabilidade dessa representação é resultante do fato de que os parâmetros das componentes senoidais variam lentamente em relação à duração da resposta impulsiva

do sistema trato vocal. Se não fosse assim, uma vez que dentro do intervalo de análise não se prevê modificação do trato vocal, os parâmetros acima deveriam ser representados de modo a variarem em função do tempo.

3.1 Estimativa dos parâmetros senoidais da fala

O problema a ser resolvido na análise e síntese é tomar uma forma de onda em um dado intervalo da fala, extrair os parâmetros que melhor representem uma porção praticamente estacionária da mesma e usá-los diretamente, ou suas versões codificadas, para reconstruir uma aproximação que seja a mais fiel possível ao sinal de fala original. Em primeira instância, a estimativa será baseada na observação de que quando a fala é sonora seu sinal pode ser considerado periódico e, portanto, seu espectro tende a apresentar raias harmônicas, definidas por uma série de componentes senoidais (Série de Fourier). Então, ao considerar a fala como perfeitamente periódica, a equação 3.1 se reduz a

$$s(n) = \sum_{l=1}^{L} A_l cos(nl\omega_0 + \phi_l) \tag{3.2}$$

na qual as freqüências das componentes senoidais são múltiplos da freqüência fundamental ω_0 e as amplitudes e fases correspondentes são dadas pelos valores complexos das amostras do espectro do sinal de fala.

O espectro de freqüências de um trecho sonoro do sinal de fala pode ser obtido utilizando-se a representação em série trigonométrica de Fourier de $s(n)$:

$$S(l\omega_0) = \frac{1}{T_0} \sum_{n=d}^{d+T_0} s(n) e^{-jnl\omega_0} \tag{3.3}$$

onde d é uma amostra qualquer do sinal de fala e $T_0 = 2\pi/\omega_0$ é o período fundamental do sinal de fala sonora, em amostras.

Note-se que $S(l\omega_0)$ será um espectro discreto de freqüências, com amostras situadas somente nas posições múltiplas de ω_0. Em termos práticos, podemos considerar que o sinal de fala sonoro é periódico somente em pequenos intervalos de tempo, o que força $s(n)$ a ser segmentado (ou janelado). Uma vez janelado, o sinal de fala apresenta no domínio da freqüência um espectro contínuo, ao invés do espectro discreto apresentado pelo sinal periódico de duração infinita. Tal espectro é obtido, não pela série, mas pela Transformada de Fourier normalizada de tempo curto do sinal de fala (em inglês: 'Short-Time Fourier Transform', conhecida pela sigla STFT). Neste caso, a expressão do espectro, considerando

um trecho arbitrário do sinal de fala fica:

$$S(\omega) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} s(n) e^{-jn\omega}. \tag{3.4}$$

Neste caso, observa-se que a magnitude de $S(\omega)$ apresenta uma série de picos que situam-se nas freqüências múltiplas de ω_0 (harmônicos do pitch). Sendo assim, a STFT do sinal de fala fornece a estimativa de amplitudes como $A_l = |S(l\omega_0)|$ e a estimativa de fases como $\phi_l = arg[S(l\omega_0)]$. A afirmativa de que a magnitude da STFT possui picos em múltiplos de ω_0 é demonstrada a seguir.

Reescrevendo a equação (3.2) através de uma representação complexa, tem-se:

$$s(n) = \sum_{l=1}^{L} \gamma_l e^{jnl\omega_0} \quad ^1 \tag{3.5}$$

onde $\gamma_l = A_l e^{j\phi_l}$.

Assim, a STFT de (3.5) fica:

$$S(\omega) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} s(n) e^{-jn\omega}$$

$$S(\omega) = \sum_{n=-N/2}^{N/2} \left(\sum_{l=1}^{L} \gamma_l \frac{e^{jnl\omega_0}}{N+1} \right) e^{-jn\omega}$$

$$S(\omega) = \sum_{l=1}^{L} \frac{\gamma_l}{N+1} \left(\sum_{n=-N/2}^{N/2} e^{j(l\omega_0 - \omega)n} \right)$$

$$S(\omega) = \sum_{l=1}^{L} \frac{\gamma_l}{N+1} \left(\sum_{m=0}^{N} (e^{j\Omega})^m (e^{-j\Omega})^{N/2} \right)$$

onde $m = n + N/2$ e $\Omega = l\omega_0 - \omega$. Além disso, sabemos que a somatória de uma progressão geométrica com razão α é dada por

$$\sum_{m=0}^{N} \alpha^m = \frac{1 - \alpha^{N+1}}{1 - \alpha}.$$

[1] O sinal real $s(n)$ em (3.2) consiste na parte real do sinal $s(n)$ em (3.5). A notação foi mantida a fim de simplificar a dedução das equações seguintes.

Então, sendo $\alpha = e^{j\Omega}$, temos que

$$S(\omega) = \sum_{l=1}^{L} \frac{\gamma_l}{N+1} \left[e^{-j\frac{\Omega N}{2}}\right] \left[\frac{1 - e^{j\Omega(N+1)}}{1 - e^{j\Omega}}\right]$$

Multiplicando e dividindo o numerador por $e^{-j\Omega(N+1)/2}$ e multiplicando e dividindo o denominador por $e^{-j\Omega/2}$, obtém-se:

$$S(\omega) = \sum_{l=1}^{L} \frac{\gamma_l}{N+1} e^{-j\frac{\Omega N}{2}} \left[\frac{e^{j\frac{\Omega(N+1)}{2}}}{e^{j\frac{\Omega}{2}}}\right] \left[\frac{e^{-j\frac{\Omega(N+1)}{2}} - e^{j\frac{\Omega(N+1)}{2}}}{e^{-j\frac{\Omega}{2}} - e^{j\frac{\Omega}{2}}}\right]$$

$$S(\omega) = \sum_{l=1}^{L} \frac{\gamma_l}{N+1} e^{-j(\frac{\Omega N}{2} - \frac{\Omega N}{2} + \frac{\Omega}{2} - \frac{\Omega}{2})} \frac{sen(\frac{\Omega(N+1)}{2})}{sen(\frac{\Omega}{2})}$$

$$S(\omega) = \sum_{l=1}^{L} \frac{\gamma_l}{N+1} \frac{sen(\frac{\Omega(N+1)}{2})}{sen(\frac{\Omega}{2})}$$

$$S(\omega) = \sum_{l=1}^{L} \gamma_l ssinc(l\omega_0 - \omega) \tag{3.6}$$

onde a função $ssinc(\Omega)$ é definida como

$$ssinc(\Omega) = \frac{sen(\frac{\Omega(N+1)}{2})}{(N+1)sen(\frac{\Omega}{2})} = ssinc(l\omega_0 - \omega).$$

Note-se que o somatório das funções "ssinc" dão ao espectro de tempo curto um aspecto de "picos e vales", como mencionado. A Figura 3.1 ilustra a superposição de duas funções *ssinc* adjacentes.

A Figura 3.2 ilustra o espectro de tempo curto de um sinal de fala sonoro janelado.

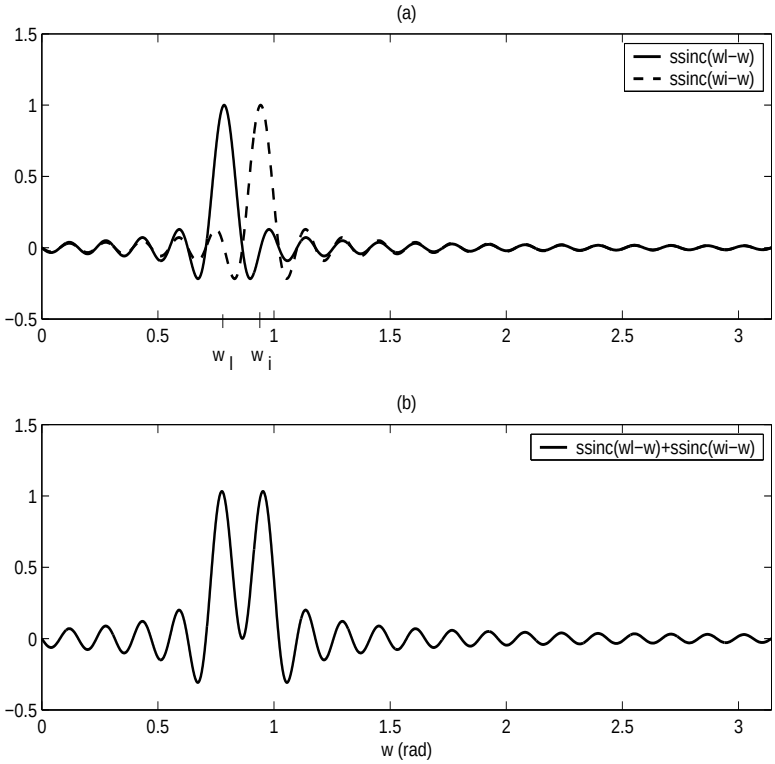

Figura 3.1: *(a)Função ssinc*$(\omega_l - \omega)$ *sobreposta à função ssinc*$(\omega_i - \omega)$ *e (b)Função ssinc*$(\omega_l - \omega) + ssinc(\omega_i - \omega)$.

Quando a fala não é periódica, o espectro ainda apresentará uma multiplicidade de picos, mas situados em freqüências que não são necessariamente harmônicas. Porém, ainda assim, tais freqüências (onde se localizam os picos) podem ser utilizadas para identificar os parâmetros das componentes senoidais. Neste caso, as freqüências correspondem às posições dos picos do espectro e as fases e amplitudes são computadas a partir das partes real e imaginária da amostra correspondente na STFT.

A análise acima assume implicitamente que a STFT é computada usando-se uma janela retangular. Além disso, vemos que dois picos vizinhos são realçados no espectro se os zeros das funções *ssinc* coincidem. Ao considerar ω_l a freqüência correspondente a um pico do espectro (no caso o pico da função $ssinc(\omega_l - \omega)$) e ω_i a freqüência correspondente ao pico vizinho (no caso o pico da função $ssinc(\omega_i - \omega)$) e considerando-se que $i = l+1$, a separação entre ω_l e ω_i (freqüência fundamental da fala sonora ω_0) deve ser tal que os zeros de cada uma das *ssinc*'s vizinhas coincidam. A Figura 3.1 ilustra o realce dos picos quando há coincidência dos pontos de zero das funções *ssinc*. O primeiro zero da função *ssinc* é obtido quando

$$sen\left[(\omega_l - \omega)\frac{N+1}{2}\right] = 0,$$

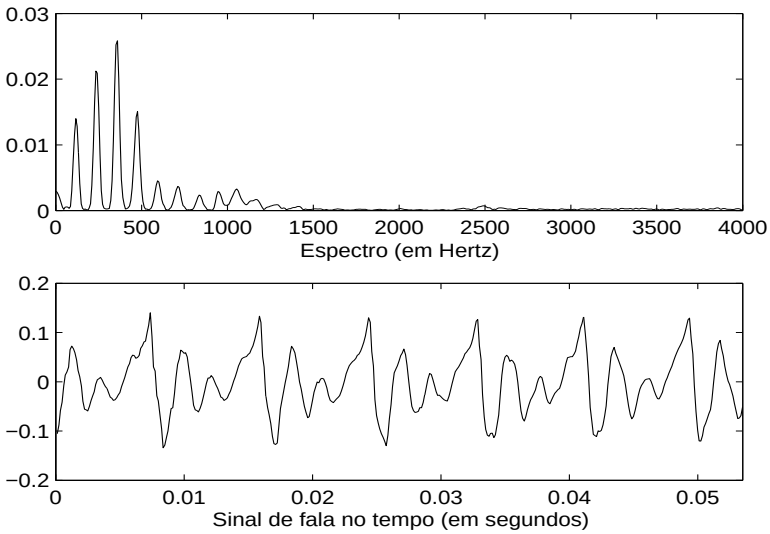

Figura 3.2: *Quadro de um sinal de fala sonoro representado nos domínios do tempo e da freqüência*

ou seja,

$$(\omega_l - \omega)\frac{N+1}{2} = \pi,$$

ou ainda:

$$\omega_l - \omega = \frac{2\pi}{N+1}.$$

Isto significa dizer que a separação entre ω_l e ω_i deve valer duas vezes $2\pi/(N+1)$. Então, para sinais de fala periódicos, devemos ter

$$|\omega_l - \omega_i| = \frac{4\pi}{(N+1)}$$

$$\omega_0 = \frac{4\pi}{(N+1)}$$

$$\frac{2\pi}{T_0} = \frac{4\pi}{(N+1)}$$

onde T_0 é o período fundamental da fala, em número de amostras. Esta expressão define a condição:

$$N + 1 = 2\,T_0. \tag{3.7}$$

A equação acima estabelece que o tamanho da janela retangular deve ser igual a duas vezes o período de pitch (em amostras) do sinal de fala para que os picos do

espectrograma sejam realçados. Porém, ainda assim, deve-se considerar o efeito dos lóbulos laterais das funções *ssinc* que interferem com os lóbulos principais de suas vizinhas. Para minimizar esse efeito, e promover a maior definição possível da posição correta dos picos do espectro de tempo curto, utiliza-se, comumente, a janela Hamming, ao invés da retangular. Contudo, sabe-se que ela reduz o efeito dos lóbulos laterais sob o preço de alargar o lóbulo principal da função *ssinc*. Deste modo, ao utilizar-se a janela Hamming, admite-se, por segurança, uma janela temporal do sinal de fala igual a duas vezes e meia o seu período de pitch. Ou seja:

$$N + 1 = 2,5\, T_0. \tag{3.8}$$

Uma vez que a largura de uma janela de análise foi especificada, a janela Hamming adaptativa ao pitch, $w(n)$, é computada e normalizada:

$$\sum_{n=-N/2}^{N/2} w(n) = 1. \tag{3.9}$$

A normalização acima garante módulo unitário para o pico da transformada de Fourier da janela Hamming, de modo que ela não interfira no valor dos picos do espectro, ou seja, tais picos terão valor exatamente igual à amplitude da onda senoidal correspondente.

O sinal janelado pode ser representado por

$$s^k(n) = \sum_k s(n).w(n - kT),$$

onde k é o número identificador da janela em análise, em relação à origem no eixo do tempo e T é o módulo da diferença, em número de amostras, entre a posição inicial do quadro k e a posição inicial do quadro $k-1$ (incremento entre quadros).

O posicionamento do sinal janelado é um fator importante. Tal posicionamento é relevante para a computação das fases do sinal de fala. Tipicamente, em processamento seqüencial de quadros, o sinal janelado é estabelecido no intervalo $0 \leq n \leq N$ [1], sendo simétrico em relação a $\frac{N}{2}$. Este posicionamento acrescenta à fase do sinal analisado uma fase linear igual a $-\omega\frac{N}{2}$. Uma vez que N é da ordem de 100 a 400 amostras (dependendo da freqüência de amostragem do sinal de fala), qualquer erro na estimativa das freqüências resulta em um grande erro na estimativa da fase e uma conseqüente degradação na qualidade da fala recon-

[1] Transformada Discreta de Fourier: $\sum_{n=0}^{N} x(n)e^{-j\frac{2\pi k}{N+1}n}$, onde $x(n)$ é o vetor a ser transformado, k é uma amostra no espectro e $N+1$ é o número de amostras do espectro e também da janela temporal. Note-se que o somatório varia de $n = 0$ até $n = N$, ao invés de variar entre $n = -\frac{N}{2}$ e $n = \frac{N}{2}$. Isto acrescenta uma fase $-\omega\frac{N}{2}$ à fase original do sinal.

struída. Apesar de estarmos falando em Transformada de Fourier e em espectro contínuo de freqüências, em termos práticos o espectro é discreto, pois é obtido por meio de alguma ferramenta computacional. Assim, o espectro tem um número de pontos que determina a sua resolução. Um erro de um ponto no espectro de freqüências, por exemplo, resulta em um erro de fase igual a $(\frac{2\pi}{M})(\frac{N}{2})$ (onde M é o número de pontos que define a resolução do espectro), que pode ser da ordem de π. Essa questão deve ser considerada na estimativa da fase e corrigida, se for o caso. Neste trabalho utilizou-se a função FFT do software Matlab® para a computação da STFT, cuja fórmula utiliza n variando de 0 a N, ao invés de $-N/2$ a $N/2$. Assim, a correção da fase foi obrigatória.

As aproximações utilizadas para definição do extrator de parâmetros descrito acima foram baseadas na suposição de fala sonora. Em nenhum momento as propriedades da fala surda foram levadas em consideração. Para fazer isso de uma maneira que resulte em amostras de amplitude descorrelacionadas, como deve ser no caso de sinal semelhante à ruído, faz-se necessário o uso do conceito da expansão de Karhunen-Loève para sinais semelhantes a ruído [TREE68]. Uma análise baseada no critério de Karhunen-Loève mostra que a representação senoidal é válida somente se as freqüências das amostras no espectro forem próximas o suficiente para garantir uma alteração suave na densidade espectral de potência. Se a largura da janela de análise for forçada a ter pelo menos 20 ms (o que determina um estreitamento mínimo da função *ssinc*), haverá, em média, um conjunto de picos separados de aproximadamente 100 Hz um do outro no espectro, o que determina uma amostragem suficientemente densa, de modo a satisfazer a condição de suavidade do espectro de potência da fala imposta pelo critério de Karhunen-Loève. Obedecida a condição de Karhunen-Loève, pode-se obter os parâmetros senoidais para a fala surda através da extração dos valores de amplitude, freqüência e fase dos picos do espectro normalmente.

Toda a análise anterior fornece uma justificativa, em parte analítica (sinais de fala sonoros) e em parte heurística (sinais de fala surdos), para a representação da forma de onda da fala em termos de amplitudes, freqüências e fases de um conjunto de ondas senoidais obtidos a partir da análise de um quadro. Como a análise da fala se desenvolve quadro a quadro, obtém-se diferentes conjuntos desses parâmetros para cada quadro ou janela temporal. O próximo desafio então é a associação das amplitudes, freqüências e fases estimadas em um quadro com as obtidas num quadro anterior, de maneira a definir conjuntos de ondas senoidais que se modifiquem suavemente no tempo.

3.2 Síntese senoidal por superposição

Se as amplitudes, freqüências e fases que são estimadas para o k-ésimo quadro são denotadas por A_l^k, ω_l^k e ϕ_l^k respectivamente, então a fala sintética para tal quadro pode ser obtida usando

$$\hat{s}^k(n) = \sum_{l=1}^{L^k} A_l^k cos(\omega_l^k n + \phi_l^k). \tag{3.10}$$

Uma vez que os parâmetros senoidais serão variantes no tempo, os pontos de transição entre os quadros sofrerão descontinuidades no sinal reconstituído, a menos que algum tipo de precaução seja tomada para garantir uma interpolação suave entre os parâmetros de quadros sucessivos. Muitos métodos têm sido desenvolvidos para resolver esse problema, porém, um método simples e satisfatório é a utilização de um interpolador de superposição (conhecido em inglês como "overlap-add interpolator") [OPPE83] [MCAU88], desde que o tamanho do quadro do sinal de fala seja suficientemente pequeno, de modo a atender o princípio de quase-estacionariedade [RABI78].

Neste caso, a forma de onda da fala sintética é obtida aplicando (3.10) nos quadros $k-1$ e k, por exemplo, a fim de gerar as formas de onda $\hat{s}^{k-1}(n)$ e $\hat{s}^k(n)$. Assim, estas são apropriadamente ponderadas e superpostas uma com a outra. Computacionalmente isto é equivalente a

$$\hat{s}(n) = \sum_k \{\omega_s(n)\hat{s}^{k-1}(n) + \omega_s(n-T)\hat{s}^k(n-T)\}, \tag{3.11}$$

onde T é o módulo da diferença, em número de amostras, entre a posição inicial do quadro k e a posição inicial do quadro $k-1$ (incremento entre quadros) e $\omega_s(n)$ é a janela de superposição de síntese, que deve ser tal que

$$\sum_k \omega_s(n - kT) \ = 1, \tag{3.12}$$

Tipicamente, janelas Triangular, Hanning e Trapezoidal têm sido usadas no processo de superposição de quadros. Neste trabalho optou-se pela janela triangular.

3.3 Resultados experimentais

Uma simulação em Matlab®, usando o sistema de análise e síntese ilustrado na Figura 3.3, foi desenvolvida com a finalidade de determinar a eficácia do modelo senoidal proposto.

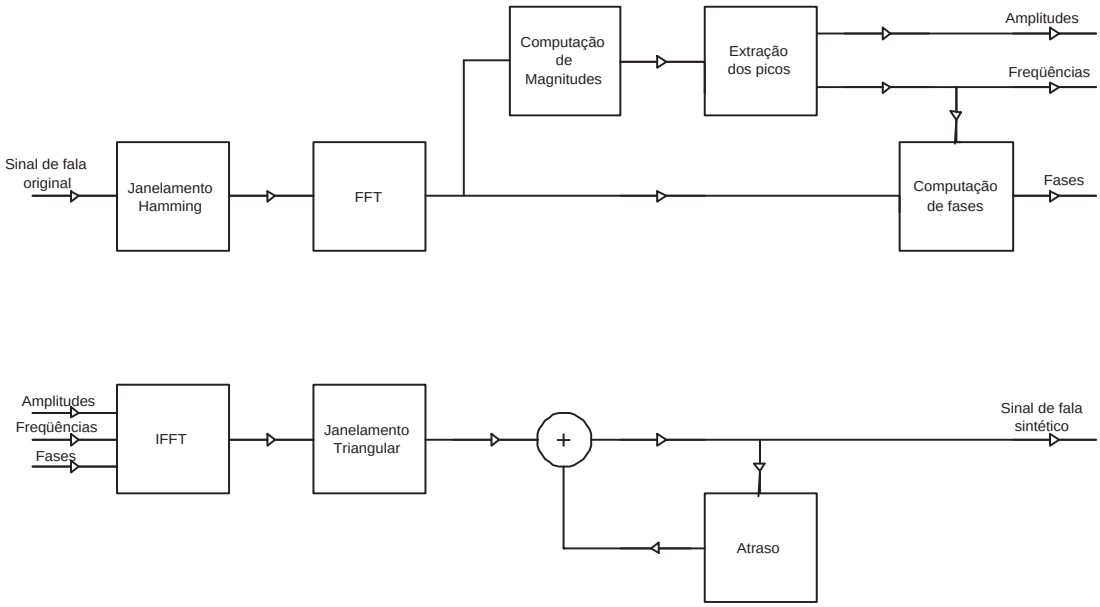

Figura 3.3: *Diagrama em blocos de um sistema de análise e síntese senoidal*

Note-se que uma FFT inversa foi utilizada para computar (3.10).

A fala processada na simulação foi amostrada a 8 kHz e analisada com intervalo de sobreposição de quadros de 10 ms. Uma FFT de 1024 pontos usando uma janela Hamming adaptativa, de largura igual a duas vezes e meia o período de pitch médio em amostras, possibilitou uma estimativa precisa dos picos, tanto para sinais de fala sonoros quanto para surdos, desde que a janela tivesse no mínimo 20 ms de largura. No caso da síntese, uma condição razoável para o tamanho do intervalo de sobreposição de quadros é que ele seja menor que 12,5 ms, uma vez que sinais de fala, em média, permanecem estacionários por até 25 ms. Neste trabalho, obedecendo o intervalo de sobreposição de quadros na analise, utilizou-se o intervalo de sobreposição de quadros da síntese igual a 10 ms. A Figura 3.4 ilustra o esquema de superposição.

Um grande banco de dados foi processado com esse sistema e nota-se que a fala sintética, ao ser ouvida, é praticamente indistinguível da original, e mesmo a sua forma de onda é muito parecida com a do sinal original. Isto sugere que a condição de quase-estacionariedade da fala foi atendida e que o uso de um

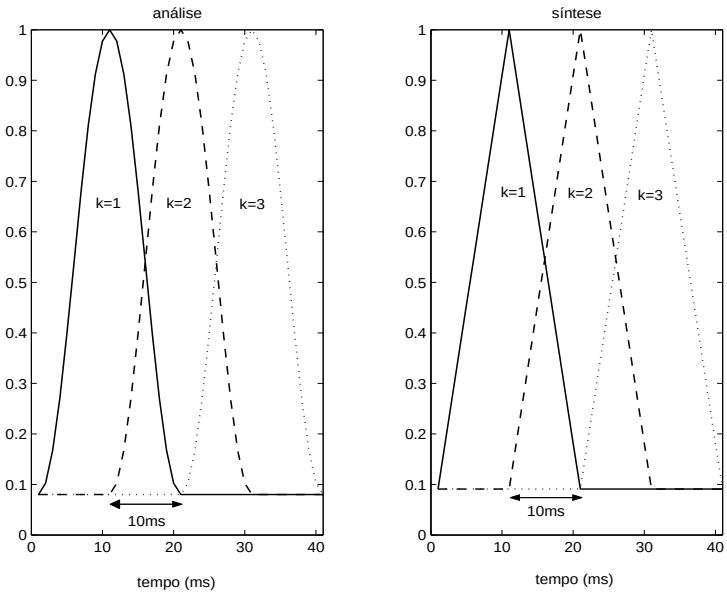

Figura 3.4: *Esquema de superposição de quadros tanto na análise quanto na síntese*

modelo paramétrico, baseado em amplitudes, freqüências e fases de um conjunto de componentes senoidais, se justifica tanto para os intervalos de fala surda quanto para sonora.

A Figura 3.5 (a) ilustra o espectrograma de um sinal de fala arbitrário original e a Figura 3.5 (b) ilustra o espectrograma do sinal sintético.

Embora o modelo senoidal tenha sido desenvolvido originalmente para um único locutor, seu modelo geral trabalha igualmente bem para sinais multi-locutores, música, fala com fundo musical e outros mais [MCAU86]. Além disso, ficou evidente que a reconstrução não é prejudicada com a presença de ruído. A fala ruidosa sintetizada, ao ser ouvida, também é praticamente indistinguível da original, graças à validade da representação de Karhunen-Loève para sinais semelhantes a ruído.

Contudo, o modelo básico de representação senoidal da fala se mostra inapropriado para a codificação da fala a baixas taxas de bit devido à grande quantidade de parâmetros a ser codificada. Com o objetivo de comprimir a taxa de dados, a classe de sinais a ser codificada deve ser restrita somente a sinais de fala. Assim, modelos mais estruturados para os parâmetros senoidais podem ser utilizados. No próximo capítulo será descrito um modelo harmônico para as freqüências das componentes senoidais. Este modelo não leva mais em consideração a freqüência de cada pico, mas sim, propõe um modelo matemático capaz de simular as freqüências dos picos por uma combinação de freqüências, envolvendo a freqüência fundamental da fala nos casos de quadros sonoros.

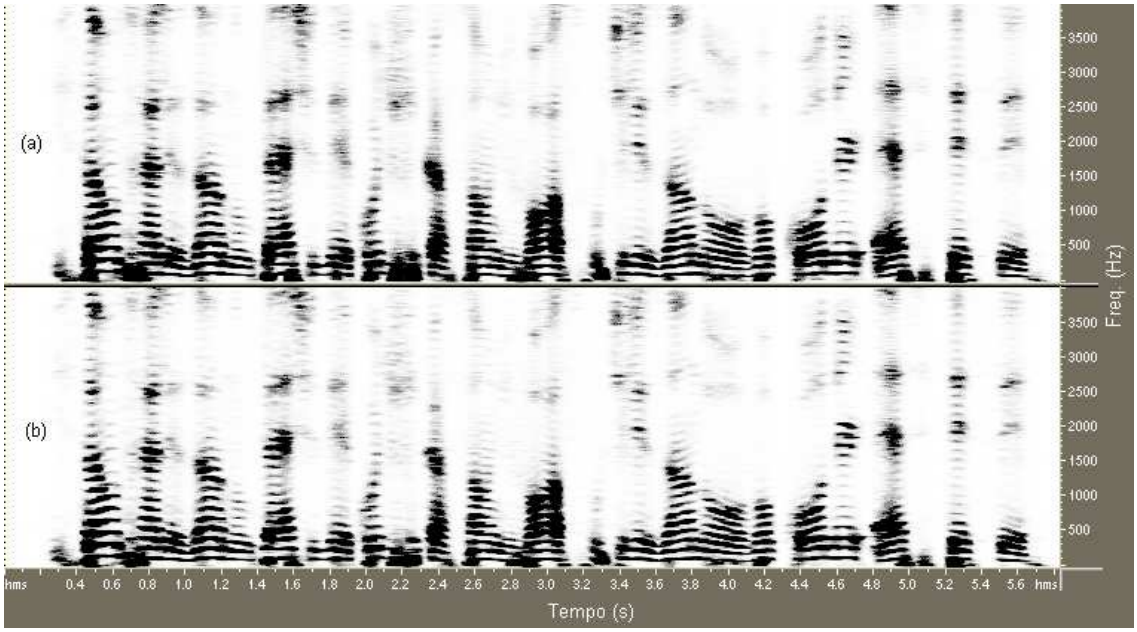

Figura 3.5: *Comparação entre os espectrogramas de sinais de fala original e sintético obtido a partir do modelo senoidal básico*

Capítulo 4

O modelo harmônico das ondas senoidais

O primeiro passo para a elaboração de um codificador de fala senoidal de baixa taxa de bits é desenvolver um modelo para as freqüências das componentes senoidais. Tal modelo deve permitir a substituição da freqüência particular de cada uma delas por um conjunto de freqüências harmonicamente relacionadas no caso de quadros sonoros. Assim o problema seria estimar a freqüência da componente fundamental de tal modo que o conjunto de componentes harmônicos seja o mais fiel possível ao conjunto de componentes medidos [MCAU90]. A freqüência fundamental estimada será, por facilidade de expressão, tratada como o pitch do locutor nos quadros sonoros (sabe-se que o pitch é, na verdade, a sensação auditiva de altura da fala). Durante fala surda a freqüência fundamental não tem significado físico, mas, com um projeto cuidadoso dos processos de estimativa e de síntese, pode-se obter uma representação efetiva da fala mesmo quando ela está no estado surdo.

Há diversos algoritmos de extração de pitch elaborados com base em informações temporais da fala, doravante denominados *algoritmos temporais* de detecção de pitch. Muitos deles são capazes de estimar o pitch de maneira precisa e confiável [KLEI93] [RABI78] [DELL99]. McAulay e Quatieri, visando uma estimativa mais eficiente quanto ao emprego dos parâmetros senoidais e além disso ainda mais precisa, apresentaram um procedimento de extração baseado nas informações da fala no domínio da freqüência. A próxima seção descreve detalhadamente esse procedimento, seguindo a pauta de McAulay e Quatieri no capítulo 4 da referência [MCAU95].

4.1 Estimativa da freqüência fundamental

Como um primeiro passo, assume-se que os quadros do sinal de fala já foram analisados em termos das componentes senoidais com as técnicas descritas no capítulo anterior. Assim, o sinal de fala $s(n)$ pode ficar representado como

$$s(n) = \sum_{l=1}^{L} A_l e^{j(n\omega_l + \theta_l)} \tag{4.1}$$

onde $\{A_l,\, \omega_l,\, \theta_l\}_{l=1}^{L}$ representa as amplitudes, freqüências e fases das L ondas senoidais que compõem o sinal de fala [1]. O objetivo é tentar substituir as componentes senoidais por um sinal arbitrário que apresente, justamente nas freqüências harmônicas, valores de amplitudes, freqüências e fases que sejam o mais próximo possível das amplitudes, freqüências e fases medidas. Este sinal poderia ser modelado como

$$\hat{s}(n; \omega_0, \phi) = \sum_{k=1}^{K} \bar{A}(k\omega_o) e^{j(nk\omega_0 + \phi_k)} \tag{4.2}$$

onde $\omega_0 = 2\pi f_0 / f_s$ é a freqüência fundamental normalizada, f_s é a freqüência de amostragem, K é o número de harmônicos do sinal na faixa espectral de voz, $\bar{A}(\omega)$ é a função de envoltória do trato vocal (envoltória do espectro) e $\phi = (\phi_1, \phi_2, ..., \phi_k)$ representa as fases dos harmônicos.

Deseja-se estimar as freqüências de pitch w_0 e as fases $(\phi_1, \phi_2, ..., \phi_k)$ de modo que $\hat{s}(n)$ seja o mais fiel possível a $s(n)$. O critério de aproximação entre $\hat{s}(n)$ e $s(n)$ utilizado é o da minimização do erro médio quadrático (MSE, do inglês: "Mean-Squared-Error"),

$$\epsilon(\omega_0, \phi) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} |s(n) - \hat{s}(n; \omega_o, \phi)|^2 \tag{4.3}$$

através de ω_0 e de ϕ. O MSE em (4.3) pode ser expandido como

$$\epsilon(\omega_0, \phi) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} \{|s(n)|^2 - 2\,\Re e[s(n)\hat{s}^*(n; \omega_o, \phi)] + |\hat{s}(n; \omega_o, \phi)|^2\} \tag{4.4}$$

O primeiro termo de (4.4) representa a potência no sinal medido e é indepen-

[1] A análise, como mencionada no capítulo anterior, está simplificada utilizando uma representação complexa das ondas senoidais e omitindo-se a notação $\Re e[.]$.

dente dos parâmetros desconhecidos. Ela é descrita por

$$P_s = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} |s(n)|^2 \tag{4.5}$$

Substituindo (4.2) no segundo termo de (4.4), tem-se a relação

$$\frac{1}{N+1} \sum_{n=-N/2}^{N/2} s(n)\hat{s}^*(n;\omega_o,\phi) = \frac{1}{N+1} \sum_{k=1}^{K} \bar{A}(k\omega_o)e^{-j\phi_k} \sum_{n=-N/2}^{N/2} s(n)e^{-jnk\omega_0} \tag{4.6}$$

Finalmente, substituindo (4.2) no terceiro termo de (4.4), tem-se a relação

$$\frac{1}{N+1} \sum_{n=-N/2}^{N/2} |\hat{s}(n;\omega_o,\phi)|^2 \cong \sum_{k=1}^{K} \bar{A}^2(k\omega_o)^1 \tag{4.7}$$

onde a aproximação é válida desde que a janela de análise satisfaça a condição $(N+1) \cong 2,5\,(2\pi/\omega_0)$, o que é garantido na prática (Relação de Parseval), definindo-se a janela de análise igual a pelo menos duas vezes e meia o período de pitch médio. Esta condição assume que o período de pitch médio já tenha sido computado, uma questão que será tratada mais adiante.

Seja

$$S(\omega) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} s(n)e^{-jn\omega} \tag{4.8}$$

a STFT normalizada (normalizada por causa da divisão por $N+1$) do sinal de fala de entrada. Utilizando (4.8) em (4.6) a expressão para o MSE em (4.4) torna-se

$$\epsilon(\omega_0,\phi) \cong P_s - 2\,\Re e\{\sum_{k=1}^{K} \bar{A}(k\omega_o)e^{-j\phi_k}S(k\omega_0)\} + \sum_{k=1}^{K} \bar{A}^2(k\omega_o) \tag{4.9}$$

Uma vez que os parâmetros fase afetam somente o segundo termo de (4.9), o MSE será minimizado, em relação à fase, ao definir-se

$$\phi_k = arg[S(k\omega_0)] \tag{4.10}$$

[1] A relação está demonstrada no Anexo A

e o MSE resultante será dado por

$$\epsilon(\omega_0) \cong P_s - 2\sum_{k=1}^{K} \bar{A}(k\omega_0)|S(k\omega_0)| + \sum_{k=1}^{K} \bar{A}^2(k\omega_0) \tag{4.11}$$

O pitch desconhecido afeta somente o segundo e o terceiro termo em (4.11), e isto pode ser combinado definindo-se

$$\rho(\omega_0) = \sum_{k=1}^{K} \bar{A}(k\omega_0)[|S(k\omega_0)| - \frac{1}{2}\bar{A}(k\omega_0)] \tag{4.12}$$

Assim, o MSE pode ser expresso como

$$\epsilon(\omega_0) = P_s - 2\rho(\omega_0) \tag{4.13}$$

Uma vez que o primeiro termo é uma constante conhecida, o MSE mínimo é obtido ao maximizar-se $\rho(\omega_0)$ através de ω_0.

No começo do capítulo pressupôs-se que os parâmetros de representação senoidal do sinal de fala de entrada são conhecidos. Neste ponto do trabalho faz-se necessário manipular (4.13) de modo a utilizá-los explicitamente. Para tal, começa-se por por substituir (4.1) em (4.5). Utilizando-se o mesmo raciocínio empregado na obtenção da relação (4.7), a potência do sinal $s(n)$ pode ser dada por

$$P_s = \sum_{l=1}^{L} A_l^2 \tag{4.14}$$

e então, como visto na dedução de (3.6), a STFT descrita em (4.8) torna-se

$$S(\omega) = \sum_{l=1}^{L} \gamma_l ssinc(l\omega_0 - \omega) \tag{4.15}$$

onde $\gamma_l = A_l e^{j\phi_l}$.

Ao assumir-se que as componentes senoidais estão bem resolvidas com a condição $N + 1 = 2,5\,T_0$, a magnitude de (4.15) pode ser aproximada por

$$|S(\omega)| \approx \sum_{l=1}^{L} A_l D(\omega_l - \omega) \tag{4.16}$$

onde

$$D(\omega_l - \omega) = \begin{cases} ssinc(\omega_l - \omega), & \text{se } |\omega_l - \omega| \leq \frac{2\pi}{N+1} \\ 0, & \text{caso contrário} \end{cases} \tag{4.17}$$

O critério de otimização então torna-se

$$\rho(\omega_0) = \sum_{k=1}^{K} \bar{A}(k\omega_0)[\sum_{l=1}^{L} A_l D(\omega_l - k\omega_0) - \frac{1}{2}\bar{A}(k\omega_0)] \tag{4.18}$$

A equação acima indica, de maneira geral, que $\rho(\omega_0)$ assume seu máximo valor quando $k\omega_0 = \omega_l$. Porém, tal indicação pode ser insuficiente, uma vez que, dependendo do valor de k, múltiplos ou submúltiplos de ω_0 podem também satisfazer a condição de maximização de $\rho(\omega_0)$. O fator decisivo é que a soma de todas as combinações resultantes da variação de l e k aponta para um máximo da função $\rho(\omega_0)$ exatamente quando ω_0 é igual ao verdadeiro pitch do sinal de fala periódico em análise.

Para facilitar a compreensão do significado desse critério, suponha-se que o sinal de fala de entrada é periódico com freqüência de pitch igual a ω_x. Então $\omega_l = l\omega_x$, $A_l = \bar{A}(l\omega_x)$. Dessa forma, quando $\omega_0 = \omega_x$, tem-se

$$\rho(\omega_x) = \frac{1}{2}\sum_{k=1}^{K}[\bar{A}(k\omega_x)]^2 = max[\rho(\omega_0)]$$

A Figura 4.1 ilustra o comportamento da função $\rho(\omega_0)$, cuja posição do maior pico identifica o valor do pitch para um sinal sonoro.

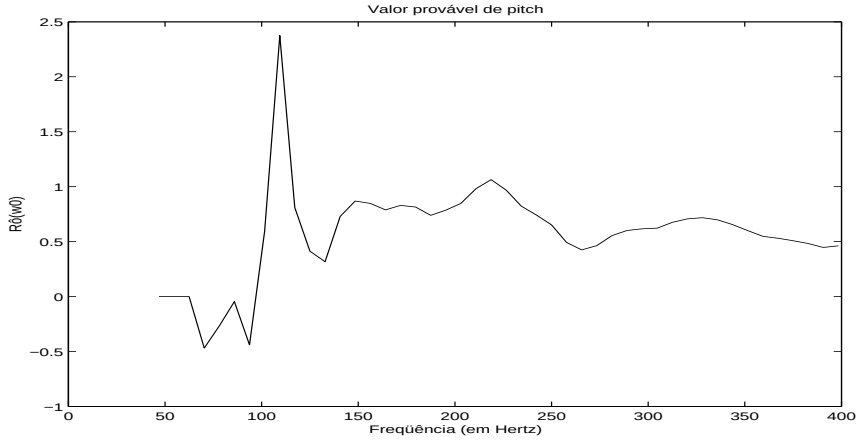

Figura 4.1: *Resultado típico de estimativa do pitch para um quadro sonoro*

A Figura 4.2 ilustra o comportamento da função $\rho(\omega_0)$ no caso de um sinal de fala surdo. Note-se que no caso de um sinal sonoro o valor máximo da função $\rho(\omega_0)$ é significativamente mais expressivo que o respectivo valor no caso de um sinal surdo.

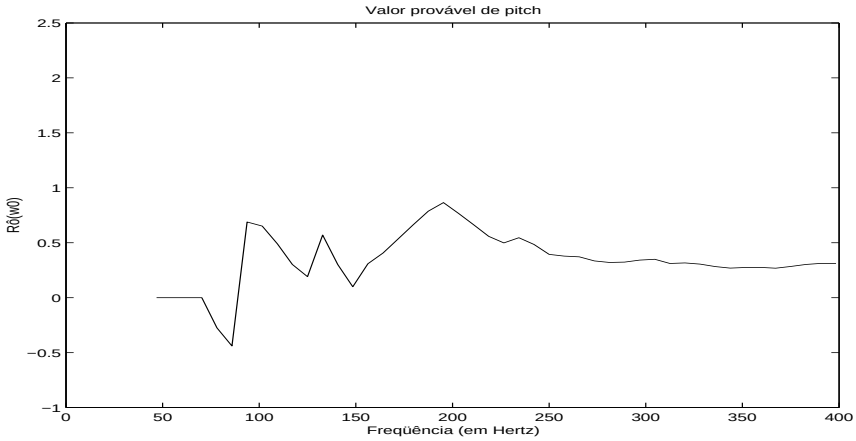

Figura 4.2: *Resultado típico de estimativa do pitch para um quadro surdo*

O critério acima é suficiente para estimativa de ω_0 com precisão satisfatória. Porém três procedimentos de otimização de seu desempenho são expostos a seguir.

4.1.1 Melhoramento do critério

Nota-se que se ω_0 for o verdadeiro pitch do sinal de fala, então haverá um harmônico dele correspondente a cada lóbulo da função $D(\omega_l - k\omega_0)$. Entretanto, no caso de ω_0 não ser o verdadeiro pitch, computa-se, para cada harmônico, somente o lóbulo que seja mais significativo para o MSE.

Cada múltiplo de um dado ω_0 define uma faixa de freqüências dentro da qual considera-se o pico mais significativo para o MSE. Assim

$$\rho(\omega_0) = \sum_{k=1}^{K} \bar{A}(k\omega_0)\{\max_{\omega_l \in F(k\omega_0)}[A_l D(\omega_l - k\omega_0)] - \frac{1}{2}\bar{A}(k\omega_0)\} \tag{4.19}$$

onde

$$F(k\omega_0) = \{\omega \mid k\omega_0 - \frac{\omega_0}{2} \leq \omega < k\omega_0 + \frac{\omega_0}{2}\}$$

Além de aumentar a confiabilidade na estimativa da freqüência de pitch, o procedimento acima proporciona ainda economia de processamento e uma certa robustez contra ruído aditivo, pois os picos pequenos devido a ruído são ignorados.

Este processo é similar ao efeito de mascaramento auditivo de tons fracos por tons intensos que estejam em sua vizinhança [KOHS02].

4.1.2 Resolução adaptativa ao pitch

Nas formulações acima está implícito que a janela de análise é de tamanho fixo igual a N+1 amostras, sendo que N+1 depende do pitch médio do sinal de fala de entrada. Isto significa que, dado um pitch médio, a largura do lóbulo principal da função "ssinc" é fixa para todos os valores candidatos ao pitch do quadro em análise. Isto é contrário ao fato de que o ouvido humano é mais sensível a erros no pitch nas freqüências baixas do que nas altas. Tal efeito pode ser levado em consideração ao definir-se a função $D(\omega_l - k\omega_0)$ como sendo de largura relativa ao pitch candidato. Assim a função $D(x)$ no k-ésimo harmônico torna-se

$$D(\omega_l - k\omega_0) = \begin{cases} \frac{sen\left[2\pi(\frac{\omega_l - k\omega_0}{\omega_0})\right]}{2\pi(\frac{\omega_l - k\omega_0}{\omega_0})}, & \text{para todo } |\omega_l - k\omega_0| \leq \frac{\omega_0}{2} \\ 0, & \text{caso contrário} \end{cases} \tag{4.20}$$

Deste modo a resolução se torna aguçada para valores baixos de pitch candidatos e, por outro lado, mais branda para os candidatos de altos valores.

4.1.3 O problema de interação com formantes

O problema de interação com formantes aparece devido ao fato de que os harmônicos próximos a elas tendem a tornar-se mais significativos no critério MSE e, conseqüentemente, conduzir a resultados ambíguos de pitch. Este efeito pode ser reduzido diminuindo-se a faixa dinâmica da variação das amplitudes das componentes senoidais. Uma maneira eficiente de fazer isso é substituir as amplitudes senoidais medidas por

$$A_l(novo) = \left(\frac{A_l}{A_{max}}\right)^{\gamma} \qquad 0 \leq \gamma \leq 1 \tag{4.21}$$

onde $A_{max} = \max\{Al\}_{l=1}^{L}$. Uma vez que o critério do MSE conduz a uma certa robustez contra ruído aditivo, torna-se desejável manter γ o mais próximo possível da unidade, introduzindo somente a compressão necessária para eliminar o problema de interação com formantes. Uma compressão acentuada faria com que os baixos níveis de amplitude devido a ruído aditivo influenciassem no critério MSE, causando distorção do resultado. O fator de compressão escolhido experimentalmente é $\gamma = 0,5$.

4.2 Estimativa da envoltória das amplitudes

A função de envoltória das amplitudes das componentes senoidais (também conhecida como função de envoltória do trato vocal), utilizada no critério MSE de estimativa do pitch como $\bar{A}(\omega)$, pode ser obtida através de várias técnicas, tais como a análise cepstral, a predição linear, etc. Porém, é desejável a utilização de um método que conduza a uma envoltória que passe exatamente pelos pontos de amplitudes medidos. Tal técnica já foi desenvolvida no trabalho *Spectral Envelope Estimation Vocoder - SEEVOC* [PAUL81].

O algoritmo SEEVOC depende do valor do pitch médio, denotado por $\bar{\omega}_0$.

A estimativa do pitch médio será abordada posteriormente e, por hora, será assumido como conhecido. O primeiro passo é procurar pelo maior valor de amplitude das componentes senoidais no intervalo $\left[\frac{\bar{\omega}_0}{2},\ \frac{3\bar{\omega}_0}{2}\right]$. Tendo encontrado a amplitude e a freqüência daquele pico, nomeado $(A_1,\ \omega_1)$, então o próximo passo é procurar no intervalo $\left[\omega_1 + \frac{\bar{\omega}_0}{2},\ \omega_1 + \frac{3\bar{\omega}_0}{2}\right]$ pelo maior pico, nomeado $(A_2,\ \omega_2)$. O processo continua procurando nos intervalos $\left[\omega_{l-1} + \frac{\bar{\omega}_0}{2},\ \omega_{l-1} + \frac{3\bar{\omega}_0}{2}\right]$ os seus respectivos maiores picos, $(A_l,\ \omega_l)$, até que o final da banda de voz seja alcançado. Se nenhum pico for encontrado em um dado intervalo de busca, então o valor do espectro no centro do intervalo é adotado e sua freqüência se torna o ponto a partir do qual continua-se o procedimento de busca.

A principal característica desse procedimento é que qualquer pico de baixa amplitude será mascarado pelo maior pico do intervalo. Além disso, o processo não é dependente de os picos serem harmônicos e nem de que o valor do pitch médio seja exato, uma vez que o processo se redefine depois que cada pico é encontrado. A envoltória SEEVOC é então obtida aplicando-se uma interpolação em segmentos constantes, de modo que o centro de cada segmento coincida com o próprio pico encontrado e suas bordas sejam $\frac{\bar{\omega}_0}{2}$ acima e $\frac{\bar{\omega}_0}{2}$ abaixo desse ponto. A Figura 4.3 ilustra a envoltória obtida pelo algoritmo SEEVOC.

4.3 Estimativa do pitch em duas etapas

O extrator de pitch MSE foi elaborado com base nos parâmetros senoidais de um sinal de fala de entrada. Isto assume implicitamente que a análise foi consumada utilizando-se uma janela Hamming de tamanho igual a duas vezes e meia o período médio de pitch. Além disso, a técnica SEEVOC também assume que uma estimativa do pitch médio esteja disponível. Por outro lado, entende-se que é necessário que o pitch já tenha sido estimado para se calcular o pitch médio. Nota-se, neste caso, a ocorrência de um dilema circular que pode ser resolvido assumindo-se um pitch médio inicial igual ao mínimo pitch possível em um sinal de

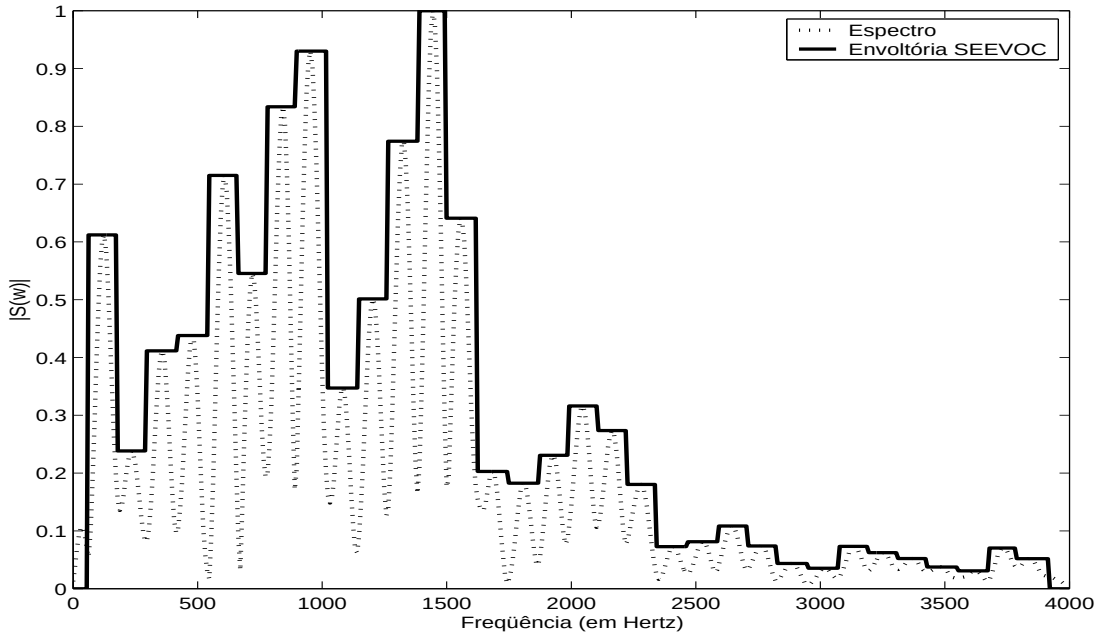

Figura 4.3: *Função envelope obtida pelo algoritmo SEEVOC*

fala qualquer, neste trabalho assumido como $f_0 = 40$ hertz. Como conseqüência, a primeira estimativa dos parâmetros senoidais, bem como do verdadeiro pitch, não é precisa. Por isso utiliza-se uma segunda etapa de cálculo dos parâmetros. Neste ponto o pitch médio utilizado não é mais o inicial, mas sim uma atualização dele, levando-se em consideração o pitch recém calculado. Desta forma, num mesmo quadro de análise os parâmetros senoidais e o pitch são calculados duas vezes. Este procedimento resolve o problema do dilema circular, mas não garante, num único quadro, o cálculo preciso do pitch e, conseqüentemente, dos parâmetros senoidais e do pitch médio. Tal garantia é obtida conforme a análise vai evoluindo, quadro a quadro, ocasionando atualização sucessiva e interativa do pitch médio, do pitch e dos parâmetros senoidais.

A fim de fornecer ao algoritmo SEEVOC um pitch médio preciso, faz-se necessário determinar um procedimento que identifique quando o pitch estimado é resultado da análise de um quadro essencialmente sonoro. Assim, o pitch médio deve ser atualizado somente quando tal situação é identificada. Este procedimento será apresentado a seguir.

4.4 Detecção de sonoridade

No contexto de modelagem senoidal da fala, a determinação de quão sonoro é um quadro é feita por uma comparação entre o espectro do sinal de fala original e o seu modelo harmônico. Deste modo, o sinal é mais sonoro quanto mais preciso for o ajuste do modelo harmônico ao espectro original. A precisão desse ajuste pode ser determinada pela relação sinal-ruído (em inglês, "Signal-to-Noise Ratio" - SNR)

$$SNR = \frac{\sum_n |s(n)|^2}{\sum_n |s(n) - \hat{s}(n; \hat{\omega}_0)|^2} \tag{4.22}$$

onde $\hat{\omega}_0$ é o pitch estimado usando os conceitos descritos anteriormente. Das equações (4.3), (4.5) e (4.13) segue que

$$SNR = \frac{P_s}{P_s - 2\rho(\hat{\omega}_0)} \tag{4.23}$$

onde agora a potência do sinal de entrada, P_s, deve ser calculada baseada nas amplitudes senoidais comprimidas, definidas em (4.21). Se o valor da SNR for alto, então o MSE é baixo, o que significa que o ajuste é bom e que o sinal de entrada é provavelmente sonoro. No caso de SNR pequeno, o MSE é grande e o ajuste é deficiente, o que indica que o sinal é provavelmente surdo. Assim, o grau de sonoridade do sinal de fala é dependente do valor de SNR. O valor exato da relação entre o valor de SNR e a probabilidade de o sinal ser sonoro é difícil de ser estabelecido, porém uma estimativa que tem se mostrado eficiente nesse contexto [MCAU95] é a seguinte:

$$P_v(SNR) = \begin{cases} 1, & SNR \geq 13dB \\ \frac{1}{9}(SNR - 4), & 4dB \leq SNR \leq 13dB \\ 0, & SNR < 4dB \end{cases} \tag{4.24}$$

onde P_v representa a probabilidade de o sinal de fala ser sonoro no quadro em análise, doravante denominada *Probabilidade de sonoridade* (em inglês, 'Voicing Probability'). Assim, o pitch médio pode ser atualizado sempre que P_v ultrapassar um determinado limiar. Neste trabalho, seguindo a sugestão de [MCAU95], adotou-se como limiar o valor $P_v = 0,8$. A fórmula de atualização do pitch médio utilizada é a seguinte:

$$\bar{\omega}_0(m) = \frac{m\bar{\omega}_0(m-1) + \omega_0}{m+1} \tag{4.25}$$

onde m é inicialmente igual a zero e vai sendo incrementado a cada atualização do pitch médio. ω_0 é o pitch recém calculado e $\bar{\omega}_0(m)$ é o pitch médio levando-se em consideração o valor de todos os pitch's até o m-ésimo pitch.

4.5 Aperfeiçoamento da estimativa do pitch

Os resultados desse trabalho indicam que o algoritmo extrator de pitch MSE é mais preciso que os algoritmos temporais, porém, em alguns casos particulares, como por exemplo no caso de interação com formantes, observa-se-lhe certa instabilidade, que, oportunamente, ocasiona estimativa ambígua. Como contribuição deste trabalho, propõe-se um algoritmo de detecção de pitch que aproveite a precisão do algoritmo apresentado, mas que incorpore-lhe a estabilidade de um algoritmo temporal. O algoritmo temporal desenvolvido nesse trabalho foi uma combinação do simples, porém eficiente, algoritmo proposto por RABINER [RABI78], com um melhoramento baseado no conceito de *pitch pulses* apresentado por KLEIJN & KROON [KLEI93].

A proposta básica é explorar o algoritmo temporal, de modo a fornecer uma freqüência de pitch que sirva como referência para o algoritmo estudado, eliminando-lhe a redundância em torno de dois ou mais valores de pitch. Deste modo, o algoritmo estudado efetua a busca da freqüência de pitch definitiva somente nas vizinhanças da freqüência de pitch proposta pelo algoritmo temporal. Foi definido um critério empírico para estabelecimento do intervalo de busca da freqüência de pitch definitiva. Utilizando-se um intervalo fixo de 140 Hz, sendo 70 Hz acima e 70 Hz abaixo da freqüência de pitch proposta pelo algoritmo temporal, obteve-se resultados satisfatórios. O passo da busca também é um fator a ser definido. Para tal, faz-se prudente levar em consideração o tipo de quantização a ser utilizada na codificação do pitch. Para este trabalho definiu-se uma quantização linear de 8 bits, o que implica em 256 níveis. A faixa de variação do pitch adotada é de 40 a 400 Hz, ou seja 360 Hz. Ao adotar essa faixa define-se um passo de 1,40625 Hz. Este foi o passo adotado neste projeto. Com base nesta proposta, o novo algoritmo é capaz de estimar o pitch com maior precisão e estabilidade.

4.6 O modelo senoidal harmônico

Observa-se que o pitch é estimado independentemente de o quadro em análise ser surdo ou sonoro. Neste caso, um fator a ser considerado é a possibilidade de o sinal ser surdo e o pitch estimado ser grande, de tal modo que se perceba certa degradação no sinal sintetizado. Isto ocorre devido ao fato de haver, nestes casos, poucos componentes senoidais para sintetizar adequadamente um sinal de fala

semelhante a ruído [TREE68]. Este problema pode ser eliminado definindo-se um pitch fixo baixo ($\approx$ 100 Hz) durante fala surda ($\omega_u = 2\pi(100/f_s)$), sempre que o pitch estimado exceda 100 Hz.

Resta agora formalizar o procedimento de ajuste das freqüências das componentes senoidais, uma vez que os quadros não são classificados somente como sonoros ou como surdos, mas com uma probabilidade de sonoridade. Assim, as freqüências das componentes senoidais são ajustadas de acordo com essa probabilidade. O procedimento exato para fazer isso é primeiramente definir uma freqüência de corte dependente da sonoridade [MAKH78], ω_c, como

$$\omega_c(P_v) = \pi P_v \tag{4.26}$$

a qual é forçada a ser sempre maior ou igual $2\pi(fc_{min}/f_s)$, onde $fc_{min} = 1500$Hertz.

Se o pitch estimado for ω_0, onde $\omega_0 > \omega_u$, então a reconstrução harmônica fica

$$\omega_k = \begin{cases} k\omega_0 & \text{para } k\omega_0 \leq \omega_c(P_v) \\ k^*(\omega_0 - \omega_u) + k\omega_u, & \text{para } k\omega_0 > \omega_c(P_v) \end{cases} \tag{4.27}$$

onde k^* é o mais alto k para o qual $k^*\omega_0 \leq \omega_c(P_v)$.

De outro modo, se o pitch estimado for $\omega_0 < \omega_u$, então a reconstrução harmônica fica

$$\omega_k = k\omega_0 \qquad \text{para qualquer } k \tag{4.28}$$

Temos portanto que

$$\hat{s}(n;\omega_0) = \sum_{k=1}^{K} \bar{A}(\omega_k) e^{j(n\omega_k + \phi_k)} \tag{4.29}$$

onde ϕ_k é a fase da STFT na freqüência ω_k. Estritamente falando, este procedimento é harmônico somente em quadros de fala fortemente sonora, onde $P_v = 1$, $\omega_c = \pi$ e ω_0 representa realmente a freqüência fundamental da fala, assegurando que todos os componentes senoidais serão múltiplos de ω_0. No caso de $P_v < 1$, nem todos os componentes senoidais serão múltiplos de ω_0 o que já é suficiente para definir que a estrutura é não harmônica.

4.7 Resultados experimentais

Como no modelo senoidal básico, mencionado no capítulo anterior, um grande banco de dados foi processado com esse sistema e nota-se que a fala sintética produzida por esse modelo, ao ser ouvida, é de boa qualidade, desde que o intervalo de sobreposição de quadros seja menor que $\approx$ 15 ms (o que também implica

em que o quadro seja menor ou igual a 30 ms, respeitando a estacionariedade do sinal de fala). Isto não somente valida o desempenho do extrator de pitch MSE, mas também mostra que se as amplitudes e fases das componentes harmônicas puderem ser codificadas eficientemente, então, no lugar das freqüências em si, codifica-se somente o pitch e a probabilidade de sonoridade.

O uso da combinação do algoritmo extrator de pitch MSE com o algoritmo temporal de extração de pitch melhorou a qualidade subjetiva da fala sintética, uma vez que a estrutura harmônica foi mais perfeitamente ajustada às freqüências das componentes senoidais.

A Figura 4.4 (a) ilustra o espectrograma de um sinal de fala arbitrário, sintetizado a partir dos parâmetros senoidais, como visto no capítulo anterior e a Figura 4.4 (b) ilustra o espectrograma do sinal sintetizado a partir dos parâmetros senoidais, porém considerando-se o modelo harmônico estudado neste capítulo.

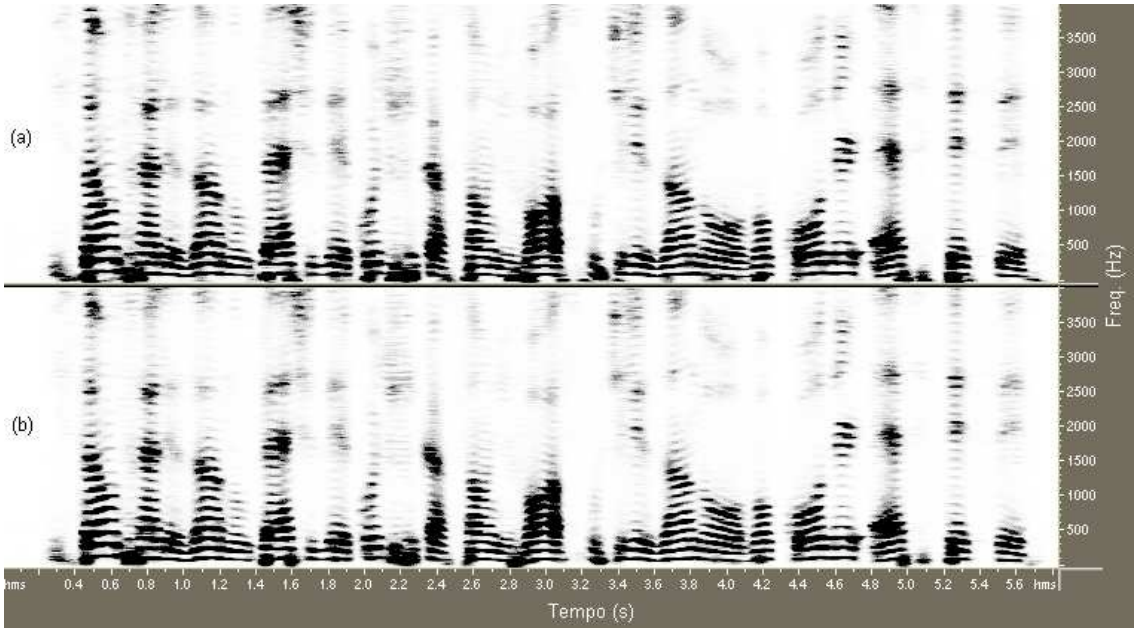

Figura 4.4: *Comparação entre os espectrogramas de sinais de fala sintéticos obtidos a partir dos modelos senoidal básico e do modelo harmônico*

O método empregado para a redução da taxa de bits na codificação das fases será abordado no capítulo seguinte.

Capítulo 5

O modelo de fase híbrida das ondas senoidais

Foi visto no capítulo anterior que pode-se obter fala sintética de boa qualidade utilizando um conjunto harmônico de componentes senoidais, desde que as amplitudes e fases sejam os valores de módulo e fase, respectivamente, provenientes das amostras complexas da STFT de cada componente harmônico. Embora o modelo harmônico tenha eliminado a necessidade de codificar as freqüências das componentes senoidais, as amplitudes e as fases de cada componente ainda representariam muitos parâmetros para uma codificação a baixas taxas de bits, tais como 2400 ou 4800 bps. Desse modo, a fim de reduzir o conjunto de parâmetros a ser codificado, torna-se evidente a necessidade de exploração de outras propriedades do mecanismo de produção da fala. Este capítulo trata da exploração das propriedades de fase do mecanismo de produção da fala, apresentando modelos que permitem simulá-la no decodificador, sem a necessidade de sua codificação e transmissão.

5.1 Modelo senoidal para fala sonora

A produção da fala sonora começa com uma seqüência de pulsos de excitação que representam o fechamento da glote, a uma taxa dada pela freqüência de pitch do locutor.

No modelo senoidal de produção da fala, o sinal de excitação é representado por uma soma de componentes senoidais. Essa excitação pode ser ainda representada por uma soma de senóides de módulo unitário que servirá de entrada primeiramente a um filtro linear representando as características do pulso glotal, em seguida a um filtro linear que representa as características do trato vocal e fi-

nalmente a um filtro que representa as características da irradiação labial. Assim, o sinal de excitação na sua forma complexa fica:

$$\hat{e}(n) = \sum_{l=1}^{L} e^{j(n-n_0)\omega_l} \tag{5.1}$$

onde n_0 corresponde a um tempo de ajuste, necessário ao sincronismo dos pulsos de pitch entre quadros vizinhos. Uma maneira de se obter tal sincronismo é garantindo que os pulsos de pitch de dois quadros consecutivos estejam em fase, numa determinada posição de referência, quando a superposição e adição do processo de síntese for executada. Sabe-se que a posição de início de cada quadro equivale exatamente à posição de início do quadro anterior somada do valor do incremento, chamada de posição de incremento. Assim, neste trabalho convencionou-se que os pulsos de pitch do início de cada quadro devem estar sincronizados com o pulso de pitch mais próximo à posição de incremento do quadro anterior. Esse procedimento exige um deslocamento temporal, n_0, dos pulsos de pitch, também conhecido como "onset time". No domínio da freqüência, o sinal de excitação fica:

$$\hat{E}(\omega) = \sum_{l=1}^{L} e^{-jn_0\omega_l} \tag{5.2}$$

Uma vez determinado o sinal de excitação, a próxima operação no modelo de produção da fala é representar as suas alterações de fase e de amplitude através de um filtro que modele simultaneamente as características do pulso glotal, do trato vocal e da irradiação labial, o qual será identificado como *filtro composto*. Seja $H_s(\omega) = |H_s(\omega)|e^{j\Phi_s(\omega)}$ a função de transferência do filtro composto, doravante denominada função de sistema. Assim, o modelo do sinal de fala no domínio da freqüência, que é na verdade a saída do filtro composto, fica:

$$\begin{aligned} \hat{S}(\omega) &= \sum_{l=1}^{L} \hat{E}(\omega_l) H_s(\omega_l) \\ \hat{S}(\omega) &= \sum_{l=1}^{L} |H_s(\omega_l)| e^{j[\Phi_s(\omega_l) - n_0\omega_l]} \end{aligned} \tag{5.3}$$

que no domínio do tempo é dado por

$$\hat{s}(n) = \sum_{l=1}^{L} |H_s(\omega_l)| e^{j[(n-n_0)\omega_l + \Phi_s(\omega_l)]} \tag{5.4}$$

Comparando a equação acima com (4.1), observa-se que as amplitudes e fases

obtidas na saída do filtro composto podem ser identificadas como:

$$\begin{aligned} A_l &= |H_s(\omega_l)| \\ \theta_l &= -n_0\omega_l + \Phi_s(\omega_l) \end{aligned} \tag{5.5}$$

onde o termo $-n_0\omega_l$ passa a ser definido como fase de excitação e $\Phi_s(\omega_l)$ como fase de sistema.

Isto identifica as amplitudes das ondas senoidais como sendo amostras da magnitude da função de sistema. Identifica ainda as fases das ondas senoidais como sendo amostras da fase da função de sistema, somadas, cada uma, ao harmônico correspondente da fase de excitação. Desta forma, o desafio de sintetizar as fases das componentes senoidais passa a ser o de determinar a fase da função de sistema e o *onset time* na etapa de decodificação.

MCAULAY & QUATIERI [MCAU95] propuseram obter a fase da função de sistema a partir da fase da função de transferência de um filtro só de pólos de fase mínima (em inglês: *all-pole minimum phase*) [1]. Por outro lado, CHANG & WANG [WWCH98] propuseram um modelo de fase de função de sistema que inclui zeros no filtro, sendo capaz de representar satisfatoriamente as características do trato vocal de uma forma geral, independentemente de locutor ou de variação temporal do trato vocal. Neste trabalho, melhores resultados foram obtidos com a superposição dessas duas fases. Tal combinação é justificada pelo fato de a fase de sistema proposta por Chang e Wang modelar o trato vocal de forma geral e a função de fase mínima acrescentar-lhe um modelamento particular a cada quadro.

Seja $|H_a(\omega_l)|e^{\Psi(\omega_l)}$ a função de transferência do modelo só de pólos (o procedimento de determinação de $H_a(\omega)$ no decodificador será apresentado no capítulo seguinte). Seja ainda $\varphi(\omega_l)$ representar a fase da função de sistema proposta por Chang e Wang, então a fase final do l-ésimo componente da função de sistema fica:

$$\Phi_a(\omega_l) = \Psi(\omega_l) + \varphi(\omega_l) \tag{5.6}$$

onde

$$\begin{aligned} \varphi(\omega_l) \;= & -tan^{-1}\left[\tfrac{g_1 sen\omega_l}{1-g_1 cos\omega_l}\right] - tan^{-1}\left[\tfrac{g_2 sen\omega_l}{1-g_2 cos\omega_l}\right] \\ & -tan^{-1}\left[\tfrac{v_1 sen\omega_l + v_2 sen2\omega_l}{1-v_1 cos\omega_l - v_2 cos2\omega_l}\right] \end{aligned} \tag{5.7}$$

com $(g_1,\ g_2) = (1,1,\ \ 1,1)$ e $(v_1,\ v_2) = (1,515,\ -0,752)$ [WWCH98].

Uma vez determinada a fase de sistema, resta ainda a determinação do *onset time*, a fim de obter-se a fase de excitação. O *onset time* pode ser obtido

[1] Sabe-se que se o filtro é só de polos e estável, isto já implica em fase mínima

empregando-se o seguinte cálculo:

$$n_0^{q+1} = n_0^q + iP_0 - T \tag{5.8}$$

onde q é o índice de um quadro qualquer; P_0 é o período de pitch do quadro $q+1$ em amostras e T é o incremento, ou seja, o intervalo de tempo, em amostras, entre dois quadros consecutivos. O índice i é o menor inteiro que, multiplicado por P_0 e adicionado ao *onset time* do quadro q, é suficiente para exceder o valor de um incremento. No quadro inicial da fala, n_0 é definido como zero.

Conclui-se então que, em sinais sonoros, combinando a fase de excitação e a fase de sistema, a fase geral das componentes senoidais da fala para o l-ésimo harmônico se torna

$$\hat{\theta}(l\omega_0) = -n_0 l\omega_0 + \Phi_a(l\omega_0) \tag{5.9}$$

Isto mostra que a reconstrução senoidal fica dependendo somente do pitch e da fase da função de transferência do filtro *all-pole*. Como mencionado anteriormente, o processo de determinação da função de transferência do filtro *all-pole* no decodificador será discutido no capítulo seguinte.

5.2 Modelo senoidal de fase para fala sonora e surda

Ao utilizar-se o modelo apresentado acima, ao invés das fases originais, a qualidade da fala sintética fica bastante próxima à da natural durante segmentos sonoros, porém a fala fica muito "assoviada" nos segmentos surdos. Por outro lado, se as fases forem substituídas por variáveis aleatórias de distribuição uniforme entre $-\pi$ e π, a fala sintética fica bastante natural durante fala surda, mas muito "chiada" durante fala sonora. Isto sugere uma generalização do modelo de fase proposto em (5.9), adicionando-lhe um componente que assuma o valor zero para segmentos sonoros e que assuma um valor aleatório com distribuição uniforme entre $-\pi$ e π ($U[-\pi, \pi]$) para segmentos de fala surda. Porém, sabe-se que este procedimento não é suficiente, devido à dificuldade de classificação entre sonoro ou surdo que determinados segmentos de fala oferecem. A solução é lançar mão do conceito de excitação mista, proposto por MARKHOUL, et al., [MAKH78], onde uma freqüência de corte dependente de sonoridade é definida. Abaixo dessa freqüência sintetiza-se fala sonora e, acima dela, fala surda. O procedimento de determinação da probabilidade de sonoridade e da própria freqüência de corte é exatamente o mesmo descrito no capítulo anterior, na seção que trata do modelo senoidal harmônico.

Fazendo com que ω_c denote a freqüência de corte dependente de sonoridade, a componente de fase aleatória pode ser definida como

$$\hat{\epsilon}(\omega) = \begin{cases} 0, & \text{se } \omega \leq \omega_c \\ U[-\pi, \pi], & \text{se } \omega \geq \omega_c \end{cases} \tag{5.10}$$

Ao incorporar-se esse componente de fase de fala surda à equação (5.9), a fase senoidal completa para o l-ésimo harmônico, independentemente da sonoridade do sinal de fala, fica

$$\hat{\theta}(l\omega_k) = -n_0\omega_k + \Phi_a(\omega_k) + \hat{\epsilon}(\omega_k) \tag{5.11}$$

onde ω_k é definida por (4.27).

5.3 Resultados experimentais

Como nos modelos anteriores, um grande banco de dados foi processado. Quando o modelo de fase sintética dependente de sonoridade foi utilizado para substituir as fases medidas no modelo senoidal harmônico, não somente a fala sintética, ao ser ouvida e comparada com o modelo do capítulo anterior, foi de boa qualidade, mas também a locução sintética se assemelhou em muito à locução original, o que implica em que mesmo com a total simulação da fase no decodificador a identificação de locutor não é prejudicada.

A Figura 5.1 (a) ilustra o espectrograma de um sinal de fala arbitrário, sintetizado a partir dos parâmetros senoidais, como visto no capítulo anterior e a Figura 5.1 (b) ilustra o espectrograma do sinal sintetizado a partir dos parâmetros senoidais, porém considerando-se o modelo harmônico estudado no capítulo anterior e o modelo de fase estudado neste capítulo.

Uma vez que os parâmetros do modelo harmônico e de fase híbrida são a freqüência de pitch, a probabilidade de sonoridade, e a função de transferência do filtro *all-pole*, e, desde que relativamente poucos bits são necessários para codificar o pitch e a probabilidade de sonoridade, a capacidade de operar este sintetizador em baixas taxas de bits passa a depender do número de bits necessários para codificar os parâmetros que representam a função de transferência do filtro *all-pole*. O desenvolvimento da estratégia dessa codificação será o assunto do capítulo seguinte.

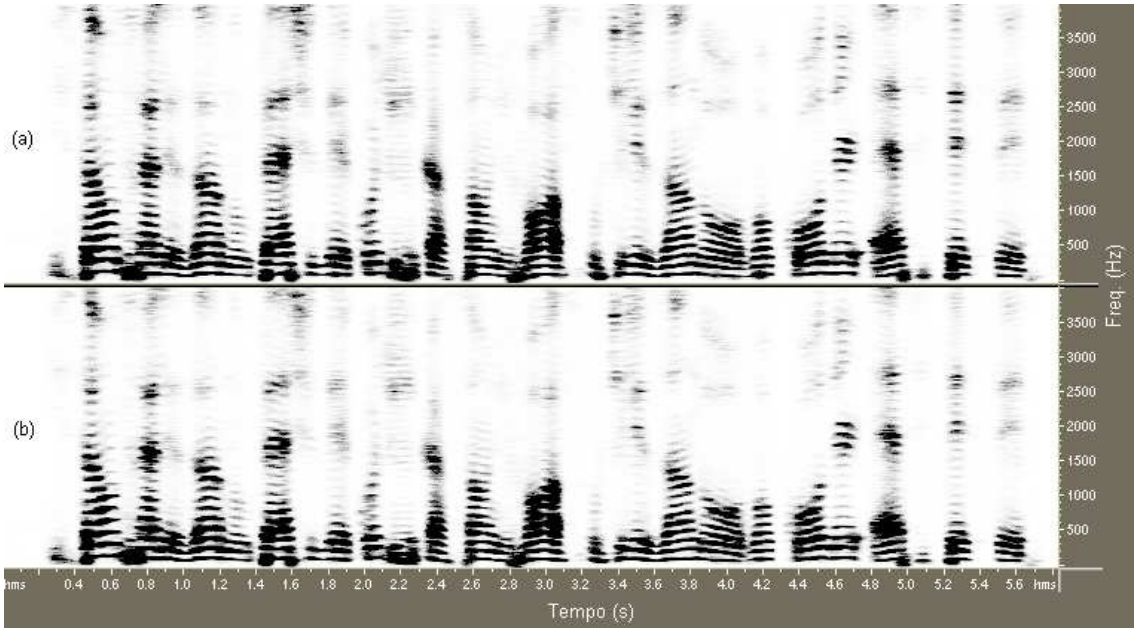

Figura 5.1: *Comparação entre os espectrogramas de sinais de fala sintéticos obtidos a partir dos modelos harmônico e de fase híbrida*

Capítulo 6

O modelo senoidal de codificação das amplitudes

Seguindo a recomendação de McAulay e Quatieri, o pitch pode ser codificado utilizando-se de $\approx 6-8$ bits e a probabilidade de sonoridade de $\approx 2-3$ bits. Deste modo a operação a baixas taxas de bits parece ser alcançável se as amplitudes puderem ser codificadas eficientemente. Porém, as amplitudes, por si só, já necessitariam de um número elevado de bits para sua correta representação. Isto porque a faixa de freqüência de voz pode abrigar, dependendo do pitch do locutor, um número expressivo de harmônicos, sabendo-se que cada harmônico teria a sua amplitude correspondente a ser codificada.

Como visto no capítulo anterior, sabe-se que as amplitudes das ondas senoidais são amostras da magnitude da função de sistema, ou seja, $A_l = |H_s(\omega_l)|$. A abordagem tomada nesse capítulo é a de definir um modelo paramétrico capaz de representar fielmente a função de sistema, para então codificar os parâmetros do modelo. Neste caso, e levando-se também em conta a fase mínima adotada no modelo de fases do capítulo anterior, tanto uma representação cepstral quanto uma do tipo *all-pole* resolveriam o problema. Porém, adotou-se neste trabalho a representação do tipo *all-pole*, devido à existência de métodos de quantização mais eficientes, como por exemplo, a transformação dos coeficientes LPC em Freqüências Espectrais Discretas (do inglês: "Line Spectral Frequency" - LSF) [ITAK75] antes de proceder à sua quantização escalar ou vetorial.

6.1 O modelo *all-pole*

Na etapa de codificação, pode-se obter uma boa aproximação da magnitude da função de sistema através de uma interpolação adequada dos parâmetros ampli-

tude. Neste caso, o método de interpolação sugerido por McAulay e Quatieri é a técnica de interpolação usando uma curva spline cúbica [UNSE93] [MCAU95]. A justificativa é a boa relação entre a complexidade de processamento e o ajuste adequado à envoltória do espectro do sinal de fala, desde que a curva spline cúbica seja ajustada ao logaritmo dos picos e não a eles diretamente.

O próximo passo é definir um método analítico para obtenção dos parâmetros do modelo *all-pole*, de modo que a sua magnitude proporcione o melhor ajuste possível à envoltória. Supondo-se que $H_a(\omega)$ represente a função de transferência do filtro *all-pole* e levando-se em consideração que esse filtro representa o trato vocal, tem-se

$$\begin{aligned} H_a(\omega) &= \frac{\sigma}{A(w;\mathbf{a})} \\ A(w;\mathbf{a}) &= 1 - \sum_{k=1}^{p} a_k e^{-jkw} \end{aligned} \tag{6.1}$$

onde σ e $\mathbf{a} = (a_1,\, a_2,\, ...\, a_p)$ são os parâmetros a serem estimados [RABI78].

O parâmetro σ pode ser considerado como o ganho do filtro, uma vez que influencia direta e proporcionalmente as amplitudes espectrais. A situação ideal é a que garanta que a energia da envoltória produzida pelo filtro (envoltória estimada) seja igual à energia da envoltória spline cúbica (envoltória medida). Inicialmente o valor de σ é admitido como unitário, a fim de obter-se apenas o traçado da envoltória estimada. Em seguida, a estimativa desse parâmetro pode ser feita relacionando-se as duas energias da seguinte forma:

$$\hat{\sigma}^2 = \frac{\frac{1}{\pi}\int_0^\pi |H_s(\omega)|^2 d\omega}{\frac{1}{\pi}\int_0^\pi |H'_a(\omega)|^2 d\omega}$$

onde $|H_s(\omega)|$ é a envoltória medida e $|H_a^{'}(\omega)|$ é a envoltória do filtro *all-pole* com ganho unitário. Então

$$\hat{\sigma} = \sqrt{\frac{\int_0^\pi |H_s(\omega)|^2 d\omega}{\int_0^\pi |H'_a(\omega)|^2 d\omega}} \tag{6.2}$$

A fim de obter-se $|H_a^{'}(\omega)|$ faz-se necessário a estimativa, em primeira instância, de seus coeficientes a_k. Esse procedimento pode ser o mesmo utilizado na obtenção dos coeficientes de um preditor linear (LPC) [RABI78], com a única diferença sendo o modo de obtenção dos coeficientes de autocorrelação. O preditor linear convencional tem o inconveniente de não poder utilizar-se de muitos coeficientes de autocorrelação sob pena de obter uma função de transferência para o filtro *all-pole* que se assemelhe ao espectro do próprio sinal de fala e não à sua envoltória. A idéia aqui é computar os parâmetros de autocorrelação a partir da

transformada inversa de Fourier do quadrado da envoltória do espectro, ao invés de uma autocorrelação temporal convencional do sinal de fala. Este procedimento se justifica por duas razões. A primeira é que a transformada de Fourier de um sinal de autocorrelação da fala é o próprio espectro de potência do sinal de fala. A segunda é que o número de coeficientes de autocorrelação pode ser arbitrário, pois quando este número tende ao infinito obtém-se um filtro cuja função de transferência é uma cópia exata da envoltória espectral medida. Então, os coeficientes de autocorrelação ficam

$$R_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_s(\omega)|^2 e^{j\omega k} d\omega \tag{6.3}$$

e os coeficientes a_k podem ser encontrados resolvendo-se a equação

$$\sum_{k=1}^{p} R_{j-k} \hat{a}_k = R_j \quad j = 1,\, 2,\, ...\, p \tag{6.4}$$

Os parâmetros estimados $\hat{\mathbf{a}}$ são, funcionalmente, os mesmos que os obtidos através de uma análise LPC no domínio do tempo. Assim, todas as técnicas usadas na análise de predição linear podem ser aplicadas aos parâmetros do modelo *all-pole*. Por exemplo, a equação matricial (6.4) é Toeplitz e sua solução pode ser obtida através do algoritmo recursivo de Levinson-Durbin. Finalmente, representações alternativas para os coeficientes $\hat{\mathbf{a}}$, tais como *coeficientes de reflexão* ou *freqüências espectrais discretas* (LSF) também podem ser utilizadas a fim de permitirem o uso de técnicas de quantização escalar e vetorial convencionalmente utilizadas em vocoders baseados na técnica de análise LPC [KULD93].

Nesta etapa, com a obtenção de $|H_a^{'}(\omega)|$, pode-se finalmente obter o ganho $\hat{\sigma}$ e, em seguida, a verdadeira função de transferência do filtro *all-pole* $H_a(\omega)$

Neste trabalho, foram testados filtros *all-pole* de diversas ordens. Em seguida, foram calculadas, a partir das amostras harmônicas da magnitude e da fase da função de transferência *all-pole* reconstruída, as amplitudes e as correspondentes parcelas da fase de sistema das componentes senoidais. Em conjunto com as técnicas de reconstrução de fase abordadas no capítulo anterior, a qualidade da fala sintética ficou boa, desde que a ordem do modelo fosse $p \geq 22$. Embora seja possível codificar modelos *all-pole* de ordem elevada adequadamente a taxas superiores a 6 Kb/s, codificar tal sistema a 2400 b/s ou mesmo a 4800 b/s seria uma tarefa difícil, pois o número de bits por parâmetro seria muito reduzido, e, certamente, a qualidade da fala sintética seria prejudicada. Isto implica em que outras propriedades do sistema da fala e audição humanas devem ser exploradas a fim de reduzir a ordem do modelo tanto quanto possível, sem introdução de

distorções na fala que produzam sons “metálicos” ou zumbidos.

Sabe-se que o ouvido humano torna-se menos sensível à resolução espectral em altas freqüências. Isto é uma característica imposta pela cóclea, situada no ouvido interno, que comporta-se como um conjunto de filtros cujas larguras das bandas de passagem aumentam progressivamente à medida que aumenta a freqüência. A filtragem realizada pela cóclea pode ser aproximada analiticamente por uma mudança adequada na escala de freqüências do sinal. Tal mudança seria da escala convencional de freqüências para uma escala perceptiva das mesmas. As já conhecidas escalas Bark e Mel [PICO93] podem representar adequadamente a escala perceptiva de freqüências. Se b representar a escala perceptiva, então a relação entre a escala perceptiva e a escala linear de freqüências pode ser escrita como $b = W(f)$, onde $W(.)$ é uma função de conversão, que pode ser chamada de "função de distorção do espectro´´ (do inglês: *warping function*). A Figura 6.1 ilustra tal função de distorção utilizando-se a escala Bark, onde $b = W(f) = 6asinh(f/600)$ [WANG92].

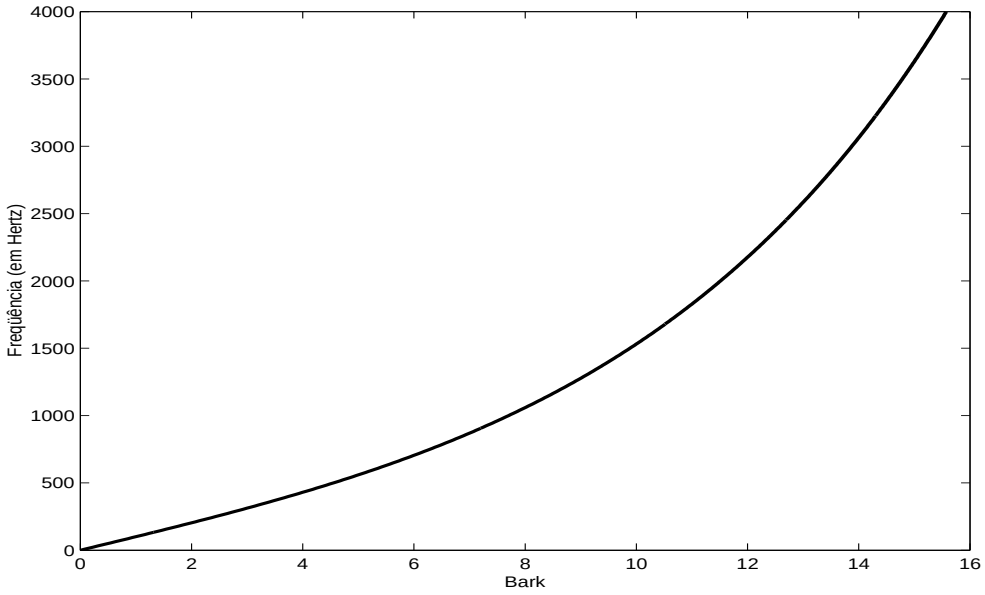

Figura 6.1: *Relação entre a escala Bark e a escala convencional de freqüências.*

Note-se que a relação é aproximadamente constante até cerca de 500 Hz, que coincide com o ponto de formação da primeira formante do sinal de fala (em média), e vai, em seguida, tornando-se progressivamente não linear à medida que a escala aumenta. Deste modo, obtém-se uma escala de freqüências distorcida (escala “subjetiva”), onde há uma resolução maior das componentes espectrais mais baixas em relação às mais altas.

A Figura 6.2 (a) ilustra o exemplo típico de uma envoltória spline cúbica ajustada às amplitudes medidas das componentes senoidais e representada na escala linear de freqüências. Ilustra ainda a dificuldade de ajuste da envoltória gerada por um filtro *all-pole* de baixa ordem, no caso ilustrado de ordem 14. A Figura 6.2 (b) mostra a representação do mesmo sinal na escala Bark, na qual é possível um ajuste razoavelmente fiel da envoltória gerada por um filtro de mesma ordem. A Figura 6.2 (c) ilustra a envoltória do modelo *all-pole* que foi ajustada na escala Bark e depois re-convertida à escala linear de freqüências.

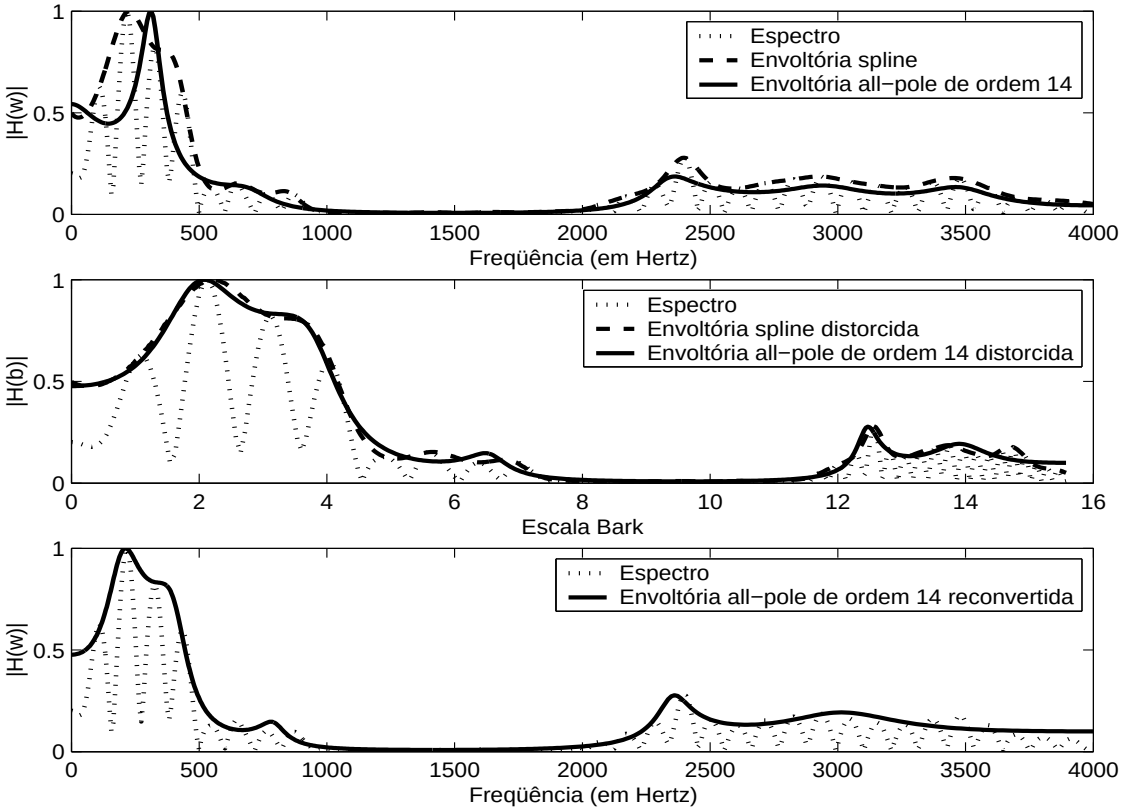

Figura 6.2: *Ajuste da envoltória de um filtro all-pole na escala "subjetiva" de freqüências.*

Comparando-se as Figuras 6.2 (a) e (c) pode-se notar que após a aplicação do procedimento de distorção da escala de freqüências a envoltória fornecida pelo filtro *all-pole* é mais fiel à envoltória spline nas baixas freqüências, onde a resolução do ouvido é maior, à custa de uma fidelidade menor nas altas freqüências, onde a resolução espectral do ouvido humano também é menor.

Desta maneira, obtém-se uma envoltória estimada que, do ponto de vista perceptual, ou seja, do ponto de vista da qualidade "subjetiva" do sinal de fala, fica muito próxima da envoltória medida.

É importante ressaltar que o procedimento acima facilita a tarefa do filtro de ajustar uma envoltória ao espectro e possibilita, mantendo a qualidade, a redução da ordem do filtro *all-pole*.

A contribuição desse trabalho na área de codificação das amplitudes está na exploração de mais uma característica do comportamento do sistema auditivo humano, a fim de reduzir ainda mais a complexidade na parametrização da envoltória, garantindo assim uma eficiência ainda maior na utilização de um filtro de ordem relativamente baixa. Tal característica refere-se à propriedade que o ouvido tem de não perceber linearmente as amplitudes dos sons. O sistema auditivo humano dá ênfase à variações, mesmo que pequenas, de sons de mais baixas amplitudes e torna-se menos sensível às mesmas variações à medida que as amplitudes aumentam. Essa variação de sensibilidade obedece a uma escala não linear, já bastante difundida no meio científico e acadêmico, conhecida como *subjective loudness*. Em português, essa expressão poderia ser interpretada como 'intensidade subjetiva' dos sons para o ouvido humano.

A proposta é que a função de envoltória $H_s(\omega)$ seja ainda submetida a uma outra conversão. Dessa vez a escala a ser convertida será a de amplitudes, ao invés da de freqüências. Assim, se $|H_{SL}|$ representar a escala de 'intensidade subjetiva', então a relação entre ela e a escala linear de amplitudes $|H|$ pode ser escrita como $|H_{SL}| = G(|H|)$, onde $G(.)$ seria uma função de conversão que, neste trabalho, foi adotada como $|H_{SL}| = \sqrt[3]{|H|}$ [HERM90]. Então, o espectro suavizado, baseado na aplicação do conceito de *subjective loudness* fica:

$$|H_{SL}(\omega)| = \sqrt[3]{|H_s(\omega)|} \tag{6.5}$$

A Figura 6.3 ilustra a suavização que é causada no espectro ao considerar-se a relação não linear que o ouvido faz entre a intensidade absoluta do som e a intensidade subjetiva.

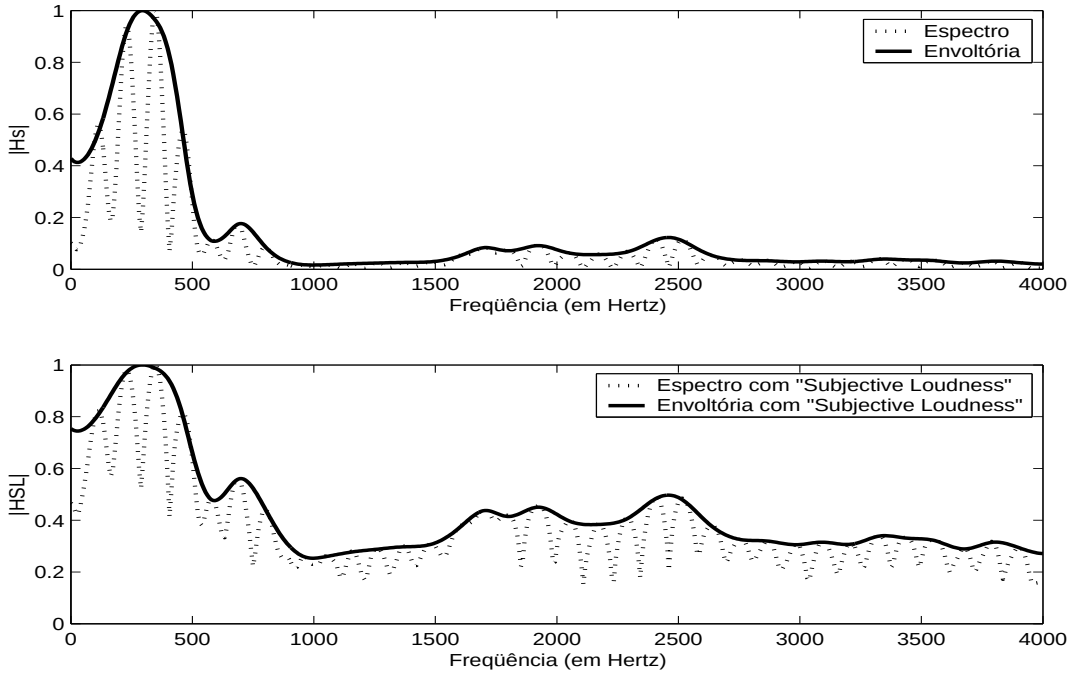

Figura 6.3: *Suavização do espectro obtida pela aplicação do conceito de Subjective Loudness.*

Baseado nos resultados práticos desse trabalho, a combinação dos procedimentos acima tornou possível a diminuição da ordem do filtro de 22 para 14. Isso significa um ganho de 8 parâmetros por quadro, tornando viável a codificação senoidal a baixas taxas de bits, praticamente sem perda de qualidade subjetiva. Uma tabela de comparação entre os resultados obtidos com e sem a aplicação do conceito de *subjective loudness* é apresentada no final desse trabalho.

A Figura 6.4 (a) ilustra o espectrograma de um sinal de fala arbitrário, sintetizado a partir dos parâmetros senoidais, considerando-se o modelo harmônico e o modelo de fases descrito nos capítulos anteriores, e a Figura 6.4 (b) ilustra o espectrograma do sinal sintetizado a partir dos parâmetros senoidais, porém considerando-se o modelo harmônico, o modelo de fase estudado no capítulo anterior e o modelo de amplitudes descrito neste capítulo.

A próxima etapa desse trabalho consiste na quantização dos parâmetros codificados. Esse assunto será abordado no capítulo seguinte.

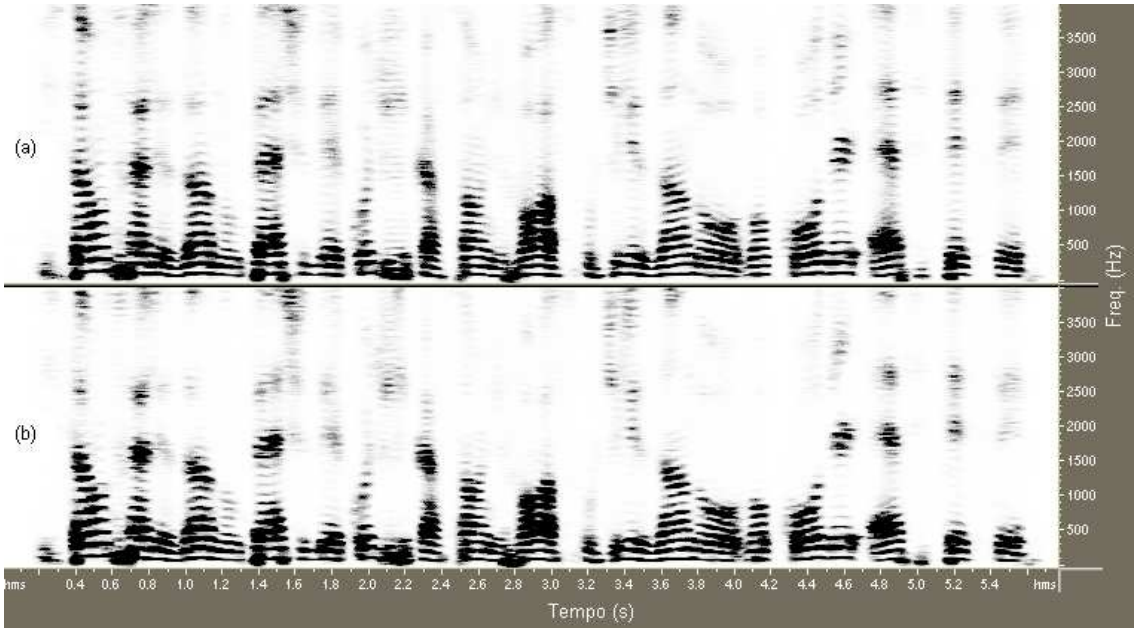

Figura 6.4: *Espectrogramas de sinais de fala sintéticos obtidos a partir do modelo de fase híbrida em comparação com o modelo de amplitudes*

Capítulo 7

Quantização dos parâmetros do modelo senoidal

Foi verificado experimentalmente que a representação de amplitudes em termos dos parâmetros de um modelo *all-pole* produz fala sintética de boa qualidade. Porém faz-se necessário considerar a etapa de quantização desses parâmetros. Ela precisa ser extremamente eficiente, a fim de manter a boa qualidade do sinal com o menor número de bits de quantização possível. Sendo assim, o objetivo fica sendo o de quantizar eficientemente os coeficientes do preditor $\hat{a}_1$, $\hat{a}_2$, ..., $\hat{a}_p$, o ganho $\hat{\sigma}$, a probabilidade de sonoridade Pv e o pitch ω_0.

Na seção *Resultados experimentais* do Capítulo 4 foi definido que o incremento máximo aceitável entre quadros sucessivos é de $15\,ms$. Isto implica em uma taxa de $66,6667$ quadros por segundo. Como visto na seção *Aperfeiçoamento da estimativa do pitch*, também no Capítulo 4, o parâmetro pitch necessita de 8 bits de quantização. Além disso, 2 bits são suficientes para codificar o parâmetro probabilidade de sonoridade. Assim, no caso de taxas de bits de 2400 b/s e 4800 b/s haveria uma disponibilidade de 36 e 72 bits por quadro, respectivamente.

Note-se que um esquema de quantização aceitável para obtenção da taxa de 4800 b/s é perfeitamente possível, pois sobrariam 62 bits para a codificação do ganho e dos 14 coeficientes do preditor. Porém, no caso de 2400 b/s a tarefa torna-se difícil, pois restariam somente 26 bits para a quantização do ganho e dos 14 coeficientes do preditor. Assim, o problema passa a ser a definição do processo de quantização do ganho e a definição de uma técnica que possibilite a quantização dos 14 coeficientes do filtro com um número reduzido de bits.

No caso do ganho, a sugestão de McAulay e Quatieri é a utilização de um preditor simples que possibilite a quantização do resíduo entre o valor original de um determinado quadro e o valor decodificado no quadro anterior. Com o objetivo de reduzir a faixa dinâmica do ganho, a codificação é aplicada sobre o logaritmo

do mesmo: $g = log\hat{\sigma}$. Assim, $g(q)$ será considerado o ganho medido para o q-ésimo quadro. No decodificador, tem-se que $\tilde{g}(q-1)$ é o ganho decodificado para o $(q-1)$-ésimo quadro. Então o ganho decodificado para o quadro q, $\tilde{g}(q)$ é dado por

$$\tilde{g}(q) = \alpha\tilde{g}(q-1) + Q[g(q) - \alpha\tilde{g}(q-1)] \tag{7.1}$$

onde α é o coeficiente de predição do ganho (sugerido por McAulay e Quatieri como ≈ 0.7) e $Q[.]$ representa uma tabela de quantização escalar linear. Este método demonstra bons resultados utilizando-se apenas 4 bits para quantização do ganho residual.

O próximo passo é a quantização dos coeficientes do preditor, mas, conforme reportado na literatura, é melhor trabalhar com uma transformação do coeficientes LPC [ITAK75]. Os coeficientes do preditor são então transformados em Freqüências Espectrais Discretas (do inglês: Line Spectral Frequencies - LSF), sendo que estas podem ser quantizadas de maneira escalar ou vetorial.

Neste ponto do trabalho o objetivo é definir uma técnica de quantização que cause a menor degradação possível à envoltória gerada pelo filtro *all-pole*.

McAulay e Quatieri sugerem a aplicação da técnica de LSFs diferenciais, descritas por SOONG & JUANG [SOON84]. Neste caso, faz-se necessário aproximadamente 3 bits para cada LSF, em média. Assim, um filtro de ordem 14 requereria aproximadamente 42 bits por quadro para alcançar um nível razoável de qualidade. Enquanto isso não é problema para um codificador de 4800 b/s, no caso de 2400 b/s faz-se necessário uma redução na taxa de quadros por segundo, o que implicaria em um incremento maior que $15\,ms$ entre quadros sucessivos e isso não é aceitável, como mencionado. Neste ponto, então, aparece uma incompatibilidade em codificar os parâmetros do modelo *all-pole* a 2400 b/s, sem servir-se de qualquer outra técnica mais apurada de quantização ou de outra propriedade da fala. Para solucionar esse problema, McAulay e Quatieri sugerem a exploração de uma propriedade das envoltórias do espectro de quadros sucessivos. Essa propriedade é a de que a envoltória não varia abruptamente de um quadro para outro, possibilitando uma interpolação entre os quadros. Uma maneira de fazer isso eficientemente é utilizando o algoritmo de interpolação de quadros de McLarnen, conhecido, em inglês, como *Frame-fill Interpolation* [MCLA78].

Uma outra abordagem para o problema da quantização seria representar os 14 coeficientes do filtro *all-pole* por dois vetores, cada um contendo 7 parâmetros LSF (LSF1 e LSF2) [SOON84] e quantizá-los vetorialmente [PALI93], providenciando-lhes um centróide para a LSF1 e outro para a LSF2, cada um com 2048 centróides (11 bits).

É importante ressaltar que neste trabalho, embora o esquema de quantização descrito anteriormente tenha sido aplicado para os parâmetros ganho, pitch e pro-

babilidade de sonoridade, a etapa de quantização das LSFs não foi implementada, devido à falta de tempo, pois não é uma tarefa simples realizar esta quantização de maneira eficiente. O capítulo seguinte trata dos resultados obtidos sem a etapa de quantização dos parâmetros LSF.

Capítulo 8

Resultados

O programa *Matlab*® foi utilizado para a implementação de diversas configurações de codificadores STC através das técnicas descritas neste trabalho.

Cada vocoder implementado possui uma combinação diferente de valores para o tamanho do incremento entre quadros, para a ordem do filtro *all-pole* e para o emprego ou não do conceito de *Subjective Loudness* na envoltória do espectro. Os valores do incremento entre quadros sucessivos assumidos foram os de 15 ms e 10 ms e a ordem do filtro *all-pole* foi alternada entre os valores 14 e 18.

Em qualquer uma das combinações acima utilizou-se um conjunto de 168 arquivos da base de dados *TIMIT Acoustic-Phonetic Speech Database* (TIMIT - Texas Instruments and the Massachusetts Institute of Technology), selecionados aleatoriamente. Uma característica importante dessa base de dados é a sua variedade fonética, uma vez que contém sentenças ricas em diversidade de fonemas e são pronunciadas por locutores selecionados entre os mais variados dialetos norte-americanos, representados por 8 regiões geográficas dos EUA.

Os arquivos da base TIMIT são originalmente amostrados a uma taxa de 16 kHz. A fim de adaptá-los aos padrões internacionais de telefonia, estes foram sub-amostrados neste trabalho para ficarem com uma freqüência de amostragem igual a 8 kHz.

A resolução espectral utilizada na simulação foi de 1024 amostras; resolução esta que está diretamente ligada ao custo de processamento mas, por outro lado, necessária para uma melhor estimativa do pitch e de seus harmônicos.

A janela de análise utilizada foi a Hamming pelas razões expostas no capítulo 3 e a janela de síntese foi a triangular devido à boa relação entre a facilidade de implementação e sua eficiência.

Os vocoders desenvolvidos são de baixa taxa de bits e podem operar a 2400 ou a 4800 bps, dependendo do esquema de quantização utilizado para os coeficientes do filtro *all-pole*, pitch, probabilidade de sonoridade e ganho. Porém, como

abordado no capítulo anterior, a etapa de quantização dos coeficientes do filtro *all-pole* não foi consumada e, por isso, os dados comparativos mostrados adiante requerem a consideração de que espera-se uma pequena degradação na qualidade da fala sintética a 2400 bps e uma degradação praticamente insignificante a 4800 bps. Deste modo, com a quantização somente do pitch, da probabilidade de sonoridade e do ganho, o sistema foi avaliado em comparação com o algoritmo MELP (Mixed Excitation Linear Prediction), que é o atualmente utilizado como padrão federal (U.S. Federal Standard) a 2400 bps nos E.U.A, em substituição ao algoritmo LPC10e. Foi ainda avaliado em comparação com o codificador padrão federal americano a 4800 bps, o CELP (Coded Excited Linear Prediction).

A avaliação subjetiva continua sendo a melhor forma de medir a qualidade da fala produzida por qualquer sistema de codificação/decodificação. No entanto, os processos padronizados de medida subjetiva, como por exemplo a escala MOS (*Mean Opinion Score*), descrita pela recomendação ITU P.800, são extremamente difíceis de serem implementados na prática, pois requerem grande consumo de tempo e de recursos financeiros. Felizmente, um grande progresso na área de pesquisa e desenvolvimento de medidas objetivas de qualidade de fala foi alcançado nos últimos 5 anos. O algoritmo PESQ (*Perceptual Evaluation of Speech Quality*), recomendação P.862 da ITU (*International Telecommunications Union*), é considerado o estado-da-arte em ferramentas para medição ponto-a-ponto de qualidade objetiva da fala.

Por meio de exaustivos testes experimentais foi demonstrado que as medidas objetivas produzidas pelo algoritmo PESQ possuem um índice de correlação estatística superior a 95% com relação ao índice subjetivo obtido na escala MOS [ANTO02]

A medida de qualidade fornecida pelo PESQ trabalha na mesma escala que a medida subjetiva MOS, ou seja, um número entre 1 e 4,5 (1-Péssimo, 2-Ruim, 3-Razoável, 4-Bom e 4,5-Ótimo). Por este motivo, esta medida também é conhecida como PESQ-MOS.

Pelas razões apresentadas anteriormente o PESQ é a ferramenta de medida de desempenho utilizada neste trabalho.

A Tabela 8.1 mostra a média dos resultados objetivos de qualidade PESQ obtidos para as várias modificações do vocoder STC apresentado neste trabalho.

Tabela 8.1: *Comparação entre as modalidades de vocoders STC*

Ordem do filtro	Incremento (ms)	Identificador	PESQ S.L.	PESQ Std.
14	10	p14q10	3,172	3,102
14	15	p14q15	3,094	3,026
18	10	p18q10	3,220	3,114
18	15	p18q15	3,132	3,103

É importante esclarecer que a sigla PESQ S.L. no cabeçalho da Tabela 8.1 significa o resultado PESQ para o vocoder STC desenvolvido com a técnica de *Subjective Loudness* e a sigla PESQ Std. (padrão) significa sem a técnica *Subjective Loudness* proposta neste trabalho. A coluna intitulada "Identificador" serve para referenciar a ordem do filtro e o tamanho do incremento na Figura 8.1.

A Figura 8.1 ilustra os resultados da Tabela 8.1 de forma gráfica.

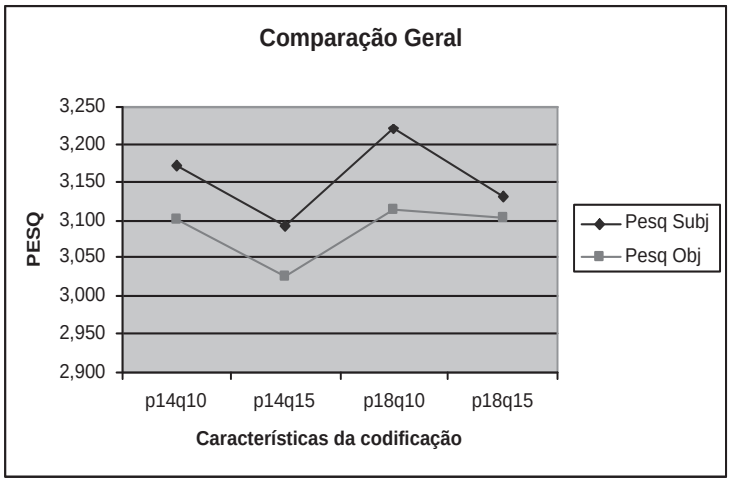

Figura 8.1: *Comparação entre as modalidades de vocoders STC*

A título de comparação, mediu-se os resultados objetivos de qualidade PESQ dos mesmos 168 arquivos para os algoritmos padrões federais americanos CELP e MELP. O resultado dos dois algoritmos esta representado na Tabela 8.2, juntamente com o resultado do vocoder STC desenvolvido neste trabalho com incremento de 10 ms, ordem do filtro *all-pole* igual a 14 e com a utilização da técnica *Subjective Loudness*.

Tabela 8.2: *Resultados PESQ dos vocoders MELP, CELP e STC*

-	MELP	CELP	STC
PESQ	2,988	3,071	3,172

A Figura 8.2 ilustra o sinal de voz original e o sinal sintetizado com o vocoder apresentado neste trabalho. A Figura 8.3 ilustra o espectrograma do sinal de voz original e o espectrograma do sinal sintetizado com o vocoder apresentado neste trabalho.

A discussão dos resultados, as conclusões e sugestões para trabalhos futuros serão abordados no próximo capítulo.

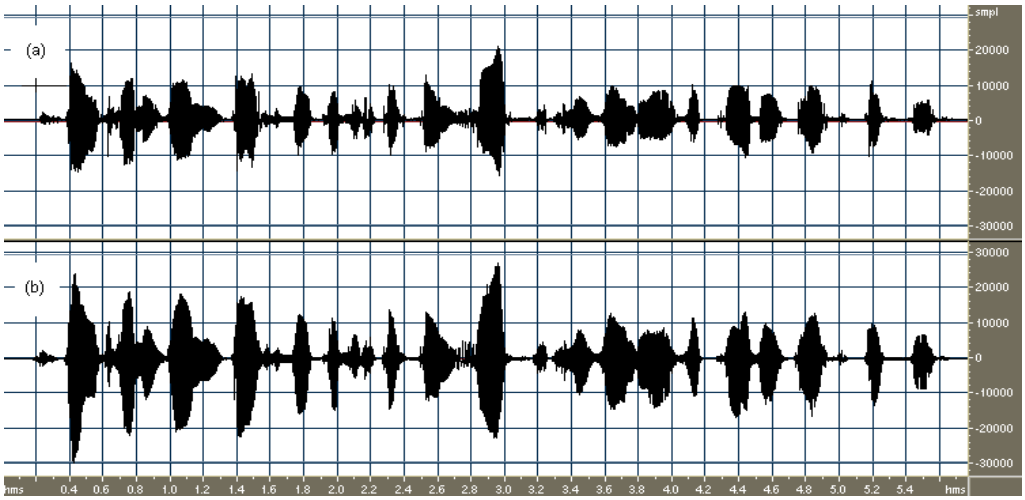

Figura 8.2: *Comparação entre os sinais de voz (a)original e (b)sintético*

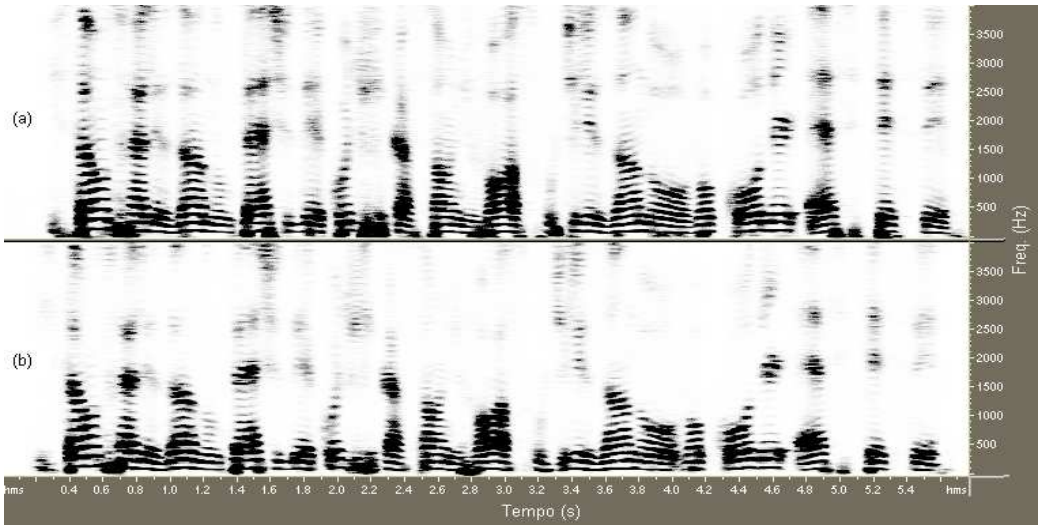

Figura 8.3: *Comparação entre os espectrogramas dos sinais de voz (a)original e (b)sintético*

Capítulo 9

Conclusão

A primeira parte deste capítulo consiste em comparar o vocoder STC com ele mesmo, levando em consideração as combinações dos fatores incremento entre quadros, ordem do filtro *all-pole* e a aplicação da técnica *Subjective Loudness*.

Baseado nos dados da Tabela 8.1, calcula-se que a modificação no fator incremento entre quadros de 15 ms para 10 ms, causa uma melhoria média no PESQ de $0,063$. A modificação na ordem do filtro de 14 para 18 causou uma melhoria média no PESQ de $0,044$ pontos. A modificação mais considerável no resultado do PESQ é justamente a da aplicação da técnica de *Subjective Loudness*, que implicou numa melhoria média de $0,068$ pontos no PESQ do vocoder STC. Este resultado é muito significativo e mostra com clareza a principal contribuição deste trabalho, pois ao contrário da diminuição do incremento ou do aumento da ordem do filtro, a aplicação da técnica de *Subjective Loudness* não implica em aumento na taxa de bits do vocoder STC.

A decisão de qual o incremento ou ainda qual a ordem do filtro a ser utilizada está vinculada à taxa de bits escolhida para a operação do vocoder STC.

A segunda parte deste capítulo é comparar o vocoder STC com seus concorrentes MELP e CELP, que trabalham a 2400 bps e 4800 bps, respectivamente.

Um vocoder STC a 2400 bps parece ser alcançável somente se o incremento entre quadros for de 15 ms e a ordem do filtro for igual a 14, pois, como discutido no capítulo anterior, restariam apenas 22 bits para a quantização das 14 LSFs. Neste caso, a título de comparação com o MELP, o vocoder STC a 2400 bps só se justificaria se o esquema de quantização degradasse o resultado do PESQ de apenas $0,106$ pontos (que é a diferença entre o PESQ do vocoder STC p14q15 e o PESQ do vocoder MELP). Isto, por si só, parece ser um desafio considerável, mas não inatingível.

O desenvolvimento do vocoder STC a 4800 bps é uma tarefa mais fácil, pois permite explorar o incremento entre quadros até 10 ms e a ordem do filtro até

18. Neste caso, a taxa de quadros por segundo seria igual a 100 e o número de bits disponíveis por quadro seria de 48. Reservando-se, como discutido no capítulo anterior, 8 bits para o pitch, 2 para a probabilidade de sonoridade e 4 para o ganho, restariam 34 bits para a quantização de 18 LSFs. Neste caso, a comparação de eficiência seria entre ele e o CELP. Assim, o vocoder STC a 4800 bps só se justificaria se o esquema de quantização a ser utilizado degradasse o resultado do PESQ de apenas $0,149$ pontos (que é a diferença entre o PESQ do vocoder STC p18q10 e o PESQ do vocoder CELP). Conforme discutido, um desafio menor que o do caso do vocoder STC a 2400 bps.

Embora o algoritmo não esteja completamente implementado, aplicando-se os métodos básicos descritos por McAulay e Quatieri e as técnicas sugeridas por este trabalho, obteve-se resultados superiores aos padrões federais norte-americanos. Ao implementar-se a um esquema de quantização para os parâmetros LSF, uma pequena degradação na qualidade é esperada, mas ainda assim espera-se uma eficiência comparável aos padrões federais norte-americanos CELP e MELP. Além disso, embora o ganho de qualidade imposto pelas técnicas de aperfeiçoamento de pitch e de síntese da fase sugeridas por este trabalho não tenham sido quantificadas, foi possível verificar nitidamente o ganho de qualidade objetiva pela estimativa do envelope do espectro baseada no conceito de *Subjective Loudness*, conforme mostrado na Tabela 8.1.

Indiscutivelmente, a elaboração da etapa de quantização dos parâmetros LSF é a sugestão de trabalho futuro mais importante à continuidade deste trabalho. Como sugestão adicional fica a análise de mais de um tipo de quantização, a saber:

- Quantização escalar utilizando a sugestão de McAulay e Quatieri de interpolar as envoltórias de espectros de quadros intercalados [MCLA78] e dividindo o espectro em duas partes LSF1 e LSF2 [SOON84]
- Quantização Vetorial
- Quantização vetorial utilizando o conceitos propostos por SOONG [SOON84] e PALIWAL [PALI93]
- Outros métodos inovadores de quantização

Há diversos estudos em andamento procurando diminuir ainda mais a taxa de bits sem perda de qualidade nos vocoders STC. Há, por exemplo, estudos relacionados com a exploração das propriedades de mascaramento temporal e freqüencial de tons no ouvido humano.

Há ainda uma demanda considerável para melhoria no procedimento de sintetização da fase.

Anexo A

Dedução da expressão (4.7)

Vimos no capítulo 4 que o sinal de fala predito pode ser representado pelas seguintes equações nos domínios do tempo ($\hat{s}(n)$) e da freqüência ($\hat{S}(k)$):

$$\hat{s}(n) = \sum_{k=1}^{K} \bar{A}(k\omega_0) e^{j\phi_k} e^{j\omega_0 kn}$$

$$\hat{S}(k) = STFT[\hat{s}(n)] = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} \hat{s}(n) e^{-j\omega_0 kn}, \;\; k = 0,\, 1,\, ...\, N$$

A expressão (4.7) está descrita a seguir:

$$\frac{1}{N+1} \sum_{n=-N/2}^{N/2} |\hat{s}(n)|^2 = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} \hat{s}(n)\hat{s}^*(n)$$

Note-se que $s(n)$ é real, então $\hat{s}(n) = \hat{s}^*(n)$. Deste modo

$$\frac{1}{N+1} \sum_{n=-N/2}^{N/2} |\hat{s}(n)|^2 = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} \underbrace{\sum_{k=1}^{K} \bar{A}(k\omega_0) e^{j\phi_k} e^{j\omega_0 kn}}_{\hat{s}(n)} \hat{s}(n)$$

$$\frac{1}{N+1} \sum_{n=-N/2}^{N/2} |\hat{s}(n)|^2 = \sum_{k=1}^{K} \underbrace{\frac{1}{N+1} \sum_{n=-N/2}^{N/2} \hat{s}(n) e^{j\omega_0 kn}}_{\hat{S}(-k)} \bar{A}(k\omega_0) e^{j\phi_k}$$

Como $\hat{S}(-k)$ é um número complexo, então $\hat{S}(-k) = \hat{S}^*(k)$. Daí

$$\frac{1}{N+1}\sum_{n=-N/2}^{N/2}|\hat{s}(n)|^2 = \sum_{k=1}^{K}\bar{A}(k\omega_0)e^{j\phi_k}\bar{A}(k\omega_0)e^{-j\phi_k}$$

$$\frac{1}{N+1}\sum_{n=-N/2}^{N/2}|\hat{s}(n)|^2 = \sum_{k=1}^{K}\bar{A}^2(k\omega_0)$$

Anexo B

New Methods for Improvement of Sinusoidal Transform Vocoders

Referências Bibliográficas

[ALME84] L. B. ALMEIDA e F. M. SILVA. *Variable-Frequency Synthesis: An Improved Harmonic Coding Scheme*, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., San Diego, CA, 1984, pp. 27.5.1-27.5.4.

[ANTO02] ANTONY W. RIX et al. *Perceptual Evaluation of Speech Quality (PESQ). The New ITU Standard for End-to-End Speech Quality Assessment*, Journal of Audio Eng. Soc., vol. 50, no. 10, October 2002, pp. 755-778.

[ATAL82] B. S. ATAL and J. R. REMDE. *A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates*, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., Paris, France, April 1982, pp. 614-617.

[DELL99] J. R. DELLER, Jr., J. H. L. HANSEN and J. G. PROAKIS. *Discrete-Time Processing of Speech Signals*, Wiley-IEEE Press, September 1999.

[FABI04] F. A. R. NASCIMENTO and F. J. FRAGA *New Methods for Improvement of Sinusoidal Transform Vocoders*, in Proc. IEEE International Conference on Multimedia and Expo (CDROM), Taipei, Taiwan, June 2004.

[FLAN66] J. L. FLANAGAN and R. M. GOLDEN. *Phase Vocoder*, in Bell Syst. Tech. J., 45, 1966, pp.1493 - 1509.

[GRIF88] D. GRIFFIN and J. S. LIM. *Multiband Excitation Vocoder*, in IEEE Trans. Acoust., Speech and Signal Proc., ASSP-36, (8), 1988, pp. 1223-1235.

[HEDE81] P. HEDELIN. *A Tone-Oriented Voice Excited Vocoder*, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., Atlanta, GA, 1981, pp. 205-208.

[HERM90] H. HERMANSKY *Perceptual linear predictive (PLP) analysis of speech*, J. Accoust. Soc. Am., vol. 87 (4), pp 1738-1752, April 1990.

[ITAK75] F. ITAKURA *Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals*, J. Accoust. Soc. Am., 57, 535(A), 1975.

[KLEI93] W. B. KLEIJN and P. KROON. *A 5.85 kb/s CELP Algorithm for Celular Applications*, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., Minneapolis, NM, 1993, pp. II596-599.

[KLEI95] W. B. KLEIJN and J. HAAGEN. *A Speech Coder Based on Decomposition of Characteristic Waveforms*, in Proc. Of ICASSP-95, Detroit, Michigan, May 16-19, 1995, pp. 508-511.

[KOHS02] S. N. KOH and G. H. CHUA *Application of Auditory Masking in Improved Multiband Excitation Model*, Applied Acoustics, 63, 2002, pp. 693-698.

[KULD93] KULDIP K. PALIWAL and BISHNU S. ATAL *Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame*, in IEEE Transactions on Speech and Audio Processing, vol. 1 (1), pp 3-14, January 1993.

[MAKH78] J. MAKHOUL, R. VISWANATHAN, R. SCHWARTZ and A. W. F. HUGGINS. *A Mixed-Source Model for Speech Compression and Synthesis*, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., Tulsa, OK, 1978, pp. 163.

[MALA79] D. MALAH. *Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals*, in IEEE Trans. Acoust., Speech and Signal Proc., ASSP-27, (2), 1979, pp. 121-133.

[MARQ88] J. S. MARQUES and L. B. ALMEIDA. *New Basis Functions for Sinusoidal Decomposition*, in Proc. EUROCON, Stockholm, Sweden, 1988.

[MCAU86] R. J. McAULAY and T. F. QUATIERI. *Speech Analysis/Synthesis Based on a Sinusoidal Representation*, in IEEE Trans. Acoust., Speech and Signal Proc., vol. 34, (4), August 1986, pp. 744-754.

[MCAU88] R. J. MCAULAY and T. F. QUATIERI. *Computationally Efficient Sine-wave Synthesis and its application to Sinusoidal Transform Coding*, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., New York, NY, April 11-14, 1988, pp. 370-373.

[MCAU90] R. J. MCAULAY and T. F. QUATIERI. *Pitch estimation and Voicing detection Based on a Sinusoidal Model*, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., Albuquerque, NM, 1990, pp. 249-252.

[MCAU95] R. J. McAULAY and T. F. QUATIERI. *Sinusoidal Coding*, in Speech Coding and Synthesis (W. B. Kleijn and K. K. Paliwal, eds.) ch. 4, Elsevier Science B. V., 1995, pp 121 - 173.

[MCLA78] E. McLARNON *A Method for Reducing the Frame Rate of a Channel Vocoder by Using Frame Interpolation*, in Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, Washington, D.C., pp 458-461, 1978.

[OPPE83] A. V. OPPENHEIM and R. W. SCHAFER. *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1983.

[PALI93] K. K. PALIWAL and B. S. ATAL *Efficient Vector Quantization of LPC Parameters at 24 bits/frame*, in Proc. IEEE Transactions on Speech and Audio Processing, Vol. 1 (1), January 1993, pp 3-14.

[PAUL81] D. B. PAUL. *Spectral Envelope Estimation Vocoder*, in IEEE Trans. on Acoust., Speech and Signal Proc., ASSP-29, 1981, pp. 786-794.

[PICO93] J. W. PICONE *Signal modeling techniques in speech recognition*, in Proc. IEEE, vol. 8 (9), pp. 1215-1247, Sept. 1993.

[PORT81] M. PORTNOFF. *Short-Time Fourier Analysis of Sampled Speech*, in IEEE Trans. Acoust., Speech and Signal Proc., ASSP-29, (3), 1981, pp. 364-373.

[RABI78] L. RABINER and R. SCHAFER. *Digital Processing of Speech*, Prentice Hall, Inc., Englewood Cliffs, NJ, 1978.

[SCHR85] M. R. SCHROEDER and B. S. ATAL. *Code-excited Linear Prediction (CELP): High-quality Speech at Very Low Bit Rates*, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., Tampa, FL, 1985, pp. 937-940.

[SOON84] F. K. SOONG and B-H. JUANG *Line Spectrum Pair (LSP) and Speech Data Compression*, in Proc. IEEE 1984 Conf. Acoust. Speech and Signal Processing, San Diego, CA, pp 1.10.1-1.10.4, 1984.

[TREE68] H. VAN TREES *Detection, Estimation and Modulation Theory, Part I*, Wiley, New York, 1968.

[UNSE93] M. UNSER, A. ADROUBI and M. EDEN *B-Spline Signal Processing*, in IEEE Trans. on Signal Procecssing, vol. 41, (2), Feb 1993, pp. 821-883. Speech and Signal Proc., Dallas, TX, 1987, pp. 51.3.1

[WANG92] S. WANG, A. SEKEY and A. GERSHO *An Objective Measure for Predicting Subjective Quality of Speech Coders*, in IEEE Journal on Selected Areas in Communications, vol. 10, (5), June 1992.

[WWCH98] W. -W. CHANG and D. -Y. WANG. *Quality Enhancement of Sinusoidal Transform Vocoders*, in IEE Proc.-Vis. Image Signal Process., vol. 145, (6), December 1998, pp 379 - 383.