TRANSFORMAÇÕES EM SINAIS DE VOZ:
MORPHING E MODIFICAÇÃO DE PITCH

Rafael Cauduro Dias de Paiva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. Luiz Wagner Pereira Biscainho, D.Sc.

Prof. Sergio Lima Netto, Ph.D.

Prof. Abraham Alcaim, Ph.D.

Profa. Mariane Rembold Petraglia, Ph.D.

RIO DE JANEIRO, RJ - BRASIL
FEVEREIRO DE 2008

PAIVA, RAFAEL CAUDURO DIAS DE

Transformações em sinais de voz: morphing e modificação de pitch [Rio de Janeiro] 2008

XIII, 111 p., 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Elétrica, 2008)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1.Processamento de sinais de fala 2.Modificação de pitch 3.Transformação de locutor

I.COPPE/UFRJ II.Título (série)

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, pelo financiamento.

Aos meus orientadores, Luiz Wagner Pereira Biscainho e Sergio Lima Netto, excelentes professores e pesquisadores extremamente competentes, pelo apoio que deram ao desenvolvimento do meu trabalho, e pela compreensão em momentos difíceis.

Aos colegas com quem convivi no LPS, pesquisadores brilhantes, entre os quais incluo Alan Tygel, Alexandre Leizor, Amaro de Lima, Bruno Bispo, Guilherme Pinto, Flávio Ávila, Leonardo Nunes, Markus Lima, Paulo Esquef, Rafael de Jesus e Wallace Martins. Por sua ajuda com o banco de dados agradeço ao Felipe Martins.

Aos amigos com quem convivi durante este tempo que morei no Rio de Janeiro e com quem muito aprendi sobre música e vida, Guilherme Pires, Luisa Francesconi, Giovanni Tristacci, Renata Gomes e Gian Matheus.

Em especial aos amigos pesquisadores Fábio Freeland e Tadeu Ferreira, que além de contribuírem em muitas discussões técnicas, sempre estiveram disponíveis para me apoiar; vocês são pessoas incríveis e desejo tudo de bom nas suas vidas.

Aos meus pais, os professores Eloiza e João Batista, por todo apoio que me deram, pela orientação em assuntos pessoais e profissionais. Vocês são e sempre serão para mim exemplos de dedicação, empenho e ética na vida profissional e pessoal.

À minha querida esposa, Juliana (*Jubee*), por todo seu amor e incentivo durante esses anos, e por que tu és — *tudo enfim que tem de belo em todo resplendor da santa natureza*[1] — a razão dos meus dias e da minha vontade de crescer. E também por ser minha amada, *pelo amor predestinada, sem a qual a vida é nada, sem a qual se quer morrer*[2] e porque sem ti *não há paz não há beleza, é só tristeza e a melancolia que não sai de mim*[3]. Te amo, e *por toda a minha vida eu vou te amar*[4].

[1]Rosa, Pixinguinha

[2]Minha Namorada, Vinicius de Moraes

[3]Chega de Saudade, Tom Jobim e Vinicius de Moraes

[4]Eu Sei Que Vou Te Amar, Tom Jobim e Vinícius de Moraes

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

TRANSFORMAÇÕES EM SINAIS DE VOZ: MORPHING E MODIFICAÇÃO DE PITCH

Rafael Cauduro Dias de Paiva

Fevereiro/2008

Orientadores: Luiz Wagner Pereira Biscainho
Sergio Lima Netto

Programa: Engenharia Elétrica

Esta dissertação apresenta técnicas de transformação de voz, que incluem modificação de *pitch* e transformação de locutor. Para isso são apresentadas ferramentas para modelagem de sinais de voz, e é proposto um algoritmo para a discriminação entre trechos sonoros e surdos.

As técnicas de modificação de *pitch* propostas usam um modelo seqüencial, baseado no algoritmo RLS (*recursive least-squares*), para a aproximação da envoltória espectral do sinal de voz, o que evita efeitos de descontinuidade e o atraso inerente ao processamento em blocos. Ainda, é proposto um sistema de modificação de *pitch* que combina o modelo seqüencial do trato vocal com o algoritmo PSOLA.

A técnica de transformação de locutor que é proposta usa uma abordagem estatística para estimar os coeficientes mel-cepstrais do locutor alvo, e usa informação de blocos passados para aumentar a naturalidade dos sinais transformados.

São apresentados resultados experimentais de sinais modificados usando os sistemas de transformação de locutor e modificação de *pitch*, com o objetivo de comprovar o funcionamento dos algoritmos propostos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

TRANSFORMATIONS ON VOICE SIGNALS:
MORPHING AND PITCH MODIFICATION

Rafael Cauduro Dias de Paiva

February/2008

Advisors: Luiz Wagner Pereira Biscainho
Sergio Lima Netto

Department: Electrical Engineering

This thesis deals with transformation techniques that include pitch modification of voice signals and voice morphing. For this purpose some tools are presented for voice modeling and for discrimination between voiced and unvoiced signals.

The pitch modification techniques proposed use a sequential scheme for the spectral envelope modeling, based on the RLS (recursive least-squares) algorithm. The sequential scheme avoids discontinuities and the inherent delay of block-processing techniques. It is also proposed a voice modification system that combines the sequential model and the PSOLA (pitch-synchronous overlap-and-add) algorithm.

It is proposed a voice morphing technique, that uses a statistical approach for the estimation of mel-cepstral coefficients of the target voice. The proposed approach also uses information of previous blocks to provide more naturality in the synthesized signal.

Some experimental results of modified signals using both the pitch modification and the voice morphing techniques are presented to show the efficiency of the proposed algorithms.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A voz é um dos principais meios de comunicação humana no dia a dia. Tanto para comunicação entre pessoas em suas interações diárias, como para a produção de arte — música, teatro, cinema — a voz tem um espaço de destaque. Esse destaque especial para o uso da voz tem guiado o desenvolvimento das tecnologias relacionadas à comunicação. Como exemplo podemos citar a importância do desenvolvimento da tecnologia de telefonia, a convergência do uso da internet para telefonia (voz sobre IP, VoIP, e ferramentas como *Skype*) e o desenvolvimento de interfaces homem/máquina que usam voz (aqui podem ser incluídos reconhecimento e síntese de voz).

Entre as ferramentas desenvolvidas para o processamento de voz, algumas estão relacionadas a transformar o conteúdo desse tipo de sinal. Entre os tipos de transformações estão as modificações de *pitch*[1] e de tempo. Outro tipo de transformação que tem recebido destaque é a transformação de locutor, chamada em alguns trabalhos de *voice morphing* ou *voice conversion*. Esses tipos de transformações têm sido aplicadas tanto como ferramentas para outros tipos de sistemas, como sintetizadores de voz, quanto como produto final.

As modificações de *pitch*/tempo são um tipo de transformação bastante difundido para edição de sinais musicais e trilhas sonoras de filmes, afinação automática de voz cantada, mudança de prosódia em sistemas de conversão texto/fala, ferramentas de auxilio para ensino de línguas, ferramentas de auxílio para compositores, etc. Transformação de locutor consiste em processar um sinal gravado com a

[1] O conceito que envolve o termo *pitch* está apresentado na Seção 2.2, e pode ser entendido como freqüência fundamental (em termos matemáticos), ou altura (em termos musicais).

voz de uma pessoa, de forma que o resultado final pareça ter sido emitido por uma outra pessoa. O grande desafio deste tipo de sistema é a obtenção de características que sejam relevantes para a representação do que consideramos ser o timbre de um locutor, e determinação de funções de mapeamento destas características acústicas, de forma a fazer uma transformaçao de timbre com qualidade. Entre as aplicações desse tipo de técnica está a adaptação de sistemas de síntese de voz, edição de voz em dublagens, recriação da voz de cantores antigos, etc.

Um exemplo bem sucedido do uso dessas técnicas foi a recriação da voz de um *castrato* no filme *Farinelli* de Gérard Corbiau [4, 5]. Neste caso, o desafio foi recriar a voz de um tipo de cantor que não existe mais, combinando as vozes de um homem e uma mulher. Outras aplicações e exemplos incluem: síntese de voz [6], transformações de expressividade de sinais de voz [7, 8], ferramentas de auxílio para pessoas com deficiência de fala [9, 10]. Entre as empresas que fabricam produtos relacionados a esse tipo de transformação estão [10]:

- Antares, que fabrica:
  - Auto-Tune®: *software* para correção automática de afinação;
  - Vocal-producer®: equipamento para correção de afinação em tempo real;
  - THROAT®: *software* que com um modelo do trato vocal consegue fazer efeitos de mudança de tamanho do trato vocal e mudança de excitação glotal, entre outros;
  - CHOIR Vocal Multiplier®: *software* para transformar uma voz em um coral;
- Celemony, que produz o Melodyne®: *software* de processamento de voz, que inclui mudança de *pitch* e tempo;
- Yamaha, que produz:
  - o Vocaloid®: sintetizador de voz cantada;
  - PLG100-VH®: placa que produz efeitos de voz que incluem mudança de gênero e vibrato;

- Voxonic e Sestek, que produzem em conjunto o *software* VOX®, para replicar o timbre de um locutor em outra língua, aplicado para dublagem em filmes e propagandas comerciais;

- Boss, que produz o Voice Transformer®: equipamento que possibilita ajuste de timbre, afinação e formantes;

- Digitech, que produz o Vovalist Live® 2 e 4: equipamento que possibilita criar vozes, como combinações de uma voz com *pitch* modificado, de acordo com a progressão harmônica de um violão ou guitarra em tempo real.

O objetivo deste trabalho é estudar técnicas de transformação de sinais de voz. Entre as técnicas estudadas estão modificação de afinação de sinais de voz, no Capítulo 5, e transformação de locutor, no Capítulo 6.

Para que esse objetivo seja alcançado, o Capítulo 2 apresenta de forma simples conceitos e características da voz; detalhes sobre o funcionamento do aparelho de fonação humana; e uma interpretação física do processo de produção de voz, que vai ser usada para os modelos do trato vocal que são descritos no capítulo seguinte. A importância desse capítulo está em apresentar ao leitor parte dos conceitos e da nomenclatura usada no texto.

O objetivo do Capítulo 3 é apresentar modelos simplificados para a descrição do processo de geração de voz. Esses modelos são ferramentas importantes, pois para que as transformações que são propostas neste texto sejam efetivas é necessário o uso de algum tipo de representação paramétrica, relacionada ao processo de produção de voz.

É comum que sinais de voz modificados apresentem algum tipo de defeito audível quando trechos sonoros e surdos não são detectados de forma adequada. Isso de deve ao fato de esses algoritmos de modificação serem projetados de forma diferente para trechos com sonoridade diferente — no caso deste texto, por exemplo, os trechos surdos não são processados. Desta forma, o objetivo do Capítulo 4 é fazer um estudo sobre ferramentas para a distinção entre trechos sonoros, surdos e de silêncio, e projetar um classificador de trechos sonoros para resolver este tipo de problema.

Sistemas de modificação de *pitch* sofrem com freqüência do atraso resultante

do processamento em blocos; esse tipo de atraso pode ser bastante perturbador para cantores que usam esse tipo de ferramenta em apresentações ao vivo — Clark [11] afirma que atrasos com mais de 5 ms em sistemas de mixagem ao vivo são perturbadores para músicos. Para resolver este tipo de problema, o objetivo do Capítulo 5 é desenvolver um algoritmo de modificação de *pitch* com pouco atraso, que possa ser implementado em tempo real. Assim, este capítulo apresenta um esquema de análise/síntese seqüencial que evita o processamento em blocos.

O objetivo do Capítulo 6 é implementar um sistema que transforma o timbre de um locutor/cantor de forma que os sinais transformados com esse sistema pareçam ter sido falados/cantados por outra pessoa. Para isso são propostas melhorias em sistemas existentes na literatura, de forma a levar em conta aspectos de continuidade da envoltória espectral para a transformação de locutor/cantor.

O Capítulo 7 conclui a dissertação, enfatizando suas principais contribuições e apontando diretrizes para possíveis extensões do trabalho desenvolvido.

Capítulo 2

Definições fundamentais e processo de produção da voz

2.1 Introdução

Em muitas áreas de processamento de sinais é possível tirar vantagem de características físicas do processo de geração do sinal a ser tratado, ou da percepção humana deste. Tanto em processamento de imagens como de áudio é possível fazer uso de modelos de percepção para codificação desses sinais. Em processamento de voz, importantes avanços foram feitos com o uso de modelos para descrever o seu processo de geração. Para a obtenção destes modelos para os sinais de voz, é preciso muitas vezes recorrer a outras áreas do conhecimento, que a princípio parecem não ter relação com engenharia, como a anatomia humana e a fonoaudiologia.

O entendimento, mesmo que simplificado, do sistema de produção de voz é de fundamental importância para o processamento adequado destes sinais. Neste capítulo são abordadas características gerais da voz, a anatomia do sistema fonatório e as suas interpretações físicas. Foram essas interpretações físicas as principais inspiradoras de sistemas de codificação e de modificação eficiente de sinais de voz. Na Seção 2.2 são abordadas algumas características e definições de conceitos comuns para descrever características de voz e música que serão usados no decorrer do texto. Na Seção 2.3 é apresentado o sistema de funcionamento do orgão de produção de fala. Na Seção 2.4 os conceitos apresentados na Seção 2.3 são usados para propor um modelo clássico para a produção de voz. Essa interpretação física para o processo

de produção de voz será usada no restante do texto, e é a base de grande parte das técnicas de processamento de voz.

2.2 Características gerais da voz

Existem algumas características básicas que tornam possível a distinção entre diferentes tipos de voz.

- *Freqüência fundamental percebida*: é chamada na literatura de *pitch*, e está relacionada à percepção humana de sinais acústicos. Embora existam modelos bastante complexos para a percepção de *pitch* [12, 13], para sinais de voz em muitos casos este é aproximado pela freqüência dos pulsos glotais. Quanto mais espaçados entre si, menor o *pitch* ou, equivalentemente, maior o período de *pitch*.

  As notas musicais estão diretamente relacionadas com um contorno geral da freqüência fundamental percebida, e alguns efeitos de modificações rápidas nesta freqüência são percebidos não como variação de nota, mas como *vibrato*. O *vibrato* é normalmente uma variação que pode se estender a 2% da freqüência fundamental, numa taxa de 5 a 10 ciclos por segundo. Entre as características específicas da voz com relação à freqüência fundamental, podemos destacar:

  - *Extensão vocal:* compreende toda a extensão de *pitch* que uma pessoa é capaz de emitir;
  - *Tessitura:* compreende uma faixa dentro da extensão vocal em que o cantor é capaz de emitir um som com características de timbre adequadas para a música.
  - *Registro vocal:* é comum cantores e profissionais da voz utilizarem o termo registro para diferenciar tipos de emissão vocal. Os dois principais registros são o *registro do peito*, que indica o tipo de emissão vocal mais usual, e o *registro da cabeça*, muitas vezes usado para designar o *falsete*. Essa nomenclatura está relacionada com a maneira como cantores sentem a sua voz ressoar, no peito ou na cabeça, ao emitir uma nota. O registro do peito é considerado como sendo o mais adequado para música, pois

a emissão adquire mais *corpo* ou *brilho*, entretanto existem músicos que preferem o registro da cabeça, devido à facilidade que este proporciona para alcançar notas mais altas.

A faixa de notas que cantores com determinados tipos de voz são capazes de emitir usando o registro do peito está ilustrada na Figura 2.1. Na figura, cada posição referente às notas *lá* (*An* na notação que foi usada, onde *n* é o número da oitava) está indicada com a sua respectiva freqüência em *hertz*. Normalmente usam-se os termos *baixo* e *alto* para diferenciar notas com freqüência percebida menor ou maior, respectivamente.

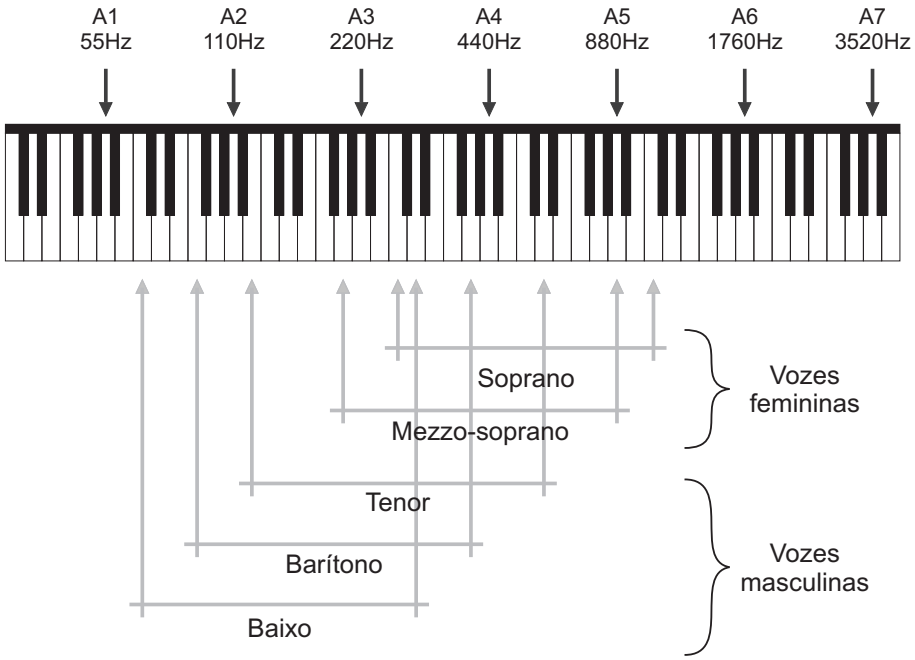
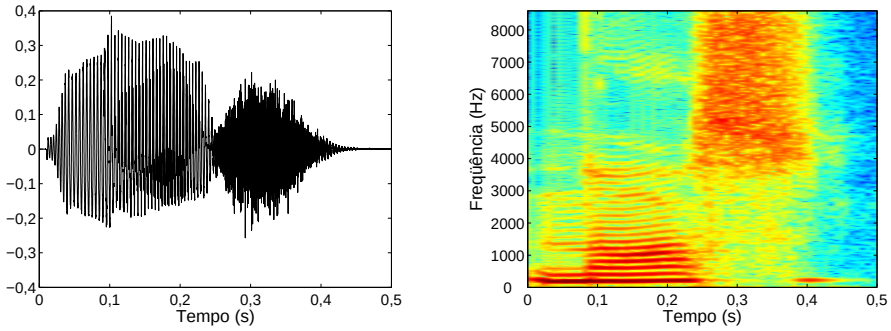


Figura 2.1: Extensão vocal de cada tipo de voz.

- *Timbre:* é definido pelo conjunto de freqüências presentes em um determinado sinal para gerar uma nota musical. Normalmente é definido pela relação de freqüências harmônicas da fundamental, apesar de também poder ser gerado por freqüências que não são exatamente harmônicas. Diferentes timbres ajudam a diferenciar fonemas, assim como a pessoa ou instrumento musical que está emitindo um som.

- *Sonoridade:* distinção entre tipos de sons emitidos, que podem ser sonoros, com características pseudo-periódicas, e surdos, com comportamento funda-

mentalmente aleatório. Sons sonoros são produzidos com vibração das pregas vocais, enquanto sons surdos são produzidos pelo ruído de escoamento turbulento do ar pelo trato vocal, sem vibração das pregas vocais. Exemplos de sons sonoros são os das vogais *(a, e, i, o, u)*, e de sons surdos são o chiado do $/s/$ e o rápido transitório do $/t/$. A Figura 2.2 mostra exemplos de sons sonoros e surdos na palavra *Nosso*, onde é possível perceber uma parte periódica corresponde a *no*, uma parte aleatória correspontente ao $/s/$ e novamente uma parte periódica correspondente ao $/o/$.



(a) (b)

Figura 2.2: Exemplo de sinais sonoros e surdos. (a) Sinal no tempo e (b) espectrograma da palavra *nosso*. Nestas figuras é possível notar uma parte periódica nos primeiros 0,25 s do sinal, correspondendo a uma parte sonora; uma parte aleatória entre 0,25 e 0,4 s, correspondendo a um trecho surdo; e um trecho sonoro entre 0,4 e 0,5 s.

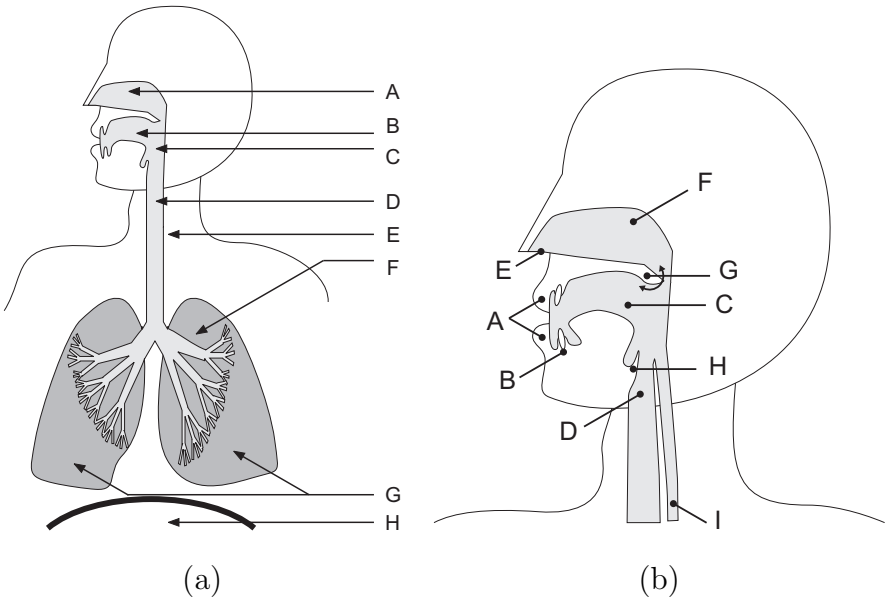
- *Intensidade (dinâmica)*: é a propriedade de determinados sons parecerem mais *fortes* ou *fracos*, e está diretamente ligada à energia do sinal. Em notação musical usam-se os termos italianos *pianissimo*, *piano*, *forte*, *fortissimo* para dar idéia da dinâmica a ser adotada na execução musical. Termos como *baixo* e *alto* devem ser evitados para descrever a intensidade, pois são normalmente atribuídos à freqüência fundamental percebida. Alguns efeitos de dinâmica são usados como recurso expressivo, entre eles esta o *tremolo*, que é uma variação cíclica na intensidade de execução de uma nota.

Observação: As 12 notas musicais na escala de temperamento igual são dispostas em intervalos geometricamente distribuídos. Cada conjunto de 12 notas é

denominado de *oitava* musical, e a 13^a nota tem o mesmo nome da 1^a nota da oitava anterior, mas com uma freqüênca 2 vezes maior. Uma vez determinada a freqüência de uma nota (*e.g.* A4 = 440 Hz), as notas adjacentes mais altas são determinadas multiplicando a freqüência por $\sqrt[12]{2}$, e as mais baixas são obtidas dividindo a freqüência de referência pelo mesmo fator.

2.3 Anatomia da fala

O modelo do processo de geração de fala pode ser dividido em 3 partes, sendo elas os foles, o vibrador e os ressonadores [2]. Uma visão geral dos órgãos de produção de fala é mostrada nas Figuras 2.3 (a) e (b) [1], e uma visão esquematizada de seu funcionamento é mostrada na Figura 2.4.



(a) (b)

Figura 2.3: (a) Visão Geral dos órgãos de produção de fala (A - cavidade nasal; B - boca; C - faringe; D - laringe; E - traquéia; F - brônquios; G - pulmões; H - diafragma); (b) Detalhe dos órgãos de produção de fala (A - lábios; B - dentes e gengivas; C - base da língua; D - pregas vocais; E - narinas; F - fossas nasais; G - véu do palato; H - epiglote). Adaptado de [1].

Na produção de voz, os foles empurram o ar através da traquéia. Na laringe ocorre um estreitamento da traquéia e conseqüente aumento da velocidade de escoamento do ar, e as pregas vocais fazem o papel do vibrador. Essa vibração tem, então,

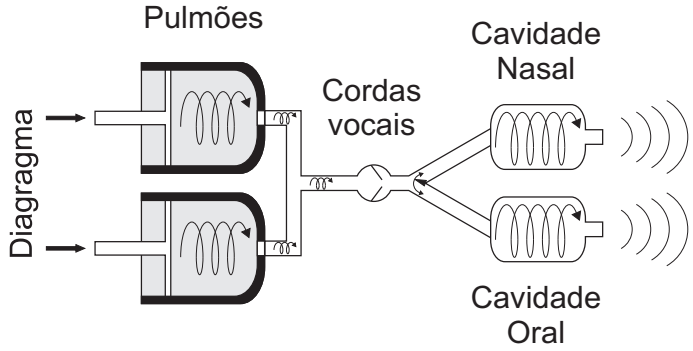


Figura 2.4: Visão esquemática da produção de fala (adaptado de [2]).

suas características modificadas pelos diversos elementos ressonadores das cavidades oral e nasal. Esses ressonadores são controlados pelos elementos de articulação da boca, faringe e laringe durante a produção de voz [2].

2.3.1 Sopro fonatório

Os foles são os responsáveis pelo que se chama de *sopro fonatório*, que é produzido pelo esvaziamento controlado do ar dos pulmões, ocorrido durante a expiração devido à sua compressão. Durante a inspiração também é possível emitir sons laríngeos, contudo esse tipo de emissão não é normal na produção de fala, e se apresenta em alguns casos patológicos.

O sopro fonatório pode ser classificado em 4 tipos [1]:

- No *sopro torácico superior* ocorre um abaixamento da caixa torácica e uma conseqüente compressão da parte superior dos pulmões.

- No *sopro abdominal* a ação de músculos do abdome produz uma retração da parede abdominal, empurrando o diafragma para cima e estreitando a caixa torácica. Nesse movimento o diafragma contém e controla a ação destes músculos, fazendo a dosagem do sopro fonatório.

- Na *respiração vertebral* a flexão e extensão da coluna vertebral torácica, com participação de toda a musculatura do tronco, leva a um arqueamento das costas e projeção do rosto para a frente, comprimindo a caixa torácica. Este tipo de expiração apresenta um contexto de esforço relativamente importante que pode levar à fadiga vocal.

- Na *respiração mista* os três mecanismos acima aparecem associados.

Os tipos de sopro fonatório acarretam diferentes formas de projeção vocal, por exemplo, o sopro abdominal é mais usado na projeção da voz para platéias, enquanto o sopro torácico superior é mais usado para a projeção da voz para a expressão normal.

2.3.2 A laringe e a vibração na produção de voz

A função principal da laringe é a de fazer o controle do caminho que o ar, ou comida, fazem durante a respiração e a deglutição. Ela faz esse controle pela abertura e fechamento da epiglote, mostrada nas Figuras 2.3 (b) e 2.5.

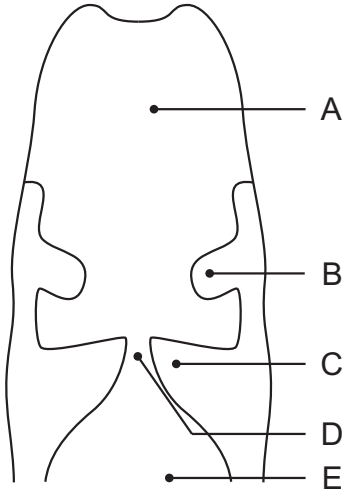


Figura 2.5: Detalhe da laringe (A - epiglote; B - prega vestibular (ou falsa corda vocal); C - prega vocal; D - glote; E - traquéia.) Adaptado de [1].

O funcionamento da laringe na produção de fala tem causado questionamentos desde muito cedo na história da ciência. Algumas das hipóteses sobre seu funcionamento datam do século II a.C. com Galiano, que compara o órgão vocal com uma flauta. Com a realização de experiências em cadáveres, Ferrein comparou em 1741 as formações da laringe com cordas de violino, sendo a corrente de ar pulmonar comparada ao arco do violino, excitando as *cordas vocais* (dando origem a esse termo), e a tensão aplicadas sobre elas o fator de controle da freqüência de vibração [1].

No século XIX apareceram teorias um pouco mais elaboradas, entre elas a teoria mioelástica, representada esquematicamente na Figura 2.6 ([3] *apud* [1]). Segundo esta teoria, as pregas vocais têm um papel passivo na vibração vocal, e sua posição de relaxamento é fechada, sendo que o pulmão exerce um papel ativo no aumento da pressão sobre elas enquanto é comprimido. Assim que a pressão exercida nas pregas vocais ultrapassa determinado limiar, elas se afastam, deixando passar uma pequena quantidade de ar. Esse escoamento em pequenas quantidades de ar a intervalos harmônicos é responsável pela vibração das pregas vocais que produz a voz.

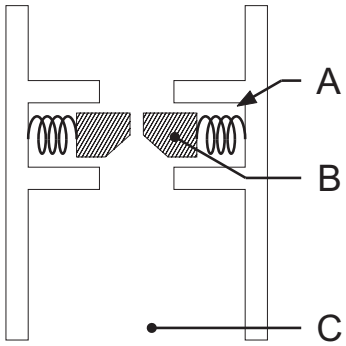


Figura 2.6: Representação de Grémy (1968) da teoria mioelástica de vibração das pregas vocais, onde: A - mola que corresponde à força de atração resultante da elasticidade da prega vocal; B - peça que representa a prega vocal; C - traquéia ([3] *apud* [1]).

Os problemas desta teoria levaram ao questionamento da ação das pregas vocais como sendo um mecanismo passivo, com o desenvolvimento de outras teorias em que as pregas vocais desempenhariam um papel ativo durante a vibração. Entre elas está a teoria neurocronáxica de Husson ([14] *apud* [1]). Mesmo assim, a hipótese mais aceita atualmente é de que as pregas têm papel passivo na produção da vibração de voz.

A Figura 2.7 mostra as fases de vibração das pregas vocais. Nessa figura é possível observar o movimento ondulatório em cada ciclo de vibração das pregas vocais, em que o fechamento da glote começa pela parte inferior e termina na parte superior da glote. Da mesma forma, durante a abertura da glote, o movimendo vai de baixo para cima. Esse comportamento da glote gera um fluxo de ar descontínuo;

cada vez que a glote abre e fecha dá-se uma variação no fluxo de ar. A freqüência de abertura e fechamento glotal determina a freqüência fundamental percebida, ou o *pitch*. A Figura 2.8 mostra uma forma de onda típica de vazão e pressão do ar durante a fonação.

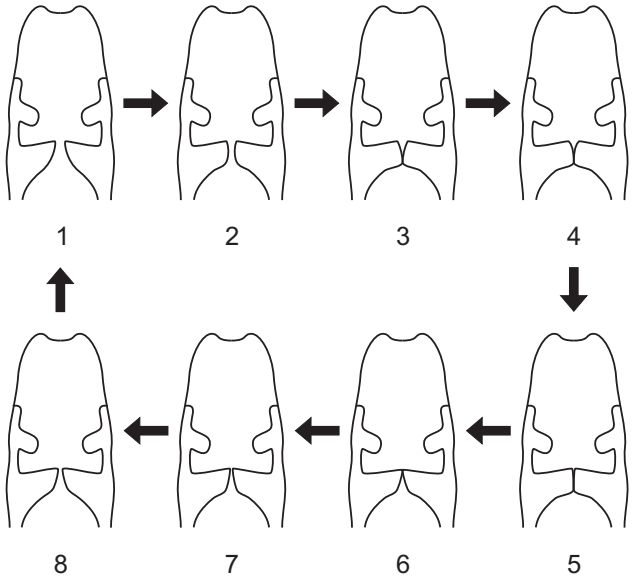


Figura 2.7: Abertura e fechamento glotal durante a fonação [2, 1].

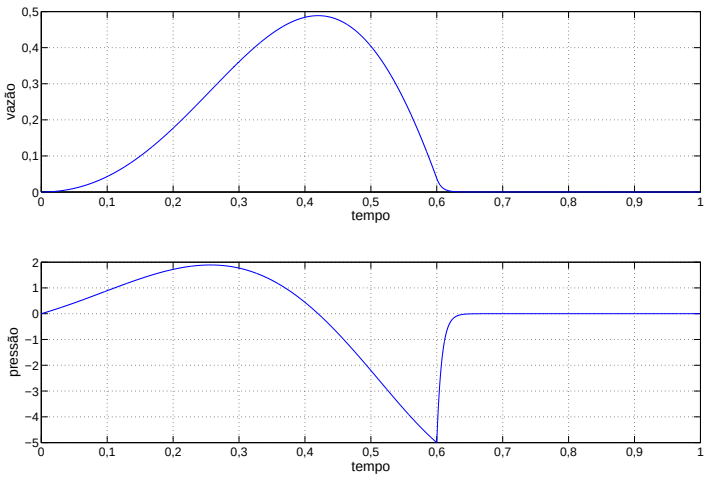


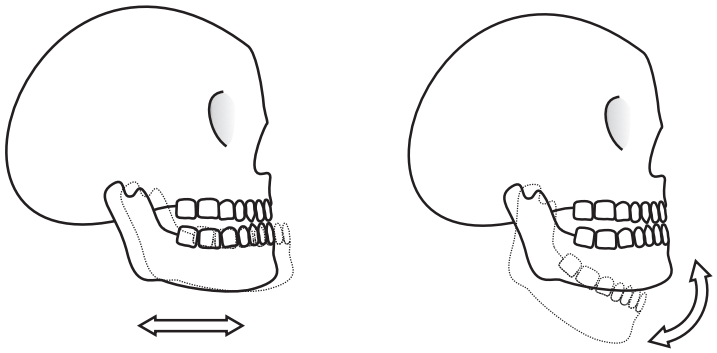
Figura 2.8: Formas de onda da vazão e pressão do ar passando pela laringe.

Influência da excitação glotal nos diferentes tipos de emissão

Existem alguns trabalhos mostrando a influência das pregas vocais em diferentes tipos de emissão. A princípio, as técnicas de processamento de voz consideravam a excitação glotal como sendo puramente impulsiva; contudo, trabalhos mais atuais tentam obter modelos mais sofisticados para a excitação glotal. Dentre estes modelos, o de Fant/Liljencrants [15] é um dos mais populares; nele é possível considerar os tempos de abertura e fechamento da glote. As formas de onda da Figura 2.8 foram obtidas usando-se este modelo. Já foi demonstrado que os parâmetros deste modelo têm uma relação direta com a sensação de esforço vocal, e os diferentes tipos de emissão (normal, suspirada, sussuro, falsete) [16, 17].

2.3.3 Os elementos de articulação e os ressonadores da voz

Uma vez gerada a vibração pelas pregas vocais, os diversos elementos articuladores do trato vocal são controlados para operar mudanças de timbre e de fonemas na emissão de voz. Na Figura 2.3 (b) é possível distinguir entre os elementos articuladores os lábios, que, juntamente com a posição da mandíbula, mostrada na Figura 2.9, controlam a forma de saída de ar do trato oral; o véu do palato e palato mole (o popular *céu da boca*), que controlam a separação do fluxo de ar entre os tratos nasal e oral; a língua, que controla o volume e forma do trato oral.

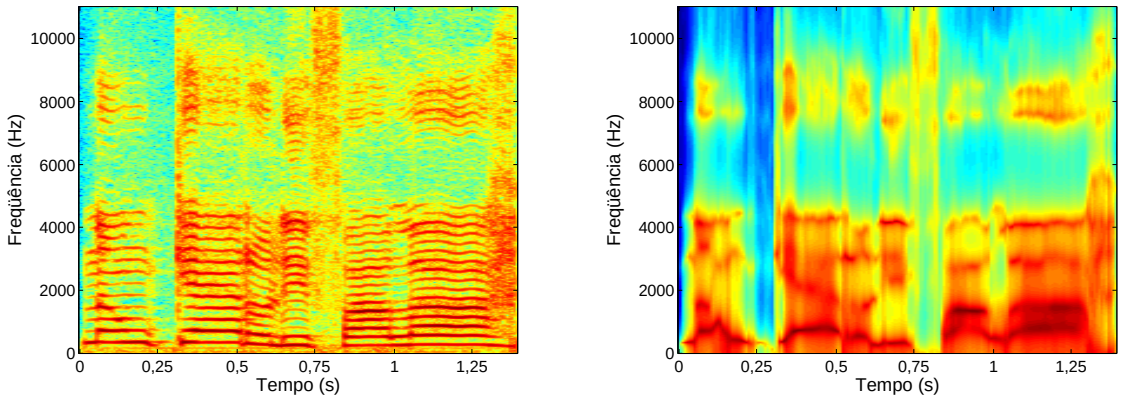


(a) (b)

Figura 2.9: Movimentos de (a) propulsão/repulsão (b) abaixamento/elevação da mandíbula [1].

É possível realizar experiências simples para demonstrar como esses elementos articuladores podem influenciar nos fonemas emitidos. Deixando os lábios em

posição de *bico* podemos emitir o fonema *u*, e ao abrir os lábios lentamente percebemos que o som emitido começa a parecer com o da vogal *a*. Ao emitir as vogais *e* e *i* observamos o levantamento da parte posterior da língua. Ao emitir a vogal



(a) (b)

Figura 2.10: (a) Espectrograma e (b) estimativa da envoltória espectral de um sinal de voz.

a e modificar o som lentamente para emitir o *an* nasalizado, podemos perceber o movimento do palato mole, e a conseqüente alteração no controle de fluxo de ar pela boca e pelo nariz.

Todos esses elementos de articulação influenciam no controle de ressonâncias, chamadas de formantes, que são responsáveis pela distinção entre fonemas e pelo timbre característico de cada pessoa. Os formantes podem ser percebidos pela visualização do espectro de um sinal de voz como picos na envoltória espectral. Dependendo da freqüência de amostragem do sinal de voz, podem-se observar de 2 a 4 dessas ressonâncias. Um exemplo de representação freqüencial de um sinal de voz com sua respectiva envoltória espectral é mostrado na Figura 2.10, onde podemos observar a variação dos formantes e da freqüência fundamental para diferentes instantes de tempo.

Existem trabalhos que mostram a variação da freqüência dos formantes para indivíduos de diferentes faixas etárias. Huber *et al.* [18] é um bom exemplo, em que se mostrou a evolução da posição dos 3 formantes do fonema /*a*/ para crianças a partir de 4 anos de idade e adultos. Este trabalho mostra para crianças uma pequena diferenciação entre sexos na posição dos formantes, que se torna mais evidente em

indivíduos com idade maior que 14 anos. A medida da posição dos 3 primeiros formantes em adultos do sexo masculino resultou em 697, 1244 e 2606 Hz, enquanto que para indivíduos do sexo feminino essas freqüências eram um pouco mais elevadas: 888, 1420, 3030 Hz.

2.4 Interpretação física: Sistema fonte-filtro

O processo de produção de voz pode ser intepretado como sendo um sistema fonte-filtro mostrado na Figura 2.11, onde os ressonadores dos tratos oral e nasal são representados por filtros digitais, e o sinal que alimenta o sistema representa a excitação glotal. A implementação do sistema é feita usando filtros recursivos IIR, com pelo menos um par de pólos para cada ressonância (10 coeficientes para sinais amostrados a 8 kHz, 30 coeficientes para sinais amostrados a 44,1 kHz).

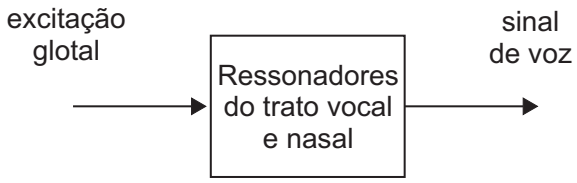


Figura 2.11: Modelagem da produção de voz por um sistema fonte-filtro.

Este tipo de representação se mostra extremamente útil para sistemas de processamento de voz. Sistemas de telefonia utilizam esse esquema para a codificação de sinais de voz com taxas muito baixas. Com esse modelo é possivel representar um janela pequena de tempo por alguns parâmetros de excitação glotal, e pelos coeficientes do modelo gerador do sinal [2].

Esse esquema básico pode ser mais elaborado, para incluir efeitos de diferentes tipos de emissão [19, 20], ou pode ter uma implementação mais simples em sistemas que devem ter alta eficiência computacional, como em codificação de sinais de voz [2].

2.5 Conclusão

Neste capítulo foram apresentados de maneira simples alguns fundamentos do processo de geração de voz, e como o entendimento deste processo pode ser

usado para gerar modelos eficientes para processamento de sinais. O resultado da Seção 2.4 é usado diretamente para gerar modelos que são usados em esquemas de análise/síntese, onde podem ser incluídos estágios de modificação e/ou compressão. Ainda foram apresentados conceitos e nomenclatura básica para facilitar a compreensão do restante do texto; entre eles está o conceito de *pitch* — cujo termo em inglês não será traduzido devido à extensão do seu conceito —, e conceitos relativos a música e canto, que podem não ser habituais para muitos pesquisadores da área de voz e processamento de sinais.

Capítulo 3

Modelagem do trato vocal

3.1 Introdução

Como visto no Capítulo 2, as características do processo de produção de voz podem ser usadas para o processamento mais eficiente deste tipo de sinal. A Seção 2.4 mostrou que a produção do sinal de voz pode ser modelada por um sistema bastante simples do tipo fonte/filtro. Entre as vantagens deste tipo de representação está sua separação em um sinal diretamente relacionado ao *pitch* e às características de sonoridade do sinal de voz e um filtro relacionado com o timbre e os fonemas emitidos. Neste capítulo, o foco reside na obtenção de modelos relacionados com o trato vocal.

Os modelos abordados nesse capítulo são: o modelo de predição linear LPC (*linear predictive coding*) obtido usando uma solução em blocos e uma solução seqüencial usando filtragem adaptativa, na Seção 3.2; o modelo LSF (*line spectral frequencies*), que deriva do modelo LPC, na Seção 3.3; o modelo baseado em transformada cepstral, na Seção 3.4; e os modelos anteriores descritos em uma escala de freqüência empenada (*frequency warping*), que pode aproximar os efeitos da percepção humana, na Seção 3.5.

Implementações dos algoritmos de processamento de sinais na escala empenada podem ser encontrados no *Toolbox* WarpTB para Matlab® [21]. Entre os algoritmos disponíveis neste *Toolbox*, estão funções de filtragem linear e modelo LPC na escala empenada (*warped*-LPC ou wLPC).

3.2 Modelo de predição linear

O problema de predição linear consiste em obter uma boa estimativa $\widehat{s}[n]$ para uma amostra de um sinal $s[n]$, a partir de uma combinação linear de suas amostras anteriores $s[n-m]$, para $m = 1, 2, \ldots, M$, ou seja,

$$\widehat{s}[n] = \sum_{m=1}^{M} a_m s[n-m], \tag{3.1}$$

onde a_m são os coeficientes de predição linear e M é a ordem do modelo. A equação (3.1) pode ser reescrita na forma matricial

$$\widehat{s}[n] = \mathbf{a}^T \mathbf{s}[n-1], \tag{3.2}$$

onde

$$\mathbf{s}[n-1] = \begin{pmatrix} s[n-1] & s[n-2] & \ldots & s[n-M] \end{pmatrix}^T \tag{3.3}$$

$$\mathbf{a} = \begin{pmatrix} a_1 & a_2 & \ldots & a_M \end{pmatrix}^T, \tag{3.4}$$

com sobrescrito T denotando transposição matricial. O erro de predição é definido como:

$$e[n] = s[n] - \widehat{s}[n], \tag{3.5}$$

de forma que este sinal pode ser entendido como o resultado da filtragem do sinal $s[n]$ por um filtro com resposta ao impulso de duração finita (*finite impulse response*, FIR) cuja função de transferência é

$$A(z) = 1 - \sum_{m=1}^{M} a_m z^{-m}. \tag{3.6}$$

Os coeficientes do modelo de predição linear a_m são obtidos de forma a minimizar o erro de predição segundo determinada função-custo. Como resultado deste processo, é obtido um conjunto de coeficientes relacionado com a parte previsível do sinal $s[n]$, enquanto o resíduo ou erro de predição $e[n]$ tende a ter comportamento puramente aleatório. Desta forma, quando for usado um número suficientemente grande de coeficientes, podemos dizer que o sinal $s[n]$ foi gerado pela filtragem de ruído branco por um filtro de resposta ao impulso de duração infinita (*infinite impulse response*, IIR), com função de transferência igual ao inverso de $A(z)$, isto é,

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{m=1}^{M} a_m z^{-m}}. \tag{3.7}$$

No caso de sinais de voz, é comum fazer com que a ordem do modelo seja pequena, por exemplo $M = 10$, para uma freqüência de amostragem de 8 kHz. Desta forma, como o filtro IIR resultante do modelo LPC possui poucos pólos, ele se torna incapaz de modelar todas as nuances do espectro do sinal de voz original $s[n]$, e podemos dizer que o modelo tende a apenas aproximar a envoltória do espectro de $s[n]$ e não seus picos individuais.

O fenômeno descrito acima pode ser facilmente observado ao se obter os modelos LPC de alta e baixa ordens para um mesmo sinal. A Figura 3.1 mostra um exemplo da magnitude do espectro de um sinal sintético, gerado pelo processo de filtragem de um sinal de excitação contendo um trem de impulsos contaminado com ruído branco por um filtro IIR de ordem 10. Ao se fazer o modelo de predição linear de ordem 10, e ao se observar a magnitude de sua resposta em freqüência, mostrada na Figura 3.2 (a), podemos constatar que o modelo com essa ordem é capaz de modelar somente a envoltória dos picos espectrais do sinal original. Já um modelo de ordem mais alta, 80 coeficientes, os picos individuais do espectro do sinal passam a ser modelados, como se vê na Figura 3.2 (c). O resultado desta simplificação é que, no caso de ordem baixa, o sinal de erro de predição resultante carrega consigo a informação de *pitch* do sinal original e assume a forma de um trem de impulsos somado a ruído branco mostrada na Figura 3.2 (b). Esse trem de impulsos tem relação com o sinal de excitação glotal mostrado na Figura 2.8.

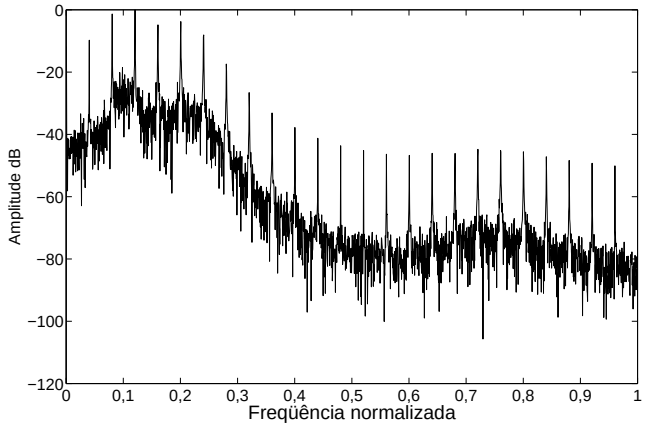


Figura 3.1: Espectro do sinal de teste.

A implicação do modelo LPC para o caso específico de sinais de voz é que os

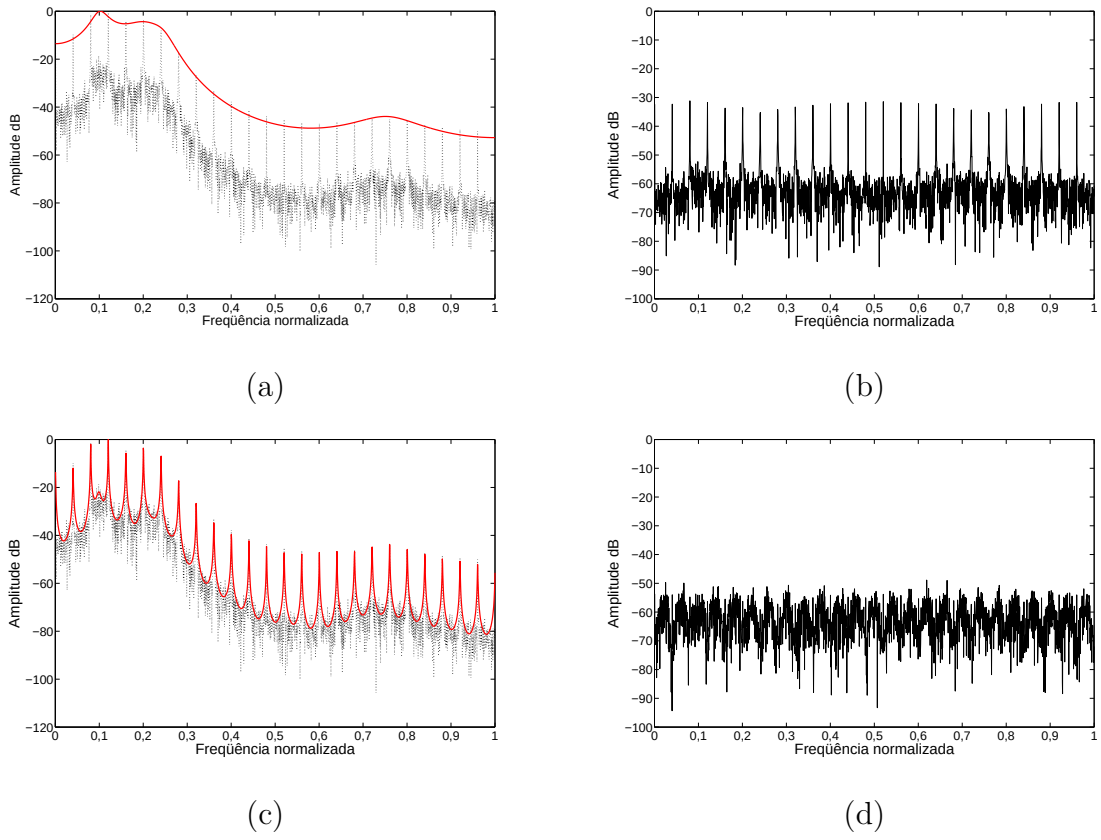


Figura 3.2: Espectro do sinal de teste e magnitude da resposta em freqüência do modelo LPC com (a) 10 e (c) 80 coeficientes; espectro do erro de predição $e[n]$ do modelo LPC com (b) 10 e (d) 80 coeficientes.

parâmetros do processo de geração destes sinais podem ser diretamente obtidos na forma de um modelo fonte/filtro, mostrado na Figura 2.11. A seguir são abordados dois tipos de solução para obtenção dos coeficientes LPC: uma solução em blocos e outra seqüencial, usando filtro adaptativo *recursive least-squares* (RLS).

3.2.1 Solução em blocos

A solução em blocos para o modelo LPC considera que o sinal a ser modelado é ergódico e estacionário no sentido amplo (*wide-sense stationary*, WSS) em um intervalo curto de tempo [2]. Usualmente essa aproximação é feita para trechos de 20 ms de sinais de voz. Desta forma, o modelo do sinal $s[n]$ é obtido em blocos $s_b[k]$ de tamanho N, tomados a cada r amostras. Para o b-ésimo bloco, o erro de predição

é dado por [22]

$$e_b[k] = s_b[k] - \widehat{s}_b[k] = s_b[k] - \mathbf{a}_b^T \mathbf{s}_b[k-1], \tag{3.8}$$

ou, na forma matricial,

$$\mathbf{e}_b = \mathbf{d}_b - \mathbf{S}_b \mathbf{a}_b, \tag{3.9}$$

onde

$$\mathbf{S}_b = \begin{pmatrix} s_b[N-1] & s_b[N-2] & \dots & s_b[N-M] \\ s_b[N-2] & s_b[N-3] & & s_b[N-1] \\ \vdots & & & \vdots \\ s_b[M] & s_b[M-1] & \dots & s_b[0] \end{pmatrix} \tag{3.10}$$

e

$$\mathbf{d}_b = \begin{pmatrix} s_b[N] & s_b[N-1] & \dots & s_b[1] \end{pmatrix}^T \tag{3.11}$$

contêm observações de $s_b[k]$.

Para obter-se a solução em blocos, pode-se construir a função-custo a seguir, dada pela norma quadrática do erro de predição:

$$\xi_b = \mathbf{e}_b^T \mathbf{e}_b = \mathbf{d}_b^T \mathbf{d}_b - 2\mathbf{a}_b^T \mathbf{S}_b^T \mathbf{d}_b + \mathbf{a}_b^T \mathbf{S}_b^T \mathbf{S}_b \mathbf{a}_b. \tag{3.12}$$

A solução será o vetor de coeficientes $\mathbf{a}_b$ que minimiza ξ_b, que pode ser encontrada fazendo-se o gradiente em relação a $\mathbf{a}_b$ da equação (3.12) igual a zero, ou seja,

$$\nabla_{\mathbf{a}_b} \xi_b = -2\mathbf{S}_b^T \mathbf{d}_b + 2\mathbf{S}_b^T \mathbf{S}_b \mathbf{a}_b = 0 \tag{3.13}$$

$$\mathbf{a}_b = \left(\mathbf{S}_b^T \mathbf{S}_b\right)^{-1} \mathbf{S}_b^T \mathbf{d}_b, \tag{3.14}$$

onde $\mathbf{S}_b^T \mathbf{S}_b$ é uma estimativa da matriz de auto-correlação do sinal $s_b[k]$, e $\mathbf{S}_b^T \mathbf{d}_b$ é uma estimativa do vetor de correlação cruzada entre o valor desejado $d_b[k] = s_b[k]$ e amostras passadas do sinal $s_b[k-l]$, para $l = 1, 2, \dots, M$.

3.2.2 Solução seqüencial

Existem alguns sistemas de processamento de voz que podem se beneficiar de um esquema seqüencial para a representação de sinais de voz [23, 24]. Entre os benefícios desse tipo de modelo está a possibilidade de processamento com pouco

atraso, e a obtenção de modelo com transições mais suaves. A solução apresentada neste texto usa um filtro adaptativo RLS (*recursive least-squares*) [25] para que seja obtido um novo modelo LPC para cada amostra do sinal analisado.

O filtro adaptativo RLS visa a minimizar a função-custo de mínimos quadrados ponderados WLS (*weigthed least-squares*)

$$\xi[n] = \sum_{i=0}^{n} \lambda^{n-i} e[i]^2 = \sum_{i=0}^{n} \lambda^{n-i} (s[i] - \hat{s}[i])^2, \tag{3.15}$$

onde $\hat{s}[n]$ é dado pela equação (3.2) e λ é o fator de esquecimento, que tem por função dar mais peso a amostras mais recentes do erro, e cujo valor deve estar na faixa $0 \ll \lambda < 1$. O vetor erro de predição para o RLS é dado na forma

$$\mathbf{e}[n] = \mathbf{d}[n] - \mathbf{S}[n-1]\mathbf{a}[n], \tag{3.16}$$

onde

$$\mathbf{S}[n-1] = \begin{pmatrix} s[n-1] & s[n-2] & \dots & s[n-M] \\ s[n-2] & s[n-3] & & s[n-M-1] \\ \vdots & & \ddots & \vdots \\ s[0] & 0 & \dots & 0 \end{pmatrix} \tag{3.17}$$

e

$$\mathbf{d}[n] = \begin{pmatrix} s[n] & s[n-1] & \dots & s[1] \end{pmatrix}^T \tag{3.18}$$

contêm observações de $s[n]$.

Desta forma, a função-custo da equação (3.15) pode ser reescrita como

$$\begin{aligned} \xi[n] &= \mathbf{e}[n]^T \Lambda[n] \mathbf{e}[n] \\ &= \mathbf{d}[n]^T \Lambda[n] \mathbf{d}[n] - 2\mathbf{a}_P[n]^T \mathbf{S}[n-1]^T \Lambda[n] \mathbf{d}[n] \\ &\quad + \mathbf{a}_P[n]^T \mathbf{S}[n-1]^T \Lambda[n] \mathbf{S}[n-1] \mathbf{a}_P[n], \end{aligned} \tag{3.19}$$

onde

$$\Lambda[n] = \begin{pmatrix} 1 & 0 & 0 & \\ 0 & \lambda & 0 & \vdots \\ 0 & 0 & \lambda^2 & \\ & \dots & & \lambda^{n-1} \end{pmatrix} \tag{3.20}$$

é uma matriz que pondera do valor da função-custo de acordo com quão antiga é uma determinada observação de $s[n]$.

A equação (3.19) é otimizada igualando-se a zero o seu gradiente, ou seja,

$$\nabla_{\mathbf{a}[n]}\xi[n] = -2\mathbf{S}[n-1]^T\Lambda[n]\mathbf{d}[n] + 2\mathbf{S}[n-1]^T\Lambda[n]\mathbf{S}[n-1]\mathbf{a}[n], \tag{3.21}$$

o que leva a

$$\begin{aligned}\mathbf{a}[n] &= \left(\mathbf{S}[n-1]^T\Lambda[n]\mathbf{S}[n-1]\right)^{-1}\mathbf{S}[n-1]^T\Lambda[n]\mathbf{d}[n] \\ &= \mathbf{R}_D[n-1]^{-1}\mathbf{p}_D[n],\end{aligned} \tag{3.22}$$

onde $\mathbf{R}_D[n-1] = \mathbf{S}[n-1]^T\Lambda[n]\mathbf{S}[n-1]$ é a matriz de auto-correlação determinística do sinal $s[n]$, e $\mathbf{p}_D[n] = \mathbf{S}[n-1]^T\Lambda[n]\mathbf{d}[n]$ é o vetor de correlação entre o valor desejado $d[n] = s[n]$ e as amostras anteriores de $s[n-k]$, para $k = 1, 2\ldots, M$.

Na prática, a equação (3.22) é computada de forma eficiente pelo cálculo recursivo de $\mathbf{R}_D[n-1]$ e $\mathbf{p}_D[n]$ de acordo com as equações [25]:

$$\mathbf{R}_D^{-1}[n-1] = \frac{1}{\lambda}\left[\mathbf{R}_D^{-1}[n-2] - \frac{\Psi[n]\Psi[n]^T}{\lambda + \Psi^T[n]\mathbf{s}[n-1]}\right], \tag{3.23}$$

$$\Psi[n] = \mathbf{R}_D^{-1}[n-2]\mathbf{s}[n-1], \tag{3.24}$$

$$\mathbf{p}_D[n] = \lambda\mathbf{p}_D[n-1] + s[n]\mathbf{s}[n-1]. \tag{3.25}$$

3.2.3 Comparação entre soluções em bloco e seqüencial

As soluções em bloco e seqüencial para o modelo LPC se relacionam diretamente. Para ilustrar melhor o fato, a equação (3.14) pode ser reescrita como

$$\mathbf{a}_b = \left(\mathbf{S}[n-1]^T\Lambda_b[n]\mathbf{S}[n-1]\right)^{-1}\mathbf{S}[n-1]^T\Lambda_b[n]\mathbf{d}[n], \tag{3.26}$$

onde

$$\Lambda_b[n] = \begin{pmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N\times N} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{pmatrix} \tag{3.27}$$

é uma matriz $n \times n$ que realiza a segmentação do sinal $s[n]$ no bloco de índice b. Desta forma, pode-se perceber que a solução seqüencial da equação (3.22) e a solução em blocos da equação (3.26) diferem somente pela matriz de pesos da solução WLS, o que pode ser visto como a aplicação de tipos diferentes de janelamento ao sinal

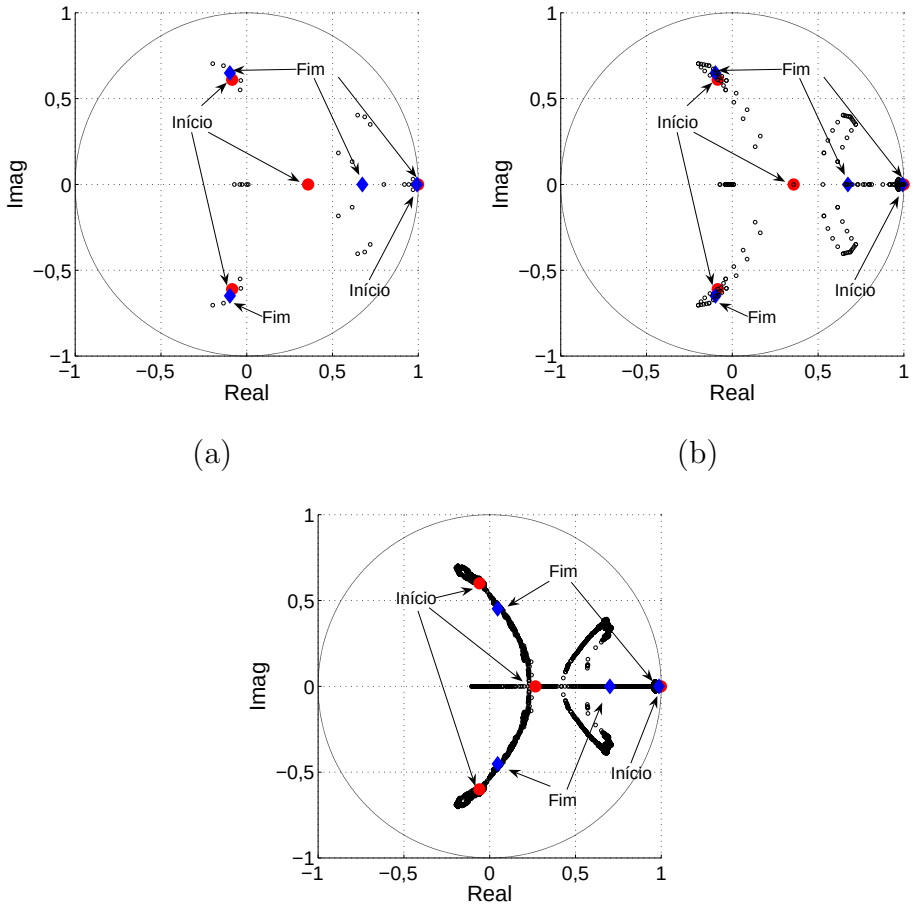
$s[n]$. No caso da solução em blocos, a matriz $\Lambda_b[n]$ pode ser entendida como uma janela retangular, enquanto que para a solução seqüencial $\Lambda[n]$ pode ser entendida como uma janela exponencial. Ambas as soluções são baseadas em uma estimativa da matriz de auto-correlação e do vetor de correlação cruzada, o que tem relação direta com a solução de Wiener [2].

A solução em blocos para o modelo LPC tem sido usada com freqüência em processamento de sinais de voz. Ela tem muitas vantagens quando se está trabalhando com codificação de sinais de voz e em sistemas que requerem uma representação compacta para esses sinais. Contudo, esse tipo de representação pode levar a alguns efeitos indesejáveis em esquemas de análise/modificação/síntese, inerentes ao processo de representação em blocos. Esses efeitos acontecem devido às descontinuidades entre blocos no estágio de síntese, que levam a defeitos audíveis.

Para contornar esses problemas, tais sistemas podem se beneficiar de um esquema seqüencial para a representação de sinais de voz. Além disso, como na abordagem seqüencial a estimativa do modelo LPC é feita para cada instante de tempo, o modelo tem transições mais suaves, o que pode levar a sinais com mais naturalidade em esquemas de análise/modificação/síntese [23].

Um exemplo comparativo entre as diversas soluções é mostrado na Figura 3.3, onde foi calculado o modelo LPC de ordem 4 para um sinal de fala com 300 ms amostrado a 44,1 kHz. Nesta figura são mostrados os pólos do modelo LPC: (a) obtido em blocos de 20 ms sem sobreposição; (b) obtido em blocos de 20 ms sem sobreposição e interpolado a cada 5 ms; (c) obtido de forma seqüencial usando RLS. As figuras mostram que o modelo obtido em blocos segue uma trajetória similar ao obtido seqüencialmente. Fica, porém, evidente que as evoluções do modelo ao longo do tempo se dão de forma muito mais suave para a solução RLS do que para a obtida em blocos, mesmo quando esta é interpolada.

A complexidade computacional para o cálculo e inversão da matriz $\mathbf{S}_b^T\mathbf{S}_b$ na obtenção dos coeficientes LPC em blocos, usando o algoritmo de Levinson-Durbin descrito em [26], é de $(M+3)(M-1)+(N-1)M$ somas, $(M+3)(M-1)+NM$ multiplicações, e $M-1$ divisões, que deve ser somada ao número de operações necessárias para calcular a auto-correlação, $M(N-1)$ somas e MN multiplicações, para cada bloco. O número de operações usadas na abordagem seqüencial, quando



(a)

(b)

(c)

Figura 3.3: Diagrama de pólos para um sinal de fala com 300 ms, onde foi usado um modelo de 4 pólos, obtido (a) em blocos de 20 ms sem sobreposição; (b) em blocos de 20 ms com coeficientes interpolados a cada 5 ms; (c) solução seqüencial.

usadas as equações (3.23), (3.24) e (3.25) para obtenção dos coeficientes LPC, é de $5M^2 + 2M$ multiplicações e $4M^2 - M$ somas para cada iteração. A complexidade computacional das duas soluções é ilutrada na Figura 3.4 para quando o modelo LPC é obtido com 10 coeficientes em blocos de 160 amostras, e 30 coeficientes em blocos de 900 amostras.

3.3 LSF

O modelo LSF (*line spectral frequencies*) é uma representação alternativa para os coeficientes LPC da equação (3.6). Este tipo de representação tem achado

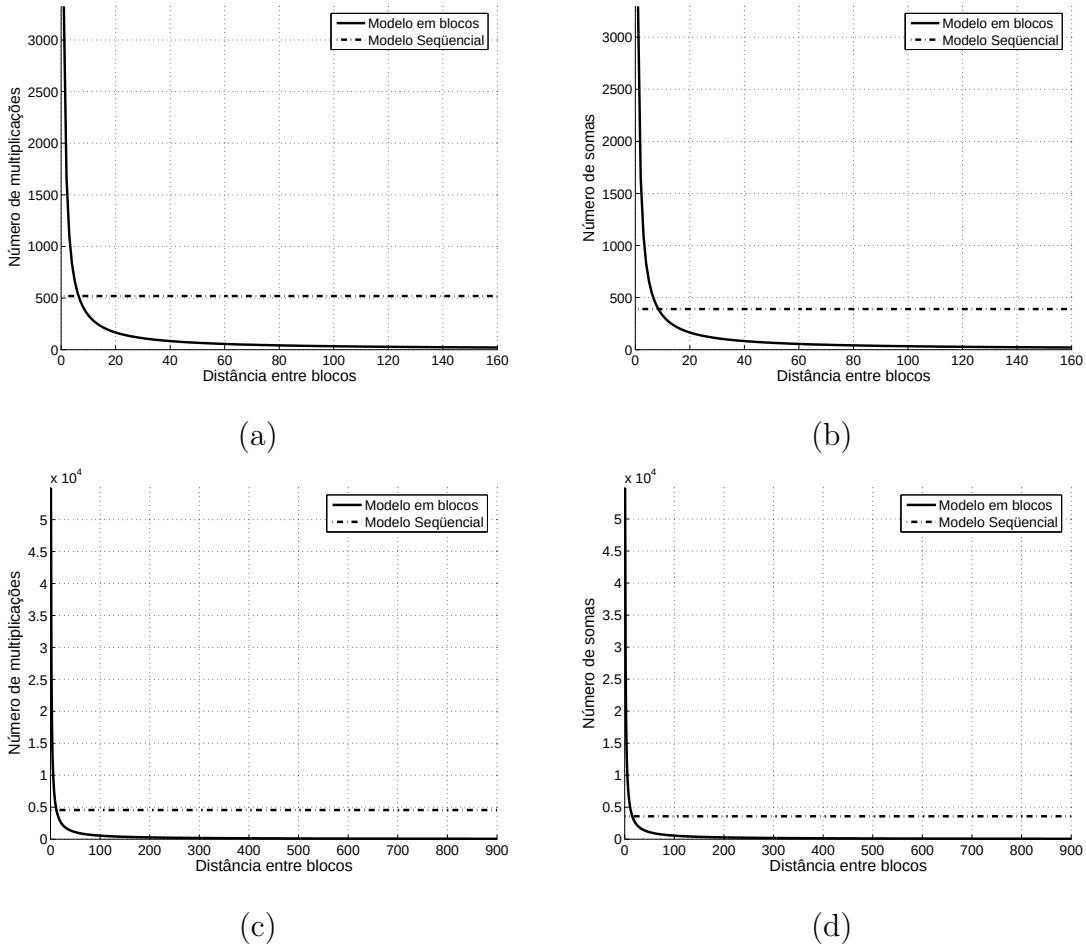


Figura 3.4: (a) e (c) Número de multiplicações e (b) e (c) número de somas por amostra necessárias para as soluções em blocos e seqüencial do modelo LPC conforme varia a distância entre blocos adjacentes, usando: (a) e (b) 10 coeficientes em janelas de 160 amostras; (c) e (d) 30 coeficientes em janelas de 900 amostras.

uso em sistemas de codificação [27] e de reconhecimento [28] de sinais de voz. Os coeficientes LSF também têm sido usados para interpolação de modelos LPC obtidos em blocos, o que vai ser feito no Capítulo 6. Para se obter os coeficientes LSF, o filtro $A(z)$ do modelo LPC é decomposto em dois polinômios

$$P(z) = A(z) + z^{-(M+1)}A(z^{-1}) \tag{3.28}$$

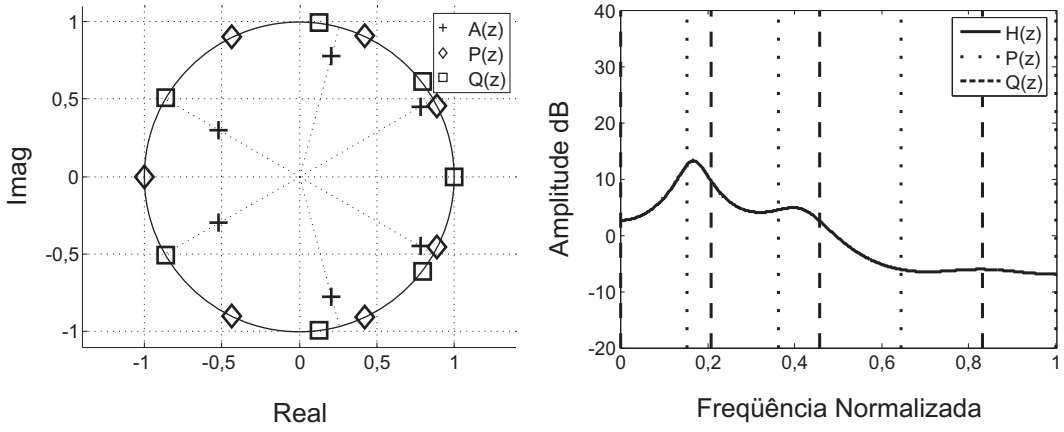
e

$$Q(z) = A(z) - z^{-(M+1)}A(z^{-1}), \tag{3.29}$$

de forma que

$$A(z) = \frac{P(z) + Q(z)}{2}, \tag{3.30}$$

onde os polinômios $P(z)$ e $Q(z)$ possuem todas as raízes com módulo igual a 1, aparecendo intercaladas na círcunferência de raio unitário quando as raízes de $A(z)$ têm módulo menor que 1, sendo que sempre existem uma raiz de $Q(z)$ em $z = 1$ e outra de $P(z)$ em $z = -1$. Os coeficientes LSF são, então, dados pelos ângulos das raízes de $P(z)$ e $Q(z)$, uma vez que com essa informação se pode reconstruir $A(z)$ de forma perfeita. Um exemplo é mostrado na Figura 3.5, que ilustra (a) o lugar das raízes de $H(z)$, $P(z)$ e $Q(z)$; (b) o módulo da resposta em freqüência de $H(z)$ e a posição angular dos coeficientes LSF, para $H(z)$ de ordem 6 com pólos em $0{,}9\angle\pm\frac{\pi}{6}$, $0{,}8\angle\pm\frac{5\pi}{12}$ e $0{,}5\angle\pm\frac{5\pi}{6}$. Com essa figura pode-se perceber que um par de coeficientes LSF se aproxima dos pontos de ressonância pronunciada de $H(z)$. Por esse motivo, é comum o uso dos coeficientes LSF para a estimativa da freqüência dos formantes de sinais de voz [2].



(a) (b)

Figura 3.5: (a) Lugar das raízes dos polinômios $A(z)$, $P(z)$ e $Q(z)$; (b) Módulo da resposta em freqüência do filtro IIR $H(z)$ e posição angular dos zeros de $P(z)$ e $Q(z)$ para 6 pólos.

O uso dos coeficientes LSF tem sido muito difundido por serem eles mais robustos à quantização que os coeficientes LPC [2, 27]. Ainda, como eles se relacionam com a posição dos formantes, apresentam vantagens em sistemas de reconhecimento de voz [28]. Além disso, os coeficientes LSF têm boas propriedades de interpolação, e são usados em codificadores de voz para suavizar as transições entre modelos LPC de blocos adjacentes.

3.4 Transformada *cepstral*

Métodos de análise cepstral têm encontrado diversas aplicações em processamento de sinais desde a sua publicação. Entre as aplicações estão análise de eco em sinais sísmicos [29], detecção de *pitch* [30], deconvolução e processamento homomórfico de sinais [31]. Adicionalmente, os coeficientes oriundos da análise cepstral têm se mostrado bastante eficientes para reconhecimento de voz [2, 32].

Sendo um sinal gerado pela passagem de um sinal de excitação por um filtro com resposta ao impulso $h[n]$, temos que a transformada de Fourier do sinal resultante é o resultado da multiplicação da transformada de Fourier do sinal pela transformada de Fourier de $h[n]$:

$$s\left[n\right] = \left(e * h\right)\left[n\right] \tag{3.31}$$

$$\mathcal{F}\left(s\right) = \mathcal{F}\left(e * h\right) = \mathcal{F}\left(e\right)\mathcal{F}\left(h\right) \tag{3.32}$$

$$S\left(\omega\right) = E\left(\omega\right)H\left(\omega\right). \tag{3.33}$$

Ao aplicar a função logarítmica à equação (3.33), temos o resultado da convolução na forma de uma soma:

$$\log S\left(\omega\right) = \log E\left(\omega\right) + \log H\left(\omega\right). \tag{3.34}$$

No caso de sinais de voz, podemos considerar a seqüência $s\left[n\right]$ como sendo gerada pela convolução de um sinal de excitação $e\left[n\right]$ por um filtro referente ao trato vocal com resposta ao impulso $h\left[n\right]$. Como visto nas seções anteriores, o espectro do sinal de excitação para trechos vozeados é periódico e tem uma forma de um trem de impulsos — logo varia rapidamente em ω — enquanto o espectro do filtro do trato vocal tem a forma de uma curva suave. Isso indica que caso fosse feita uma análise *'espectral'* do espectro do sinal de voz, existiriam componentes de *'baixas freqüências'* relativas à influência do trato vocal (que varia lentamente em ω) e componentes de *'altas freqüências'* relativas ao sinal de excitação (que varia rapidamente em ω) [2]. Na prática é usada a transformada de Fourier inversa do espectro para calcular o cepstro de um sinal

$$\mathcal{F}^{-1}\left(\log S\left(\omega\right)\right) = \mathcal{F}^{-1}\left(\log E\left(\omega\right)\right) + \mathcal{F}^{-1}\left(\log H\left(\omega\right)\right). \tag{3.35}$$

Essa é a idéia da análise cepstral, uma análise *'espectral'* do espectro do sinal. Por ser

um tipo de análise da freqüência às avessas, as sílabas do termo *freqüência* aparecem trocadas na nomenclatura do cepstro, que é então chamada *qüefrência*.

Os coeficientes cepstrais são definidos de acordo com a equação (3.36). Na prática eles podem ser calculados usando a transformada rápida de Fourier (*fast Fourier transform*, FFT) para blocos do sinal de voz. A Figura 3.6 mostra o procedimento para o cálculo do cepstro usando FFT.

$$c = \mathcal{F}^{-1}\left(\log S\left(\omega\right)\right). \tag{3.36}$$

(a)

(b)

Figura 3.6: Implementação da transformada cepstral em blocos, (a) transformada direta e (b) transformada inversa.

A transformada cepstral mostrada na Figura 3.6 apresenta alguns detalhes que devem ser comentados. Uma vez que o resultado da FFT é complexo, é necessário definir a operação de logaritmo complexo. Em casos onde não é necessária a reconstrução do sinal, o procedimento da Figura 3.6 é simplificado de forma que somente o módulo do espectro do sinal é analisado. Os coeficientes oriundos deste tipo de análise são chamados de cepstro real. Essa simplificação é útil no caso de sistemas de detecção de *pitch*. Contudo, no caso em que se deseja realizar transformação ou filtragem no domínio do cepstro, é necessário definir uma transformada com reconstrução perfeita. Isso pode ser feito aproveitando-se o fato de o logaritmo de um número complexo com módulo A e fase ϕ ser

$$\log_e Ae^{j\phi} \;\; = \;\; \log_e A + j\phi. \tag{3.37}$$

Assim, é possível separar o cálculo dos coeficientes cepstrais em uma parte relativa ao módulo e outra relativa à fase da FFT.

Em relação à fase, é comum limitar seu valor ao intervalo $-\pi < \phi \leq \pi$. Essa restrição causa descontinuidades em $j\phi$, e força o aparecimento de altas qüefrências no resultado do cepstro, o que pode ser resolvido usando-se algoritmos de desdobramento da fase (*phase unwraping*). Uma discussão completa deste assunto é encontrada em [31].

3.4.1 Relação entre coeficientes cepstrais e LPC

Os problemas relativos à implementação da transformada cepstral tornam interessante relacionar os coeficientes LPC com os coeficientes cepstrais. Esse tipo de relação é interessante também porque os coeficientes cepstrais calculados desta forma não levam consigo tanta influência do sinal de excitação da voz (com exceção da influência do decaimento espectral relativo à forma do pulso glotal apresentado na Seção 2.3.2).

Uma forma de calcular os coeficientes cepstrais a partir dos coeficientes LPC é usando os seus pólos, uma vez que

$$\begin{aligned} \log H(z) &= \log\left(\frac{A}{\prod_{m=1}^{M}(1 - z_m z^{-m})}\right) \\ &= \log A - \sum_{m=1}^{M} \log\left(1 - z_m z^{-m}\right), \end{aligned} \tag{3.38}$$

onde z_m são os pólos de $H(z)$. Considerando que a transformada de Fourier pode ser calculada analisando-se a transformada Z na circunferência de raio unitário $|z| = 1$, e considerando que os pólos de $H(z)$ são de fase mínima, podemos calcular os coeficientes cepstrais como [31]:

$$c[n] = \begin{cases} \log A, & n = 0, \\ \sum_{m=1}^{M} \frac{z_m^n}{n}, & n > 0. \end{cases} \tag{3.39}$$

Para calcular os coeficientes LPC a partir dos coeficientes cepstrais, basta calcular a auto-correlação, e usar a solução para os coeficientes LPC em blocos da equação (3.14):

$$\mathbf{a}_b = \mathbf{R}^{-1}\mathbf{p}, \tag{3.40}$$

onde a função de auto-correlação, que compõe a matriz de auto-correlação

$$\mathbf{R} = \begin{bmatrix} r[0] & r[1] & \dots & r[M-1] \\ r[1] & r[0] & & r[M-2] \\ \vdots & & \ddots & \\ r[M-1] & r[M-2] & \dots & r[0] \end{bmatrix} \tag{3.41}$$

e o vetor de correlação cruzada

$$\mathbf{p} = \begin{bmatrix} r[1] & r[2] & \dots & r[M] \end{bmatrix}^T, \tag{3.42}$$

é definida a partir dos coeficientes cepstrais $c[n]$ da forma:

$$\mathbf{r} \;=\; \mathrm{FFT}^{-1}\left(|e^{\mathrm{FFT}(\mathbf{c})}|^2\right); \tag{3.43}$$

aqui, $\mathbf{r}$ é um vetor com a função de auto-correlação $r[k]$, para $k = 0, \dots, N$, e $\mathbf{c}$ é um vetor com os coeficientes cepstrais. Em uma implementação rápida, a matriz $\mathbf{R}$ pode ser invertida usando-se o algoritmo de Levinson-Durbin [26].

3.5 Modelos usando escala de freqüência empenada

Muito da tecnologia atual de áudio e voz leva em conta características do sistema auditivo humano. Dentre as características mais usadas está o uso de escalas perceptivas. Estas escalas normalmente são aproximadas por meio de funções não-lineares da escala linear em hertz, e apresentam uma resolução melhor em baixas freqüências do que em altas. Exemplos de escalas perceptivas são as escalas mel [2, 33], bark [2, 33, 34, 35] e ERB (*equivalent rectangular bandwidth*) [34, 35].

Uma maneira simples de levar em conta esse tipo de modelo é pelo uso de técnicas de processamento usando uma escala de freqüência empenada, *frequency-warped signal processing* [36]. Esse tipo de técnica faz o mapeamento de filtros e transformadas por meio da transformação bilinear

$$\hat{z}^{-1} = D(z) = \frac{z^{-1} - \rho}{1 - \rho z^{-1}}, \tag{3.44}$$

que pode ser facilmente invertida usando-se

$$z^{-1} = \frac{\hat{z}^{-1} + \rho}{1 + \rho \hat{z}^{-1}}, \tag{3.45}$$

onde ρ é o fator de empenamento na freqüência. O filtro que implementa a transformação bilinear é mostrado na Figura 3.7.

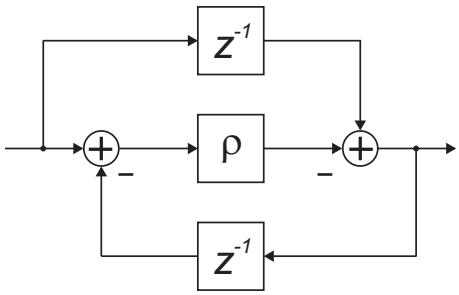
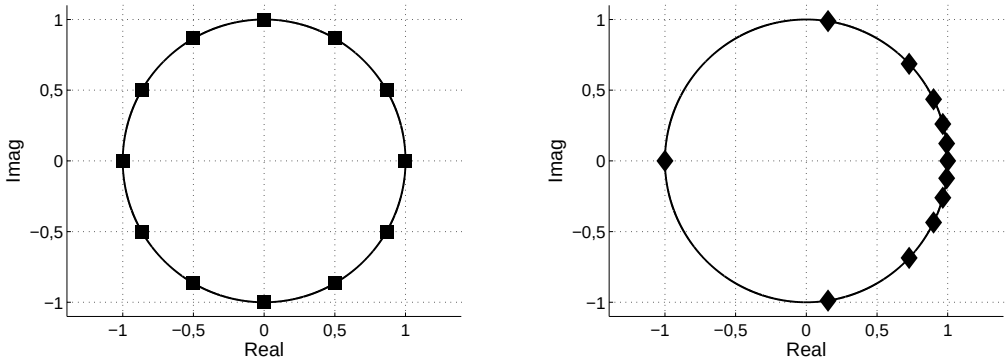


Figura 3.7: Filtro passa-tudo da transformação bilinear da equação (3.44).

Esse tipo de técnica tem sido bastante difundido, e encontra-se na literatura o uso de transformadas e bancos de filtros empenados [37, 38, 39, 40, 41, 42, 43], o projeto de filtros em escalas perceptivas [36, 44], e a sua aplicação associada à obtenção do modelo LPC [45, 46, 47].

A Figura 3.8 (a) mostra a posição de pólos igualmente espaçados no círculo de raio unitário no domínio da freqüência empenada $\hat{z}$, e a Figura 3.8(b) mostra a posição destes pólos no domínio da freqüência linear z. É importante notar que o mapeamento dado pelas equações (3.44) e (3.45) não altera a estabilidade dos filtros, uma vez que pólos estáveis em $\hat{z}$ dão origem a pólos estáveis em z e vice-versa.



(a) (b)

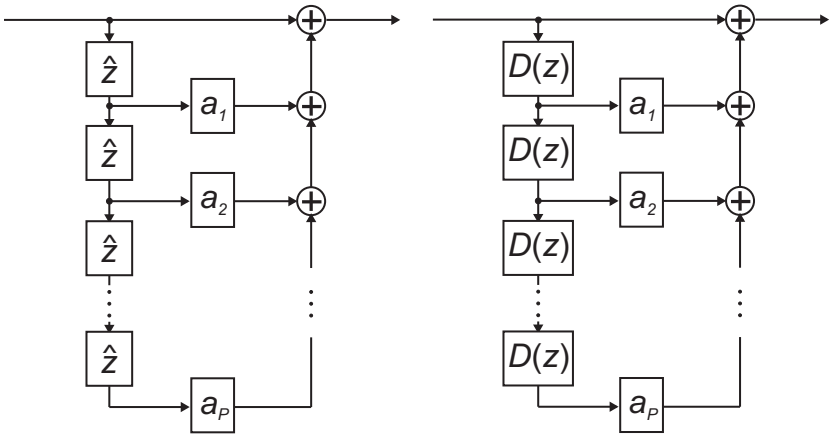
Figura 3.8: Mapeamento do (a) plano empenado $\hat{z}$ no (b) plano z, quando o fator de distorção da freqüência é $\rho = 0{,}6267$.

O significado do mapeamento bilinear apresentado na equação (3.44) é que

é possível projetar um filtro em um domínio de freqüências empenadas $\hat{z} = \mathrm{e}^{-j\hat{\omega}} = \mathrm{e}^{-jA(\omega)}$, onde

$$A(\omega) = \arg(D(z)) \tag{3.46}$$

é uma função que determina o mapeamento entre a escala de freqüência empenada $\hat{\omega}$ e a escala de freqüência linear, ou em radianos por amostra, ω, e implementar o filtro usando a transformação bilinear da equação (3.44), o que equivale a trocar os blocos de atraso unitário pelo filtro da Figura 3.7. A Figura 3.9 mostra a representação de um filtro FIR cujos coeficientes a_k, $k = 1, \ldots, p$, foram projetados no domínio das freqüências empenadas, wFIR, e o modo como é feita a sua implementação prática.



(a) (b)

Figura 3.9: (a) Filtro projetado no domínio de freqüências empenadas que possui somente zeros; (b) Implementação de um filtro usando a transformação bilinear.

3.5.1 Fator de empenamento

O fator de empenamento ótimo, que aproxima determinada escala de freqüência, pode ser obtido de acordo com a rotina de otimização demonstrada por Abel e Smith [35]:

$$\hat{\rho} = \frac{\mathbf{s}^T\mathbf{V}\mathbf{d}}{\mathbf{s}^T\mathbf{V}\mathbf{s}}, \tag{3.47}$$

onde

$$\mathbf{s}(k) = \text{sen}\left(\frac{b(\omega_k)+\omega_k}{2}\right), \tag{3.48}$$

$$\mathbf{d}(k) = \text{sen}\left(\frac{b(\omega_k)-\omega_k}{2}\right) \tag{3.49}$$

e $\mathbf{V}$ é uma matrix diagonal, que faz a ponderação da solução WLS da equação (3.47), cujos elementos são obtidos fazendo-se

$$v(k) = \frac{1}{1+\rho^2-2\rho\cos(\omega_k)}. \tag{3.50}$$

Na primeira iteração do algoritmo, $\mathbf{V}$ é inicializada com uma matriz identidade, nos próximos passos o valor obtido em (3.47) é usado em (3.50).

Foram obtidos valores de ρ ótimos para diversas taxas de amostragem, para aproximar a escala mel, originando a aproximação dada pela equação abaixo [33]:

$$\text{Pitch(mel)} = 1127{,}0148\log\left(1+\frac{f(\text{Hz})}{700}\right). \tag{3.51}$$

Com esses valores foi aproximada uma fórmula para obtenção do fator ρ usando a ferramenta *curve fitting tool* do Matlab®:

$$\hat{\rho}_{\text{mel}}\left(f_s\right) = -4{,}014{f_s}^{-0{,}095} + 2,08. \tag{3.52}$$

3.5.2 Modelo de predição linear

A obtenção dos coeficientes LPC da escala empenada (wLPC) pode ser feita usando os coeficientes de auto-correlação calculados de acordo com a Figura 3.10 [46, 47]. Uma vez calculados os coeficientes de auto-correlação, basta aplicar o algoritmo de Levinson-Durbin [26] para obter os coeficientes LPC.

Para aplicar o filtro inverso do modelo wLPC, é necessária uma modificação no filtro,

$$H(\hat{z}) = \frac{1}{1-\sum_{m=1}^{M} a_m \hat{z}^{-m}}, \tag{3.53}$$

uma vez que, quando aplicada a transformação bilinear $\hat{z} = D\left(z\right)$, o filtro resultante possui uma recursão com atraso zero, que o torna não implementável. As soluções para a implementação de filtros recursivos com escala de freqüência empenada são discutidas em [44].

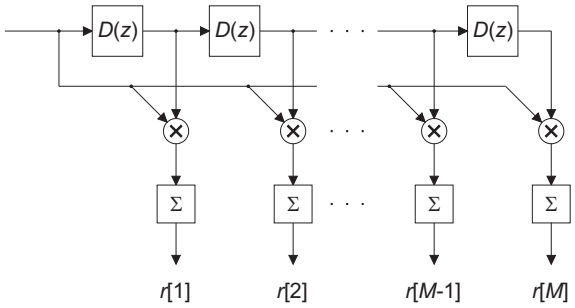


Figura 3.10: Cálculo dos coeficientes de auto-correlação usando escala de freqüência empenada.

3.5.3 Transformada *mel-cepstral*

A transformada cepstral pode ser definida para escalas perceptivas usando técnicas de processamento de sinais em escalas empenadas. Grande parte dos trabalhos que usam o cepstro usam também a escala mel, com implementação que não possui reconstrução perfeita. Na sua implementação mais simples, os coeficientes mel-cepstrais são calculados fazendo o mapeamento do resultado da primeira FFT na escala mel. Este procedimento é ilustrado na Figura 3.11.

Figura 3.11: Cálculo dos coeficientes mel-cepstrais [2].

Alternativas para o método da Figura 3.11 incluem a substituição da transformada de Fourier por uma transformada de Fourier usando escala de freqüência empenada, descrita nos artigos [40, 41, 48]. Contudo, a matriz usada para fazer o cálculo da transformada de Fourier de tempo discreto se aproxima da singularidade, por isso podem ocorrer problemas ao fazer o cálculo da transformada inversa.

Para o cálculo dos coeficientes mel-cepstrais, podemos calcular os coeficientes wLPC usando o método descrito na Seção 3.5.2 (com ρ escolhido para aproximar a escala mel da equação (3.52)), e aplicar a relação descrita na Seção 3.4.1. Outro método é descrito em [49], onde os coeficientes LPC são primeiro calculados, em um segundo passo é aplicada a transformação bilinear para distorcer o modelo para a escala mel, e no passo final os coeficientes mel-cepstrais são calculados a partir do

modelo ARMA (auto-regressivo com média móvel) obtido. A vantagem do método anterior é que o modelo LPC já é obtido em uma escala perceptiva, sendo assim seus pólos distribuídos de acordo com uma resolução que se aproxima da resolução da audição humana para freqüências, ao contrário do que acontece no método descrito em [49].

3.6 Conclusão

Este capítulo apresentou detalhes de modelos do trato vocal que são usados para a implementação das técnicas descritas no restante deste texto. A discussão sobre o modelo LPC na Seção 3.2 apresentou este tipo de modelo, enfatizando seu significado na análise de sinais de voz. Técnicas para a obtenção do modelo LPC foram apresentadas, e são usadas para modificação de *pitch* no Capítulo 5. Um estudo comparativo entre duas soluções para o modelo LPC mostra as vantagens do modelo seqüencial para sistemas de análise/modificação/síntese, uma vez que possui transições suaves entre modelos e evita descontinuidades, e do modelo em blocos, que possibilita uma representação compacta importante para compressão de voz. Uma representação alternativa dos coeficientes LPC é apresentada na Seção 3.3, e será usada no Capítulo 6 para interpolação do modelo LPC na síntese de sinais modificados. Ainda é apresentada na Seção 3.4 a transformada cepstral, que tem sido muito usada em sistemas de reconhecimento de voz e detecção de *pitch*, e cujas aplicações muitas vezes aparecem combinadas com o uso da escala mel. Ainda é apresentada na Seção 3.5 uma visão alternativa dos modelos apresentados anteriormente, com o uso de técnicas de processamento na escala de freqüência empenada (*frequency warped signal processing*). Com o uso deste tipo de técnica é possível redefinir os modelos de voz de forma a levar em conta aspectos da percepção humana. Desta forma é possível obter um modelo de predição linear na escala empenada wLPC, e com ele fazer uma implementação da transformada mel-cepstral. A transformada mel-cepstral como é definida na Seção 3.5.3 será usada para fazer a transformação de locutor no Capítulo 6.

Capítulo 4

Discriminação de trechos sonoros e surdos em sinais de voz

4.1 Indrodução

Como apresentado no Capítulo 2, os fonemas utilizados na produção de voz podem ser classificados em sonoros ou surdos de acordo com a vibração ou não, respectivamente, das pregas vocais. Muitas técnicas de processamento de sinais apresentadas neste trabalho fazem uso de um modelo físico que leva em consideração o processo de geração apresentado na Seção 2.4. Portanto, para que esse modelo possa ser bem aproveitado é necessário que se faça a distinção entre trechos do sinal de voz contendo silêncio, fonemas sonoros e fonemas surdos, que serão tratados de forma diferenciada pelo processamento subseqüente.

Para a distinção entre fonemas sonoros e surdos, será feita uma análise de parâmetros que podem ser facilmente extraídos de sinais de voz, e que são indicados na literatura do assunto como sendo relevantes para essa discriminação. Entre essas características a serem extraídas do sinal estão a energia do sinal, a taxa de cruzamentos por zero [2], os coeficientes de auto-correlação [50], os coeficientes LPC e energia do erro de predição [51]. Entre outras técnicas para discriminação de sonoridade estão o uso de modelo harmônico [52] e de *wavelets* [53].

Depois de feita uma análise teórica de cada uma dessas características, uma análise estatística é apresentada para destacar a relevância de cada parâmetro, tanto isoladamente como quando relacionado a outros parâmetros.

Ao final do capítulo será feito o projeto de um classificador de sinais sonoros, surdos e de silêncio, levando em conta as características de cada variável analisada nas seções anteriores, e com ênfase na detecção de trechos sonoros.

4.2 Energia do sinal

Uma das primeiras características que podem ser observadas para a discriminação entre sons sonoros e surdos de sinais de voz é a energia do sinal. Ao se observar gráficos de sinais de voz, percebem-se trechos de baixa energia e sem periodicidade aparente, trechos de silêncio e trechos de maior amplitude em que se pode observar uma forte periodicidade. Esses trechos periódicos e de maior energia são trechos sonoros, e sua discriminação com relação aos trechos surdos e de silêncio é que tem maior importância neste trabalho.

A energia do bloco k, sendo $x_k[n]$ a n-ésima amostra do bloco k, pode ser calculada como [51]:

$$E_k = \sum_{n=0}^{N-1} x_k[n]^2, \tag{4.1}$$

onde N é o número de amostras do bloco.

Em decibéis, a energia fica como:

$$E_{k\text{dB}} = 10 \log_{10}(\epsilon + E_k) \quad \text{dB}, \tag{4.2}$$

onde ϵ é uma constante de valor pequeno para evitar que se faça o logaritmo de zero.

Apesar de a energia do sinal ser um bom indicativo para a discriminação entre techos sonoros e surdos, ela pode ser uma estimativa não muito robusta. Isso é devido ao fato de a energia do sinal de voz variar de acordo com diferentes ganhos de microfones e amplificadores ou diferenças entre circuitos de A/D, isso sem se considerar, no caso de voz cantada, os recursos usados para obter expressividade pelo artista, que pode cantar forte ou fraco, assim como modificar sua distância ao microfone. Pode-se tornar a energia um parâmetro mais robusto para a discriminação entre sons sonoros e surdos, normalizando-a por uma estimativa da energia média entre blocos adjacentes:

$$E_{k\text{dB}} = 10 \ \log_{10}\left(\epsilon + \frac{E_k}{E_{\text{av}}}\right) \quad \text{dB}, \tag{4.3}$$

onde

$$E_{\text{av}} = E_{\text{av}}\,(1-a) + E_k\ a \tag{4.4}$$

é a energia média entre blocos adjacentes, sendo $0 < a < 1$ um fator de esquecimento. A energia normalizada tem como vantagem sobre a estimativa dada por (4.2) o fato de funcionar de forma adaptativa, adequando-se de forma automática ao sinal que está sendo analisado.

A complexidade computacional dos algoritmos de extração das características baseadas na energia do sinal será mostrada mais adiante na Tabela 4.1.

4.3 Taxa de cruzamentos por zero

A taxa de cruzamentos por zero é um dos métodos mais populares para a discriminação entre sons sonoros e surdos [2]. Ela é calculada de forma bastante simples; para o bloco k, a taxa de cruzamentos por zero do sinal com amostras $x_k[n]$ é

$$ZC_k = \frac{1}{N}\sum_{n=0}^{N-1}\frac{|\text{sinal}(x_k[n]) - \text{sinal}(x_k[n-1])|}{2}, \tag{4.5}$$

onde

$$\text{sinal}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0. \end{cases} \tag{4.6}$$

Ao se observar um sinal senoidal, pode-se perceber que ele cruza zero 2 vezes por período. Sinais periódicos com maior riqueza espectral, como é o caso de sinais de voz, podem cruzar o zero mais do que 2 vezes por período, contudo existe uma tendência a que isso não aconteça com uma freqüência tão grande quanto acontece com sinais aleatórios. Essa característica pode ser evidenciada no exemplo ilustrado na Figura 4.1. Nesta figura, a taxa de cruzamentos por zero foi calculada para um sinal composto de uma senóide contaminada com ruído branco aditivo gaussiano. No caso de o ruído ter amplitude zero, a tendência do sinal seria ter 2 cruzamentos por zero a cada período. À medida que a energia do ruído aumenta com relação à energia da senóide, a tendência é que aumente a taxa de cruzamento por zeros.

O exemplo da Figura 4.1 ilustra como a taxa de cruzamento por zeros pode ser usada para discriminação entre sons sonoros e surdos; contudo, ela também

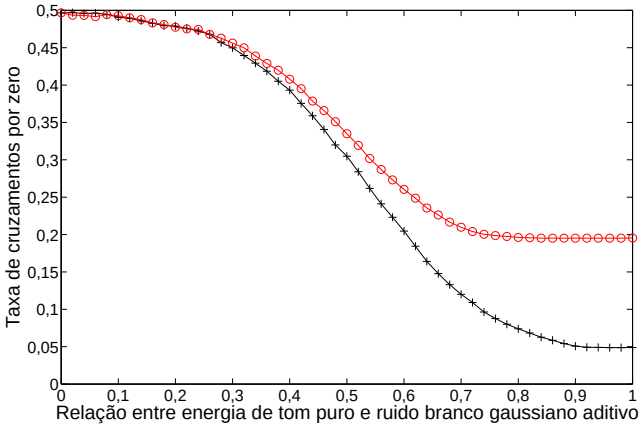


Figura 4.1: Curva da taxa de cruzamentos por zero pela variação da energia de um tom puro em relação à energia de um sinal de ruído branco. A curva com marcas '+' foi gerada com um tom de período de 200 amostras, e a com marcas 'o' com um tom de período de 800 amostras.

mostra que esse tipo de métrica é sensível a ruído de fundo e a variações no *pitch* do sinal analisado.

4.4 Características baseadas na auto-correlação

Além da energia do sinal, a auto-correlação pode dar um bom indicativo para a discriminação entre sons sonoros e surdos. O motivo disto é a característica 'passa-baixas' dos sons sonoros e 'passa-altas' dos sons surdos. A característica 'passa-baixas' nos diz que o sinal vozeado não varia muito com relação às suas amostras adjacentes, de forma que os primeiros coeficientes de auto-correlação tendem a ser significativamente maiores que os demais. A auto-correlação para uma distância τ pode ser calculada como:

$$r_\tau = \sum_{n=0}^{N-1} x_k[n] x_k[n-\tau]. \tag{4.7}$$

Segundo [50], grande parte da energia dos sinais sonoros está compreendida abaixo dos 1500 Hz; assim, podemos esperar que as auto-correlações para valores de τ menor do que metade do período de 1500 Hz (3 amostras para 8 kHz e 15 amostras para 44,1 kHz de taxa de amostragem) sejam maiores que zero. Assim,

para a discriminação sonoro/surdo/silêncio, pode-se utilizar a medida:

$$Ac_{P,k} = \frac{\sum_{p=1}^{P} r_p}{r_0}, \tag{4.8}$$

que utiliza uma média normalizada das auto-correlações com distâncias de $p = 1$ a $p = P$.

Outra métrica que pode ser aliada a sistemas detecção de *pitch* é comparar a auto-correlação correspondente ao período de *pitch* do trecho analisado com o componente com atraso zero r_0. Este coeficiente pode ser encontrado pela busca do valor máximo de auto-correlação num intervalo admissível para o período de *pitch*.

A complexidade computacional do cálculo de coeficientes de auto-correlação está ilustrada mais adiante na Tabela 4.1. Nela é feita uma comparação entre o cálculo dos coeficientes de auto-correlação individuais e o método rápido com $\tau = 1, \ldots, N$ usando FFT. O método rápido para o cálculo dos coeficientes de auto-correlação é ilustrado na equação a seguir:

$$r = \text{iFFT}\left(\left|\text{FFT}\left(x_k\right)\right|^2\right). \tag{4.9}$$

4.5 Predição linear

Os coeficientes de predição linear (LPC) de um sinal de áudio e o modo como eles aproximam a envoltória do espectro do sinal modelado foram apresentados no Capítulo 3. Como já foi dito na Seção 4.4, os sinais sonoros têm concentração de energia em baixas feqüências e os sinais surdos, em altas. Uma vez que estes sinais podem ser discriminados pela faixa de freqüências onde a energia se concentra, também os pólos do modelo LPC vão se concentrar em faixas que permitem a discriminação entre sons sonoros e surdos. Pela aplicação de um sistema treinado, e usando uma métrica para verificar a distância entre os coeficientes LPC, é possível fazer essa discriminação [54]. Uma maneira simples de aplicar esse conceito seria usar o primeiro coeficiente LPC [51].

Outra característica a ser obtida da predição linear é a energia do erro de predição. Ela pode ser entendida como uma medida da não-uniformidade do espectro do sinal modelado. Como sinais sonoros normalmente têm uma estrutura de formantes bem definida, eles acabam por gerar um erro de predição menor quando comparados ao sons surdos [51].

A discussão sobre a complexidade computacional da obtenção dos coeficientes LPC foi feita no Capítulo 3.

4.6 Estimativa do chão de ruído

Uma característica bastante marcante nas partes periódicas de sinais de áudio de alta qualidade é a diferença entre os picos harmônicos que formam a parte periódica do sinal com relação ao chão de ruído, que pode ser gerado tanto pelo ruído causado pela respiração como pelo ruído de fundo. Uma forma de usar essa informação para discriminação entre sons sonoros e surdos é fazer uma estimativa do chão de ruído [55], e comparar esta estimativa com o máximo pico do espectral do bloco analisado.

A estimativa do chão de ruído para um bloco k do sinal $x[n]$ pode ser feita seguindo os passos abaixo [55]:

1. Cálculo do módulo da DFT do bloco:

$$X_k[l] = \left| \sum_{n=0}^{N-1} x_k[n] e^{\frac{-2j\pi l}{N}} \right|. \tag{4.10}$$

2. Remoção de zeros eventuais usando um filtro de média móvel de 3 coeficientes:

$$X'_k[l] = \frac{X_k[l-1] + X_k[l] + X_k[l+1]}{3}. \tag{4.11}$$

3. Cálculo do inverso de X'_k:

$$R_k[l] = \frac{1}{X'_k[l]}. \tag{4.12}$$

4. Uso de um filtro de média móvel de N_f amostras para suavizar R_k e obter o inverso da estimativa do chão de ruído:

$$R'_k[l] = \frac{1}{N_f} \sum_{c=-\frac{N_f}{2}}^{\frac{N_f}{2}+1} R_k[l-c]. \tag{4.13}$$

5. Cálculo do inverso de R'_k:

$$X_k^{SS}[l] = \frac{1}{R'_k[l]}. \tag{4.14}$$

Desta forma, X_k^{SS} é uma estimativa do espectro estocástico no bloco x_k. O motivo para trabalhar com o inverso da transformada de Fourier do sinal é evitar que os valores elevados dos picos da parte harmônica do sinal influenciem muito na estimativa. Desta forma, um bom indicativo para a discriminação entre sons sonoros e surdos é a medida da diferença entre o chão de ruído X_k^{SS} e os picos de X_k. Uma maneira simples de se aplicar este método é comparar os dois espectros no ponto de máximo de X_k, normalizando pela energia do bloco da equação (4.1), ou seja:

$$d_k = \frac{X_k[i] - X_k^{SS}[i]}{E_k}, \quad i \mid X_k[i] = \max_l \left(X_k[l]\right). \tag{4.15}$$

A complexidade da estimativa do espectro estocástico pode ser reduzida calculando-se $X_k^{SS}[i]$ somente para o valor de i que maximiza $X_k[i]$.

A Tabela 4.1 compara a complexidade computacional envolvida no cálculo dos diversos parâmetros discutidos anteriormente.

Tabela 4.1: Complexidade computacional para extração de cada parâmetro.

Parâmetro	Somas	Multiplicações/divisões	$\log(x)$	FFT†
E	$N+2$	$N+1$	1	0
EM	$N+3$	$N+3$	1	0
Zc	N	2	0	0
r_τ	$N-1$	$N+1$	0	0
r_τ, $\tau = 1 \ldots N$	0	N	0	2
X_k^{SS}	$(3+N_f)\,N$	$(5+N_f)\,N+1$	0	1

† O algoritmo de raiz 2 que implementa a FFT utiliza $N \log_2 N$ somas e $\frac{N}{2}\log_2 N - \frac{3}{2}N + 2$ multiplicações complexas [56].

4.7 Avaliação dos parâmetros para discriminação

Para se testar os parâmetros para discriminação sonoro/surdo/silêncio, foi usada uma base de dados com 10 frases foneticamente balanceadas [57] faladas por uma pessoa do sexo masculino e uma do sexo feminino. Os sinais de teste utilizados foram amostrados a 44,1 kHz, com resolução de 24 bits, e possuem alta relação sinal-ruído. O pré-processamento dos dados foi feito para garantir que cada frase tivesse

média zero e variância unitária. Uma vez obtido o banco de dados, os trechos sonoros, surdos e de silêncio foram marcados, primeiro usando um discriminador simples, com uma posterior depuração manual do resultado do classificador.

Os resultados apresentados a seguir incluem testes para determinar quão robustos os parâmetros de disciminação sonoro/surdo/silêncio são com relação a variações de razão sinal ruído, assim como variações de amplitude do sinal analisado. Para isso foram criados 5 bancos de dados a partir do banco original com as características a seguir:

- **Banco 1:** Banco original com alta razão sinal-ruído (*signal-to-noise ratio* SNR);
- **Banco 2:** Frases contaminadas com ruído branco gaussiano a 20 dB de SNR;
- **Banco 3:** Replicação do banco em 4 vezes, e variação de ganho de 0,1 a 10. Para isso, todas as frases deste banco foram colocadas seqüencialmente em um vetor, e as amostras das frases foram multiplicadas por uma senóide com freqüência fundamental de 1 Hz. Essa multiplicação tem por objetivo simular a variação de distância do locutor ao microfone, ou eventuais variações de ganho dos transdutores;
- **Banco 4:** Replicação do banco em 4 vezes, sendo uma réplica com alta SNR e as demais contaminadas com ruído branco gaussiano a 140, 80 e 20 dB de SNR (como os sinais usados neste texto são de alta qualidade, não se optou por SNRs muito baixas);
- **Banco 5:** Variação de amplitude, de acordo com o Banco 3, e de SNR, de acordo com o Banco 4.

Os parâmetros de discriminação foram calculados para janelas de 1024 amostras, com um salto entre janelas adjacentes de 128 amostras. A seguir está a lista dos parâmetros testados, com as legendas usadas nas figuras e tabelas de resultados.

- **E** - energia do bloco em dB calculada na equação (4.2);
- **EM** - energia do bloco normalizada pela energia média de blocos adjacentes calculada na equação (4.3) com $a = 0{,}0025$;

- **SS** - diferença normalizada entre o máximo pico da FFT e o chão de ruído calculada na equação (4.15) com $N_f = 50$;
- **Zc** - Taxa de cruzamento por zeros calculada na equação (4.5);
- **Ac1, 5, 10, 15** - Soma dos $n = 1, 5, 10, 15$ primeiros coeficientes de auto-correlação normalizados pela energia do bloco, calculados na equação (4.8);
- **AcM** - Máximo coeficiente de auto-correlação normalizado pela energia do bloco;
- **Lp1** - Primeiro coeficiente LPC obtido usando a função do Matlab®, com 10 coeficientes;
- **ELp** - Resíduo do LPC calculado com os coeficientes obtidos no item anterior.

O primeiro teste feito com os parâmetros foi calcular o coeficiente de auto-correlação entre as variáveis. Este coeficiente é um indicativo direto de como diferentes parâmetros estão relacionados entre si. Um coeficiente de correlação elevado indica que os dois parâmetros testados adicionam informação redundante para a discriminação entre classes. Nas Tabelas 4.2, 4.3, 4.4, 4.5, 4.6 são mostrados esses coeficientes para o Bancos 1 a 5. As tabelas mostram que, como era de se esperar, os dois parâmetros relacionados a energia do bloco e os parâmetros extraídos da auto-correlação do bloco são bastante correlacionados entre si. Também os parâmetros de taxa de cruzamento por zeros (Zc), e a diferença entre o máximo pico da FFT e o chão de ruído (SS) estão bem correlacionados aos parâmetros de auto-correlação.

As Tabelas 4.2, 4.3, 4.4, 4.5 e 4.6 mostram os resultados do cálculo do coeficiente de auto-correlação entre cada parâmetro analisado e o rótulo de classe.

Para testar a taxa de acerto de cada parâmetro, foram obtidas estimativas das funções de densidade de probabilidade (*probabilily density function* PDF) $p(x_j|\mathcal{C}_i)$ (verossimilhança) usando janelas de Parzen [58], onde C_i representa as classes sonoro, surdo e silêncio. A estimativa das PDFs é, então, dada por

$$p(x|\mathcal{C}_i) = \frac{1}{n_{\mathcal{C}_i}} \sum_{x_k \in \mathcal{C}_i} \frac{1}{h_n} \varphi\left(\frac{x - x_k}{h_n}\right), \tag{4.16}$$

onde

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \tag{4.17}$$

Tabela 4.2: Coeficientes de correlação entre as variáveis testadas e rótulo de classe (Banco 1).

	E	EM	SS	Zc	Ac1	Ac5	Ac10	Ac15	AcM	Lp1	ELp
E	1	0,99	0,13	-0,17	0,09	0,10	0,12	0,11	0,12	-0,33	-0,56
EM	0,99	1	0,11	-0,16	0,09	0,09	0,10	0,09	0,12	-0,33	-0,55
SS	0,13	0,11	1	-0,72	0,66	0,75	0,80	0,84	0,77	0,30	-0,53
Zc	-0,17	-0,16	-0,72	1	-0,97	-0,95	-0,92	-0,90	-0,66	-0,12	0,71
Ac1	0,09	0,09	0,66	-0,97	1	0,91	0,87	0,85	0,62	0,05	-0,63
Ac5	0,10	0,09	0,75	-0,95	0,91	1	0,98	0,95	0,66	0,22	-0,66
Ac10	0,12	0,10	0,80	-0,92	0,87	0,98	1	0,99	0,69	0,29	-0,67
Ac15	0,11	0,09	0,84	-0,90	0,85	0,95	0,99	1	0,70	0,32	-0,65
AcM	0,12	0,12	0,77	-0,66	0,62	0,66	0,69	0,70	1	0,28	-0,50
Lp1	-0,33	-0,33	0,30	-0,12	0,05	0,22	0,29	0,32	0,28	1	0,03
ELp	-0,56	-0,55	-0,53	0,71	-0,63	-0,66	-0,67	-0,65	-0,50	0,03	1
Sonoro	0,71	0,70	0,53	-0,64	0,57	0,62	0,64	0,62	0,45	-0,08	-0,74
Surdo	-0,13	-0,12	-0,69	0,81	-0,74	-0,84	-0,85	-0,84	-0,63	-0,28	0,62
Silêncio	-0,83	-0,82	0,07	-0,05	0,08	0,11	0,11	0,12	0,11	0,42	0,30

Tabela 4.3: Coeficientes de correlação entre as variáveis testadas e rótulo de classe para sinais com SNR = 20dB (Banco 2).

	E	EM	SS	Zc	Ac1	Ac5	Ac10	Ac15	AcM	Lp1	ELp
E	1	0,98	0,63	-0,88	0,86	0,74	0,74	0,73	0,81	-0,81	-0,96
EM	0,98	1	0,60	-0,85	0,83	0,71	0,71	0,70	0,80	-0,81	-0,94
SS	0,63	0,60	1	-0,82	0,77	0,85	0,87	0,89	0,87	-0,33	-0,72
Zc	-0,88	-0,85	-0,82	1	-0,98	-0,93	-0,93	-0,92	-0,89	0,66	0,92
Ac1	0,86	0,83	0,77	-0,98	1	0,90	0,90	0,90	0,86	-0,69	-0,88
Ac5	0,74	0,71	0,85	-0,93	0,90	1	0,99	0,98	0,84	-0,45	-0,85
Ac10	0,74	0,71	0,87	-0,93	0,90	0,99	1	0,99	0,85	-0,43	-0,85
Ac15	0,73	0,70	0,89	-0,92	0,90	0,98	0,99	1	0,85	-0,41	-0,83
AcM	0,81	0,80	0,87	-0,89	0,86	0,84	0,85	0,85	1	-0,56	-0,86
Lp1	-0,81	-0,81	-0,33	0,66	-0,69	-0,45	-0,43	-0,41	-0,56	1	0,67
ELp	-0,96	-0,94	-0,72	0,92	-0,88	-0,85	-0,85	-0,83	-0,86	0,67	1
Sonoro	0,74	0,71	0,81	-0,90	0,86	0,92	0,93	0,92	0,81	-0,47	-0,83
Surdo	-0,30	-0,29	-0,56	0,47	-0,37	-0,64	-0,63	-0,62	-0,46	-0,01	0,46
Silêncio	-0,67	-0,65	-0,47	0,70	-0,76	-0,54	-0,54	-0,55	-0,58	0,65	0,60

é uma janela exponencial cuja largura é controlada pelo fator h_n, x_k são os pontos de treinamento pertencentes a classe $\mathcal{C}_i$ e $n_{\mathcal{C}_i}$ é o número total de pontos de treinamento da classe $\mathcal{C}_i$. As PDFs foram calculadas em 100 pontos linearmente espaçados entre os valores máximo e mínimo de cada parâmetro x_j.

Uma vez obtidas as PDFs dos parâmetros testados, elas foram usadas para

Tabela 4.4: Coeficientes de correlação entre as variáveis testadas e rótulo de classe para sinais com variação de amplitude (Banco 3).

	E	EM	SS	Zc	Ac1	Ac5	Ac10	Ac15	AcM	Lp1	ELp
E	1	0,96	0,10	-0,13	0,08	0,08	0,09	0,09	0,10	-0,26	-0,45
EM	0,96	1	0,10	-0,14	0,08	0,08	0,10	0,09	0,11	-0,29	-0,48
SS	0,10	0,10	1	-0,72	0,66	0,75	0,80	0,84	0,77	0,30	-0,54
Zc	-0,13	-0,14	-0,72	1	-0,97	-0,95	-0,92	-0,90	-0,66	-0,12	0,71
Ac1	0,08	0,08	0,66	-0,97	1	0,91	0,87	0,85	0,62	0,05	-0,63
Ac5	0,08	0,08	0,75	-0,95	0,91	1	0,98	0,95	0,66	0,22	-0,66
Ac10	0,09	0,10	0,80	-0,92	0,87	0,98	1	0,99	0,69	0,29	-0,67
Ac15	0,09	0,09	0,84	-0,90	0,85	0,95	0,99	1	0,70	0,31	-0,65
AcM	0,10	0,11	0,77	-0,66	0,62	0,66	0,69	0,70	1	0,28	-0,50
Lp1	-0,26	-0,29	0,30	-0,12	0,05	0,22	0,29	0,31	0,28	1	0,03
ELp	-0,45	-0,48	-0,54	0,71	-0,63	-0,66	-0,67	-0,65	-0,50	0,03	1
Sonoro	0,57	0,62	0,53	-0,64	0,57	0,62	0,64	0,62	0,45	-0,08	-0,74
Surdo	-0,11	-0,11	-0,69	0,81	-0,75	-0,84	-0,85	-0,84	-0,63	-0,27	0,62
Silêncio	-0,66	-0,72	0,06	-0,05	0,08	0,11	0,11	0,12	0,11	0,43	0,30

Tabela 4.5: Coeficientes de correlação entre as variáveis testadas e rótulo de classe para sinais com variação de SNR (Banco 4).

	E	EM	SS	Zc	Ac1	Ac5	Ac10	Ac15	AcM	Lp1	ELp
E	1	1	0,19	-0,25	0,22	0,19	0,20	0,19	0,22	-0,20	-0,47
EM	1	1	0,18	-0,25	0,21	0,18	0,19	0,18	0,22	-0,21	-0,46
SS	0,19	0,18	1	-0,74	0,67	0,78	0,82	0,86	0,80	0,01	-0,56
Zc	-0,25	-0,25	-0,74	1	-0,97	-0,92	-0,90	-0,89	-0,76	0,32	0,78
Ac1	0,22	0,21	0,67	-0,97	1	0,87	0,85	0,83	0,72	-0,36	-0,71
Ac5	0,19	0,18	0,78	-0,92	0,87	1	0,98	0,96	0,73	-0,09	-0,68
Ac10	0,20	0,19	0,82	-0,90	0,85	0,98	1	0,99	0,75	-0,04	-0,69
Ac15	0,19	0,18	0,86	-0,89	0,83	0,96	0,99	1	0,76	-0,03	-0,68
AcM	0,22	0,22	0,80	-0,76	0,72	0,73	0,75	0,76	1	-0,09	-0,60
Lp1	-0,20	-0,21	0,01	0,32	-0,36	-0,09	-0,04	-0,03	-0,09	1	0,46
ELp	-0,47	-0,46	-0,56	0,78	-0,71	-0,68	-0,69	-0,68	-0,60	0,46	1
Sonoro	0,70	0,69	0,60	-0,67	0,61	0,69	0,71	0,69	0,54	-0,11	-0,65
Surdo	-0,15	-0,15	-0,65	0,62	-0,52	-0,76	-0,78	-0,76	-0,57	-0,15	0,50
Silêncio	-0,78	-0,78	-0,08	0,20	-0,24	-0,07	-0,08	-0,07	-0,09	0,33	0,32

classificar cada ponto do banco de dados de treinamento pelo critério de máxima verossimilhança. Desta forma foi gerada a Tabela 4.7.

Pela comparação dos índices de acertos indicados na tabela, é possível destacar que a energia do sinal teve um bom resultado para a detecção de silêncio, mesmo quando comparada com a energia normalizada. Apesar deste resultado é possível

Tabela 4.6: Coeficientes de correlação entre as variáveis testadas e rótulo de classe para sinais com variação de amplitude e SNR (Banco 5).

	E	EM	SS	Zc	Ac1	Ac5	Ac10	Ac15	AcM	Lp1	ELp
E	1	0,95	0,15	-0,19	0,16	0,14	0,15	0,15	0,17	-0,17	-0,37
EM	0,95	1	0,15	-0,20	0,17	0,15	0,16	0,15	0,18	-0,17	-0,40
SS	0,15	0,15	1	-0,74	0,67	0,78	0,82	0,86	0,80	0,01	-0,56
Zc	-0,19	-0,20	-0,74	1	-0,97	-0,92	-0,90	-0,89	-0,76	0,32	0,78
Ac1	0,16	0,17	0,67	-0,97	1	0,87	0,85	0,83	0,72	-0,36	-0,71
Ac5	0,14	0,15	0,78	-0,92	0,87	1	0,98	0,96	0,73	-0,09	-0,68
Ac10	0,15	0,16	0,82	-0,90	0,85	0,98	1	0,99	0,75	-0,04	-0,69
Ac15	0,15	0,15	0,86	-0,89	0,83	0,96	0,99	1	0,76	-0,03	-0,68
AcM	0,17	0,18	0,80	-0,76	0,72	0,73	0,75	0,76	1	-0,09	-0,60
Lp1	-0,17	-0,17	0,01	0,32	-0,36	-0,09	-0,04	-0,03	-0,09	1	0,46
ELp	-0,37	-0,40	-0,56	0,78	-0,71	-0,68	-0,69	-0,68	-0,60	0,46	1
Sonoro	0,54	0,59	0,60	-0,67	0,61	0,70	0,71	0,69	0,54	-0,11	-0,65
Surdo	-0,12	-0,13	-0,65	0,62	-0,52	-0,76	-0,78	-0,76	-0,57	-0,15	0,50
Silêncio	-0,60	-0,66	-0,08	0,20	-0,24	-0,07	-0,08	-0,07	-0,09	0,33	0,32

que ajustes no parâmetro a usado para calcular a energia média na equação (4.4) tornem essa métrica mais robusta. Esses ajustes podem ser feitos otimizando o valor de a, de preferência com um banco de dados que tenha já na gravação variações de energia.

Os parâmetros de auto-correlação foram os que indicaram resultados melhores para a detecção de trechos sonoros e a distinção sonoro/surdo. Um destaque especial deve ser dado à observação de que os índices de acerto para esta métrica se mostraram bastante robustos, mesmo para sinais contaminados com ruído e com variação de amplitude.

Outro resultado interessante é que algumas métricas apresentaram um ganho de eficiência consideravelmente grande quando os sinais foram contaminados com ruído de 20 dB. Este é o caso dos parâmetros AcM e Ac1 quando usados para detecção de silêncio, e dos parâmetros de auto-correlação quando usados para detecção de trechos sonoros.

A explicação para as variações em taxa de acerto com a variação de SNR do sinal pode ser retirada da observação das suas respectivas PDFs, mostradas nas Figuras 4.2 e 4.3. Nelas é possível observar uma variação bastante brusca na forma das PDFs de trechos de silêncio para os parâmetros Zc, ELp e baseados na auto-

Tabela 4.7: Taxas de acerto usando critério da máxima verossimilhança para cada parâmetro.

Taxa de acerto %	E	EM	SS	Zc	Ac1	Ac5	Ac10	Ac15	AcM	Lp1	ELp
Banco 1											
Total	78	25	13	21	72	69	67	61	7	12	15
Detecção Silêncio	96	88	21	22	80	73	71	70	47	67	15
Detecção Sonoro	81	33	33	32	79	83	83	79	32	28	33
Distinção Sonoro/surdo †	78	23	22	22	91	94	94	88	23	22	22
Banco 2											
Total	75	27	35	10	87	76	77	79	68	31	11
Detecção Silêncio	95	91	77	25	96	80	81	83	89	83	86
Detecção Sonoro	79	33	53	10	91	96	96	96	78	32	11
Distinção Sonoro/surdo †	76	23	45	11	90	95	96	95	74	37	13
Banco 3											
Total	63	21	13	19	70	68	67	62	9	9	13
Detecção Silêncio	88	72	29	21	79	72	71	69	39	54	13
Detecção Sonoro	69	33	35	32	78	82	84	80	32	27	33
Distinção Sonoro/surdo †	65	23	25	21	90	94	95	89	24	22	22
Banco 4											
Total	76	14	13	13	73	70	69	63	10	10	12
Detecção Silêncio	94	45	26	13	80	74	73	71	36	22	12
Detecção Sonoro	80	33	34	31	82	86	86	81	32	33	33
Distinção Sonoro/Surdo †	77	22	23	22	91	94	94	88	23	22	22
Banco 5											
Total	61	19	14	13	72	70	69	64	11	10	12
Detecção Silêncio	84	63	26	13	80	74	73	71	36	22	12
Detecção Sonoro	69	33	35	30	82	86	86	82	33	30	33
Distinção Sonoro/surdo †	65	23	25	22	90	94	94	89	24	22	22

† Total de erro considerando somente blocos que não foram marcados como sendo de silêncio.

correlação.

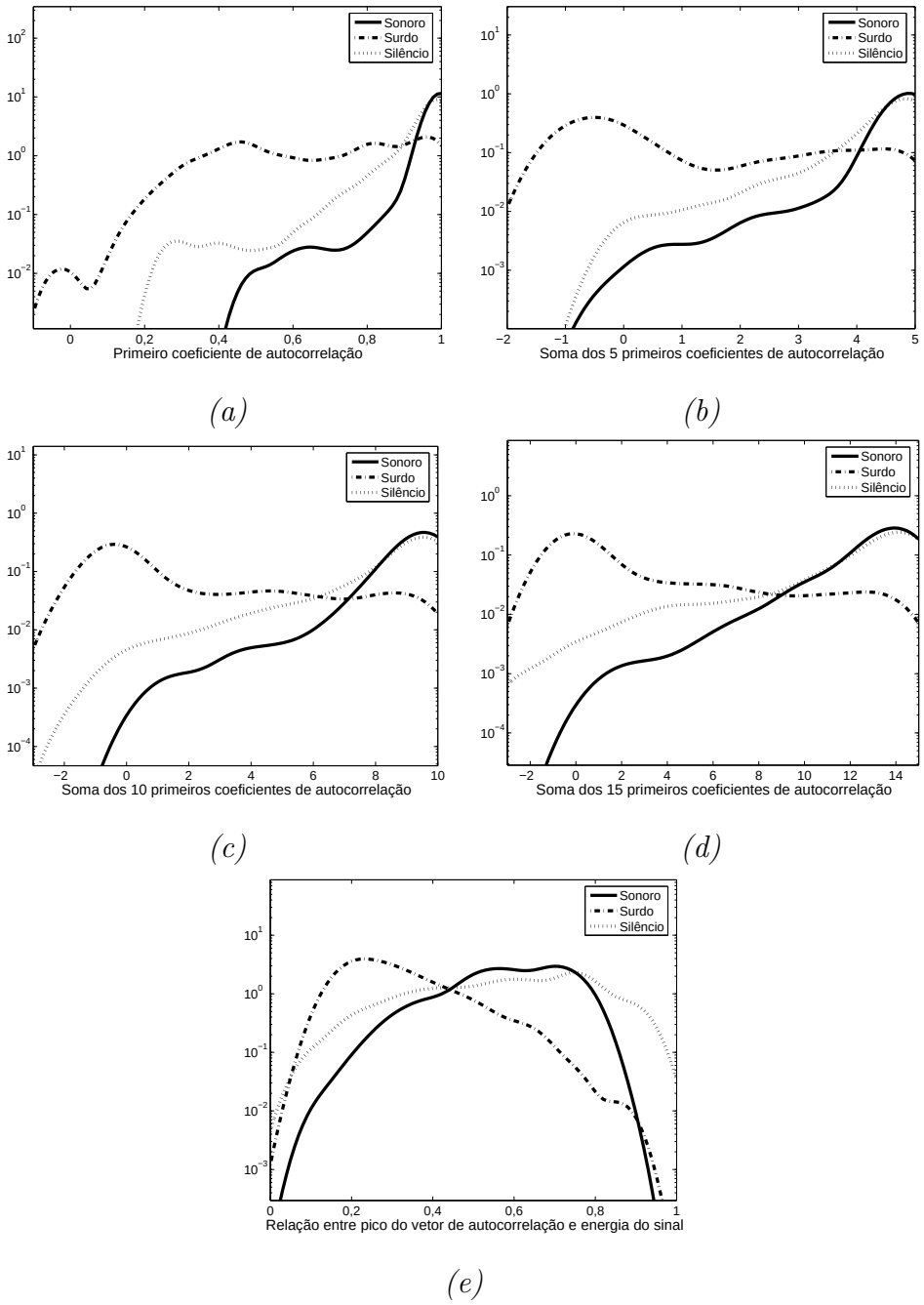


Figura 4.2: Funções de densidade de probabilidade para sinais sonoros, surdos e de silêncio com alta SNR.

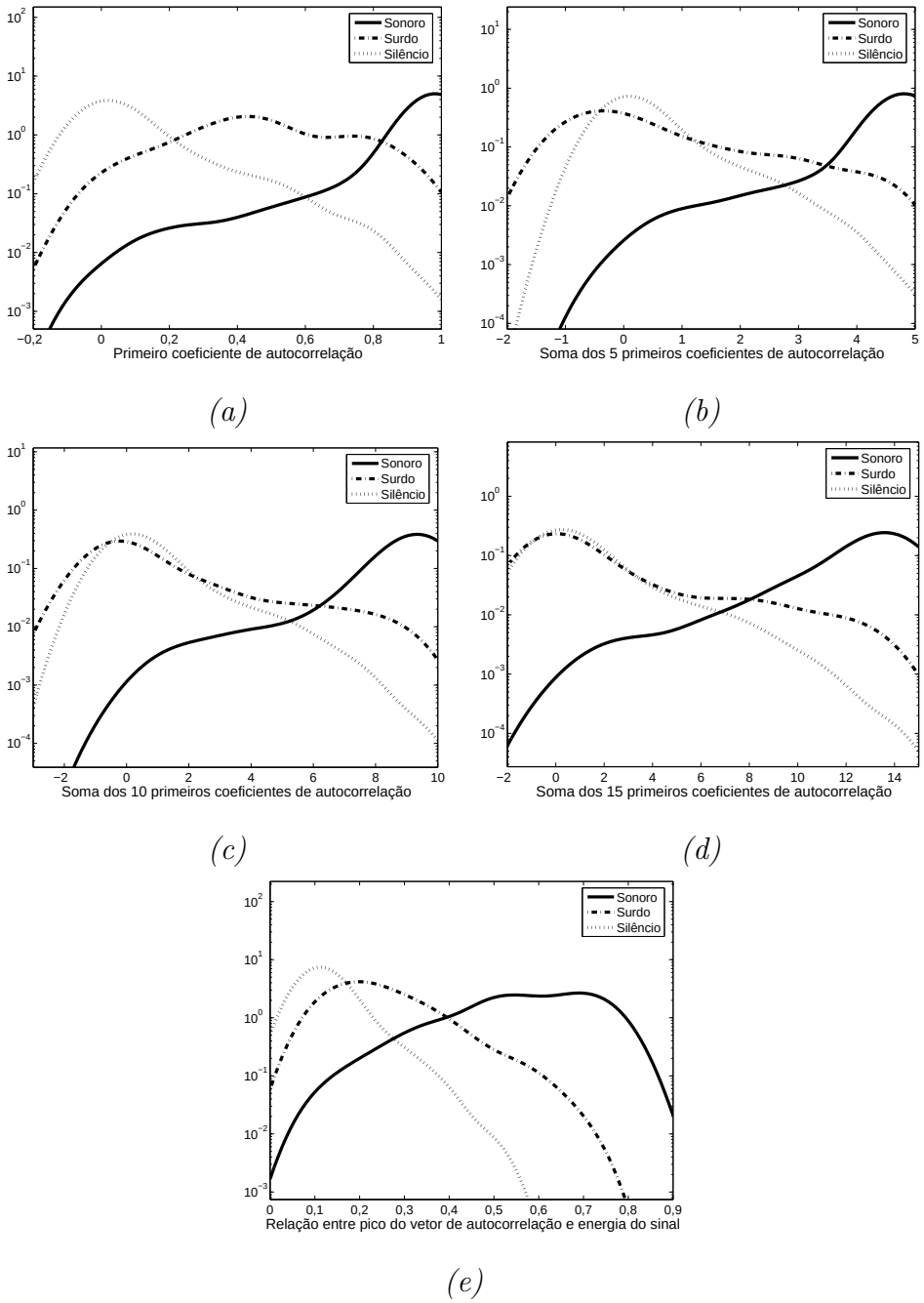


Figura 4.3: Funções de densidade de probabilidade para sinais sonoros, surdos e de silêncio com 20dB de SNR.

4.7.1 Avaliação de grupos de parâmetros para discriminação

Feita a análise da seção anterior, é possível escolher a energia do bloco, E, e a soma dos 10 primeiros coeficientes de auto-correlação, Ac10, como parâmetros de discriminação. Esta seção tem por objetivo analisar o resultado nas taxas de

acerto usando os outros parâmetros em conjunto com E e Ac10. Para isso será projetado um discriminante linear de Fisher [58]. O objetivo do discriminante linear de Fisher é descobrir um mapeamento linear para um dado vetor de parâmetros de D dimensões para um valor escalar, de forma que seja maximizada a separação entre classes. A equação (4.18) mostra o mapeamento de um vetor $\mathbf{f}$ em um escalar y pelo produto escalar entre $\mathbf{f}$ e $\mathbf{w}$:

$$y = \mathbf{w}^T \mathbf{f}. \tag{4.18}$$

A solução ótima para o vetor de mapeamento $\mathbf{w}$ é encontrada pela maximização da função-custo a seguir:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}, \tag{4.19}$$

onde $\mathbf{S}_B$ é a matriz de dispersão entre classes e $\mathbf{S}_W$ é a matriz de dispersão intra-classes. Para duas classes $\mathcal{C}_i$ com $n_{\mathcal{C}_i}$ elementos, $i = 1, 2$:

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)^T (\mathbf{m}_1 - \mathbf{m}_2) \tag{4.20}$$

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \tag{4.21}$$

$$\mathbf{m}_i = \frac{1}{n_{\mathcal{C}_i}} \sum_{\mathbf{f} \in \mathcal{C}_i} \mathbf{f} \tag{4.22}$$

$$\mathbf{S}_i = \sum_{\mathbf{f} \in \mathcal{C}_i} (\mathbf{f} - \mathbf{m}_i)^T (\mathbf{f} - \mathbf{m}_i). \tag{4.23}$$

A maximização da equação (4.19) é feita de forma a maximizar a dispersão entre classes de y, dada por $\mathbf{w}^T \mathbf{S}_B \mathbf{w}$, e minimizar a dispersão intra-classes de y, dada por $\mathbf{w}^T \mathbf{S}_W \mathbf{w}$. A solução para duas classes é encontrada fazendo-se

$$\mathbf{w} = \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2). \tag{4.24}$$

A solução do discriminante de Fisher descreve deste modo um hiperplano de separação entre classes. Em alguns casos o uso do discriminante de Fisher pode não ser suficiente para a discriminação entre classes, especialmente quando a superfície de separação entre classes não descreve um hiperplano simples no domínio de $\mathbf{f}$. Para que o discriminante de Fisher seja capaz de descrever superfícies mais complexas, a dimensão de $\mathbf{f}$ pode ser aumentada com o uso de uma função *kernel*. Exemplos de função *kernel* são polinômios e funções de base radial. A equação (4.25) ilustra

funções *kernel* polinomiais de ordem 2 e 3:

$$\mathbf{f}_2 = \begin{pmatrix} \mathbf{f} \\ \mathbf{f}(1)^2 \\ \mathbf{f}(1)\mathbf{f}(2) \\ \vdots \\ \mathbf{f}(1)\mathbf{f}(\mathcal{D}) \\ \mathbf{f}(2)^2 \\ \mathbf{f}(2)\mathbf{f}(3) \\ \vdots \\ \mathbf{f}(\mathcal{D})^2 \end{pmatrix} \quad \text{e} \quad \mathbf{f}_3 = \begin{pmatrix} \mathbf{f} \\ \mathbf{f}_2 \\ \mathbf{f}(1)^3 \\ \mathbf{f}(1)^2\mathbf{f}(2) \\ \vdots \\ \mathbf{f}(1)^2\mathbf{f}(\mathcal{D}) \\ \mathbf{f}(1)\mathbf{f}(2)^2 \\ \mathbf{f}(1)\mathbf{f}(2)\mathbf{f}(3) \\ \vdots \\ \mathbf{f}(\mathcal{D})^3 \end{pmatrix}, \tag{4.25}$$

onde $\mathbf{f}_k$ representa a função *kernel* usando polinômio de ordem k.

A Tabela 4.8 mostra o resultado de taxa de acerto dos classificadores usando os parâmetros E e Ac combinados com os outros parâmetros e usando diferentes funções *kernel*. O banco de dados usado para gerar estes resultados foi o banco original contaminado com 20 dB de SNR, e com variação de amplitude de 0,1 a 10 vezes. Este procedimento visa a tornar o sistema robusto a variações de amplitude e de SNR. Para sinais com SNR maior que 20 dB o classificador vai passar o sinal por um pré-processamento que adiciona ruído a ele para obter um sinal resultante com essa SNR.

Os resultados da Tabela 4.8 são separados em detecção de trechos sonoros, detecção de silêncio e discriminação sonoro/surdo. Com esses resultados vão ser projetados 2 classificadores diferentes. O primeiro detecta trechos sonoros em um único estágio. O segundo faz um primeiro estágio de detecção de silêncio, para depois discriminar os trechos em sonoro/surdo. Uma vez projetados estes classificadores, o que obtiver melhor desempenho na detecção de trechos sonoros vai ser escolhido.

Como resultado da taxa de acerto mostrada na Tabela 4.8, foi escolhido o parâmetro Ac15 como sendo significativo para detecção de trechos sonoros e discriminação sonoro/surdo. Os passos subseqüentes do projeto do classificador envolvem seqüencialmente a escolha dos parâmetros que apresentam melhor taxa de acerto em conjunto com os parâmetros já escolhidos. Desta forma foram escolhidos os parâmetros SS e Zc, e a taxa de acerto com todos esses parâmetros aparece na Ta-

Tabela 4.8: Taxas de acerto usando E e Ac10.

Acerto %	**f**	EM	SS	Zc	Ac1	Ac5	Ac15	AcM	Lp1	ELp
Função *kernel*	polinômio de ordem 1									
*	79,9	76,8	94,2	90,0	89,6	94,2	94,5	90,7	79,9	88,2
†	84,3	81,7	80,5	88,6	90,3	80,7	80,4	86,4	91,1	86,9
‡	79,0	70,0	95,0	89,6	87,9	94,6	95,1	90,6	73,6	87,1
Função *kernel*	polinômio de ordem 2									
*	73,1	71,1	83,7	86,5	87,6	94,1	93,5	81,3	80,2	81,0
†	81,4	69,2	82,2	90,6	91,8	81,9	81,4	88,8	89,7	88,3
‡	87,9	62,6	94,8	92,8	91,9	95,6	95,8	92,0	87,7	90,6
Função *kernel*	polinômio de ordem 3									
*	80,0	79,0	88,9	86,3	86,3	91,0	89,9	84,9	81,2	84,4
†	79,8	83,1	79,3	89,4	91,5	80,7	80,6	86,2	89,1	86,0
‡	81,1	60,7	91,4	86,6	85,9	93,0	91,9	87,1	78,1	84,3

* Detecção de trechos sonoros. † Detecção de silêncio. ‡ Discriminação entre sonoro e surdo considerando somente blocos que não foram marcados como sendo de silêncio.

bela 4.9. O passo seguinte seria a escolha de AcM, mas como este parâmetro não resultou numa melhora muito significativa na taxa de acerto, foram escolhidos os parâmetros E, SS, Zc, Ac10 e Ac15 e uma função *kernel* linear para disciminação sonoro/surdo e detecção de trechos sonoros.

Tabela 4.9: Taxas de acerto usando E, Ac10, Ac15, SS e Zc.

Acerto %	**f**	EM	Ac1	Ac5	AcM	Lp1	ELp
Função *kernel*	polinômio de ordem 1						
*	96,6	92,7	96,5	96,5	96,6	95,7	96,5
‡	95,9	93,6	95,8	95,7	95,9	95,8	95,7
Função *kernel*	polinômio de ordem 2						
*	90,5	81,8	90,5	93,7	86,7	83,9	86,9
‡	94,5	89,8	93,1	95,0	91,0	94,8	92,6
Função *kernel*	polinômio de ordem 3						
*	95,3	86,7	94,3	95,6	95,6	91,8	95,5
‡	95,7	89,8	95,7	96,0	96,0	95,4	95,8

* Detecção de trechos sonoros. ‡ Discriminação entre sonoro e surdo considerando somente blocos que não foram marcados como sendo de silêncio.

Para detecção de silêncio, as etapas de projeto seguiram os mesmos passos que para a detecção de trechos sonoros, e discriminação sonoro/surdo. Com a Tabela 4.8 foi escolhido o parâmetro Ac1. Nos passos seguintes foi testada a taxa de acerto, e foram escolhidos os parâmetros E, Zc, Ac1, Ac10 e Lp1 usando como função *kernel*

um polinômio de ordem 2. O resultado da taxa de acertos para estes parâmetros é mostrado na Tabela 4.10.

Tabela 4.10: Taxas de acerto usando E, Ac10, Ac1, Lp1, Zc e AcM.

Acerto %	**f**	EM	SS	Ac5	Ac15	ELp
Função *kernel*	polinômio de ordem 1					
†	93,5	92,5	92,6	92,3	92,6	93,5
Função *kernel*	polinômio de ordem 2					
†	94,4	94,0	94,3	94,3	94,2	94,4
Função *kernel*	polinômio de ordem 3					
†	94,3	93,1	94,2	94,3	94,3	94,3

† Detecção de silêncio.

O primeiro estágio do classificador é a normalização dos parâmetros para que cada um tenha média zero e desvio-padrão unitário. Os valores de média e desvio-padrão estão na Tabela 4.11.

Tabela 4.11: Média e desvio padrão dos parâmetros usados para discriminação sonoro/surdo/silêncio.

	E	EM	SS	Zc	Ac1	Ac5	Ac10	Ac15	AcM	Lp1	ELp
Média	60	-8,7	0,13	0,17	0,74	3,2	6,2	8,9	0,45	-0,55	-37
Desvio Padrão	14	12	0,099	0,17	0,35	2,2	4,1	5,9	0,22	0,31	7,6

Os resultados dos dois classificadores usados estão ilustrados nas Tabelas 4.12 e 4.13.

Tabela 4.12: Resultado do classificador com 1 estágio.

Rótulo de treinamento	Rótulo identificado	
%	Silêncio/Surdo	Sonoro
Silêncio	13,6	0,1
Surdo	18,2	1,1
Sonoro	2,2	64,8

O resultado final de acerto de detecção de trechos sonoros foi de 96,58% para o classificador com 1 estágio e 96,35% para o de 2 estágios. Devido à simplicidade do classificador de 1 estágio, ele deve ser escolhido quando se deseja somente a detecção

Tabela 4.13: Resultado do classificador com 2 estágios.

Rótulo de treinamento	Rótulo identificado		
%	Silêncio	Surdo	Sonoro
Silêncio	12,9	0,7	0,1
Surdo	4,3	14,2	0,9
Sonoro	0,6	2,1	64,3

de trechos sonoros. Os coeficientes do classificador de 1 estágio são

$$\mathbf{w}_V = \begin{pmatrix} 0{,}48 & 0{,}522 & -0{,}412 & 0{,}387 & 0{,}421 \end{pmatrix}^T, \tag{4.26}$$

e o classificador final é

$$y = \mathbf{w}_V^T \mathbf{f}_V + 0{,}5760, \tag{4.27}$$

onde

$$\mathbf{f}_V = \begin{pmatrix} \mathrm{E} & \mathrm{SS} & \mathrm{Zc} & \mathrm{Ac10} & \mathrm{Ac15} \end{pmatrix}^T, \tag{4.28}$$

sendo o trecho classificado como sonoro quando $y > 0$ e surdo/silêncio em caso contrário.

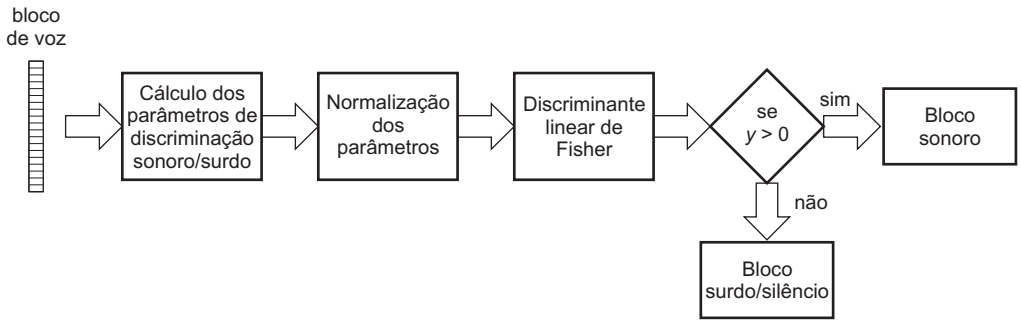


Figura 4.4: Determinação de sonoridade de blocos de voz.

A Figura 4.4 mostra como é feita a classificação de trechos sonoros. Nesta figura estão ilustrados o cálculo dos parâmetros para discriminação, a normalização dos parâmetros usando a Tabela 4.11 e o cálculo do discriminante linear de Fisher usando a equação (4.27).

4.8 Conclusão

Este capítulo apresentou uma análise de parâmetros relevantes para a discriminação entre trechos sonoros, surdos e de silêncio. Os resultados deste capítulo são importantes para a implementação do resto deste trabalho, uma vez que as técnicas de transformação de sinais de voz descritas nos Capítulos 5 e 6 serão aplicadas somente a trechos sonoros.

O estudo apresentado nas Seções 4.2, 4.3, 4.4 e 4.5 é importante para demonstrar como certos parâmetros que são bastante conhecidos e difundidos na literatura técnica podem ser usados para discriminação sonoro/surdo/silêncio. Ainda foi apresentado um novo tipo de parâmetro baseado na estimativa do espectro do ruído de fundo presente no sinal na Seção 4.6. Este parâmetro tem por objetivo auxiliar na detecção de trechos sonoros que têm algumas características bastante similares a trechos surdos, como é o caso dos fonemas /*z*/ (da palavra casa) e /*j*/ (da palavra gente). Esse tipo de fonema apresenta uma taxa de cruzamentos por zero bastante elevada, e é facilmente confundido com um fonema surdo.

Ainda foi escolhido um banco de dados com frases foneticamente balanceadas, e com uma variação bastante grande de SNR e amplitude. Essas variações têm por objetivo garantir que o classificador seja robusto a esses tipos de variações — que podem levar a um funcionamento precário em outros classificadores. As variações de SNR e amplitude usadas no banco de dados têm como objetivo simular condições diferentes de gravação, e técnicas de expressividade que são usadas por alguns artistas. Algumas dificuldades são enfrentadas para produzir o banco de dados, uma vez que a determinação de quais trechos são sonoros, surdos ou de silêncio é bastante difícil. Por isso é importante que o banco de dados seja suficientemente rico estatisticamente, de forma a garantir que erros eventuais de marcação dos trechos não leve ao um erro significativo no treinamento do classificador.

Capítulo 5

Modificação de *pitch*/tempo de sinais de voz

5.1 Introdução

Dos tipos de modificação operados em sinais de voz, as modificações de *pitch*/tempo estão entre as mais comuns. Esse tipo de modificação tem achado utilidade em diversos produtos comerciais, entre os quais estão: sistemas de edição de sinais musicais; trilhas sonoras de filmes; modificação de prosódia em sistemas de conversão texto/fala; sistemas automáticos de correção de afinação; efeitos de coral; etc.

As primeiras implementações de modificação de *pitch*/tempo usavam modificação de velocidade de reprodução de gravações musicais. Uma das primeiras implementações que era capaz de modificar *pitch* e tempo de maneira independente era totalmente mecânica [59, 60], e algumas implementações modernas continuam usando uma abordagem inspirada em variações de velocidade [61]. Apesar de sua simplicidade, esse tipo de implementação tem sido abandonado para sinais de voz, uma vez que as variações de velocidade causam um deslocamento da posição dos formantes da voz. Esse tipo de defeito é extremamente desagradável para fatores grandes de modificação de *pitch*, e pode ser percebido como sendo uma voz não humana por corresponder à impresão causada por uma modificação do tamanho do trato vocal.

Técnicas bem sucedidas para modificação de *pitch* surgiram pelo uso de mo-

delos de predição linear [62, 22], uma vez que com esses modelos é possível operar transformações que mantêm as características do trato vocal. Outros métodos fazem uso da transformada rápida de Fourier; entre eles está o *phase vocoder*. Este tipo de sistema está sujeito a distorção de fase, chamada na literatura de *phasiness* [63].

Outros tipos de técnicas que preservam as características do trato vocal são baseados em sobreposição-e-soma de maneira síncrona ao *pitch* (*pitch-synchonous overlap-and-add*, PSOLA). Esse tipo de método é chamado de não paramétrico, por não usar um modelo explícito para a envoltória espectral do sinal de voz, e possui algumas variantes: o PSOLA aplicado no domínio do tempo (*time-domain*, TD-PSOLA); aplicado no domínio da freqüência (*frequency-domain*, FD-PSOLA); usando um modelo de predição linear (*linear-prediction*, LP-PSOLA) [64]; e em conjunto com modelo de componentes senoidais (SINOLA) [65].

Neste capítulo serão abordados algoritmos de modificação de *pitch* usando o modelo LPC seqüencial apresentado na Seção 3.2.2; usando o PSOLA no domínio do tempo [64]; e usando o LP-PSOLA com o modelo LPC seqüencial. Parte do texto deste capítulo é baseado nos trabalhos publicados em congressos da área pelo autor deste texto [23, 24], com os resultados associados.

5.2 Modificação de *pitch*/tempo usando LPC

O sistema de modificação de *pitch*/tempo desta seção utiliza o esquema de análise/síntese da Figura 5.1. Neste esquema é usado o modelo LPC, discutido na Seção 3.2, com o objetivo de obter um filtro que aproxima a envoltória espectral do sinal de voz, e é considerado independente do sinal de excitação. Neste esquema o sinal de excitação para o sinal modificado é aproximado como sendo um trem de impulsos, de forma que fica simples controlar o período de *pitch* do sinal modificado.

O modelo LPC deste sistema usa a solução seqüencial com RLS apresentada na Seção 3.2.2. O objetivo de usar a solução seqüencial para o modelo LPC é obter um sistema que tenha pouco atraso para possibilitar a implementação em tempo real, e também obter um sistema com transições mais suaves, o que pode levar a um sinal modificado mais natural. A ordem do modelo LPC para este sistema deve ser grande o suficiente para aproximar a envoltória espectral do sinal de voz, contudo

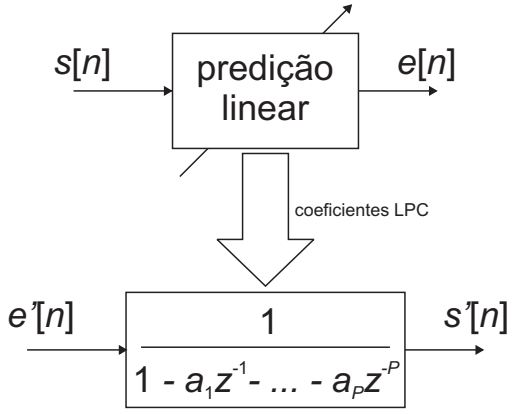


Figura 5.1: Esquema de modificação de *pitch* usando o modelo LPC.

não deve modelar a freqüência fundamental do sinal analisado. Na prática, a ordem do modelo LPC está compreendida na faixa entre $10 \leq M \leq 15$ para sinais com freqüência de amostragem baixa (8 kHz$\leq f_s \leq$16 kHz), e deve ser mais alta para sinais de alta qualidade ($f_s \geq$44,1 kHz). Como os sinais usados neste trabalho são de alta qualidade ($f_s =$44,1 kHz), foram usados $M = 30$ coeficientes.

A Figura 5.2 mostra como o sinal de excitação para o sistema da Figura 5.1 é determinado. No primeiro bloco é determinado o período de *pitch* $p[n]$. As técnicas que podem ser usadas para este bloco incluem métodos baseados em auto-correlação [2], análise cepstral [30], o método YIN [66] e aproximação de mínimos quadrados no domínio da freqüência [67], entre outros [68]. Contudo estes métodos

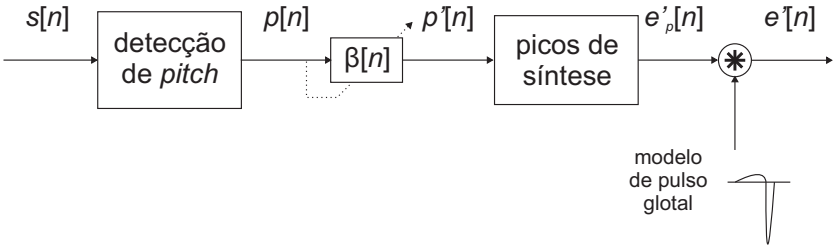


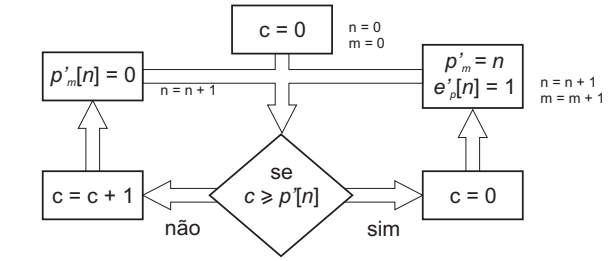
Figura 5.2: Determinação do sinal de excitação modificado.

são todos baseados em estimativas de *pitch* para blocos relativamente grandes (entre 20 e 50 ms), o que leva a um atraso do tamanho de um bloco em sistemas em tempo real. Para contornar esses problemas é possível determinar o *pitch* de um sinal de voz determinando-se os instantes de fechamento glotal deste sinal. Técnicas para

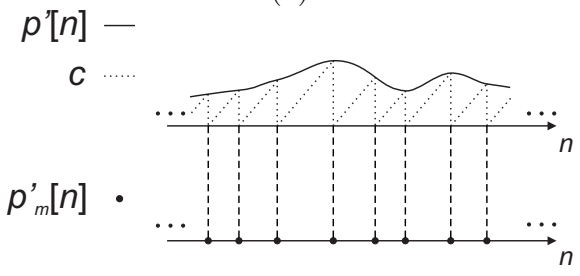
detecção dos instantes de fechamento glotal incluem o uso de *wavelets* [53, 69, 70], norma de Frobenius [71] e atraso de grupo [72]. Em conjunto com esses parâmetros é necessário o uso de técnicas heurísticas e programação dinâmica [73]. Para a determinação do período de *pitch* neste bloco foi usada a implementação para Matlab® de Naylor *et al.* [73] disponível para Matlab em uma rotina do Voicebox [74], e alguns sinais foram gravados usando um sinal do eletro-glotógrafo *EGGs for Singers*®.

No segundo bloco, o período de *pitch* do sinal original $p[n]$ é multiplicado por um fator $\beta[n]$. No caso mais simples, este fator é escolhido como uma constante. Em sistemas mais elaborados, $\beta[n]$ pode ser variável no tempo, de forma a fazer afinação automática de uma melodia ou modificar a prosódia de uma frase falada.

Com a informação do período de *pitch* desejado $p'[n] = p[n]\beta$ é possível determinar marcas de *pitch*, ou instantes de fechamento glotal, p'_m, para o sinal modificado. Essas marcas devem estar dispostas de forma que o sinal modificado tenha o período de *pitch* próximo a $p'[n]$; para isso é usado o algoritmo descrito na Figura 5.3 (a). Neste algoritmo, um contador c é usado para determinar os



(a)



(b)

Figura 5.3: Determinação das marcas de *pitch* do sinal modificado.

instantes de fechamento glotal. Este contador é incrementado a cada instante de tempo e comparado com o período de *pitch* desejado $p'[n]$. A cada vez que o contador atinge um valor maior ou igual a $p'[n]$, o contador é zerado e uma marca de *pitch*

é alocada a este instante: $p'_m = n$ e $e'_p[n] = 1$. Métodos mais complexos podem ser usados neste bloco. Uma vez que a posição ideal das marcas de *pitch* pode não coincidir com um instante de amostragem n, métodos de atraso fracionário podem ser usados para gerar o sinal $e'_p[n]$ [75]; esse procedimento se torna mais importante à medida que se usam sinais com menores taxas de amostragem. Também alguns efeitos de *aliasing* podem ocorrer devido à forma como é gerado o sinal $e'_p[n]$. Para contornar esses problemas, algumas técnicas descritas em [76] podem ser adaptadas para trem de impulsos. Como a taxa de amostragem dos sinais utilizados é bem mais alta que o período de *pitch* dos sinais típicos de voz, técnicas mais aprimoradas para gerar as marcas de *pitch* não foram necessárias neste trabalho.

No último bloco da Figura 5.2, é feita a sobreposição-e-soma de pulsos glotais, de acordo com as marcas de *pitch* determinadas no bloco anterior, para a determinação do sinal de excitação modificado $e'[n]$. Essa operação pode ser representada pela convolução do sinal $e'_p[n]$ com um modelo de pulsos glotais. Apesar de existirem modelos bastante refinados de excitação glotal, como o de Fant/Liljencrants [15], o modelo da envoltória espectral usando LPC não considera esse tipo de excitação, e acaba por modelar o decaimento espectral relativo à excitação glotal. Por isso foi usado como modelo de excitação glotal um pulso, com comprimento de aproximadamente dois períodos de *pitch*, do resíduo de predição linear.

Uma vez obtido o modelo LPC no instante n no estágio de análise da Figura 5.1 e determinado o sinal de excitação modificado $e'[n]$, é possível calcular o sinal modificado $s'[n]$

$$s'[n] = e'[n] - \mathbf{a}_P^T[n]\mathbf{s}'[n-1], \tag{5.1}$$

onde

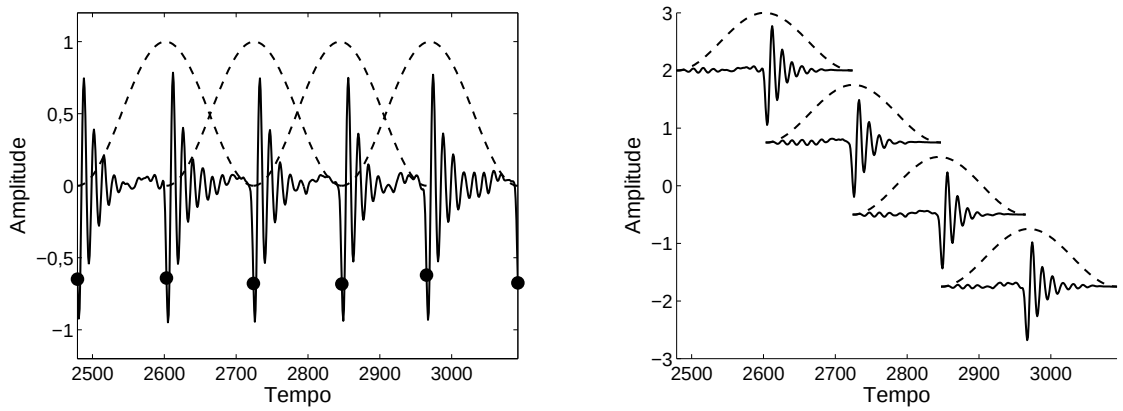
$$\mathbf{s}'[n-1] = (s'[n-1]\ s'[n-2]\ \ldots\ s'[n-P])^T. \tag{5.2}$$

5.3 Modificação de *pitch*/tempo usando PSOLA

As próximas seções apresentam alguns detalhes de dois tipos diferentes de técnicas baseadas na sobreposição-e-soma de maneira síncrona ao *pitch* PSOLA, o TD-PSOLA e o LP-PSOLA. O TD-PSOLA é apresentado a título de ilustração e para melhor compreensão do LP-PSOLA.

5.3.1 TD-PSOLA

O PSOLA aplicado no domínio do tempo (TD-PSOLA) está entre as implementações mais populares deste tipo de técnica. A característica principal das técnicas baseadas no PSOLA é de decompor o sinal de entrada $s[n]$ em segmentos de análise, sobrepor e somar cada segmento para a síntese final do sinal modificado. A Figura 5.4 (a) mostra um sinal sintético, com suas marcas de *pitch*, e decomposto em segmentos de maneira síncrona ao *pitch* [64].



(a) (b)

Figura 5.4: (a) Sinal de voz com suas marcas de *pitch* e janelas para decomposição; (b) segmentos decompostos do sinal em (a).

A decomposição de um sinal $s[n]$ em segmentos é feita por:

$$s_m = h[n]s[n - p_m], \tag{5.3}$$

onde $h[n]$ é uma janela com valor máximo em $n = 0$, p_m é a m-ésima marca de *pitch* do sinal $s[n]$, e $s_m[n]$ é o m-ésimo segmento do sinal $s[n]$. O efeito de multiplicar o sinal $s[n]$ por uma janela $h[n]$ é de suavizar o espectro de s_m. Quanto menor a janela, menor é sua resolução freqüencial [64]. A Figura 5.5 mostra esse efeito para a transformada de Fourier do sinal na Figura 5.4(a), usando janelas de tamanho diferente. Nesta figura podemos notar que ao usar uma janela de 2 períodos de *pitch* síncrona a p_m, $\mathcal{F}(s_m)$ aproxima a envoltória espectral do sinal $s[n]$, ao passo que segmentos oriundos de janelas não-síncronas ao *pitch*, ou de tamanho maior que 2, carregam consigo a informação do período de *pitch*.

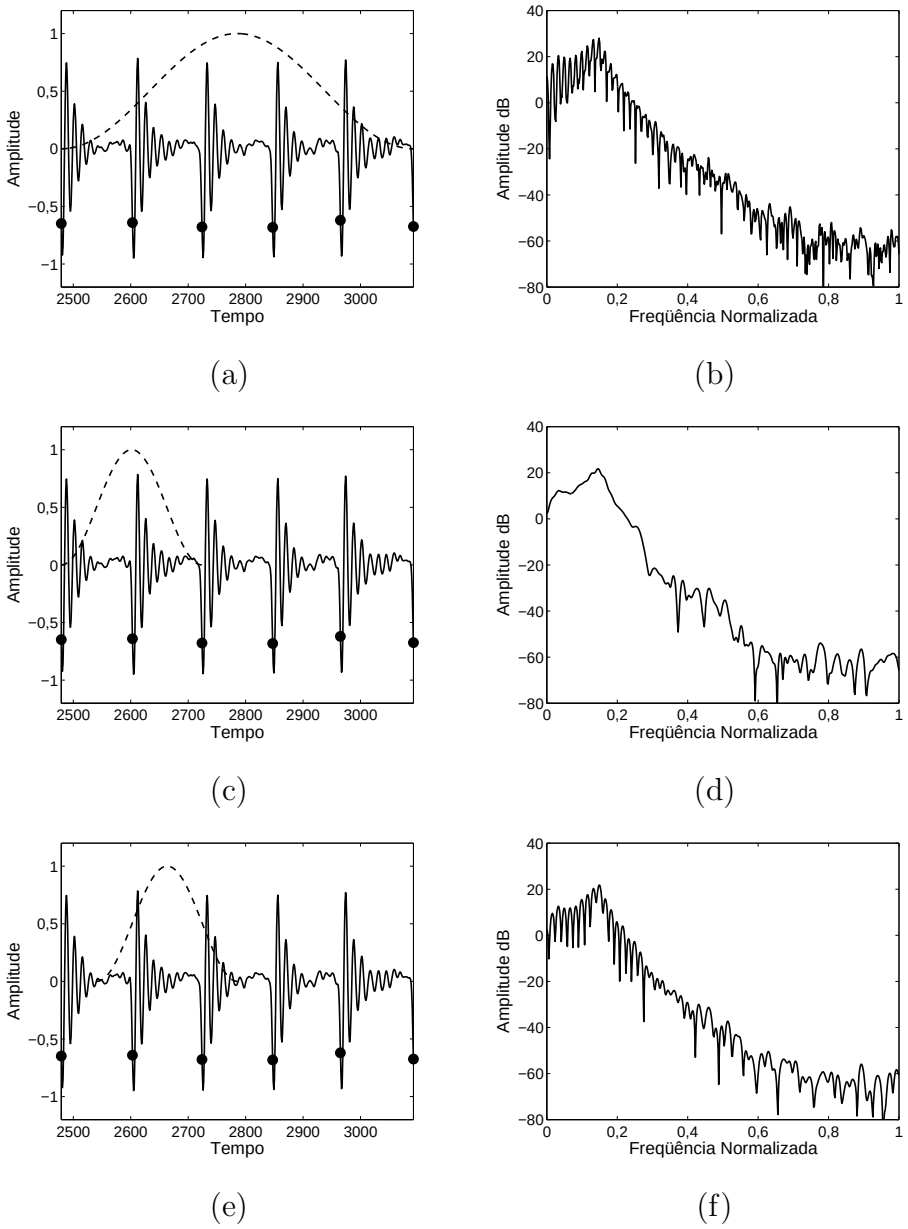


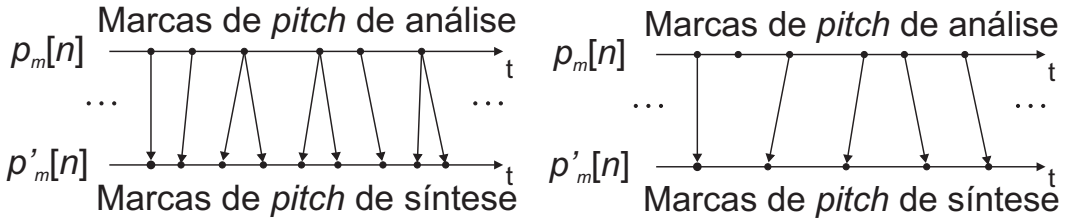
Figura 5.5: Exemplo ilustrando o efeito do janelamento de um sinal $s[n]$, onde: (a), (c) e (e) mostram o sinal de teste (linha cheia) com as janelas que foram usadas para segmentação (linha tracejada); (b), (d) e (f) mostram o espectro de segmentos do janelados do sinal de teste de acordo com as janelas em (a), (c) e (e), respectivamente. Foram usadas janelas com: (a) e (b) 5 períodos de *pitch*; (c) e (d) 2 períodos de *pitch* aplicada de maneira síncrona aos instantes de fechamento glotal; (e) e (f) 2 períodos de *pitch* aplicada de maneira assíncrona.

Esse resultado indica que para a decomposição em segmentos no PSOLA deve ser usada uma janela com resolução freqüencial relativamente alta, de forma

a preservar as características do sinal de voz, embora ao mesmo tempo a resolução não deva ser tão boa a ponto de aproximar o período de *pitch* do sinal original. O resultado das Figuras 5.5(d) e (e) também mostra a importância de se obterem marcas de *pitch* confiáveis, uma vez que ao aplicar uma janela que não é síncrona ao *pitch*, o segmento s_m pode carregar a informação de *pitch* do sinal original. Caso esses efeitos ocorram, o sinal modificado pode apresentar um efeito de 'rouquidão'.

O primeiro passo para aplicar o algoritmo PSOLA é determinar as marcas de *pitch* do sinal $s[n]$ que correspondem aos instantes de fechamento glotal em sinais de voz. As técnicas usadas para determinação dos instantes de fechamento glotal foram discutidas na Seção 5.2. No segundo passo, é necessário estabelecer as marcas de *pitch* do sinal a ser sintetizado p'_m, que são obtidas de acordo com o período de *pitch* desejado $p'[n]$ pelo algoritmo da Figura 5.3 (a).

Uma vez obtidas as marcas de *pitch* de análise p_m e síntese p'_m, é necessário achar correspondências entre elas, que vão ser usadas para o estágio de síntese final. As correspondências entre p_m e p'_m são relativamente simples de se obter no caso de modificação de *pitch*, sendo necessário apenas achar a marca de *pitch* de análise que mais se aproxime de cada marca de *pitch* de síntese. A Figura 5.6 mostra um exemplo com marcas de *pitch* de análise e síntese, e as suas correspondências representadas por setas.



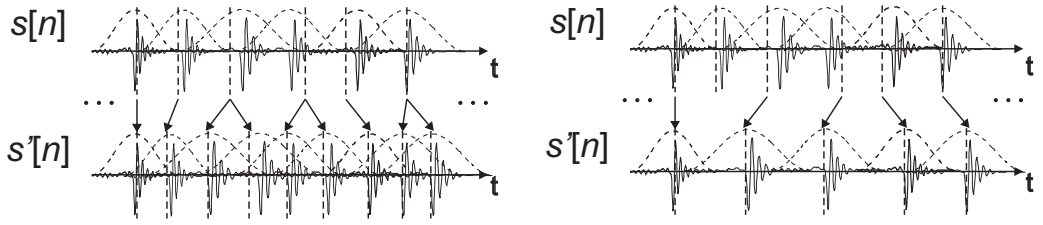
(a) (b)

Figura 5.6: Correspondência entre marcas de *pitch* de análise e síntese para (a) $p'[n] < p[n]$ e (b) $p'[n] > p[n]$.

No estágio de síntese final, os segmentos decompostos s_m devem ser somados no sinal modificado de acordo com as posições que correspondem às novas marcas de *pitch* p'_m, de acordo com

$$s'[n + p'_m + k] = s'[n + p'_m + k] + s'_m[k], \tag{5.4}$$

onde s'_m é um segmento s_m com correspondência $p_m \rightarrow p'_m$, e k varia na região de suporte da janela usada para segmentar o bloco, h_m. A Figura 5.7 mostra de forma esquemática esse procedimento.



(a) (b)

Figura 5.7: Exemplo ilustrativo de modificação de *pitch* usando o TD-PSOLA: (a) $p'[n] < p[n]$; (b) $p'[n] > p[n]$.

5.3.2 LP-PSOLA

Uma alternativa para a modificação de *pitch*/tempo é combinar o esquema de análise/síntese da Figura 5.1 com a técnica do PSOLA. Neste esquema, mostrado na Figura 5.8, um preditor é usado para modelar a envoltória espectral do sinal de entrada $s[n]$, e o PSOLA é usado para modificar o *pitch* ou tempo do sinal de excitação $e[n]$. Já que o erro de predição é o melhor sinal para reconstruir o sinal $s[n]$ usando o filtro do modelo LPC, esse esquema tem vantagens com relação ao sistema apresentado na Seção 5.2. O uso do erro de predição também apresenta vantagens, porque caso a ordem do modelo LPC não seja grande o suficiente para modelar a envoltória espectral de $s[n]$, a informação que falta modelar será preservada no sinal $e[n]$.

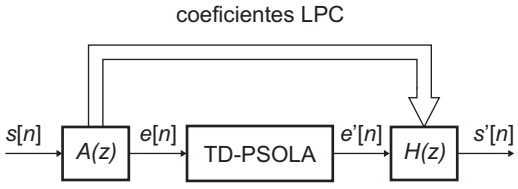
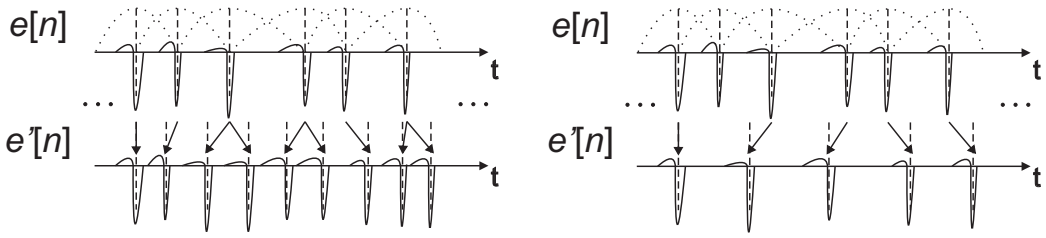


Figura 5.8: Modificação de *pitch*/tempo usando o LP-PSOLA.

O esquema de modificação de $e[n]$ segue os mesmos passos que o TD-PSOLA para o sinal $s[n]$. A Figura 5.9 ilustra como os segmentos do sinal $e[n]$ são decompos-

tos, e depois é feita a sobreposição e adição para a síntese final do sinal modificado $s'[n]$.



(a) (b)

Figura 5.9: Exemplo ilustrativo de modificação de *pitch* de $e[n]$ usando o LP-PSOLA: (a) $p'[n] < p[n]$; (b) $p'[n] > p[n]$.

5.4 Resultados experimentais

A seguir são mostrados alguns resultados de dois dos sistemas descritos acima[1]. Os sinais de teste usados para esta seção são todos sinais de voz cantada, de boa qualidade, amostrados a 44,1 kHz, 16 bits.

Nos dois sistemas que foram implementados foi usado o filtro adaptativo RLS para obter o modelo LPC de forma seqüencial, conforme descrito na Seção 3.2.2, com 30 coeficientes e fator de esquecimento igual a $\lambda = 0{,}9999$, cuja escolha foi feita com base em testes auditivos informais. A escolha de um fator tão próximo de 1 pode ser justificada, uma vez que os sinais musicais usados possuem maior grau de estacionaridade, mantendo suas características estatisticas por períodos mais longos, do que a voz falada. Os sinais modificados foram gerados com um fator de modificação de *pitch* constante ao longo do sinal, e foi de $\beta[n] = \frac{1}{2}$ ou $\beta[n] = 2$, que equivale em termos musicais a uma transposição para uma oitava mais aguda ou mais grave, respectivamente.

5.4.1 Resultados usando LPC

Para testar o sistema da Seção 5.2, foi usada a gravação de uma voz masculina cantada. Os resultados das Figuras 5.10(c) e 5.10(d) foram gerados com

[1]Exemplos disponíveis em http://www.lps.ufrj.br/∼rcdpaiva/mest/modpitch/

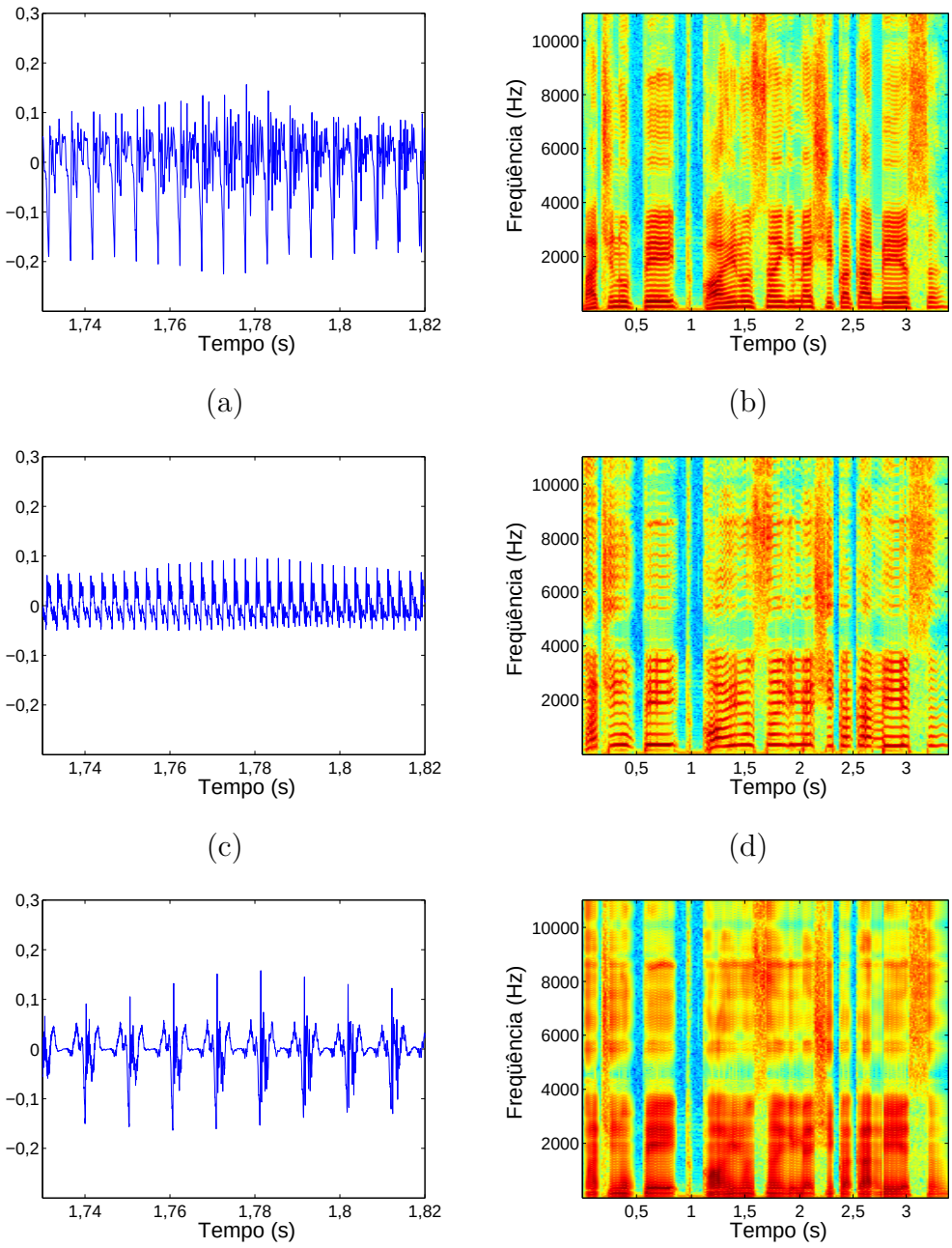


Figura 5.10: Trecho (a) do sinal original, e dos sinais modificados com (c) $\beta[n] = \frac{1}{2}$ e (e) $\beta[n] = 2$; (b), (d) e (f) Espectrogramas dos sinais (a), (c) e (e), respectivamente.

$\beta[n] = \frac{1}{2}$, e os resultados das Figuras 5.10(e) e 5.10(f) foram gerados com $\beta[n] = 2$. As Figuras 5.10(a), 5.10(c) e 5.10(e) mostram trechos dos sinais 5.10(a) original e 5.10(c) e 5.10(e) modificados; as Figuras 5.10(b), 5.10(d) e 5.10(f) mostram os espectrogramas dos sinais das Figuras 5.10(a), 5.10(c) e 5.10(e), respectivamente.

Os espectrogramas dos sinais modificados na Figura 5.10 mostram que a

envoltória espectral do sinal original foi mantida nos sinais modificados. Isso indica que as posições dos formantes nos sinais modificados se mantiveram, de acordo com o esperado.

5.4.2 Resultados usando LP-PSOLA

Para testar o sistema da Seção 5.3.2 foram usadas duas gravações. O primeiro sinal consiste da gravação de uma voz masculina cantada. O cantor da gravação é um barítono, cantor profissional de ópera. Os resultados para essa gravação são mostrados na Figura 5.11.

O segundo sinal consiste da gravação de uma voz feminina cantada, com mais trechos de surdos que a primeira gravação. A cantora da gravação é uma cantora profissional de *pop-rock*. Os resultados para essa gravação são mostrados na Figura 5.12.

O fator de modificação de *pitch* usado para gerar o resultados das Figuras 5.11 e 5.12 foram (c) e (d) $\beta[n] = \frac{1}{2}$, (e) e (f) $\beta[n] = 2$. As Figuras 5.11 e 5.12(a), (c) e (e) mostram trechos dos sinais (a) original e (c) e (e) modificados; as Figuras 5.11 e 5.12(b), (d) e (f) mostram os espectrogramas dos sinais (a), (c) e (e), respectivamente.

Os espectrogramas dos sinais modificados nas Figuras 5.11 e 5.12 mostram que a envoltória espectral do sinal original foi mantida nos sinais modificados. Como foi dito para os resultados na Seção 5.4.1, isso indica que as posições dos formantes nos sinais modificados se mantiveram, de acordo com o esperado.

O resultado de testes auditivos informais revelaram que o LP-PSOLA apresentado na Seção 5.3.1 tem resultados bastante superiores aos do sistema usando modelo simples de excitação glotal apresentado na Seção 5.2. Um ponto de superiodidade do LP-PSOLA é que, por usar o sinal do erro de predição como excitação para o modelo LPC de síntese, ele mantém a informação da envoltória espectral no caso de trechos onde a ordem do modelo seja insuficiente. Outro motivo é que o sistema da Seção 5.2 usa um único modelo de pulso glotal para o sinal inteiro, e isto pode levar a falta de naturalidade em alguns trechos do sinal sintetizado.

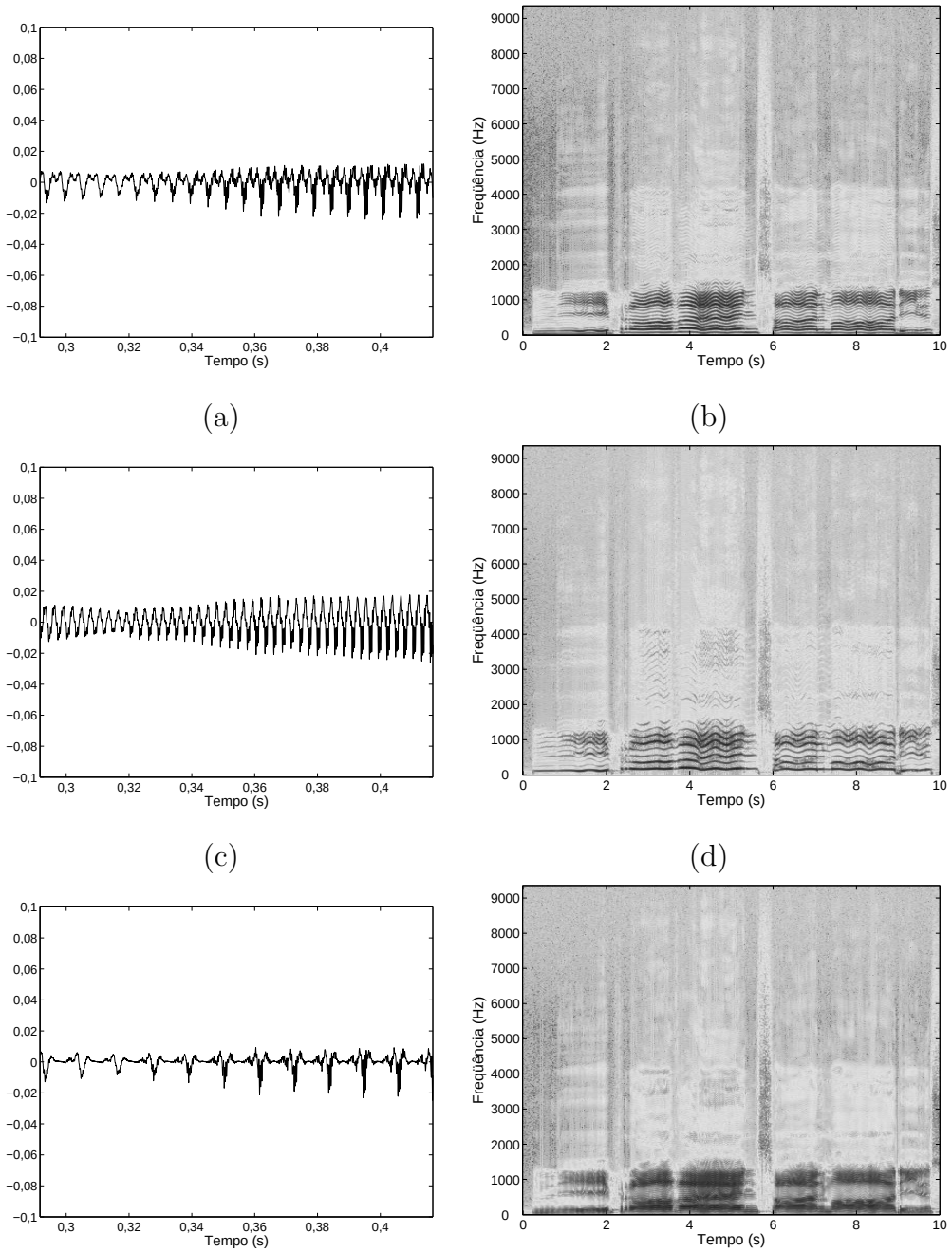


Figura 5.11: Resultados do LP-PSOLA para voz masculina. (a) e (b) sinal original; (b) e (c) sinal modificado com $\beta[n] = \frac{1}{2}$; (b) e (c) sinal modificado com $\beta[n] = \frac{1}{2}$; (a), (c) e (e) Trechos dos sinais; (a), (c) e (e) Espectrogramas dos sinais (a), (c) e (e), respectivamente.

5.4.3 Comparação entre TD-PSOLA e LP-PSOLA

A Figura 5.13 ilustra a diferença entre o TD e LP-PSOLA, para fatores de modificação de *pitch* grandes. Um sinal de voz feminina foi modificado com $\beta = 2$.

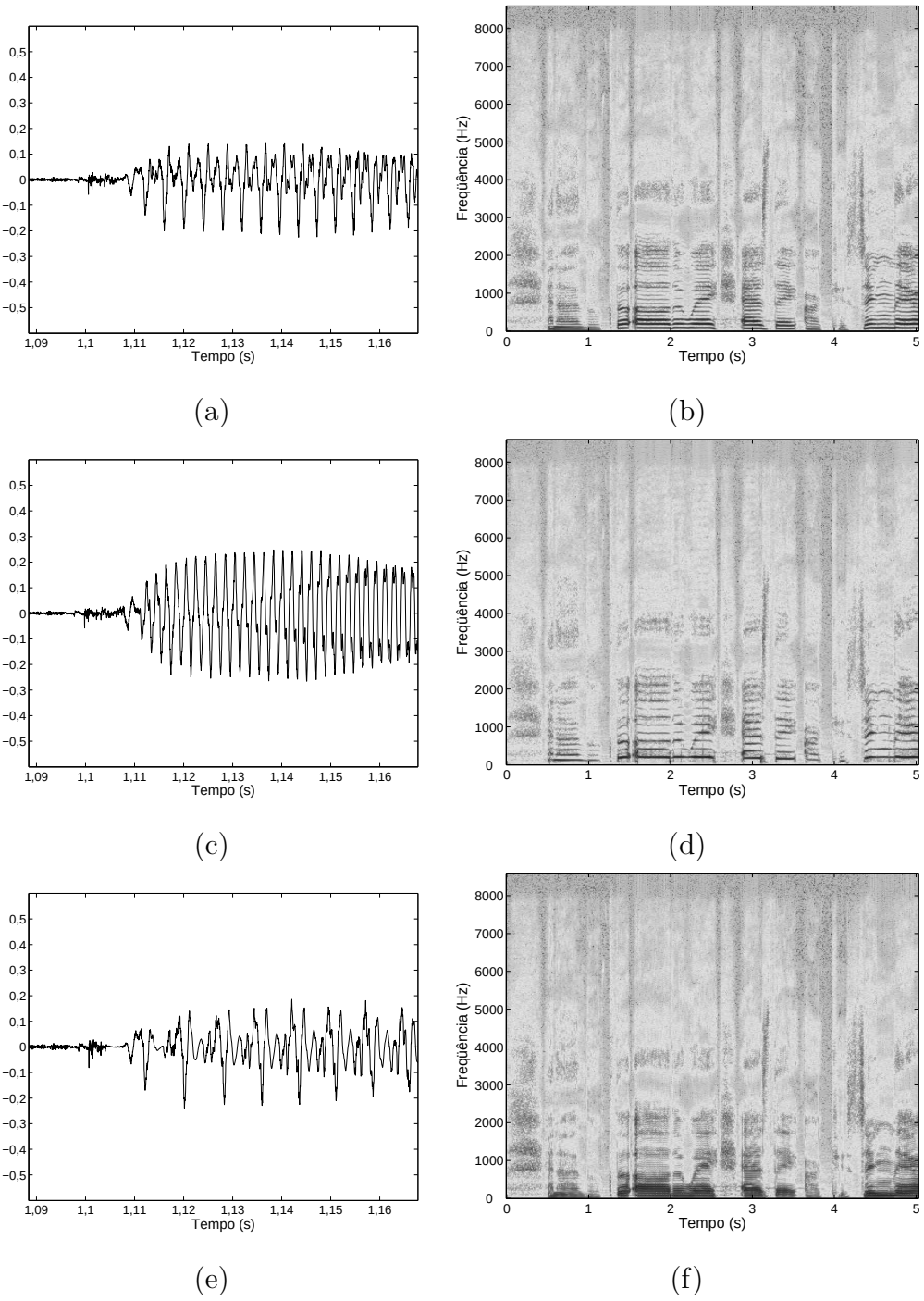
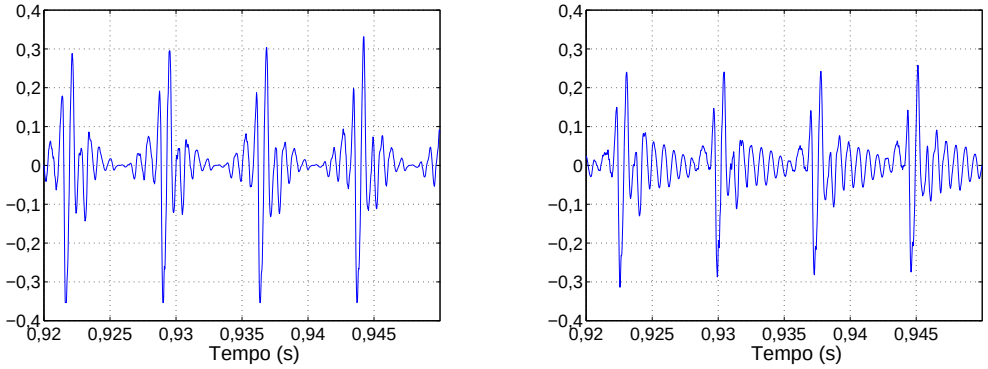


Figura 5.12: Resultados do LP-PSOLA para voz feminina. (a) e (b) sinal original; (b) e (c) sinal modificado com $\beta[n] = \frac{1}{2}$; (b) e (c) sinal modificado com $\beta[n] = \frac{1}{2}$; (a), (c) e (e) Trechos dos sinais; (a), (c) e (e) Espectrogramas dos sinais (a), (c) e (e), respectivamente.

No resultado desta figura, o sinal modificado usando TD-PSOLA apresenta uma diminuição da energia do sinal entre 2 marcas de *pitch* de síntese consecutivas. No

cado do LP-PSOLA, o sinal modificado apresenta uma interpolação entre marcas de *pitch* consecutivas, evitando assim a descontinuidade de energia presente na Figura 5.13(a). A diferença entre os resultados é devida ao modelo IIR de filtro usado para filtrar o sinal de excitação modificado. As diferenças de energia entre marcas de *pitch* consecutivas podem levar a sinais modificados com 'rouquidão'.



(a) (b)

Figura 5.13: Sinal com *pitch* modificado com $\beta = 2$ usando: (a)TD-PSOLA; (b) LP-PSOLA com RLS.

5.5 Conclusão

Neste capítulo foram apresentados métodos para modificação de *pitch* e tempo de sinais de voz. Entre os métodos apresentados estão algoritmos que foram previamente publicados pelo autor deste texto, um deles usando o modelo LPC com solução seqüencial e modelo de pulso glotal como um simples pulso do resíduo de predição [23], e outro método que mistura a técnica PSOLA com o sistema anterior [24]. Para o entendimento adequado do método usando o LP-PSOLA, o TD-PSOLA foi apresentado, e uma breve discussão foi feita sobre esse tema.

Os resultados experimentais mostraram que é possível fazer modificação de *pitch* e tempo com alta qualidade, preservando as características do trato vocal, usando o modelo LPC seqüencial apresentado na seção 3.2.2. O modelo seqüencial foi aplicado nestes algoritmos por levar a estimativas da envoltória espectral que variam suavemente, como foi mostrado na Seção 3.2.3. Sendo assim, os arquivos sintetizados apresentaram boa qualidade, sem transições anormais. A qualidade dos

sinais modificados usando o LP-PSOLA da Seção 5.3.2 é bastante superior à dos sinais modificados usando o sistema da Seção 5.2. A qualidade inferior do sistema da Seção 5.2 pode ser atribuída à escolha do modelo de excitação glotal inadequado para todos os instantes de tempo. Uma melhoria significativa poderia ser feita caso o modelo de pulso glotal fosse variável, sendo uma amostra do resíduo de predição tomada a cada 20∼50 ms.

Os sistemas das Seções 5.2 e 3.2.2 foram projetados para a diminuição do atraso quando a modificação de *pitch* é feita em tempo real. Para isso, o sistema da Seção 5.2 apresenta uma vantagem sobre o sistema da Seção 5.3.2, uma vez que, não usando o PSOLA, não precisa de marcas de *pitch* muito precisas — pode-se dizer que a maior dificuldade do sistema LP-PSOLA é a obtenção de marcas de *pitch* em tempo real.

Ainda foram mostradas as vantagens do LP-PSOLA sobre o TD-PSOLA na Seção 5.4.3 quando é usado um fator de modificação de *pitch* grande. Ainda podemos acrescentar que o LP-PSOLA apresenta maior flexibilidade de extensão para outras aplicações. É possível modificar as características do trato vocal modificando o modelo LPC, por exemplo, para transformar a voz de alguém em outra pessoa, como será verificado no próximo capítulo desta dissertação.

Capítulo 6

Transformação de locutor

6.1 Introdução

Sistemas de transformação de locutor, *voice morphing* ou *voice conversion* na literatura técnica, têm sido objeto de estudo de diversos trabalhos. Entre as inúmeras aplicações desse tipo de sistema, uma das mais citadas é a adaptação de sistemas de síntese de voz para um novo locutor. Sistemas de síntese de voz ou sistemas de transformação texto/fala (*text-to-speech*, TTS) precisam de uma base de dados bastante grande do locutor cuja voz vai ser sintetizada, que deve ser segmentada em unidades de síntese. A base de dados para síntese é, portanto, de custo elevado, demandando um extenso tempo de trabalho. Uma alternativa para gerar um sistema TTS com vários locutores é se basear somente em um locutor, e adaptá-lo para outros locutores. Assim é possível usar somente um banco de dados grande, para a síntese do locutor principal, e usar um sistema de transformação de voz com bases de dados menores e muito mais simples de construir [77, 78, 79]. A redução de banco de dados também pode ser conseguida em sistemas de reconhecimento de voz que utilizam técnicas de redução ou adaptação de locutor [80]. Em ambientes virtuais, onde usuários criam uma nova identidade para interagir com outros usuários (os *avatars*), é interessante o uso de transformação de locutor para modificar a voz do usuário original em uma outra voz [81]. Outra aplicação de sistemas de transformação de locutor é a de aumento de qualidade e inteligibilidade da voz de pessoas com deficiência de fala, como pacientes que foram submetidos a cirurgia de laringectomia [10, 82]. Aplicações de transformação de locutor têm se

mostrado comercialmente interessantes, e já existem muitas patentes referentes a esse tipo de sistema [79, 80, 81, 83, 84].

Entre os sistemas de transformação de locutor, encontramos duas classes principais de transformação. A primeira consiste em soluções que fazem transformação de um locutor (ou cantor) em outro com conhecimento prévio da frase a ser transformada. Esse tipo de sistema tem como aplicação sistemas de *Karaokê* em que a pessoa que está cantando assume o timbre de um outro cantor (possivelmente com a voz mais parecida com a de um cantor famoso) [85]. A dificuldade desse tipo de sistema é fazer, em tempo real, o alinhamento temporal entre uma frase, que está sendo cantada no momento do processamento, e outra, gravada previamente. Para isso esse tipo de sistema deve usar esquemas como *dynamic time warping* (DTW) para alinhamento entre duas frases [2] ou esquemas mais complexos com modelo de Markov escondido (*hidden Markov model*, HMM) para alinhamento texto/fala [86, 87]. Existem também sistemas mais simples que não fazem o alinhamento temporal, pois consideram que a pessoa que está cantando em tempo real canta no mesmo ritmo que o cantor original [84].

O segundo tipo de sistema de transformação de sinais de voz é treinado previamente, com uma base de dados de dois locutores, de forma que a transformação de locutor possa ser feita independentemente da frase que está sendo emitida. Para isso é necessária uma base de dados com frases que contenham todos os fonemas que são emitidos pelos dois locutores, que devem ser alinhadas, para poderem ser treinadas. Os métodos para fazer a conversão de locutores incluem uso de quantização vetorial (*vector quantization*, VQ) [88], clusterização suave usando soma ponderada dos centróides dos *clusters* [89] ou modelo de misturas gaussianas (*Gaussian mixture model*, GMM) [20, 77, 90], análise de componentes principais (*principal component analysis*, PCA) [91] e transformadas *wavelets* [92, 93]. O sistema de transformação de locutor implementado neste capítulo é deste segundo tipo.

Este capítulo começa apresentando características relevantes para a distinção entre locutores na Seção 6.2. A determinação destas características é importante para que se defina que tipo de parâmetros devem ser modificados para obter uma transformação de locutor de qualidade. Na Seção 6.3 é apresentado de forma simples um esquema genérico de transformação de locutor para dar uma visão geral de como

a transformação é operada. Na Seção 6.4 é mostrado um sistema de transformação simples, usando quantização vetorial. Este sistema é apresentado para mostrar de forma intuitiva por que outros sistemas são superiores. A Seção 6.5 mostra rapidamente como é feita a análise de componentes principais, e qual a sua interpretação quando aplicada a coeficientes cepstrais que serão usados nas seções seguintes. A Seção 6.6 começa apresentando, resumidamente, métodos de clusterização suave, que têm por objetivo resolver os problemas inerentes do método baseado em quantização vetorial. Logo após é mostrado o método de mapeamento dos coeficientes mel-cepstrais usado neste capítulo, que leva em consideração a continuidade da envoltória espectral para sintetizar o sinal modificado. Na Seção 6.7 é apresentado o sistema de transformação de locutor desenvolvido nesta dissertação. Nesta seção são detalhados os sistemas de treinamento e transformação, e é descrita a forma de geração do banco de dados usado para treinamento. Na Seção 6.8 são apresentados alguns resultados, e são discutidos efeitos da variação de alguns parâmetros sobre a qualidade de transformação de locutor.

6.2 Características individuais de locutores

Entre as características que nos permitem distinguir diferentes locutores estão fatores associados com a prosódia e o estilo de expressão — que dependem de aspectos sócio-culturais, dialeto, assim como da comunidade em que o locutor está inserido — e fatores acústicos, que estão relacionados a aspectos físicos e fisiológicos, e que também podem estar relacionados ao estado emocional do locutor. Entre os principais fatores acústicos, podemos citar [94]:

- fatores relativos ao *pitch*:
  - *pitch* médio;
  - padrões tempo/freqüência do *pitch*;
  - variação de *pitch*;
- fatores relativos à excitação glotal:
  - forma de onda de excitação glotal;

  - decaimento espectral;

- fatores relativos ao trato vocal:
  - envoltória espectral;
  - valores absolutos das freqüências dos formantes;
  - padrões tempo/freqüência dos formantes;
  - envoltória espectral média;
  - largura de banda dos formantes.

Como o foco deste capítulo é a transformação de timbre, fatores relacionados à prosódia não serão processados. Os fatores que são usados neste texto são o *pitch* médio e os fatores relativos ao trato vocal.

6.3 Visão geral do esquema de transformação de locutor

No esquema apresentado na Figura 6.1, o sinal de voz do locutor-fonte $s_s[n]$, depois de passar por um filtro de pré-ênfase $H_{\text{pre}}(z)$, é analisado usando um modelo de predição linear, obtendo-se a informação da envoltória espectral dada pelo modelo wLPC $H_s(\hat{z})$, descrito na Seção 3.5.2, e o resíduo de predição que está relacionado à excitação glotal. A parte da transformação consiste em modificar o *pitch* do erro de predição, e aplicar uma função de mapeamento entre o modelo wLPC do locutor original $H_s(\hat{z})$ e o modelo do locutor-alvo $\hat{H}_t(\hat{z})$.

O estágio de modificação de *pitch* usa o TD-PSOLA (*time-domain pitch-synchonous overlap-and-add*) descrito na Seção 5.3.1, e é aplicado de forma que o sinal sintetizado tenha *pitch* médio igual ao *pitch* médio do locutor-alvo. Para isso é aplicado um fator de modificação de *pitch* conforme a equação

$$\beta_{s,t} = \frac{\overline{p}_t}{\overline{p}_s}, \tag{6.1}$$

onde $\overline{p}_t$ e $\overline{p}_s$ são os períodos de *pitch* médio do locutor-alvo e locutor-fonte, respectivamente. A modificação de *pitch* por um fator constante é importante para que seja mantida uma relação geométrica do *pitch* de trechos diferentes e, conseqüentemente,

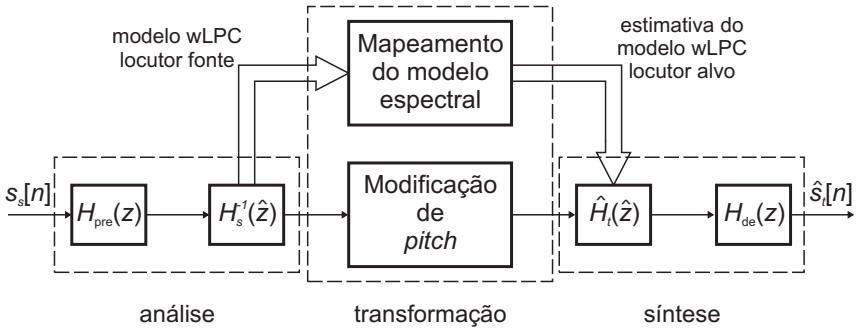


Figura 6.1: Sistema de transformação de locutor com: bloco de análise, que envolve pré-ênfase, determinação do modelo wLPC; bloco de transformação, que envolve mapeamento do modelo wLPC do locutor-fonte no locutor-alvo, e modificação de *pitch*; bloco de síntese, que envolve utilização do modelo wLPC do locutor-alvo, e de-ênfase.

para que a mesma linha melódica seja mantida. Supondo uma melodia com 3 notas, com freqüências f_o, $f_o 2^{\frac{4}{12}}$ e $f_o 2^{\frac{-5}{12}}$, caso o *pitch* dessa melodia seja modificado com um fator constante β a melodia terá notas com freqüências f'_o, $f'_o 2^{\frac{4}{12}}$ e $f'_o 2^{\frac{-5}{12}}$, onde $f'_o = \frac{f_o}{\beta}$, e a razão entre as freqüências que compõem a melodia não é modificada, logo percebemos a mesma melodia em uma tonalidade diferente. Outras opções são possíveis para a escolha de $\beta_{s,t}$. Considerando que a melodia do cantor-fonte esteja esteja afinada de acordo com a escala temperada, a melodia com *pitch* modificado continua nesta escala se

$$\beta_{s,t} \approx 2^{\frac{k}{12}}, \tag{6.2}$$

sendo k um número inteiro. Outra possibilidade é quando o cantor-fonte canta com um acompanhamento musical que não será modificado, logo o sinal modificado com o timbre do cantor-alvo não deve estar em outra tonalidade e $\beta_{s,t}$ deve ser aproximado para modificação em oitavas, ou seja

$$\beta_{s,t} \approx 2^k, \tag{6.3}$$

sendo k novamente um número inteiro. O fator utilizado ao longo deste trabalho é dado pela equação (6.1).

O mapeamento do modelo wLPC usa um estágio de representação intermediário, onde são calculados os coeficientes mel-cepstrais do locutor original $\mathbf{\Theta_s}$

e do locutor-alvo $\mathbf{\Theta_t}$ a partir dos coeficientes wLPC, de acordo com a Seção 3.4.1. O objetivo deste estágio de representação intermediária é obter coeficientes cuja medida de distância seja mais significativa para a representação de diferentes fonemas, e neste sentido os coeficientes mel-cepstrais se mostram bastante superiores aos coeficientes wLPC.

Como último estágio de transformação, o sinal do erro de predição com *pitch* modificado é processado por $\hat{H}_t(\hat{z})$ e pelo filtro de de-ênfase $H_{\text{de}}(z)$ para gerar o sinal modificado $\hat{s}_t[n]$. Este último estágio garante que a modificação de *pitch* terá as mesmas vantagens do LP-PSOLA (*linear-prediction pitch-synchonous overlap-and-add*) descritas na Seção 5.4.3.

A importância dos filtros de pré- e de-ênfase devem ser destacadas. O filtro de pré-ênfase normalmente é um filtro do tipo passa altas $H_{\text{pre}}(z) = 1 - \alpha z^{-1}$. Como o decaimento espectral decorrente da excitação glotal tem característica passa-baixas, esse filtro tende a inverter a resposta em freqüência da excitação, e por isso permite uma melhor estimativa dos formantes da voz. O filtro de de-ênfase é simplesmente um filtro IIR, que é o inverso do filtro de pré-ênfase $H_{\text{de}}(z) = H_{\text{pre}}^{-1}(z)$.

6.4 Quantização vetorial

Entre as técnicas mais simples de transformação de locutor estão as baseadas em quantização vetorial [88]. Neste tipo de algoritmo, o conjunto de vetores de representação espectral $\mathbf{\Theta}$ — no caso deste trabalho, coeficientes mel-cepstrais — de cada locutor é agrupado em Q grupos, e cada vetor de coeficientes é aproximado por um valor correspondente ao 'centro' de cada grupo.

O agrupamento dos vetores de representação espectral é feito usando um algoritmo de classificação não supervisionada, *K-means* [58], que encontra centróides μ_i em torno dos quais os vetores de representação espectral se agrupam em classes $\mathcal{C}_{s,i}$ e $\mathcal{C}_{t,i}$, para as bases de dados do locutor-fonte e do locutor-alvo respectivamente. A classificação nas classes $\mathcal{C}_i$ é feita de forma que

$$\mathbf{\Theta}_m \in \mathcal{C}_i, \quad \min_l D\left(\mathbf{\Theta}_m, \mu_l\right) = i, \tag{6.4}$$

onde $D\left(\mathbf{\Theta}_m, \mu_l\right)$ é uma medida de distância entre o vetor de representação espectral $\mathbf{\Theta}_m$ no bloco m, e o centróide da l-ésima classe μ_l.

Na segunda parte do treinamento é preciso fazer o alinhamento temporal entre as frases dos locutores, para que se consiga achar a correspondência entre as classes dos locutores fonte e alvo. A correspondência entre classes dos dois locutores $i \rightarrow j$ é estabelecida quando a maior parte dos blocos do locutor-fonte pertencentes à classe $\mathcal{C}_{s,i}$ estão alinhados a blocos do locutor-alvo pertencentes à classe $\mathcal{C}_{t,j}$. Desta forma o mapeamento é feito de maneira abrupta, uma vez que os vetores de coeficientes mel-cepstrais são sempre aproximados pelos centróides de cada classe.

6.5 Análise de componentes principais

Análise de componentes principais (*principal-component analysis*, PCA) tem sido usada em reconhecimento de voz, e em adaptação de sistemas de reconhecimento e síntese de voz, normalmente em conjunto com técnicas baseadas em HMM [95]. A idéia central de PCA é reduzir a dimensão de um conjunto de dados cujas variáveis estejam relacionadas [96, 97].

Seja $\mathbf{x}$ um vetor de dimensão P, cujas componentes sejam variáveis aleatórias, com matriz de covariância $\mathbf{\Sigma}$. As componentes principais de $\mathbf{x}$ são dadas por combinações lineares dos elementos de $\mathbf{x}$, que podem ser representadas pelo produto interno $\tilde{x}_j = \boldsymbol{\alpha}_j^T \mathbf{x}$, onde $\boldsymbol{\alpha}_j$ é o vetor de mapeamento na j-ésima componente principal. A primeira componente principal pode ser encontrada obtendo-se $\boldsymbol{\alpha}_1$ que maximiza a variância de $\tilde{x}_1 = \boldsymbol{\alpha}_1^T \mathbf{x}$, $\mathrm{var}\left[\tilde{x}_1\right] = \boldsymbol{\alpha}_1^T \mathbf{\Sigma} \boldsymbol{\alpha}_1$, com a restrição de que $\left|\boldsymbol{\alpha}_1\right|^2 = \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = 1$. Para isso, pode-se usar um multiplicador de Lagrange e maximizar

$$\boldsymbol{\alpha}_1^T \mathbf{\Sigma} \boldsymbol{\alpha}_1 - \lambda \left(\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 - 1 \right). \tag{6.5}$$

Diferenciando a equação (6.5) em relação a $\boldsymbol{\alpha}_i$ e igualando o resultado a zero, encontramos

$$\mathbf{\Sigma} \boldsymbol{\alpha}_1 - \lambda \boldsymbol{\alpha}_1 = \left(\mathbf{\Sigma} - \lambda \mathbf{I}_P \right) \boldsymbol{\alpha}_1 = 0, \tag{6.6}$$

onde $\mathbf{I}_P$ é uma matriz identidade de ordem P. Assim, $\boldsymbol{\alpha}_1$ é um autovetor e λ é o autovalor correspontende da matriz $\mathbf{\Sigma}$. As outras componentes principais são encontradas de maneira similar, com a restrição de que as compomentes principais $\tilde{x}_j$ sejam descorrelacionadas entre si. Uma maneira de restringir a solução de forma que

isso aconteça é fazer com que os vetores de mapeamento nas componentes principais sejam ortogonais, ou seja,

$$\boldsymbol{\alpha}_i^T \boldsymbol{\alpha}_j = 0, \quad i \neq j. \tag{6.7}$$

Dada esta restrição, podemos encontrar as outras componentes principais, de maneira similar à primeira, e o resultado é que as outras componentes principais também são dadas pelos autovetores da matriz de auto-correlação $\boldsymbol{\Sigma}$. A variância de cada componente principal é obtida da forma

$$\mathrm{E}\left[\left(\tilde{x}_j - \overline{\tilde{x}}_j\right)^2\right] = \lambda_j, \tag{6.8}$$

onde λ_j é o autovalor correspondente ao j-ésimo autovetor α_j. Os autovetores são então ordenados de forma decrescente, tais que

$$\lambda_i < \lambda_j, \quad i > j. \tag{6.9}$$

Sendo assim, podemos decompor um vetor $\mathbf{x}$ de dimensão P em p componentes principais, na forma

$$\tilde{\mathbf{x}}_p = \mathbf{A}_p^T \mathbf{x}, \tag{6.10}$$

onde $p \leq P$,

$$\tilde{\mathbf{x}}_p = \left[\begin{array}{cccc} \tilde{x}_1 & \tilde{x}_2 & \ldots & \tilde{x}_p \end{array} \right]^T, \tag{6.11}$$

e

$$\mathbf{A}_p = \left[\begin{array}{cccc} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \ldots & \boldsymbol{\alpha}_p \end{array} \right] \tag{6.12}$$

e $\mathbf{x}$ pode ser reconstruído a partir das componentes principais:

$$\hat{\mathbf{x}} = \mathbf{A}_p \tilde{\mathbf{x}}_p. \tag{6.13}$$

6.5.1 Componentes principais do cepstro

Usando PCA sobre os coeficientes cepstrais, ou mel-cepstrais, é possível conseguir uma redução de dimensionalidade significativa, o que pode ser útil no caso de reconhecimento de voz [32] e transformação de locutor. Como na sua definição o

cepstro inclui a operação de logaritmo, a reconstrução dos coeficientes cepstrais $\boldsymbol{\Theta}$, pode ser entendida como

$$\hat{\boldsymbol{\Theta}} = \mathbf{A}_p \tilde{\boldsymbol{\Theta}}_p = \sum_{j=1}^{p} \tilde{\Theta}_j \boldsymbol{\alpha}_j, \tag{6.14}$$

cuja transformada inversa é dada por

$$\begin{aligned} \mathrm{e}^{\mathcal{F}(\hat{\boldsymbol{\Theta}})} &= \mathrm{e}^{\mathcal{F}\left(\sum_{j=1}^{p} \tilde{\Theta}_j \boldsymbol{\alpha}_j\right)} \\ &= \prod_{j=1}^{p} \mathrm{e}^{\mathcal{F}\left(\tilde{\Theta}_j \boldsymbol{\alpha}_j\right)} \\ &= \prod_{j=1}^{p} \mathrm{e}^{\tilde{\Theta}_j \log A_j(\omega)} \\ &= \prod_{j=1}^{p} \mathrm{e}^{\log A_j^{\tilde{\Theta}_j}(\omega)} \\ &= \prod_{j=1}^{p} A_j^{\tilde{\Theta}_j}(\omega), \end{aligned} \tag{6.15}$$

sendo

$$\log A_j(\omega) = \mathcal{F}(\boldsymbol{\alpha}_j) \tag{6.16}$$

um filtro relacionado à j-ésima componente principal de $\hat{\boldsymbol{\Theta}}$.

Desta forma, podemos dizer que o mapeamento dos coeficientes cepstrais, ou mel-cepstrais, de um filtro $H(z)$ nas suas componentes principais pode ser interpretado como a decomposição deste filtro em modos, ou subfiltros, relacionados ao processo de produção de voz:

$$H(z) \approx \prod_{j=1}^{p} A_j^{\tilde{\Theta}_j}(z), \tag{6.17}$$

ou, no caso do mel-cepstro,

$$H(\hat{z}) \approx \prod_{j=1}^{p} A_j^{\tilde{\Theta}_j}(\hat{z}). \tag{6.18}$$

A Figura 6.2 ilustra o resultado da decomposição das equações (6.17) e (6.18). Caso a matriz de mapeamento nas componentes principais seja calculada para um único locutor, esses filtros podem estar relacionados à estrutura física e aos modos de articulação deste locutor.

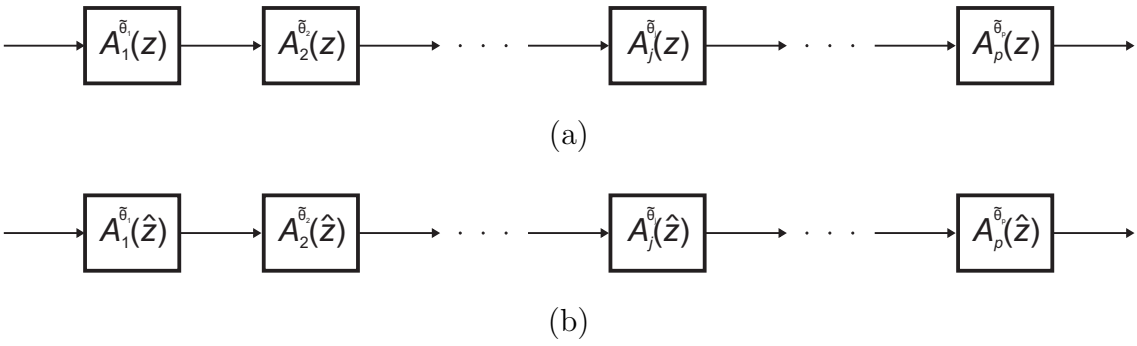


Figura 6.2: Interpretação das componentes principais aplicadas ao (a) cepstro e (b) mel-cepstro como uma decomposição de um filtro $H(z)$, ou $H(\hat{z})$, em subfiltros.

6.6 Clusterização suave

Diversos sistemas de transformação de locutor utilizam o que se chama de clusterização suave (*soft-clustering*), para evitar os efeitos indesejáveis causados pela quantização vetorial (entre eles as transições abruptas entre blocos causadas pela aproximação em centróides). Uma forma de implementar a clusterização suave é fazer o mapeamento entre $\mathbf{\Theta}_s$ e $\mathbf{\Theta}_t$ como sendo uma soma dos centróides $\mu_{t,i}$, encontrados de acordo com a Seção 6.4, ponderada pelo inverso da distância de $\mathbf{\Theta}_s$ aos centróides $\mu_{s,i}$ [89].

Outros sistemas utilizam um modelo de mistura gaussiana (*gaussian mixture model*, GMM) [77], com uma abordagem mais estatística para o problema de conversão de voz. O modelo GMM é dado por

$$p\left(\mathbf{x}\right) = \sum_{i=1}^{g} P\left(\mathcal{C}_i\right) \mathrm{N}\left(\mathbf{x}, \mu_i, \mathbf{\Sigma}_i\right), \tag{6.19}$$

onde g é o número de misturas de gaussianas, $P\left(\mathcal{C}_i\right)$ é a probabilidade *a priori* da classe $\mathcal{C}_i$ e

$$\mathrm{N}\left(\mathbf{x}, \mu, \mathbf{\Sigma}\right) = \frac{1}{\left(2\pi\right)^{\frac{p}{2}} \left|\Sigma\right|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}\left(\mathbf{x}-\mu\right)^T \Sigma^{-1} \left(\mathbf{x}-\mu\right)} \tag{6.20}$$

denota uma distribuição normal com média μ e matrix de covariância $\mathbf{\Sigma}$, onde p é o tamanho do vetor $\mathbf{x}$. Os parâmetros da equação (6.19) podem ser obtidos usando-se um algoritmo *expectation-maximization*, EM [98].

Na abordagem de Stylianou *et al.* [77, 90], é obtido um modelo GMM para o locutor-fonte de forma independente, e depois de se obter o alinhamento entre frases, uma função de mapeamento linear é determinada para cada centróide obtido

no modelo. A função de mapeamento é dada, então, por uma soma do resultado destas funções lineares, ponderadas pela probabilidade *a posteriori* de cada classe $\mathcal{C}_i$. Já na abordagem utilizada por Kain [20], o modelo GMM é obtido para a densidade conjunta dos parâmetros espectrais dos locutores alvo e fonte, e a função de mapeamento é obtida calculando-se o valor esperado de $\mathbf{\Theta}_{t,m}$ dado $\mathbf{\Theta}_{s,m}$.

Como visto na Seção 6.2, entre as características acústicas que nos permitem distinguir entre diferentes locutores está a evolução da envoltória espectral, ou o padrão tempo/freqüência dos formantes. Em sistemas de reconhecimento de voz, é uma prática comum usar o *delta-cepstro* (a diferença entre coeficientes cepstrais dos blocos atual e anterior) [2], como uma forma de adicionar este tipo de informação dinâmica para o sistema de reconhecimento de voz. Para obter uma função de mapeamento que leve em conta a evolução das características espectrais dos locutores, a função de mapeamento proposta por Kain [20] foi estendida para fazer a estimativa de $\mathbf{\Theta}_{t,m}$ no bloco m levando em conta os coeficientes mel-cepstrais do bloco anterior $\mathbf{\Theta}_{t,m-1}$. Ainda, como são necessários muitos coeficientes mel-cepstrais para poder reconstruir de forma satisfatória o modelo wLPC, foram usadas somente as componentes principais do cepstro, de acordo com a Seção 6.5. Desta forma, considerando-se o modelo de mistura de gaussianas

$$p\left(\tilde{\mathbf{\Theta}}_m\right) = \sum_{i=1}^{g} P\left(\mathcal{C}_i\right) \mathrm{N}\left(\tilde{\mathbf{\Theta}}_m, \mu_i, \mathbf{\Sigma}_i\right), \tag{6.21}$$

onde

$$\tilde{\mathbf{\Theta}}_m = \left[\begin{array}{cc} \tilde{\mathbf{\Theta}}_{\mathrm{est},m}^T & \tilde{\mathbf{\Theta}}_{t,m}^T \end{array} \right]^T \tag{6.22}$$

e

$$\tilde{\mathbf{\Theta}}_{\mathrm{est},m} = \left[\begin{array}{ccc} \tilde{\mathbf{\Theta}}_{t,m-1}^T & \tilde{\mathbf{\Theta}}_{s,m}^T & \tilde{\mathbf{\Theta}}_{s,m-1}^T \end{array} \right]^T. \tag{6.23}$$

é obtida a função de mapeamento entre os coeficientes mel-cepstrais do locutor-fonte e locutor-alvo $\mathrm{F}_{st}\left(\widehat{\tilde{\mathbf{\Theta}}}_{t,m-1}, \tilde{\mathbf{\Theta}}_{s,m}, \tilde{\mathbf{\Theta}}_{s,m-1}\right)$ como o valor esperado de $\tilde{\mathbf{\Theta}}_{t,m}$ [98, 99]

$$\begin{aligned} \widehat{\tilde{\mathbf{\Theta}}}_{t,m} &= \mathrm{F}_{st}\left(\widehat{\tilde{\mathbf{\Theta}}}_{t,m-1}, \tilde{\mathbf{\Theta}}_{s,m}, \tilde{\mathbf{\Theta}}_{s,m-1}\right) \\ &= \mathrm{E}\left[\tilde{\mathbf{\Theta}}_{t,m} | \tilde{\mathbf{\Theta}}_{\mathrm{est},m}\right] \\ &= \sum_{i=1}^{g} P\left(\mathcal{C}_i | \tilde{\mathbf{\Theta}}_{\mathrm{est},m}\right) \left[\nu_i + \Gamma_i \tilde{\mathbf{\Theta}}_{\mathrm{est},m}\right], \end{aligned} \tag{6.24}$$

onde $\mathrm{E}\left[a|b\right]$ denota o valor esperado de a condicionado a b,

$$P\left(\mathcal{C}_i|\tilde{\mathbf{\Theta}}_m\right) = \frac{P\left(\mathcal{C}_i\right)\mathrm{N}\left(\tilde{\mathbf{\Theta}}_m, \mu_i, \mathbf{\Sigma}_i\right)}{\sum_{j=1}^{g} P\left(\mathcal{C}_j\right)\mathrm{N}\left(\tilde{\mathbf{\Theta}}_m, \mu_j, \mathbf{\Sigma}_j\right)} \tag{6.25}$$

é a probabilidade *a posteriori* da classe $\mathcal{C}_i$, e

$$\Gamma_i = \mathbf{\Sigma}_i^{t,\mathrm{est}}\left(\mathbf{\Sigma}_i^{\mathrm{est,est}}\right)^{-1} \tag{6.26}$$

$$\nu_i = \mu_i^t - \mathbf{\Sigma}_i^{t,\mathrm{est}}\left(\mathbf{\Sigma}_i^{\mathrm{est,est}}\right)^{-1}\mu_i^{\mathrm{est}} \tag{6.27}$$

$$\mathbf{\Sigma}_i = \begin{bmatrix} \mathbf{\Sigma}_i^{\mathrm{est,est}} & \mathbf{\Sigma}_i^{t,\mathrm{est}} \\ \mathbf{\Sigma}_i^{\mathrm{est},t} & \mathbf{\Sigma}_i^{t,t} \end{bmatrix} \tag{6.28}$$

$$\mu_i = \begin{bmatrix} \mu_i^t \\ \mu_i^{\mathrm{est}} \end{bmatrix}. \tag{6.29}$$

Como somente os trechos sonoros foram processados neste trabalho, e como a função de mapeamento da equação (6.24) utiliza informação do bloco anterior, quando há uma transição entre trechos sonoros e surdos é necessário usar um outro modelo GMM para estimar $\tilde{\mathbf{\Theta}}_{t,m}$. A função de mapeamento é, então, simplificada na forma

$$\begin{aligned} \widehat{\tilde{\mathbf{\Theta}}}_{t,m} &= \bar{\mathrm{F}}_{st}\left(\tilde{\mathbf{\Theta}}_{s,m}\right) \\ &= \mathrm{E}\left[\tilde{\mathbf{\Theta}}_{t,m}|\tilde{\mathbf{\Theta}}_{s,m}\right] \\ &= \sum_{i=1}^{g} P\left(\bar{\mathcal{C}}_i|\tilde{\mathbf{\Theta}}_{s,m}\right)\left[\bar{\nu}_i + \bar{\mathbf{\Gamma}}_i\tilde{\mathbf{\Theta}}_{s,m}\right], \end{aligned} \tag{6.30}$$

onde

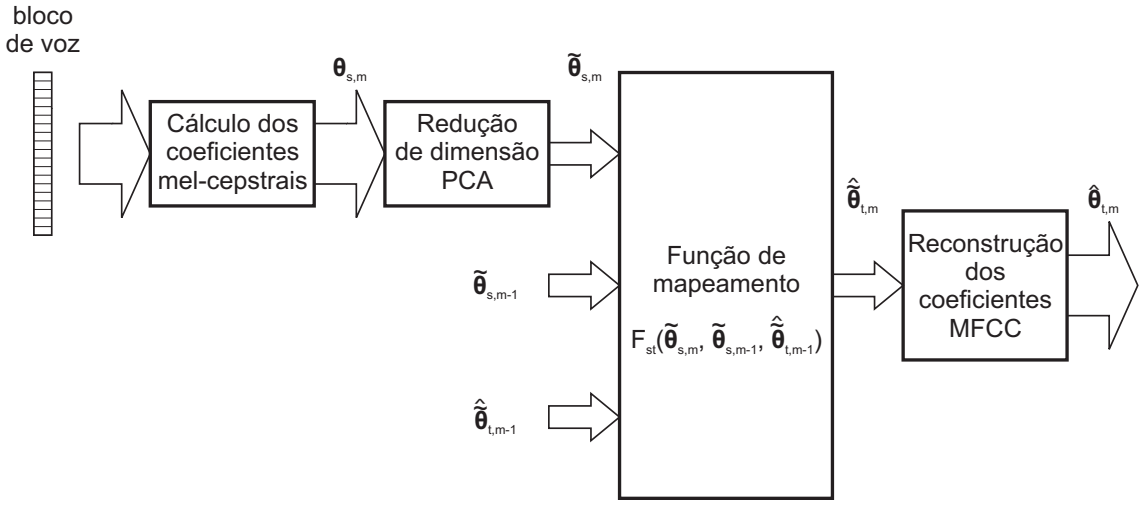
$$\bar{\Gamma}_i = \bar{\mathbf{\Sigma}}_i^{t,s}\left(\bar{\mathbf{\Sigma}}_i^{s,s}\right)^{-1} \tag{6.31}$$

$$\bar{\nu}_i = \bar{\mu}_i^t - \bar{\mathbf{\Sigma}}_i^{t,s}\left(\bar{\mathbf{\Sigma}}_i^{s,s}\right)^{-1}\bar{\mu}_i^s \tag{6.32}$$

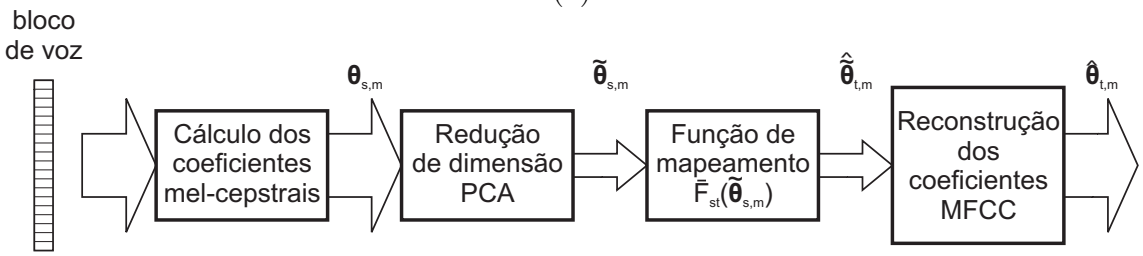
$$\bar{\mathbf{\Sigma}}_i = \begin{bmatrix} \bar{\mathbf{\Sigma}}_i^{s,s} & \bar{\mathbf{\Sigma}}_i^{t,s} \\ \bar{\mathbf{\Sigma}}_i^{s,t} & \bar{\mathbf{\Sigma}}_i^{t,t} \end{bmatrix} \tag{6.33}$$

$$\bar{\mu}_i = \begin{bmatrix} \bar{\mu}_i^t \\ \bar{\mu}_i^s \end{bmatrix}. \tag{6.34}$$

A Figura 6.3 mostra como os coeficientes mel-cepstrais do locutor-fonte $\mathbf{\Theta}_{s,m}$ são transformados quando: (a) os blocos m e $(m-1)$ são sonoros; (b) o bloco m é sonoro e o $(m-1)$ é surdo.



(a)



(b)

Figura 6.3: Diagrama da função de mapeamento entre os coeficientes mel-cepstrais do locutor-fonte e o locutor-alvo quando: (a) os blocos atual m e anterior $(m-1)$ são sonoros; (b) o bloco atual m é sonoro e o anterior $(m-1)$ é surdo. Nestas figuras estão o cálculo dos coeficientes MFCC, o cálculo das componentes principais, a função de mapeamento $\mathrm{F}_{st}\left(\widehat{\tilde{\boldsymbol{\Theta}}}_{t,m-1}, \tilde{\boldsymbol{\Theta}}_{s,m}, \tilde{\boldsymbol{\Theta}}_{s,m-1}\right)$ ou $\bar{\mathrm{F}}_{st}\left(\tilde{\boldsymbol{\Theta}}_{s,m}\right)$ e a reconstrução dos coeficientes MFCC estimados $\widehat{\boldsymbol{\Theta}}_{t,m}$ a partir de suas componentes principais $\widehat{\tilde{\boldsymbol{\Theta}}}_{t,m}$.

6.7 Sistema proposto

6.7.1 Banco de dados

O banco de dados usado neste sistema de transformação de locutor consiste nas gravações de duas vozes masculinas, de cantores profissionais de ópera. O primeiro é um barítono e o segundo é um tenor, portanto o tipo de timbre dos dois cantores é significativamente diferente. Os sinais foram gravados com microfone condensador Shure SM81®[1], com resposta em freqüência praticamente plana entre

[1]http://www.shure.com

20 e 20.000 Hz; um sinal auxiliar do eletroglotógrafo *EGGs for singers*®[2] foi usado para a obtenção de marcas de *pitch*.

Os sinais gravados foram 40 frases foneticamente balanceadas retiradas do trabalho de Alcaim *et al.* [57] (os primeiros 4 grupos de frases do artigo), das quais 30 foram usadas para treinar o sistema (3 primeiros grupos de frases) e 10 foram usadas para teste (quarto grupo de frases). No processo de gravação das frases, os cantores improvisaram a melodia livremente. Dessa forma, a generalidade dos exemplos foi extremamente alta.

6.7.2 Estágio de treinamento

No sistema que foi implementado foram usados os coeficientes mel-cepstrais obtidos a partir dos coeficientes wLPC, de acordo com a Seção 3.5.3. Foram obtidos inicialmente 100 coeficientes cepstrais para cada bloco, sendo os blocos segmentados com 1024 amostras e espaçados de 512 amostras. Foram usados 20 coeficientes wLPC, com $\rho = 0{,}627$ (valor calculado pela equação (3.52) para aproximar a escala mel quando $f_s = 44{,}1\,\text{kHz}$). Foram extraídas as componentes principais dos coeficientes mel-cepstrais individualmente para cada locutor, e usadas 30 componentes para a transformação de locutor.

A Figura 6.4 mostra como foi feito o treinamento da função de transformação (a) na primeira iteração; (b) ao longo do treinamento incremental. No primeiro estágio do treinamento é feito o alinhamento temporal entre os blocos dos locutores fonte e alvo. O alinhamento é feito usando uma implementação de Ellis [100] do algoritmo *dynamic time warping* (DTW), usando 40 coeficientes mel-cepstrais na primeira iteração. Nos estágios de treinamento incremental foram usados vetores

$$\mathbf{X}_m = \begin{bmatrix} \boldsymbol{\Theta}_{x,m} \\ \widehat{\boldsymbol{\Theta}}_{y,m} \\ \boldsymbol{\Theta}_{x,m} \end{bmatrix} \tag{6.35}$$

$$\mathbf{Y}_m = \begin{bmatrix} \boldsymbol{\Theta}_{y,m} \\ \boldsymbol{\Theta}_{x,m} \\ \widehat{\boldsymbol{\Theta}}_{y,m} \end{bmatrix} \tag{6.36}$$

[2] http://www.eggsforsingers.eu

para o alinhamento temporal, onde os vetores $\mathbf{\Theta}_x$ e $\widehat{\mathbf{\Theta}}_x$ são os coeficientes mel-cepstrais do locutor x e do locutor y transformado no locutor x, respectivamente. Aqui se deixou de falar em locutor-fonte e locutor-alvo, pois uma vez obtidas as GMMs, as funções de mapeamento entre os locutores x e y são facilmente obtidas usando os resultados das equações (6.24) e (6.30). O objetivo dos estágios de treinamento incremental é utilizar a informação dos coeficientes mel-cepstrais transformados para melhorar o alinhamento temporal entre as frases dos dois locutores; contudo, não se deixou de usar os coeficientes mel-cepstrais originais para evitar que no caso em que se obtém um modelo errado por causa do alinhamento inicial, o modelo seja polarizado para uma solução errada nas iterações seguintes.

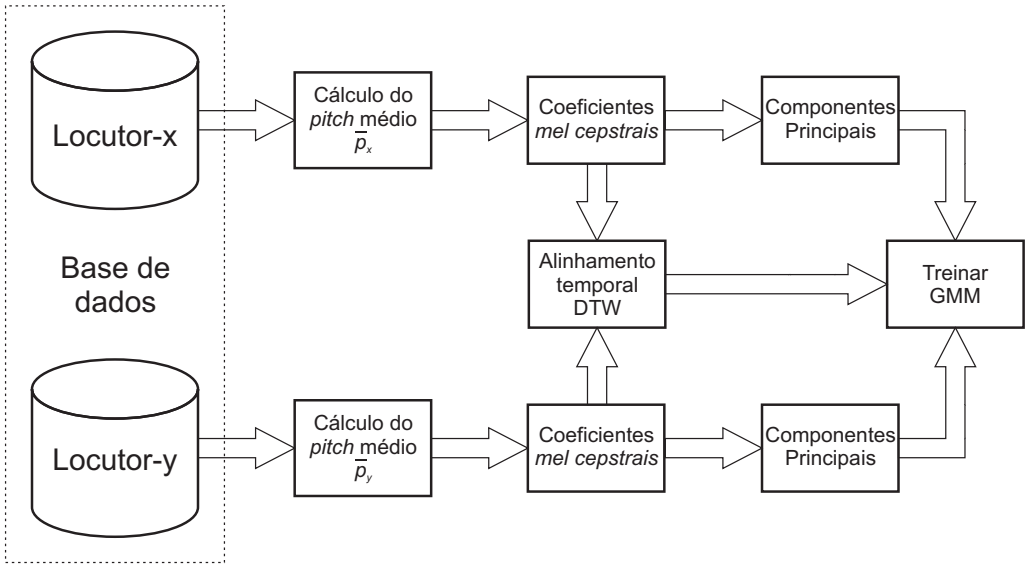
Com os blocos dos locutores alinhados, são treinados os modelos GMM das funções de mapeamento das equações (6.24) e (6.30). Estes modelos foram obtidos usando uma implementação para Matlab® do algoritmo EM disponível no *Statistical Pattern Recognition Toolbox* [101]. Devido a problemas de condicionamento das matrizes do algoritmo, é importante adicionar às matrizes de covariância do modelo após cada iteração uma matriz diagonal $\epsilon\,\mathbf{I}$, onde $\mathbf{I}$ é uma matriz-identidade e $\epsilon \approx 0{,}001$ é uma constante [20].

6.7.3 Estágio transformação

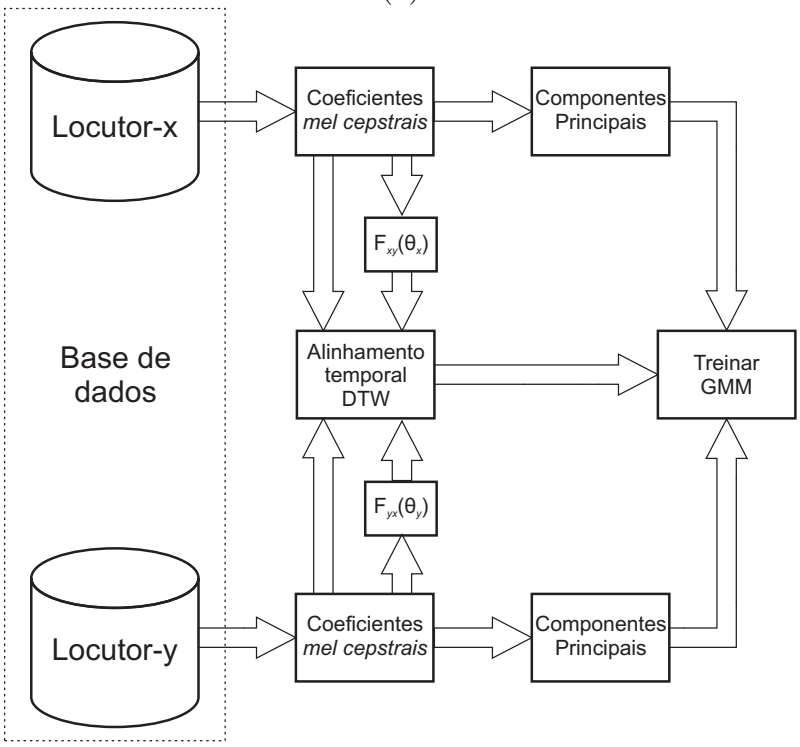
No estágio de síntese, os coeficientes mel-cepstrais calculados pelas funções de transformação das equações (6.24) e (6.30) são usados para obter o modelo wLPC em cada bloco. O sinal de excitação com *pitch* modificado é, então, filtrado por $\widehat{H}_t\,(\hat{\omega})$ — modelo wLPC estimado para o locutor-alvo — e pelo filtro de de-ênfase. Para garantir que o modelo tenha transições suaves entre blocos, os modelos wLPC são interpolados usando os coeficientes LSF, devido a suas boas propriedades de interpolação. Para isso, os blocos de 1024 amostras espaçados de 512 amostras que são usados para análise são divididos em sub-blocos de 128 amostras. Assim a interpolação entre os coeficientes LSF dos blocos m e $(m-1)$ é feita fazendo-se

$$\bar{\mathbf{a}}_{\mathrm{lsf},4m+k} = \frac{\mathbf{a}_{\mathrm{lsf},m}(4-k) + \mathbf{a}_{\mathrm{lsf},m+1}k}{4}, \tag{6.37}$$

onde $\mathbf{a}_{\mathrm{lsf},m}$ é um vetor com os coeficientes LSF do bloco m, $\bar{\mathbf{a}}_{\mathrm{lsf},4m+k}$ é um vetor com os coeficientes LSF interpolados do sub-bloco $4m+k$ e $k = 0, \ldots, 3$, sendo



(a)



(b)

Figura 6.4: Esquema de treinamento da função de mapeamento (a) treinamento simples; (b) treinamento incremental.

que quando $k = 0$, o sub-bloco com índice $4m + k$ está alinhado com o bloco de análise com índice m. A Figura 6.5 ilustra o modo como é feita a interpolação. A filtragem subseqüente é feita calculando-se para cada sub-bloco os coeficientes wLPC interpolados, e usando os estados do filtro wIIR do sub-bloco $4m+k-1$ como

condição inicial para o bloco $4m+k$ em conjunto com a função *wfilter* implementada no WarpTB [21].

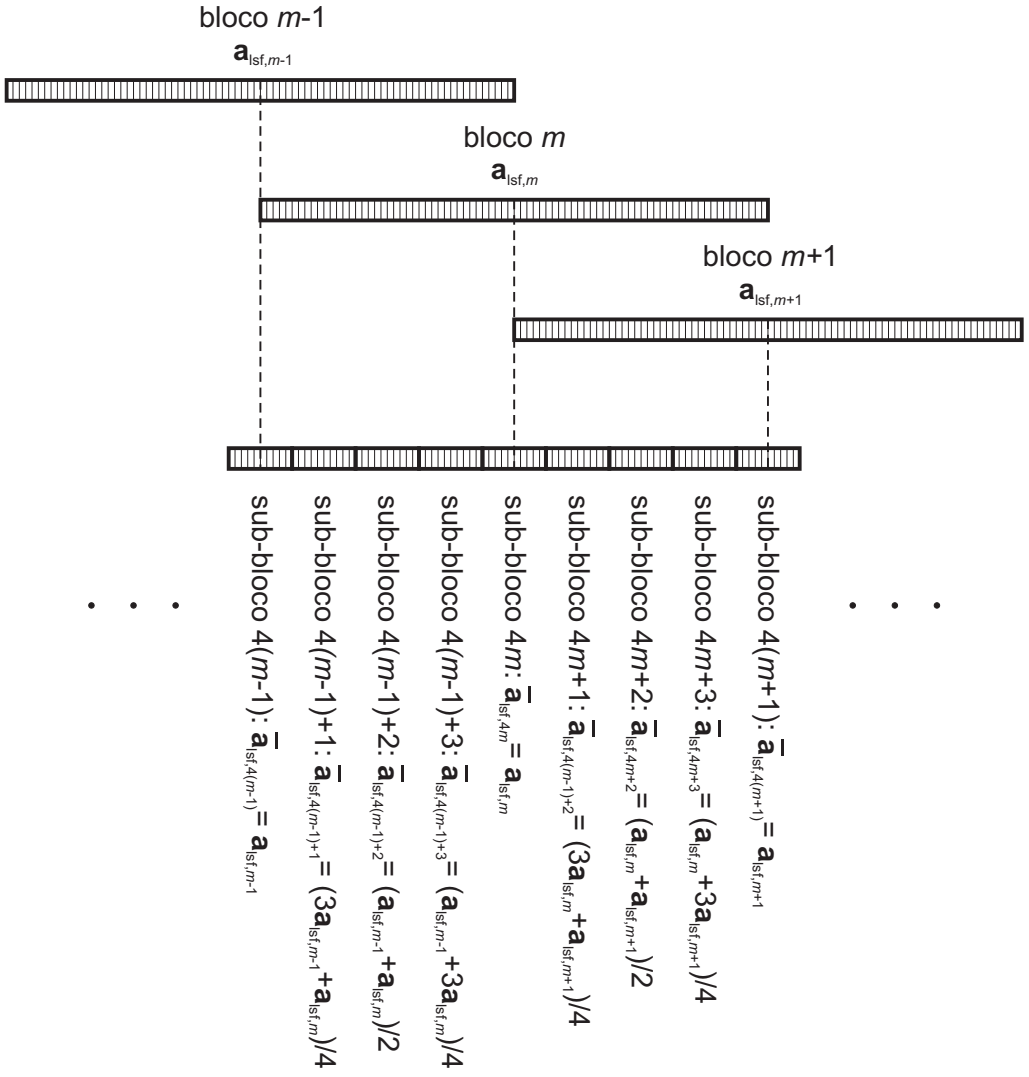


Figura 6.5: Interpolação dos coeficientes LSF correspondentes aos blocos de análise em sub-blocos de síntese.

6.8 Resultados experimentais

A seguir são mostrados os sonogramas dos resultados da transformação de locutor para uma das frases de teste '*A paixão dele é a natureza*', tanto para a transformação do tenor para o barítono, como do barítono para o tenor. Foram usados para este exemplo $g = 100$ misturas de gaussianas, $p = 30$ componentes principais e $M = 20$ coeficientes wLPC, que foram mapeados em 100 coeficientes

cepstrais[3].

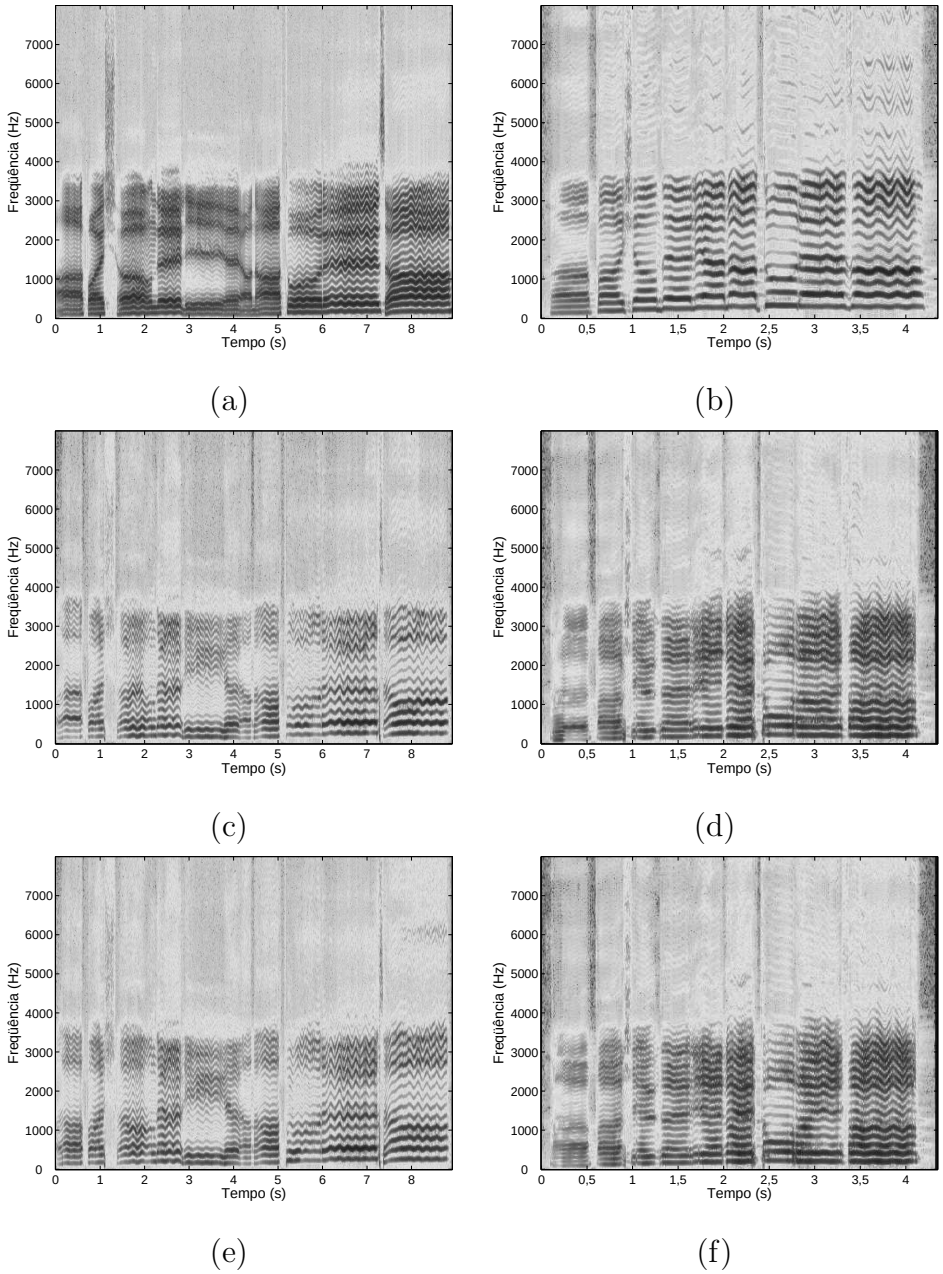


Figura 6.6: Espectrogramas dos sinais originais do (a) barítono; (b) tenor; espectrogramas dos sinais transformados (c) e (e) barítono → tenor; (d) e (f) tenor → barítono; (c) e (d) na primeira iteração; (e) e (f) na segunda iteração. Nas figuras pode-se notar que o desenho de *pitch* dos sinais modificados permanece inalterado, e que a envoltória espectral se aproxima da envoltória espectral do cantor-alvo.

Testes auditivos informais mostraram que os sinais modificados tinham o

[3]Exemplos disponíveis em http://www.lps.ufrj.br/∼rcdpaiva/mest/morph/ex/

timbre do locutor-alvo. Em alguns casos houve fonemas que foram trocados, o que é um efeito provável das dificuldades de alinhamento dos sinais. O que se observou com relação à ordem do modelo wLPC é que para a taxa de amostragem de 44,1 kHz, 12 coeficientes não eram suficientes para a transposição de locutor[4], e com 30 coeficientes os sinais sintetizados começavam a apresentar certa rouquidão[5] — que pode ser atribuída à ordem excessiva do modelo, quando este começa a modelar não só a envoltória do espectro, mas alguns picos individuais relativos ao *pitch*.

6.9 Avaliação objetiva

Uma alternativa de avaliação objetiva é apresentada por [102] para a medida da redução da distância entre os parâmetros espectrais dos locutores fonte e alvo depois da transformação:

$$R = \left(1 - \frac{D\left(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t\right)}{D\left(\mathbf{\Theta}_s, \mathbf{\Theta}_t\right)}\right) \times 100\%, \tag{6.38}$$

onde $D\left(\mathbf{\Theta}_X, \mathbf{\Theta}_Y\right)$ é uma medida de distância entre $\mathbf{\Theta}_X$ e $\mathbf{\Theta}_Y$, que pode ser a distância euclidiana. A distância medida sobre os coeficientes mel-cepstrais é um bom indicativo de similaridade, uma vez que é equivalente à distância da magnitude das respostas em freqüência das envoltórias espectrais transformada e alvo. Como nesta equação está a razão entre as distâncias da envoltória espectral, depois da transformação $D\left(\widehat{\mathbf{\Theta}}_t, \mathbf{\Theta}_t\right)$ e antes da transformação $D\left(\mathbf{\Theta}_s, \mathbf{\Theta}_t\right)$, ela dá um indicativo do aumento de similaridade depois do mapeamento da envoltória espectral.

Contudo, existem aspectos que tornam esse tipo de avaliação pouco robusto. A Tabela 6.1 mostra a avaliação objetiva do sistema usando informação dinâmica, com informação de blocos anteriores, e do sistema sem informação dinâmica. Pela tabela é possível constatar que a transformação de locutor usando informação dinâmica tem uma nota pior que quando esta infomação não é usada, apesar de testes auditivos informais terem mostrado claramente que esse sistema é superior. Esse resultado pode ser atribuído ao fato de essa métrica não levar em consideração a evolução da envoltória espectral. Outro problema de se avaliar sistemas de transformação com

[4]Exemplos disponíveis em http://www.lps.ufrj.br/∼rcdpaiva/mest/morph/ex12lp/

[5]Exemplos disponíveis em http://www.lps.ufrj.br/∼rcdpaiva/mest/morph/ex30lp/

esse tipo de métrica é que o alinhamento temporal dos dados nem sempre é perfeito. Essa é uma dificuldade que foi intensificada pelo tipo de banco de dados que usado neste trabalho, uma vez que na gravação da base de dados os cantores improvisaram livremente as melodias. Outro ponto interessante é que métodos que usam a estimativa de $\widehat{\mathbf{\Theta}}_t$ sem informação de blocos passados, apesar de terem apresentado resultados inferiores em alguns teste auditivos informais, acabam alcançando melhores resultados nessa avaliação objetiva.

Tabela 6.1: Avaliação objetiva dos sistemas de transformação de locutor.

Nota %	primeira iteração	segunda iteração
com informação dinâmica	62,2871	63,3767
sem informação dinâmica	69,9175	70,1090

6.10 Conclusão

Este capítulo apresentou um método de transformação de locutor que muda o timbre de uma frase falada por uma pessoa, de forma a parecer que a frase modificada foi falada por outra pessoa.

O capítulo começa fazendo na Seção 6.2 uma análise de fatores que caracterizam o que consideramos ser o timbre da voz de uma pessoa, e que nos permitem dintinguir auditivamente entre diferentes vozes. São essas as características usadas nas seções subseqüentes para transformar o timbre de um sinal de voz de forma que ele pareça ter sido emitido por outro locutor. Na Seção 6.3 é apresentado o esquema geral para transformação de locutor que vai ser usado no decorrer do capítulo. Esta seção destaca aspectos importantes com relação ao fator de modificação de *pitch* que deve ser aplicado, e ainda apresenta as razões para utilizar os coeficientes mel-cepstrais para a transformação da envoltória espectral.

Na Seção 6.4 é apresentado um sistema simples de transformação de locutor usando quantização vetorial. Este sistema é apresentado para justificar as melhorias que são implementadas nos esquemas seguintes.

A Seção 6.5 apresenta ferramentas que foram usadas para a implementação

do trabalho. A Seção 6.5 apresenta, de forma simplificada, o método para obtenção das componentes principais usada em PCA. Nesta seção é também feita uma interpretação do uso de PCA sobre o mel-cepstro. Nesta interpretação as bases de mapeamento nas componentes principais são interpretadas como filtros típicos, que podem estar relacionados à estrutura física do locutor, o que leva a uma interpretação que relaciona as componentes principais do cepstro a uma decomposição de um filtro ou modelo LPC em modos, ou subfiltros. Na Seção 6.6 foram apresentados de maneira resumida alguns algoritmos de clusterização suave, e foi proposta uma nova solução que aplica PCA e usa a informação de blocos anteriores para fazer a transformação de locutor. O sistema implementado é então descrito na Seção 6.7, onde são dados os detalhes necessários à implementação realizada neste trabalho de mestrado.

Resultados experimentais são mostrados e comentados na Seção 6.8, onde é possível destacar os pontos fortes e fracos do sistema implementado. Um problema enfrentado neste trabalho foi a dificuldade de alinhamento das frases. Uma vez que os cantores improvisavam melodias completamente diferentes para cada frase, o algoritmo de alinhamento teve o desempenho prejudicado. Em alguns exemplos de frases, um cantor mantém um fonema durante aproximadamente 1 segundo, enquanto o outro canta o mesmo fonema em 0,2 segundo. Entre os pontos fortes estão o uso da informação de blocos anteriores para fazer a transformação, o que leva a sinais sintetizados mais naturais, e com menos defeitos. Ainda, com o uso de PCA é possível uma redução bastante significativa da complexidade computacional do algoritmo, uma vez que de foram usadas somente 30 componentes de 100 coeficientes mel-cepstrais.

Capítulo 7

Conclusões

Este trabalho apresentou técnicas e alguns fundamentos necessários ao desenvolvimento de sistemas que operam transformações em sinais de voz. Como resultado final, ele apresentou aplicações de processamento de áudio e voz que incluem modificação de *pitch* de sinais de voz e transformação de locutor. Para isso foi necessário fazer um estudo de ferramentas que tornam estas aplicações possíveis.

As ferramentas apresentadas incluem formas de modelagem do trato vocal e um algoritmo de detecção de trechos sonoros e surdos de sinais de voz. As ferramentas de modelagem do trato vocal são de importância fundamental para a implementação dos algoritmos de transformação que foram propostos. Entre essas ferramentas foi apresentada uma comparação de diferentes tipos de modelo LPC, associada a uma publicação do autor [23]. Adicionalmente, foi apresentada a aplicação de modelos do trato vocal aproximando escalas perceptivas, com os quais foi proposto um método para calcular os coeficientes mel-cepstrais a partir dos coeficientes wLPC (*warped linear-prediction coefficients*) que, até onde o autor tem conhecimento, não foi proposto na literatura.

O estudo que foi feito com relação à detecção de trechos sonoros e surdos surgiu da observação que os algoritmos implementados no texto apresentam defeitos quando a detecção não é feita de maneira correta, e que nem sempre os algoritmos baseados somente em taxa de cruzamento por zeros e energia têm desempenho eficiente. Os algoritmos de modificação de *pitch* e transformação de locutor são implementados somente em partes sonoras de sinais de voz, e muitas vezes eles forçam uma periodicidade artificial em trechos surdos, que acabam por ser processados,

levando a defeitos audíveis. Por isso foram analisadas métricas comuns na literatura para a detecção de trechos sonoros, e foi proposta uma métrica que usa uma estimativa do ruído de fundo que, até onde o autor tem conhecimento, não consta na literatura. Uma vez apresentadas estas métricas, foi projetado um classificador simples de trechos sonoros e surdos, que foi usado como ferramenta na parte de transformação de sinais de voz.

A primeira das aplicações apresentada é a modificação de *pitch* de sinais de voz. O foco do autor no projeto destes sistemas é chegar a algoritmos que possam ser implementados em tempo real e com pouco atraso. É bastante comum que produtos que fazem modificação de *pitch* tenham um atraso que atrapalha o desempenho de artistas. Por isso foi feita uma implementação de um algoritmo de modificação de *pitch* que faz o modelo do trato vocal de forma seqüencial — que poderia dar origem a produtos de baixo custo — e um algoritmo que faz a modificação de *pitch* combinando a solução do algoritmo anterior com o PSOLA (*pitch-synchronous overlap-and-add*), que é uma técnica bastante difundida para estes sistemas. Os resultados destas aplicações deram origem a publicações em um congresso nacional [23] e um internacional [24].

A segunda aplicação tratada nesta dissertação é a transformação de locutor/cantor. Nesta aplicação foi desenvolvido um sistema que faz com um sinal falado/cantado por uma pessoa pareça ter sido falado/cantado por outra. O grande desafio nesta aplicação é obter modelos de representação de sinais de voz que estejam relacionados ao que consideramos o timbre de uma pessoa, e a obtenção de funções de mapeamento que possam fazer a transformação no domínio destes modelos de representação de forma natural. Para isso foram propostos aprimoramentos em métodos conhecidos na literatura. O primeiro deles foi a aplicação de análise de componentes principais, uma vez que, do modo como o sistema foi implementado, o treinamento dos modelos de misturas gaussianas era interrompido por problemas de condicionamento de matrizes. Esses problemas foram resolvidos usando as componentes principais do mel-cepstro para treinamento. Ainda, com o uso de PCA (*principal component analysis*), foi possível fazer uma interpretação do significado das componentes principais do cepstro como sendo relativas à estrutura física de locutores/cantores. Para a implementação da função de mapeamento entre locu-

tores/cantores, foi proposto um método onde a informação de blocos anteriores é usada para melhorar a continuidade dos sinais modificados. Esse procedimento mostrou uma melhora significativa na naturalidade dos sinais transformados. Na revisão bibliográfica feita pelo autor não foi encontrada menção à interpretação que foi feita da PCA sobre o cepstro nem do uso de informação de blocos anteriores para a transformação de locutor.

Extensões do trabalho podem incluir o projeto de classificadores de trechos sonoros que utilizem outros tipos de ruído, como ruído veicular, para aplicações de celular e comunicações. Isso tornaria o classificador mais robusto para ambientes diferentes do ambiente idealizado com ruído branco de fundo.

Para a aplicação de modificação de *pitch*, um problema encontrado neste tipo de algoritmo é que, mesmo tendo boa qualidade final, os sinais modificados com fator de modificação de *pitch* muito grandes (em torno de $\beta = 2$) ou muito pequenos (em torno de $\beta = 0{,}5$) aparentam outro tipo de emissão ou registro vocal — para uma voz masculina modificada com $\beta = 0{,}5$, o sinal modificado tem som de *falsete*. Esse efeito ocorre provavelmente devido à não inclusão de um modelo de excitação glotal mais elaborado, e também pelo fato de as características do trato vocal não serem totalmente desacopladas do *pitch* como é assumido nos modelos de envoltória espectral. Extensões do trabalho podem incluir esse tipo de informação, realizando uma investigação sobre modificações na envoltória espectral em conjunto com a modificação de *pitch*. Adicionalmente é interessante um estudo comparativo entre os sistemas de modificações de *pitch* usando modelo LPC em blocos e seqüencial.

Na parte de transformação de locutor/cantor, foram encontradas dificuldades no treinamento do sistema decorrentes do modo como o banco de dados foi gravado. Como os cantores improvisaram melodias para as frases, o alinhamento temporal das frases ficou bastante dificultado. Além disto, é possível que pequenas diferenças nas posições dos formantes, quando os cantores cantam com *pitch* diferente, levem a um efeito de suavização dos formantes, o que pode levar a defeitos nos sinais modificados. Outro fator que merece atenção é a avaliação objetiva de sistemas de transformação de locutor. Para a avaliação objetiva seriam necessários avaliadores de qualidade de áudio e voz sem referência, e avaliadores de similaridade entre os sinais transformados e os sinais da voz-alvo. Isso representa um desafio bastante

grande, uma vez que é interessante que o avaliador de similaridade leve em conta não só a envoltória espectral dos sinais transformados bloco a bloco, mas o seu modo de evolução. Isso indica as seguintes extensões ao trabalho:

- Base de dados:
  - gravação de uma base de dados com frases cantadas com auxílio de partitura e metrônomo, o que diminuiria os efeitos decorrentes das variações da envoltória espectral com o *pitch*, e facilitaria bastante o alinhamento temporal das frases;
  - investigação sobre o tamanho mínimo para a base de dados a ser usada;
- Métodos objetivos de avaliação:
  - investigação de avaliação de qualidade sem referência, para avaliar os defeitos audíveis resultantes da transformação de locutor/cantor;
  - investigação de medidas de similaridade entre o sinal transformado e a envoltória espectral da voz-alvo.

Referências Bibliográficas

[1] HUCHE, F. L., ALLALI, A., *A Voz: A Anatomia e Fisiologia dos Órgãos da Voz e da Fala*, v. 1. 3 ed. Artmed, 2005.

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

[3] GREMY, F., "Considérations sur l'énergie mécanique du larynx durant la phonation", *Journal Physiol.*, v. 52, pp. 555–567, 1960.

[4] CORBIAU, G., "Farinelli", Filme, Sony Pictures Classics, 1994.

[5] DEPALLE, P., GARCIA, G., RODET, X., "The recreation of a castrato voice, Farinelli's voice". In: *Proc. of the WASPAA'95 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 15–18, New Paltz, EUA, Outubro 1995.

[6] BONADA, J., SERRA, X., "Synthesis of the singing voice by performance sampling and spectral models", *IEEE Signal Processing Magazine*, v. 24, n. 2, pp. 67–79, Março 2007.

[7] FABIG, L., JANER, J., "Transforming singing voice expression - The sweetness effect". In: *Proc. of the DAFx'04 - 7th International Conference on Digital Audio Effects*, Nápoles, Italia, Outubro 2004.

[8] LOSCOS, A., BONADA, J., "Emulating rough and growl voice in spectral domain". In: *Proc. of the DAFx'04 - 7th International Conference on Digital Audio Effects*, Nápoles, Italia, Outubro 2006.

[9] BONADA, J., LOSCOS, A., "Esophageal voice enhancement by modeling radiated pulses in frequency domain". In: *121st Audio Engineering Society Convention*, San Francisco, EUA, Outubro 2006. Preprint 6952.

[10] LOSCOS, A., *Spectral processing of the singing voice*. Tese de D.Sc., Universitat Pompeu Fabra, Barcelona, Espanha, 2007.

[11] CLARK, R., "Designing the delay out of digital mixing systems". In: *Anais do AES'07 - 5o Congresso de Engenharia de Áudio da AES Brasil*, pp. 151–157, São Paulo, Brasil, Maio 2007.

[12] HARTMANN, W. M., "Pitch, periodicity, and auditory organization", *Journal of the Acoustical Society of America*, v. 100, n. 6, pp. 3491–3502, Dezembro 1996.

[13] MEDDIS, R., HEWITT, M., "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification", *Journal of the Acoustical Society of America*, v. 89, n. 6, pp. 2866–2882, Junho 1991.

[14] HUSSON, R., *Physiologie de la Phonation*. Masson, 1962.

[15] FANT, G., LILJENCRANTS, J., LIN, Q., *A four-parameter model of glottal flow*, Internal Report STL-QPSR 26 4, Dept. for Speech, Music and Hearing - Royal Institute of Technology, Estocolmo, Suécia, 1985.

[16] KLATT, D. H., KLATT, L. C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *Journal of the Acoustical Society of America*, v. 87, n. 2, pp. 820–857, Fevereiro 1990.

[17] FU, Q., MURPHY, P., "Robust glottal source estimation based on joint source-filter model optimization", *IEEE Transactions on Audio, Speech and Language Processing*, v. 14, n. 2, pp. 492–501, Março 2006.

[18] HUBER, J. E., STATHOPOULOS, E. T., CURIONE, G. M., *et al.*, "Formants of childrem, women, and men: The effects of vocal intensity variation", *Journal of the Acoustical Society of America*, v. 106, n. 3, pp. 1532–1542, Setembro 1999.

[19] LU, H. L., *Toward a High-Quality Singing Synthesizer with Vocal Texture Control*. Tese de D.Sc., Stanford University, Palo Alto, EUA, Julho 2002.

[20] KAIN, A. B., *High Resolution Voice Transformation*. Tese de D.Sc., Oregon Health and Science University, Portland, EUA, Outubro 2001.

[21] HÄRMÄ, A., KARJALAINEN, M., "WarpTB - Matlab Toolbox for Warped DSP", Toolbox for Matlab, 2000, Homepage: http://www.acoustics.hut.fi/software/warp/ (acesso em 17 de janeiro de 2008).

[22] MAKHOUL, J., "Linear prediction: A tutorial review", *Proceedings of the IEEE*, v. 63, n. 4, pp. 561–580, Abril 1975.

[23] PAIVA, R. C. D., BISCAINHO, L. W. P., NETTO, S. L., "A sequential system for voice pitch modification". In: *Anais do AES'07 - 5o Congresso de Engenharia de Áudio da AES Brasil*, pp. 11–16, São Paulo, Brasil, Maio 2007.

[24] PAIVA, R. C. D., BISCAINHO, L. W. P., NETTO, S. L., "On the application of RLS adaptive filtering for voice pitch modification". In: *Proc. of the DAFx'07 - 10th International Conference on Digital Audio Effects*, pp. 27–32, Bordeaux, França, Setembro 2007.

[25] DINIZ, P. S. R., *Adaptive Filtering: Algorithms and Practical Implementations*. 2 ed. Kluwer, 2002.

[26] STROBACH, P., *Linear Prediction Theory: A mathematical basis for adaptive systems*. Springer-Verlag, 1990.

[27] SOONG, F. K., JUANG, B. H., "Line spectrum pairs (LSP) and speech data compression". In: *Proc. of the ICASSP'84 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 9, pp. 37–40, Março 1984.

[28] PALIWAL, K. K., "A study of LSF representation for speaker-dependent and speaker-independent HMM-based speech recognition systems". In: *Proc. of the ICASSP'90 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 2, pp. 801–804, Albuquerque, EUA, Abril 1990.

[29] OPPENHEIM, A. V., SCHAFER, R. W., "From frequency to quefrency: A history of the cepstrum", *IEEE Signal Processing Magazine*, v. 21, n. 5, pp. 95–106, Setembro 2004.

[30] NOLL, A. M., "Cepstrum pitch determination", *Journal of the Acoustical Society of America*, v. 41, n. 2, pp. 293–309, Fevereiro 1967.

[31] OPPENHEIM, A. V., SCHAFER, R. W., *Discrete-Time Signal Processing*. Prentice Hall, 1989.

[32] YNOGUTI, C. A., VIOLARO, F., "On the use of principal component analysis over mel cepstral coefficients", *Telecomunicações, Revista do Instituto Nacional de Telecomunicações*, v. 5, n. 2, pp. 13–17, Dezembro 2002.

[33] KIM, H. G., MOREAU, N., SIKORA, T., *MPEG-7 audio and beyond audio content indexing and retrieval*. Wiley, 2005.

[34] SPANIAS, A., PAINTER, T., ATTI, V., *Audio signal processing and coding*. Wiley, 2006.

[35] SMITH, J. O., ABEL, J. S., "Bark and ERB bilinear transforms", *IEEE Transactions on Speech and Audio Processing*, v. 7, n. 6, pp. 697–708, Novembro 1999.

[36] HÄRMÄ, A., KARJALAINEN, M., SAVIOJA, L., *et al.*, "Frequency-warped signal processing for audio applications", *Journal of Audio Engineering Society JAES*, v. 48, n. 11, pp. 1011–1031, Novembro 2000.

[37] OPPENHEIM, A., JOHNSON, D., STEIGLITZ, K., "Computation of spectra with unequal resolution using the fast Fourier transform", *Proceedings of the IEEE*, v. 59, n. 2, pp. 299–301, Fevereiro 1971.

[38] BRACCINI, C., OPPENHEIM, A., "Unequal bandwidth spectral analysis using digital frequency warping", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 59, n. 4, pp. 299–301, Fevereiro 1974.

[39] CHO, N. I., MITRA, S. K., "Warped discrete cosine transform and its application in image compression", *IEEE Transactions on Circuits and Systems for Video Technology*, v. 10, n. 8, pp. 1364–1373, Dezembro 2000.

[40] MAKUR, A., MITRA, S. K., "Warped discrete-Fourier transform: Theory and applications", *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, v. 48, n. 9, pp. 1086–1093, Setembro 2001.

[41] FRANZ, S., MITRA, S. K., SCHMIDT, J. C., *et al.*, "Warped discrete Fourier transform: A new concept in digital signal processing". In: *Proc. of the ICASSP'02 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 2, pp. 1205–1208, Orlando, EUA, Maio 2002.

[42] SHANKAR, B. M. R., MAKUR, A., "Allpass delay chain-based IIR PR filterbank and its application to multiple description subband coding", *IEEE Transactions on Signal Processing*, v. 50, n. 4, pp. 814–823, Abril 2002.

[43] FELDBAUER, C., KUBIN, G., "Critically sampled frequency-warped perfect reconstruction filterbank". In: *ECCTD'03 - European Conference on Circuit Theory and Design*, Cracóvia, Polônia, Setembro 2003.

[44] KARJALAINEN, M., HÄRMÄ, A., LAINE, U. K., "Realizable warped IIR filters and their properties". In: *Proc. of the ICASSP'97 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 3, pp. 1205–1208, Munique, Alemanha, Abril 1997.

[45] HÄRMÄ, A., *Peceptual aspects and warped techniques in audio coding*. Tese de M.Sc., Helsinki University of Technology, Helsinque, Finlândia, Maio 1997.

[46] HÄRMÄ, A., LAINE, U. K., "A comparison of warped and conventional linear predictive coding", *IEEE Transactions on Speech and Audio Processing*, v. 9, n. 5, pp. 579–588, Julho 2001.

[47] STRUBE, H. W., "Linear prediction on a warped frequency scale", *Journal of the Acoustical Society of America*, v. 68, n. 4, pp. 1071–1076, Outubro 1980.

[48] MAKUR, A., "Fast computation of WDFT and its application in image compression". In: *Proc. of the TENCON'06 - IEEE Region 10 Conference*, pp. 1–4, Hong Kong, Novembro 2006.

[49] MERWE, C. J. V. D., PREEZ, J. A. D., "Calculation of LPC-based cepstrum coefficients using mel-scale frequency warpping". In: *Proc. of the COMSIG'91*

- *South African Symposium on Communications and Signal Processing*, pp. 17–21, Pretoria, África do Sul, Agosto 1991.

[50] CAO, Y., SRIDHARAN, S., MOODY, M., "Voiced/unvoiced/silence classification of noisy speech in real time audio signal processing". In: *5th Australian Regional Audio Engineering Society Convention*, Sydney, Austrália, Março 1995. Preprint 4045.

[51] ATAL, B. S., RABINER, L. R., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 24, n. 3, pp. 201–212, Junho 1976.

[52] FISHER, E., TABRIKIAN, J., DUBNOV, S., "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model", *IEEE Transactions on Audio, Speech and Language Processing*, v. 14, n. 2, pp. 502–510, Março 2006.

[53] JANER, L., BONET, J. J., LLEIDA-SOLANO, E., "Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms". In: *Proc. of the ICSLP'96 - IEEE International Conference on Spoken Language Processing*, v. 2, pp. 1209–1212, Filadélfia, EUA, Outubro 1996.

[54] RABINER, L. R., SAMBUR, M., "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 25, n. 4, pp. 338–343, Agosto 1977.

[55] LAURENTI, N., POLI, G. D., MONTAGNER, D., "A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds", *IEEE Transactions on Audio, Speech and Language Processing*, v. 15, n. 2, pp. 531–541, Fevereiro 2007.

[56] DINIZ, P. S. R., SILVA, E. A. B., NETTO, S. L., *Processamento Digital de Sinais: Projeto e análise de Sistemas*. 1 ed. Bookman, 2004.

[57] ALCAIM, A., SOLEWICZ, J. R., MORAES, J. A. D., "Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no por-

tuguês falado no Rio de Janeiro", *Revista da Sociedade Brasileira de Telecomunicações*, v. 7, n. 1, pp. 23–41, Dezembro 1992.

[58] DUDA, R. O., HART, P. E., STORK, D. G., *Pattern Classification*. 2 ed. Wiley Interscience, 2000.

[59] LAROCHE, J., "Time and pitch scale modification of audio signals". In: Kahrs, M., Brandenburg, K. (eds.), *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Publishers, pp. 279–309, 2002.

[60] FAIRBANKS, G., EVERITT, W. L., JAEGER, R. P., "Method for time or frequency compression-expansion of speech", *Transactions of the IRE Professional Group on Audio*, v. 2, n. 1, pp. 7–12, Janeiro 1954.

[61] HAGHPARAST, A., PENTTINEN, H., VÄLIMÄKI, V., "Real-time pitch-shifting of musical signals by a time-varying factor using normalized filtered correlation time-scale modification (NFC-TSM)". In: *Proc. of the DAFx'07 - 10th International Conference on Digital Audio Effects*, pp. 7–13, Bordeaux, França, Setembro 2007.

[62] MOORER, J. A., "The use of linear prediction of speech in computer music applications", *Journal of Audio Engineering Society*, v. 27, n. 3, pp. 134–140, Março 1979.

[63] LAROCHE, J., DOLSON, M., "Phase-vocoder: About this phasiness business". In: *Proc. of the WASPAA'97 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, EUA, Outubro 1997.

[64] MOULINES, E., LAROCHE, J., "Non-parametric techniques for pitch-scale and time-scale modification of speech", *Speech Communication*, v. 16, n. 2, pp. 175–205, Fevereiro 1995.

[65] PEETERS, G., *Modèles et modification du signal sonore adaptés à ses caractéristiques locales*. Tese de D.Sc., Université Paris 6, Paris, França, Julho 2001.

[66] CHEVEIGNÉ, A., KAWAHARA, H., "YIN, a fundamental frequency estimator for speech and music", *Journal of the Acoustical Society of America*, v. 111, n. 4, pp. 1917–1930, Abril 2002.

[67] CHOI, A., "Real-time fundamental frequency estimation by least-square fitting", *IEEE Transactions on Speech and Audio Processing*, v. 5, n. 2, pp. 201–205, Março 1997.

[68] GERHARD, D., *Pitch Extraction and Fundamental Frequency: History and Current Techniques*, Report, Dept. of Computer Science, University of Regina, Regina, Canadá, Novembro 2003. Technical Report TR-CS 2003-06.

[69] JANER, L., "Modulated Gaussian wavelet transform based speech analyser (MGWTSA) pitch detection algorithm (PDA)". In: *Proc. of the EUROSPEECH'95 - Fourth European Conference on Speech Communication and Technology*, pp. 401–404, Madri, Espanha, Setembro 1995.

[70] KADAMBE, S., BOUDREAUX-BARTELS, G. F., "Application of the wavelet transform for pitch detection of speech signals", *IEEE Transactions on Information Theory*, v. 38, n. 2, pp. 917–924, Março 1992.

[71] MA, C., KAMP, Y., WILLEMS, L. F., "A Frobenius norm approach to glottal closure detection from the speech signal", *IEEE Transactions on Speech and Audio Processing*, v. 2, n. 2, pp. 258–265, Abril 1994.

[72] BROOKES, M., NAYLOR, P. A., GUDNASON, J., "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech", *IEEE Transactions on Audio, Speech and Language Processing*, v. 14, n. 2, pp. 456–466, Março 2006.

[73] NAYLOR, P. A., KOUNOUDES, A., GUDNASON, J., *et al.*, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm", *IEEE Transactions on Audio, Speech and Language Processing*, v. 15, n. 1, pp. 34–43, Janeiro 2007.

[74] BROOKES, M., "Voicebox", Toolbox for Matlab, 2003, Homepage: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (acesso em 6 de janeiro de 2008).

[75] LAAKSO, T. I., VÄLIMÄKI, V., KARJALAINEN, M., *et al.*, "Splitting the unit delay - Tools for fractional delay filter design", *IEEE Signal Processing Magazine*, v. 13, n. 1, pp. 30–60, Janeiro 1996.

[76] VÄLIMÄKI, V., HUOVILAINEN, A., "Oscillator and filter algorithms for virtual analog synthesis", *Computer Music Journal*, v. 30, n. 2, pp. 19–31, 2006.

[77] STYLIANOU, Y., CAPPÉ, O., MOULINES, E., "Statistical methods for voice quality transformation". In: *Proc. of the EUROSPEECH'95 - Fourth European Conference on Speech Communication and Technology*, pp. 447–450, Madri, Espanha, Setembro 1995.

[78] KAIN, A., MACON, M., "Personalizing a speech synthesizer by voice adaptation". In: *Proc. of the SSW'98 - Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 225–230, Blue Mountains, Austrália, Novembro 1998.

[79] JUNQUA, J. C., PERRONNIN, F., KUHN, R., *et al.*, "Voice personalization of speech synthesizer", US Patent 6970820, United States Patent and Trademark Office - USPTO, Novembro 2005, Assignee: Matsushita Electric Industrial Co., Ltd.

[80] SCHALK, T. B., "Method for reducing database requirements for speech recognition systems", US Patent 5845246, United States Patent and Trademark Office - USPTO, Dezembro 1998, Assignee: Voice Control Systems, Inc.

[81] BERESIN, E., PUGH, J., "Voice avatars for wireless multiuser entertainment services", US Patent 6987514, United States Patent and Trademark Office - USPTO, Janeiro 2006, Assignee: Nokia Corporation.

[82] YE, H., YOUNG, S., "Quality-enhanced voice morphing using maximum likelihood transformations", *IEEE Transactions on Audio, Speech and Language Processing*, v. 14, n. 4, pp. 1301–1312, Julho 2006.

[83] GIBSON, B. C., LUPINI, P. R., SHPAK, D. J., "Targeted vocal transformation", US Patent 6336092, United States Patent and Trademark Office - USPTO, Janeiro 2002, Assignee: Ivl Technologies Ltd.

[84] MATSUMOTO, S., "Karaoke apparatus converting singing voice into model voice", US Patent 5621182, United States Patent and Trademark Office - USPTO, Abril 1997, Assignee: Yamaha Corporation.

[85] CANO, P., LOSCOS, A., BONADA, J., *et al.*, "Voice morphing system for impersonating in Karaoke applications". In: *Proc. of the ICMC'00 - International Computer Music Conference 2000*, Berlim, Alemanha, 2000.

[86] RABINER, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, v. 77, n. 2, pp. 257–286, Fevereiro 1989.

[87] LOSCOS, A., CANO, P., BONADA, J., "Low-delay singing voice alignment to text". In: *Proc. of the ICMC'99 - International Computer Music Conference*, Beijing, China, 1999.

[88] ABE, M., NAKAMURA, S., SHIKANO, K., *et al.*, "Voice conversion through vector quantization". In: *Proc. of the ICASSP'88 - IEEE International Conference on Acoustics, Speech, and Signal Processing 1988*, Nova Iorque, EUA, Abril 1988.

[89] TURK, O., ARSLAN, L. M., "Robust processing techniques for voice conversion", *Computer Speech and Language*, v. 20, n. 4, pp. 441–467, Outubro 2006.

[90] STYLIANOU, Y., CAPPÉ, O., MOULINES, E., "Continuous probabilistic transform for voice conversion", *IEEE Transactions on Speech and Audio Processing*, v. 6, n. 2, pp. 131 –142, Março 1998.

[91] MAKKI, B., SEYEDSALEHI, S. A., SADATI, N., *et al.*, "Voice conversion using nonlinear principal component analysis". In: *Proc. of the CIISP'07 - IEEE Symposium on Computational Intelligence in Image and Signal Processing 2007*, pp. 336–339, Honolulu, EUA, Abril 2007.

[92] ORPHANIDOU, C., MOROZ, I. M., ROBERTS, S. J., "Wavelet-based voice morphing", *WSEAS Transactions on Systems*, v. 10, n. 3, pp. 3297–3302, Dezembro 2004.

[93] GUIDO, R. C., VIEIRA, L. S., JR., S. B., *et al.*, "A neural-wavelet architecture for voice conversion", *Neurocomputing*, v. 71, n. 1 - 3, pp. 174–180, Agosto 2007.

[94] KUWABARA, H., SAGISAKA, Y., "Acoustic characteristics of speaker individuality: Control and conversion", *Speech Communication*, v. 16, n. 2, pp. 165–173, Fevereiro 1995.

[95] KUHN, N., NGUYEN, P., JUNQUA, J. C., *et al.*, "Eigenfaces and eigenvoices: Dimensionality reduction for specialized pattern recognition". In: *Proc. of the WMSP'98 - IEEE Workshop on Multimedia Signal Processing*, pp. 71–76, Redondo Beach, EUA, Dezembro 1998.

[96] JOLLIFE, I. T., *Principal Component Analysis*. Springer-Verlag, 1986.

[97] BELL, A. J., "Information theory, independent-component analysis, and applications". In: Haykin, S. (ed.), *Unsupervised Adaptive Filtering*, v. 1, capítulo 3, Wiley Interscience, pp. 237–264, 2000.

[98] GHAHRAMANI, Z., JORDAN, M. I., "Supervised learning from incomplete data via an EM approach". In: Cowan, J. D., Tesauro, G., Alspector, J. (eds.), *Advances in Neural Information Processing Systems*, v. 6, pp. 120–127, 1994.

[99] KAMBHATLA, N., *Local models and Gaussian mixture models for statistical data processing*. Tese de D.Sc., Oregon Health and Science University, Beaverton, EUA, 1996.

[100] ELLIS, D., "Dynamic Time Warp (DTW) in Matlab", Matlab Code, 2003, Homepage: http://www.ee.columbia.edu/∼dpwe/resources/matlab/dtw/ (acesso em 30 de janeiro de 2008).

[101] FRANC, V., SCHLESINGER, M. I., HLAVAC, V., "Statistical Pattern Recognition Toolbox", Toolbox for Matlab, 2000, Homepage:

http://cmp.felk.cvut.cz/cmp/software/stprtool/index.html (acesso em 30 de janeiro de 2008).

[102] ZHAO, L., GAO, Y., "Voice conversion adopting SOLAFS". In: *Proc. of the SNPD'07 - Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing 2007*, v. 1, pp. 543–548, Qingdao, China, Julho 2007.