MARCOS ODEBRECHT JÚNIOR

CONVERSÃO DO CONTORNO DE *PITCH* POR DIVISÃO DE COMPONENTES PARA APLICAÇÃO EM SISTEMAS DE CONVERSÃO DE VOZ

FLORIANÓPOLIS

2009

UNIVERSIDADE FEDERAL DE SANTA CATARINA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

CONVERSÃO DO CONTORNO DE *PITCH* POR DIVISÃO DE COMPONENTES PARA APLICAÇÃO EM SISTEMAS DE CONVERSÃO DE VOZ

Dissertação submetida à
Universidade Federal de Santa Catarina
como parte dos requisitos para a
obtenção do grau de Mestre em Engenharia Elétrica.

MARCOS ODEBRECHT JÚNIOR

Florianópolis, Abril de 2009.

CONVERSÃO DO CONTORNO DE *PITCH* POR DIVISÃO DE COMPONENTES PARA APLICAÇÃO EM SISTEMAS DE CONVERSÃO DE VOZ

Marcos Odebrecht Júnior

'Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Engenharia Elétrica, Área de Concentração em *Circuitos e Instrumentação Eletrônica*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina.'

Prof. Rui Seara, Dr.
Orientador

Prof[a]. Kátia Campos de Almeida, Ph.D.
Coordenadora do Programa de Pós-Graduação em Engenharia Elétrica

Banca Examinadora:

Prof. Rui Seara, Dr.
Presidente

Prof. Sidnei Noceti Filho, D.Sc.

Prof. Fernando Santana Pacheco, Dr.

AGRADECIMENTOS

Ao Prof. Rui Seara pela orientação exemplar e primorosa atenção na revisão do texto da dissertação.

Aos membros da banca examinadora pelas criteriosas observações e valiosos comentários que contribuíram para o refinamento da versão final do trabalho.

Aos amigos de mestrado Arthur de Oliveira Jaekel, Cristiano Ferreira, Eduardo Almeida, André Pires Nóbrega Tahim e Juan Rodrigo Velásquez López.

Aos integrantes do LINSE pela inestimável colaboração no processo de avaliação subjetiva requerido para o desenvolvimento deste trabalho. Em especial, à Izabel Christine Seara e ao Elton Luiz Fontão pelo olhar aguçado na revisão do texto e da formatação do trabalho.

Aos professores do Departamento de Engenharia Elétrica da Universidade Regional de Blumenau.

À Prof[a]. Cíntia Soares pela constante ajuda durante o desenvolvimento do trabalho.

Aos amigos e familiares por compartilharem todos os momentos de mais esta importante etapa no eterno processo de aprendizado. Em especial à Beatriz Odebrecht pelo carinho e apoio sempre presentes.

À Anamélia Sant'Anna pelo valioso incentivo e incansável compreensão sem os quais este trabalho não teria se concretizado.

Resumo da Dissertação apresentada à UFSC como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

CONVERSÃO DO CONTORNO DE *PITCH* POR DIVISÃO DE COMPONENTES PARA APLICAÇÃO EM SISTEMAS DE CONVERSÃO DE VOZ

Marcos Odebrecht Júnior

Abril/2009

Orientador: Prof. Rui Seara, Dr.
Área de Concentração: Comunicações e Processamento de Sinais.
Palavras-Chave: Algoritmo MOMEL, conversão de voz, conversão do contorno de *pitch*, INTSINT, prosódia.
Número de Páginas: 54.

RESUMO: Esta dissertação propõe uma nova técnica de conversão do contorno de *pitch* para aplicação em sistemas de conversão de voz. O principal objetivo deste trabalho é possibilitar a aplicação do método proposto aos mais diferentes tipos de sistemas de conversão de voz sem que para tanto seja necessário adaptar ou criar um novo banco de sinais de fala. A abordagem proposta considera o algoritmo MOMEL (*modelling melody*) para dividir o contorno de *pitch* levando em conta os componentes macroprosódico e microprosódico, sendo que cada um deles é convertido separadamente. A contribuição do componente macroprosódico, obtida pela interpolação dos dados usando a codificação INTSINT (*international transcription system for intonation*), é então convertida utilizando um modelo de misturas gaussianas (GMM); enquanto, a contribuição do componente microprosódico é convertida por seleção de segmentos de contorno de *pitch*. Os problemas inerentes à avaliação de desempenho dos sistemas de conversão de voz são discutidos e um parâmetro denominado índice de desempenho é modificado para permitir uma avaliação objetiva da conversão do contorno de *pitch*. O desempenho do método proposto é confrontado com dois dos métodos mais utilizados na literatura: conversão utilizando normalização gaussiana (GN) e GMM. O desempenho dos diferentes métodos considerados são avaliados através de dois testes subjetivos: de preferência e de similaridade. Os resultados obtidos ratificam a medida adotada, indicando uma preferência pelo método proposto através da melhoria significativa de desempenho frente aos demais métodos avaliados. A flexibilidade da nova abordagem possibilita ampla gama de aplicações nos mais variados tipos de sistemas de conversão de voz.

Abstract of Dissertation presented to UFSC as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

PITCH CONTOUR CONVERSION BY COMPONENT DIVISION FOR APPLICATION IN VOICE CONVERSION SYSTEMS

Marcos Odebrecht Júnior

April/2009

Advisor: Prof. Rui Seara, Dr.
Area of Concentration: Communication and Signal Processing.
Keywords: INTSINT, MOMEL algorithm, pitch contour conversion, prosody, voice conversion.
Number of Pages: 54.

ABSTRACT: This research work proposes a new strategy to obtain the pitch contour conversion for application in voice conversion systems. The main goal of this work is to improve the current pitch contour conversion techniques aiming its use in the different types of voice conversion systems, without requiring any changes in the recorded speech corpus. The new approach uses different conversion strategies taking into account both macroprosodic and microprosodic components, which are obtained from the pitch contour by using the modelling melody (MOMEL) algorithm. The macroprosodic component contribution is obtained by the interpolation of the international transcription system for intonation (INTSINT) codification, which in turn is converted applying a Gaussian mixture model (GMM). In addition, the microprosodic component contribution is converted by selecting segments from a pitch contour codebook. Drawbacks inherent to the evaluation of voice conversion systems are also discussed. A parameter, termed performance index, is adapted here to measure the pitch contour conversion performance. The proposed approach is compared with two other techniques from the literature (Gaussian normalization and GMM) for performance. Two different subjective tests are carried out to assess the perceptual performance of the considered methods: preference and similarity tests. Experimental results confirm the performance and flexibility of the proposed approach as compared with other techniques from the open literature.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos e Acrônimos

α_i	Peso do i-ésimo componente do modelo de misturas gaussianas
$\hat{\mathbf{y}}$	Contorno de *pitch* convertido
$\mathbf{x}$	Contorno de *pitch* fonte
$\mathbf{y}$	Contorno de *pitch* alvo
$\mathcal{L}$	Função de verossimilhança
μ	Valor médio
Σ	Matriz de covariância
σ	Desvio padrão
θ	Conjunto de parâmetros do modelo GMM
$\tilde{f}(n)$	Transformada IDCT de $F(k)$ de tamanho fixo
$D(\hat{\mathbf{y}}, \mathbf{x})$	Distância convertido-fonte
$D(\hat{\mathbf{y}}, \mathbf{y})$	Distância convertido-alvo
$D(\mathbf{x}, \mathbf{y})$	Distância fonte-alvo
$E[\cdot]$	Valor esperado
$F(k)$	Transformada DCT de $f(n)$ de tamanho $K = N$
$f(n)$	Segmento de contorno de *pitch* de tamanho N
F_{i}	Freqüência do i-ésimo formante
m	Número de misturas do modelo de misturas gaussianas
$N(\cdot)$	Função de distribuição de probabilidade gaussiana
$p(\cdot)$	Função de distribuição de probabilidade

Hz	Hertz
kbps	Kilobit por segundo
B	Símbolo INTSINT *Bottom*
D	Símbolo INTSINT *Downstepped*
DCT	*Discrete cosine transform*
DTW	*Dynamic time warping*
EM	*Expectation maximization*
GMM	*Gaussian mixture model*
H	Símbolo INTSINT *Higher*
IDCT	*Inverse discrete cosine transform*
IHMD	*Inverse harmonic mean distortion*
INTSINT	*International transcription system for intonation*
L	Símbolo INTSINT *Lower*
M	Símbolo INTSINT *Mid*
$\text{MACRO}_{\text{GMM}}$	Conversão do componente macroprosódico do contorno de *pitch* utilizando GMM
MACRO_{GN}	Conversão do componente macroprosódico do contorno de *pitch* utilizando GN
MOMEL	*Modelling melody*
MSE	*Mean-square error*
PB	Português brasileiro
PI	*Performance index*
S	Símbolo INTSINT *Same*
$\text{SELEÇÃO}_{\text{GMM}}$	Conversão do componente macroprosódico do contorno de *pitch* utilizando GMM em conjunto com a conversão do componente microprosódico por seleção de segmentos de contorno de *pitch*

T	Símbolo INTSINT *Top*
TTS	*Text-to-speech*
U	Símbolo INTSINT *Upstepped*

Capítulo 1

Introdução

Dentre as diversas informações contidas em um sinal de fala, o conteúdo da mensagem é a de primordial importância. Todavia, outras informações podem também ser transmitidas através do sinal de fala, a saber: entonação, emoção e também identidade do locutor, sendo esta última de fundamental importância em comunicação oral.

Um locutor, com identidade conhecida pelo ouvinte, pode ser facilmente identificado pela pronúncia de uma única palavra ou até mesmo de uma única sílaba. Devido a esta capacidade de reconhecimento e identificação, é possível manter uma conversação simultânea com diferentes pessoas.

A manipulação da característica de identidade percebida de um locutor através de seu sinal de fala é o objetivo da conversão de voz. Em linhas gerais, a voz de um locutor, denominado locutor alvo, deve ser imitada a partir de um sinal de fala de outro locutor, denominado locutor fonte.

Os sistemas de conversão de voz possuem duas etapas distintas. Na primeira, fase treinamento, dado um conjunto de elocuções dos locutores fonte e alvo, uma função de conversão é estimada. O desafio dessa etapa é identificar um conjunto de características que diferencia o sinal de voz dos dois locutores. Na segunda etapa (conversão propriamente dita), parâmetros do locutor alvo são estimados a partir dos parâmetros do locutor fonte e da função de conversão obtida na fase de treinamento.

Idealmente, um sistema completo de conversão de voz é capaz de converter o sinal de fala de qualquer locutor fonte para o de qualquer locutor alvo. Ainda, a conversão deve ser independente das características prosódicas e emotivas dos sinais de fala utilizados seja no treinamento ou na conversão. Não obstante, tanto o treinamento quanto a conversão devem ser realizados independentemente do conteúdo da mensagem e das características do(s) ambiente(s) no(s) qual(is) as elocuções são obtidas.

Os diversos parâmetros que podem ser extraídos do sinal de fala para serem manipulados por um sistema de conversão de voz podem ser divididos em segmentais e suprasegmentais (ou prosódicos). No primeiro, enquadram-se as formas de parametrizar os segmentos do sinal de fala. Como parte das características prosódicas, o sinal de fala exibe freqüência fundamental, duração e intensidade, sendo suas correspondentes qualidades percebidas denominadas, respectivamente, altura, quantidade e volume (*pitch, length* e *loudness*). Apesar da distinção existente entre freqüência fundamental e *pitch*, os termos acústicos e perceptuais são utilizados como equivalentes.

1.1 Aplicações de Sistemas de Conversão de Voz

Uma das motivações para realização deste trabalho é a existência de um considerável número de aplicações de sistemas de conversão de voz. Dentre as possíveis áreas de aplicação temos [1]–[6]:

Personalização da conversão texto fala. A qualidade dos sistemas de conversão texto fala (TTS - *text-to-speech*) vem aumentando consideravelmente nos últimos anos, especialmente, com o franco desenvolvimento de técnicas de síntese concatenativa. No entanto, tais sistemas geram fala com características prosódicas e voz idênticas às do falante considerado no desenvolvimento do sistema TTS. Para obter fala com diferentes características, um novo banco de fala deve ser obtido. Todavia essa tarefa requer grande esforço durante a gravação, manipulação e marcação dos dados. Uma solução plausível para esse problema consiste em utilizar a técnica de conversão de voz como um módulo de pós-processamento para obtenção de fala com identidade alterada [7].

Tradução automática. Assim como em sistemas TTS, sistemas de tradução automática podem utilizar a conversão de voz como um sistema de pós-processamento. O sinal de fala resultante do sistema de tradução automática pode ser convertido por um sistema de conversão de voz de modo a facilitar a conversação entre locutores de diferentes nacionalidades [5], [6], [8].

Ensino de idiomas. Acredita-se que o tempo necessário para o aprendizado de idiomas estrangeiros possa ser sensivelmente reduzido utilizando sistemas de conversão de voz. Nessa aplicação, com o professor como locutor fonte e o estudante como locutor alvo, seria possível gerar um sinal com a identidade percebida do aluno e correta dicção (proveniente da fala do professor) [5], [9]–[11].

Identificação de locutor. Com o objetivo de melhorar os sistemas de reconhecimento, identificação e verificação de locutores é possível simular ataques contra esses sistemas

empregando sinais de fala convertidos [6]. Ainda, o melhor entendimento da iteração dos diferentes mecanismos envolvidos na produção do sinal de fala, devido ao desenvolvimento de novos sistemas de conversão de voz, pode trazer substanciais melhorias à normalização de locutores.

Tratamento médico. Pacientes com doenças degenerativas do aparelho fonador podem ser beneficiados por sistemas de conversão de voz. A voz degradada pode ser convertida para uma outra de melhor inteligibilidade.

Entretenimento. Este é provavelmente o campo com maior possibilidade de aplicação para sistemas de conversão de voz. A voz original de uma celebridade pode ser mantida mesmo após a dublagem de uma entrevista, peça, filme, etc. para um idioma estrangeiro, sem a necessidade do locutor original dominar tal idioma e, até mesmo, sem a participação dele no processo de dublagem. Personagens interpretados por atores que perderam características importantes da fala devido ao envelhecimento ou a doenças seriam beneficiados por técnicas de conversão de voz. Com o objetivo de adaptar uma obra à exibição a diferentes faixas etárias, elocuções indesejadas podem ser substituídas sem perda de qualidade e sem que seja necessário levar o locutor original novamente ao estúdio, técnica denominada *looping* [12]. Um sistema avançado de *karaoke* pode valer-se da conversão de voz para aproximar a voz do usuário a qualquer voz disponibilizada pelo sistema de conversão de voz empregado. Seguindo essa linha, uma canção inédita, composta por um(a) cantor(a) falecido(a), pode ser gravada por outro(a) cantor(a) e a obra, após passar por um sistema de conversão de voz, pode vir a ser percebida como se o(a) cantor(a) original houvesse realizado tal gravação [4]. A indústria de jogos eletrônicos valeria-se desses sistemas no desenvolvimento de jogos com a possibilidade de gerar falas para diferentes personagens ou alterando a entonação das elocuções conforme as necessidades. Não obstante, o usuário pode ter sua voz inserida no jogo, bastando gravar um banco de treinamento [6].

1.2 Revisão Bibliográfica

A conversão de características prosódicas do sinal de fala é o campo menos estudado da conversão de voz. A conversão da flutuação da freqüência fundamental, denominada contorno de *pitch*, do sinal de fala durante a elocução, tido como uma das principais características prosódicas, tem levado atualmente a um grande esforço de pesquisa. As freqüências de *pitch* são obtidas a partir do período de *pitch* que é a diferença de tempo entre dois fechamentos consecutivos da glote, durante a produção de um sinal de fala vozeado.

Esta seção discute as principais técnicas de conversão do contorno de *pitch*, como

parte integrante de um sistema de conversão de voz, destacando suas principais características e limitações. Iniciando pela normalização gaussiana, seguindo pela modelagem por gráfico de dispersão e o modelo de misturas gaussianas. Em seguida, são apresentadas abordagens com aplicações mais específicas. Por fim, comentários referentes aos métodos são expostos.

1.2.1 Normalização Gaussiana

A normalização gaussiana é a técnica mais utilizada para a alteração do contorno de *pitch* em aplicações de conversão de voz. Ela modifica a média e o desvio padrão do contorno de *pitch* do locutor fonte em direção a média e o desvio padrão do contorno de *pitch* do locutor alvo [3], [5], [7], [8], [13]–[23].

Considerando que as freqüências de *pitch* de ambos locutores possuem função distribuição de probabilidade normal, cada valor de freqüência de *pitch* é modificada por

$$\hat{\mathbf{y}} = \frac{(\mathbf{x} - \mu_{\mathbf{x}})}{\sigma_{\mathbf{x}}}\sigma_{\mathbf{y}} + \mu_{\mathbf{y}} \tag{1.1}$$

onde $\mathbf{x} = [x(1)\, x(2)\, \ldots\, x(N)]^{\mathrm{T}}$, $\mathbf{y} = [y(1)\, y(2)\, \ldots\, y(N)]^{\mathrm{T}}$, e $\hat{\mathbf{y}} = [\hat{y}(1)\, \hat{y}(2)\, \ldots\, \hat{y}(N)]^{\mathrm{T}}$ são, respectivamente, os vetores de contorno de *pitch* fonte, alvo e convertido. As variáveis $\mu_{\mathbf{x}}$, $\sigma_{\mathbf{x}}$, $\mu_{\mathbf{y}}$, e $\sigma_{\mathbf{y}}$ representam a média e o desvio padrão de $\mathbf{x}$ e $\mathbf{y}$, respectivamente [15].

A principal vantagem da conversão utilizando normalização gaussiana é sua simplicidade e facilidade na obtenção dos dados de treinamento, uma vez que necessita-se apenas de valores médios e desvios padrão das freqüências de *pitch* de cada um dos locutores. Todavia, por alterar apenas a média e o desvio padrão do contorno de *pitch*, a normalização gaussiana não realiza modificações detalhadas na estrutura de entonação, isto é, não é capaz de alterar o formato da curva que representa o contorno de *pitch* [5], [24]. Uma variação dessa técnica é obtida com a modificação da mediana e a distância inter-quartis, ao invés da média e desvio padrão, por serem as primeiras consideradas mais robustas para o problema em questão [25]. Entretanto, os resultados obtidos são bastante semelhantes.

1.2.2 Modelagem por Gráfico de Dispersão

A modelagem por gráfico de dispersão tem como objetivo reduzir as restrições impostas pelo processo de normalização gaussiana, mais especificamente a obtenção de uma função de mapeamento linear e a suposição de que as freqüências de *pitch* têm função distribuição de probabilidade normal [15].

Após extrair do banco de dados de treinamento, para ambos os locutores, o valor médio das freqüências de *pitch* de cada fonema, obtém-se uma função de n-ésima ordem que melhor represente a função distribuição de probabilidade conjunta do valor médio da freqüência de *pitch* de cada fonema dos dois locutores. Uma abordagem baseada nesse método, adaptada para a conversão de voz, é discutida em [24]. Os resultados indicam um desempenho ligeiramente superior quando comparados com os resultados da normalização gaussiana.

1.2.3 Modelo de Misturas Gaussianas

Como extensão da conversão por normalização gaussiana, o modelo de misturas gaussianas (GMM - *Gaussian mixture model*) considera uma combinação linear de funções de mapeamento [24]. O GMM é baseado na segmentação dos dados em m misturas com função distribuição de probabilidade normal dada por

$$p(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x}; \mu_i; \Sigma_i) \tag{1.2}$$

onde i, μ, Σ e α representam, respectivamente, o número, o valor médio, o desvio padrão e o peso de cada mistura. Por fim, $N(\mathbf{x}; \mu_i; \Sigma_i)$ caracteriza uma distribuição normal unidimensional. Por definição, no modelo GMM a soma do peso de todas as misturas deve ser unitária, ou seja, $\sum_i \alpha_i = 1$. Os parâmetros α, μ e σ do modelo são obtidos com o algoritmo de maximização do valor esperado[1] (EM - *expectation maximization*) [26], [27].

A probabilidade condicional $P(\mathcal{C}_i|\mathbf{x})$ da variável aleatória $\mathbf{x}$ ter sido gerada pelo componente i pode ser calculada com a aplicação da regra de Bayes [26]. Assim,

$$P(\mathcal{C}_i|\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}; \mu_i; \Sigma_i)}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}; \mu_j; \Sigma_j)} \,. \tag{1.3}$$

A função de conversão é obtida com a minimização do erro quadrático médio (MSE - *mean-square error*) entre os valores das freqüências de *pitch* convertidas e alvo. Para tanto, são obtidas estimativas de um modelo GMM de densidade conjunta fonte e alvo, com função densidade de probabilidade $p(\mathbf{z}) = p(\mathbf{x},\mathbf{y})$, com $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$. O valor esperado de $\mathbf{y}$ dado $\mathbf{x}$

[1] O objetivo do algoritmo EM é maximizar a função de verossimilhança do conjunto de parâmetros que define o modelo de misturas gaussianas. O algoritmo é constituído de duas partes, na primeira é calculada a verossimilhança esperada para, então, na segunda, maximizar essa função.

($E[\mathbf{y}|\mathbf{x}]$) pode ser estimado com a seguinte regressão [2]:

$$\begin{aligned}\mathcal{F}(\mathbf{x}) = E[\mathbf{y}|\mathbf{x}] &= \int_{-\infty}^{+\infty} \mathbf{y} p(\mathbf{y}|\mathbf{x})\, dy \\ &= \sum_{i=1}^{M} [\mu_i^{\mathbf{y}} + \boldsymbol{\Sigma}_i^{\mathbf{yx}} (\boldsymbol{\Sigma}_i^{\mathbf{xx}})^{-1} (\mathbf{x} - \mu_i^{\mathbf{x}})] P(\mathcal{C}_i|\mathbf{x})\end{aligned} \tag{1.4}$$

onde $\boldsymbol{\Sigma}_i$ é a matriz de covariância conjunta do *i*-ésimo componente da gaussiana, definido como

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{\mathbf{xx}} & \boldsymbol{\Sigma}_i^{\mathbf{xy}} \\ \boldsymbol{\Sigma}_i^{\mathbf{yx}} & \boldsymbol{\Sigma}_i^{\mathbf{yy}} \end{bmatrix} \tag{1.5}$$

e $\boldsymbol{\Sigma}_i^{\mathbf{xx}}$ representa a covariância de $\mathbf{x}$ de tal forma que

$$\boldsymbol{\Sigma}_i^{\mathbf{xx}} = E\{[\mathbf{x} - \mu_i^{\mathbf{x}}][\mathbf{x} - \mu_i^{\mathbf{x}}]^{\mathrm{T}}\} \tag{1.6}$$

ainda, $\boldsymbol{\Sigma}_i^{\mathbf{xy}}$ representa a covariância cruzada de $\mathbf{x}$ e $\mathbf{y}$, dada por

$$\boldsymbol{\Sigma}_i^{\mathbf{xy}} = E\{[\mathbf{x} - \mu_i^{\mathbf{x}}][\mathbf{y} - \mu_i^{\mathbf{y}}]^{\mathrm{T}}\}. \tag{1.7}$$

Por fim, $\Sigma_i^{\mathbf{yx}}$ é o transposto de $\boldsymbol{\Sigma}_i^{\mathbf{xy}}$, $\boldsymbol{\Sigma}_i^{\mathbf{yy}}$ é a matriz de covariância de $\mathbf{y}$ e o parâmetro $\boldsymbol{\mu}$ do modelo GMM de densidade conjunta é

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_i^{\mathbf{x}} \\ \mu_i^{\mathbf{y}} \end{bmatrix}. \tag{1.8}$$

Dessa forma, (1.3) é reescrita como

$$P(\mathcal{C}_i|\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}; \mu_i^{\mathbf{x}}; \Sigma_i^{\mathbf{xx}})}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}; \mu_j^{\mathbf{x}}; \Sigma_j^{\mathbf{xx}})}. \tag{1.9}$$

A Equação (1.4) não faz qualquer consideração a respeito da função distribuição de probabilidade das freqüências de *pitch* do locutor alvo visto que o particionamento é realizado a partir de observações de ambos locutores. Além disso, como a dimensão do espaço de parâmetros dobra, o esforço computacional requerido pelo algoritmo EM será consideravelmente maior [7]. Dessa forma, o modelo GMM pode ser visto como um modelo oculto de Markov (HMM - *hidden Markov model*) simplificado que possui distribuição de probabilidade de transição de estado normal, com todos os estados conectados (modelo ergódico) e todas as probabilidades de transição que levam a um estado com mesmos valores [3]. A consideração de que cada mistura é gaussiana está de acordo com o teorema do limite central,

quando dispõe-se de dados de treinamento suficientes[2].

Resultados da aplicação do modelo GMM na conversão do contorno de *pitch* indicam um desempenho superior quando comparado com a normalização gaussiana e semelhantes aos obtidos com modelagem por gráfico de dispersão [24]. Além disso, por ser considerada uma técnica mais refinada, quando comparada com a normalização gaussiana, o modelo GMM é utilizado como medida de comparação de desempenho no desenvolvimento de novas técnicas [28].

1.2.4 Seleção de Contorno

O objetivo da seleção de contorno é utilizar um dos contornos de *pitch* observado no treinamento para substituir o contorno de *pitch* da elocução a ser convertida. Inicialmente proposto em [15] e refinado em [24], a seleção de contorno emprega um contorno que é produzido pelo locutor alvo, ao contrário das abordagens introduzidas anteriormente que buscam a modificação do contorno de *pitch* da elocução a ser convertida. Métodos baseados na seleção de contorno fundamentam-se na consideração de que ambos locutores podem gerar contornos de *pitch* similares para diferentes elocuções.

Enquanto em [15] o contorno de *pitch* de menor distância DTW[3] (*dynamic time warping*) da elocução inteira é transplantado, em [24] uma interpolação linear de todos os contornos observados no treinamento é realizada.

Como vantagem frente as técnicas apresentadas anteriormente, a seleção de contorno é capaz de utilizar um contorno de *pitch* que foi produzido pelo locutor alvo ao invés de um contorno de *pitch* manipulado (artificial). Por outro lado, o desempenho das técnicas baseadas em seleção de contorno é altamente dependente do tamanho do banco de dados de treinamento, especialmente quando se busca alterar o padrão de entonação da elocução. Em aplicações específicas, com restrições quanto à variabilidade do contorno de *pitch* ou com vocabulário limitado, a seleção de contorno se torna mais atrativa [4].

Uma extensão natural da seleção de contorno é utilizar segmentos do contorno de *pitch* ao invés do contorno de *pitch* da elocução inteira. Dependendo da aplicação, os segmentos podem ser definidos como segmentos vozeados [4], [24], sílabas [28] ou fonemas, como será discutido posteriormente no presente trabalho.

Independente de como o segmento de contorno de *pitch* é definido, o segmento a ser convertido é comparado com todos os segmentos do locutor fonte observados no trei-

[2]De acordo com o teorema do limite central, seria necessário uma quantidade de dados de treinamento infinita.

[3]O DTW é um algoritmo baseado em programação dinâmica que, neste caso, fornece uma medida de similaridade entre os contornos de *pitch* fonte e alvo, mesmo que desalinhados [29].

namento. Em seguida, obtém-se de um banco de fala paralelo[4], o segmento de contorno de *pitch* produzido pelo locutor alvo que coincide com o selecionado. Para possibilitar as abordagens que utilizam segmentos do contorno de *pitch*, é necessário um banco de sinais de fala paralelo corretamente alinhado.

1.2.5 Transformação Prosódica Estilística

A transformação prosódica estilística[5,6] é aplicada à conversão de voz entre dois idiomas com o objetivo principal de limitar o fator de modificação do contorno de *pitch*, evitando assim, a obtenção de valores possivelmente incorretos que poderiam introduzir distorções no sinal sintetizado [6].

A transformação, além de alterar a média e o desvio-padrão, também modifica a inclinação do contorno de *pitch*, através de uma reta estimada para modelar sua evolução global na elocução.

Além dos resultados mostrados em [6] serem dependentes de diferentes ajustes de parâmetros, que devem ser testados e alterados repetidas vezes conforme a conversão desejada, a transformação é incapaz de realizar alterações locais no contorno de *pitch*. Devido às peculiaridades da aplicação, o contorno alvo não é considerado em [6].

1.2.6 Comentários

Apesar dos resultados nem sempre satisfatórios, o método da normalização gaussiana é bastante utilizado e aceito como um método capaz de modificar de forma generalizada as principais características prosódicas da elocução [3]. Porém, detalhes na estrutura do contorno de *pitch* não podem ser modelados, dado que apenas a média e o desvio padrão são modificados. Dessa forma, o resultado será o contorno de *pitch* do locutor fonte com a média e o desvio padrão do contorno de *pitch* do locutor alvo.

Eliminando algumas das restrições impostas pela normalização gaussiana, dentre elas a obtenção de uma função de conversão linear e a suposição de que os dados têm função distribuição de probabilidade normal, tanto a modelagem por gráfico de dispersão quanto o modelo de misturas gaussianas conduzem a um refinamento na conversão do contorno de *pitch* para a conversão de voz. Todavia, sem melhorias expressivas.

[4] Por banco de fala paralelo entende-se que as elocuções são as mesmas para todos os locutores.

[5] Título original: *Stylistic Prosody Transformation*.

[6] A estilística é um ramo da lingüística responsável pelo estudo das variações de estilo empregadas em diferentes contextos.

Uma limitação inerente ao modelo GMM refere-se à consideração fundamental de que os vetores de observação são independentes. Essa simplificação torna o modelo adequado a aplicações em que se acredita que o aspecto seqüencial das observações, ou seja, o índice temporal, seja irrelevante, o que não é verdadeiro quando se procura converter características prosódicas do sinal de fala.

Buscando a utilização de um contorno de *pitch* produzido pelo locutor alvo, disponível no banco de treinamento, a seleção de contorno tem se mostrado um método promissor. Diferentes abordagens podem ser desenvolvidas baseadas em variações dessa técnica de seleção em função da aplicação em questão e dos dados disponíveis para realizar o treinamento.

Para concluir a revisão dos métodos estudados, alguns pontos importantes que diferenciam as abordagens citadas e a forma como seus resultados são apresentados precisam ser salientados:

Medida de distância. A dificuldade em adotar uma métrica capaz de estimar a distância ou distorção entre dois contornos de *pitch* resulta na incapacidade de confrontar os resultados das diferentes técnicas existentes. O maior problema recai sobre a inexistência de um único contorno de *pitch* alvo, dado que um mesmo locutor dificilmente pronuncia repetidas vezes a mesma elocução com as mesmas características prosódicas. Entretanto, a adoção de uma medida de distância possibilita refinar a implementação de diferentes técnicas sem que testes subjetivos tenham de ser realizados repetidas vezes.

Aplicações. As peculiaridades de cada método, em que nem sempre o contorno de *pitch* do locutor alvo é conhecido, limitam as possíveis aplicações dos sistemas de conversão do contorno de *pitch* e, novamente, a capacidade de comparar resultados entre diferentes abordagens.

Resultados subjetivos. A importância da realização de testes subjetivos recai não somente na validação da medida de distância adotada, como também na ratificação dos resultados obtidos. A realização dos testes subjetivos se torna indispensável visto que é possível obter uma melhora de desempenho subjetivo, mesmo com a piora dos resultados objetivos. Dentre os fatores responsáveis por essas aparentes inconsistências temos a incapacidade de definir um único contorno de *pitch* alvo, visto que o mesmo locutor pode gerar diferentes padrões de entonação para a mesma elocução. Assim, o principal objetivo dos sistemas de conversão de contorno de *pitch* é fornecer um contorno que seja considerado como aceitável em testes subjetivos para que, em conjunto com um sistema completo de conversão de voz, o resultado geral seja ainda satisfatório.

1.3 Objetivos

Dentre os fatores determinantes na escolha da área de pesquisa podemos citar: a diversidade de aplicações dos sistemas de conversão de voz e a relativa escassez de abordagens para o tratamento das características supra-segmentais propostas na literatura (quando comparada com a quantidade de métodos propostos para a conversão de características segmentais).

Observando as limitações dos métodos de conversão do contorno de *pitch* apresentados, bem como a dificuldade de avaliação do desempenho e comparação dos resultados, os objetivos gerais e específicos serão detalhados de modo a contornar, sempre que possível, as limitações dos atuais sistemas de conversão do contorno de *pitch* para aplicação em sistemas de conversão de voz.

1.3.1 Objetivos Gerais

Temos como principal objetivo desenvolver um método de conversão do contorno de *pitch* para aplicações em sistemas de conversão de voz capaz de sobrepujar, tanto em testes objetivos quanto subjetivos, os métodos mais utilizados na literatura pesquisada. Para tanto, faz-se necessário empregar uma medida de distância adequada para avaliação objetiva bem como o desenvolvimento de testes subjetivos para validar a abordagem proposta.

Uma vez que os sistemas de conversão de voz possuem ampla gama de possíveis aplicações o método de conversão do contorno de *pitch* deve ser estruturado de forma a ser facilmente incorporado aos mais diferentes tipos de sistemas de conversão de voz. Portanto é desejável que o método de conversão do contorno de *pitch* seja independente de dados existentes apenas em bancos de dados específicos para conversão texto-fala e também independente de parâmetros de simulação que necessitem ser testados e reconfigurados.

1.3.2 Objetivos Específicos

Para possibilitar o desenvolvimento do sistema de conversão do contorno de *pitch* proposto e alcançar os objetivos gerais é necessário detalhar os objetivos específicos como segue:

Banco de dados. Extrair, diretamente do sinal de fala, parâmetros adequados para o problema em questão e armazená-los para obter um banco de dados apropriado para o desenvolvimento de um sistema de conversão de contorno de *pitch*.

Treinamento da função de conversão. De posse do banco de dados, utilizar um conjunto adequado de ferramentas e procedimentos para obter uma função de conversão capaz de alterar o contorno de *pitch*.

Conversão do contorno de *pitch*. Empregar métodos de análise e síntese de sinais de fala que sejam capazes de alterar o contorno de *pitch*, mantendo níveis aceitáveis de qualidade no sinal sintetizado.

Comparação objetiva. Definir uma métrica adequada para a comparação objetiva do método proposto com duas das técnicas disponíveis na literatura pesquisada, mais especificamente, a conversão do contorno de *pitch* utilizando normalização gaussiana e o modelo de misturas gaussianas.

Testes subjetivos. Desenvolver testes subjetivos apropriados para ratificar a análise objetiva realizada.

Facilidade de integração. Evitar, sempre que possível, parâmetros que precisem ser testados e reconfigurados, facilitando assim a integração com um sistema completo de conversão de voz. Além do mais, extrair, sempre que possível, todos os dados necessários diretamente do sinal de voz, não restringindo as possíveis aplicações para a abordagem proposta.

Todas as hipóteses estudadas e testadas no decorrer do estudo são baseadas em elocuções do português falado no Brasil (ver banco de dados apresentado no Apêndice A).

1.4 Estrutura da Dissertação

A dissertação está organizada como segue. O Capítulo 1 introduziu o conceito de conversão de voz, suas aplicações, os principais métodos de conversão do contorno de *pitch*, o foco da pesquisa e os objetivos gerais e específicos. Foram ainda elucidadas as principais dificuldades encontradas na avaliação de desempenho dos diferentes métodos de conversão do contorno de *pitch*. Importantes definições relativas ao processo de produção da fala e uma breve introdução ao sistema sonoro do português brasileiro são discutidas no Capítulo 2. A descrição detalhada do método proposto, suas limitações, resultados e observações pertinentes são apresentados no Capítulo 3. Os testes subjetivos e os resultados obtidos com o método proposto são mostrados no Capítulo 4. Por fim, as conclusões, comentários e sugestões para trabalhos futuros fazem parte do Capítulo 5. Os detalhes das elocuções do banco de sinais de fala empregado no desenvolvimento do banco de dados, testes, avaliações objetivas, subjetivas e informais estão detalhadas no Apêndice A.

Capítulo 2

Produção da Fala

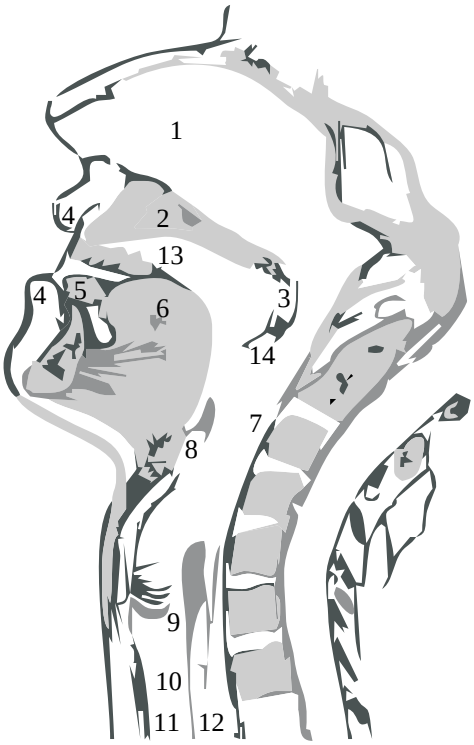
O sinal de fala é caracterizado por uma forma de onda de relativa complexidade devido a suas características dinâmicas. Entretanto, o estudo das técnicas de processamento de sinais de fala pode ser facilitado através do conhecimento dos principais mecanismos envolvidos no seu processo de produção. Para tanto, este capítulo apresenta uma descrição introdutória do aparelho fonador humano e o sistema sonoro do português brasileiro com o objetivo de elucidar fenômenos comumente observados nos bancos de sinais de fala.

2.1 Aparelho Fonador Humano

O ser humano não possui órgãos fonadores específicos para esse fim. Os diferentes órgãos utilizados na produção de fala foram adaptados para essa função relativamente tarde na história da evolução humana. Os pulmões, a laringe e a cavidade nasal são partes integrantes do mecanismo respiratório. A língua, os dentes, a glote, dentre outros elementos, desempenham funções no processo digestório [30]–[32]. Como nenhum órgão realiza função específica à fonação, para a área de interesse deste trabalho, utiliza-se o termo *aparelho fonador* para descrever os órgãos, direta ou indiretamente, relacionados à produção de fala.

O aparelho fonador é dividido em três sistemas: respiratório ou subglotal, fonatório ou laringeal e articulatório ou supralaringeal. O sistema subglotal é formado pelos pulmões, músculos pulmonares, tubos brônquios e traquéia. A laringe e seus constituintes formam o sistema fonatório. O sistema articulatório é formado pelas cavidades nasais, cavidade oral, língua, palato, nariz, dentes e lábios. Para evidenciar a localização das estruturas que constituem o aparelho fonador humano um diagrama simplificado é ilustrado na Figura 2.1.

Dentre as diferentes funções desempenhadas pelas estruturas que constituem o aparelho fonador humano, podemos destacar no escopo da produção de sinais de fala, os seguintes elementos:



1 - cavidade nasal
2 - palato duro
3 - véu palatino
4 - lábios
5 - dentes
6 - língua
7 - faringe
8 - epligote
9 - pregas vocais
10 - laringe
11 - traquéia
12 - esôfago
13 - alvéolo
14 - úvula
pulmões e brônquios (não representados)

Figura 2.1: Diagrama simplificado do aparelho fonador humano.

Pulmões. Órgãos respiratórios que fornecem a corrente de ar, matéria-prima da fonação.

Brônquios. Tubos que conectam os pulmões à traquéia.

Traquéia. Faz a conexão da laringe com os brônquios. A cavidade formada pelos brônquios e a traquéia atua como um ressoador de baixa freqüência.

Laringe. Situa-se entre a traquéia e a língua, sendo constituída por uma série de cartilagens revestidas por uma membrana mucosa que é movimentada pelos músculos da laringe[1]. As dobras da membrana mucosa dão origem às pregas vocais. Na laringe encontra-se também a epiglote[2].

Faringe. Interliga as cavidades nasais e oral com a laringe. Suas cavidades[3] funcionam como uma caixa de ressonância de tamanho variável.

Cavidades Nasais. Cavidades paralelas entre as narinas e a faringe. Também assume papel de ressoador.

[1]Músculos cricoaritenóideos posterior e lateral.

[2]A epiglote é responsável por evitar a comunicação do aparelho respiratório com o aparelho digestório durante a deglutição.

[3]A faringe é formada pela naso, oro e laringofaringe.

Boca. Graças, sobretudo, ao movimento da língua e do maxilar, a boca pode variar de forma e volume, alterando assim o fluxo de ar e, conseqüentemente, gerar diferentes segmentos de fala. Na boca, encontra-se ainda a úvula que, mesmo não possuindo movimentação própria, pode ser vista como um importante articulador, pois se movimenta em conjunto com a língua. A úvula tem função de impedir ou permitir a passagem de ar pelas cavidades nasais através, respectivamente, de seu levantamento ou abaixamento. Além disso, a úvula também pode vibrar. Por fim, os lábios constituem a terminação do aparelho fonador, sendo capazes de produzir movimentos bastante distintos.

No processo de produção de fala, o ar expelido dos pulmões por via dos brônquios, penetra na traquéia e chega à laringe, onde, ao atravessar a glote, costuma encontrar o primeiro obstáculo à sua passagem. Nesse ponto, o fluxo de ar pode encontrá-la aberta ou fechada. Se estiver aberta, o ar força a passagem através das pregas vocais retesadas, fazendo-as vibrar e produzir o som musical característico dos segmentos de fala vozeados (ou sonoros). Se estiver fechada, com as pregas vocais relaxadas, o ar escapa da laringe sem vibrações, formando assim os segmentos de fala denominados não vozeados (ou surdos) [30].

A iteração das pregas vocais com a corrente expiratória produz um sinal aproximadamente periódico, com freqüência fundamental definida pelas características constituintes das pregas vocais. Dentre os fatores que influenciam as características das pregas vocais, podemos citar: idade, sexo e peculiaridades do desenvolvimento individual, sendo que na puberdade a laringe e as pregas vocais experimentam um considerável crescimento, definindo importantes características relativas à identidade do locutor [31].

A Figura 2.2 apresenta um esboço da movimentação das pregas vocais para um período do sinal de excitação, ou seja, um período de *pitch*. Em (a) e (b), a pressão do ar na traquéia força as pregas vocais até conseguir desobstruir a passagem. Nas partes (c) e (d), o ar flui através das pregas, forçando o surgimento de nova constrição, porém, dessa vez, na parte inferior das pregas vocais. Por fim, em (e) e (f), o sinal de excitação empurra a obstrução para a parte de cima, iniciando um novo ciclo do processo, ou seja, o próximo período de *pitch*.

As cavidades supraglóticas atuam como ressoadores, impondo importantes alterações ao sinal de excitação. Os formantes, freqüências de ressonância do trato vocal, são resultado das diferentes configurações possíveis dessas cavidades ressoadoras em conjunto com os demais aparatos do aparelho fonador. O valor da freqüência dos quatro primeiros e principais formantes (F_1, F_2, F_3 e F_4) tem relação com as seguintes configurações:

i) O deslocamento da língua no plano vertical basicamente define o valor do primeiro formante (F_1).

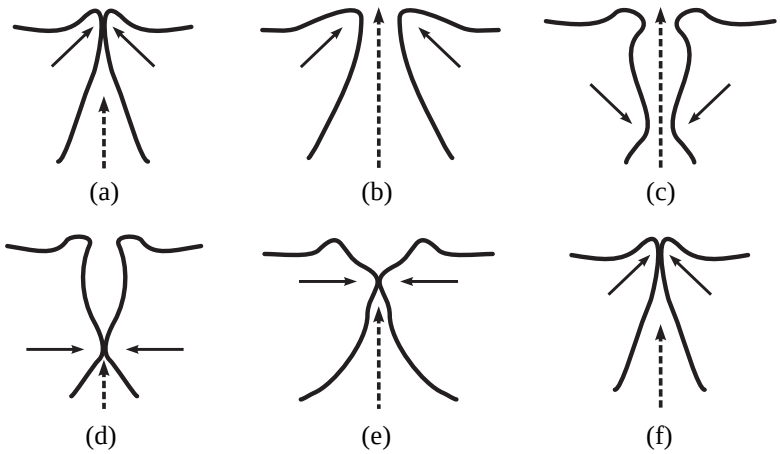


Figura 2.2: Esboço do movimento das pregas vocais.

ii) A freqüência do segundo formante (F_2) é determinada pelo deslocamento da língua no plano horizontal.

iii) O grau de obstrução formado entre língua e faringe define o valor do terceiro formante (F_3).

iv) A freqüência do quarto formante (F_4) é função da posição vertical da laringe.

O valor da freqüência dos formantes pode também sofrer pequenas alterações em função da posição dos lábios [32]. Quando o palato mole se encontra abaixado, ocorre acoplamento do trato vocal com as cavidades nasais. Esse acoplamento resulta na interação das freqüências de ressonâncias. Além disso, é possível ocorrer freqüências de anti-ressonâncias ou anti-formantes. Os anti-formantes são freqüências de ressonância próximas das freqüências dos formantes que causam a redução de amplitude desses formantes devido à perda de energia causada pelo acoplamento entre o trato vocal e as cavidades nasais [32].

2.2 Sistema Sonoro do Português Brasileiro

A informação contida na fala pode ser representada pela concatenação de um conjunto finito de elementos, denominados fones. Cada língua possui seu próprio conjunto de fones.

Vislumbrando o entendimento dos mecanismos presentes no processo de produção dos sinais de fala e da relação entre os órgãos do sistema fonador, apresentados na Seção 2.1, as seções seguintes introduzem importantes conceitos da Fonologia do português brasileiro (doravante PB).

2.2.1 Vogais

As vogais se diferenciam dos demais segmentos de fala pelo fato de o sinal de excitação que as produz não experimentar nenhuma obstrução. Ou seja, não existe contato entre articuladores ativos e passivos [32], [33].

O número de vogais do PB varia em função da posição da vogal na palavra: existem sete vogais tônicas orais e cinco nasais, cinco pretônicas, quatro postônicas não-finais e três postônicas em final de palavra. Quando em posição tônica, as vogais criam sete oposições do tipo s[i]lo, s[e]co, s[ɛ]co, s[a]co, s[ɔ]co, s[o]co e s[u]co, conforme Tabela 2.1.

Tabela 2.1: Vogais em posição tônica

	Não-arredondadas			Arredondadas		
altas	/i/				/u/	
médias		/e/		/o/		(2º grau)
médias			/ɛ/	/ɔ/		(1º grau)
baixa			/a/			
	anterior		central	posterior		

Nas sílabas átonas, o sistema vocálico de sete vogais fica reduzido devido à perda de um traço distintivo, fenômeno denominado neutralização. A neutralização é caracterizada pela junção de dois fonemas em uma única unidade fonológica. Ex.: caf[ɛ] - caf[e]teira, b[ɛ]lo - b[e]leza e s[ɔ]l - s[o]laço.

Quando as vogais médias pretônicas assimilam a altura da vogal alta da sílaba imediatamente seguinte, ocorre a harmonia vocálica[4]. Na ocorrência da harmonia vocálica, são encontradas variantes como p[e]pino - p[i]pino e c[o]ruja - c[u]ruja. Uma situação semelhante se repete com o /e/ e /o/ pretônicos em hiato com um /a/ tônico, como nos infinitivos *voar* e *passear*. O /i/ e o /u/ tendem a substituir o /e/ e o /o/, respectivamente, resultando em pronúncias como [vuar] e [pasiar]. Essas atrofias ou hipertrofias dos elementos do sistema vocálico são comumente referenciadas por flutuações [34].

Conforme [34], [35], quando em posição postônica, as vogais podem ser subdivididas em não-final e final. Quando em posição não-final, a neutralização ocorre apenas entre as vogais posteriores /o/ e /u/, sendo que as anteriores permanecem inalteradas.

[4]As variações que não causam a neutralização constituem objeto de estudo do modelo variacionista e são comumente denominadas flutuações.

2.2.2 Consoantes

Diferentemente das vogais, durante o processo de produção das consoantes, o sinal de excitação sofre obstrução total ou parcial em um ou mais pontos do sistema fonador. O resultado dessa obstrução é o *ruído* característico das consoantes, em contraste com os segmentos provenientes das vogais. O ruído proveniente das consoantes pode ser caracterizado por um som aperiódico, tanto contínuo quanto plosivo, e apresenta uma energia acústica consideravelmente menor do que a das vogais.

As consoantes do PB podem ser separadas em labiais, anteriores e posteriores, conforme Tabela 2.2 [35].

Tabela 2.2: Consoantes do português brasileiro

Labiais	/p/	/b/	/f/	/v/	/m/		
Anteriores	/t/	/d/	/s/	/z/	/n/	/l/	/ɾ/
Posteriores	/k/	/g/	/ʃ/	/ʒ/	/ɲ/	/ʎ/	/ɾ/

Consoantes pré-vocálicas ocorrem em fase inicial da obstrução da passagem do ar, dominando a fase inicial em que se desfaz a obstrução e é superado o impedimento bucal à passagem de corrente expiratória [35], [36]. Por fim, na consoante pós-vocálica, a articulação se concentra na fase de fechamento, e a abertura da boca, que produz a vogal silábica, se reduz ou anula, sem solução de continuidade, para criar o elemento consonântico de travamento da sílaba [36].

As consoantes também podem ser classificadas em função da região e maneira como são articuladas, ou seja, o ponto e o modo de articulação, respectivamente. Quando classificadas quanto ao modo de articulação, são divididas em oclusivas e constritivas (fricativas, laterais ou vibrantes). Quando a classificação é realizada de acordo com o ponto de articulação, as consoantes podem ser classificadas como bilabiais, labiodentais, linguodentais, alveolares, palatais e velares. Ainda, considerando o papel das pregas vocais, as consoantes podem ser denominadas vozeadas (sonoras) ou não vozeadas (surdas). Por fim, as consoantes podem ser orais ou nasais, dependendo do papel das cavidades bucal e nasais.

2.2.3 Semivogais

Semivogais são as vogais assilábicas i̯ e u̯ em encontros vocálicos, formando ditongos decrescentes e ditongos crescentes. Os ditongos decrescentes são formados pela seqüência de uma vogal e uma semivogal, podendo ser orais ou nasais. Ditongos crescentes são constituído de uma semivogal seguida de uma vogal e são sempre orais [37].

2.3 Conclusões

O entendimento das possíveis combinações dos articuladores no processo de produção de fala, das limitações fisiológicas do aparelho fonador e da forma como o sistema sonoro do português brasileiro se organiza são de grande valia no desenvolvimento de sistemas de processamento de sinais de fala.

As flutuações das realizações dos elementos do sistema sonoro, além das variantes resultantes da harmonia vocálica, são responsáveis por grandes alterações na forma como diferentes locutores realizam uma determinada elocução. Em função desses fenômenos, importantes considerações a respeito do alinhamento das elocuções do banco de treinamento serão apresentadas no Capítulo 3.

Capítulo 3

Abordagem Proposta

Este capítulo apresenta o método desenvolvido para conversão do contorno de *pitch* para aplicações em conversão de voz. Para possibilitar a conversão do contorno de *pitch* é necessário extrair todos os dados importantes das elocuções para o desenvolvimento do banco de treinamento. Em seguida, as elocuções são alinhadas para que seja possível realizar a comparação objetiva dos resultados obtidos com aqueles obtidos pelos diferentes métodos de conversão do contorno de *pitch* implementados. Para tanto, é introduzido o parâmetro índice de desempenho obtido a partir de duas medidas de distância calculadas entre os contornos de *pitch* fonte, alvo e convertido. A abordagem utilizada no desenvolvimento deste trabalho é discutida após avaliar o desempenho dos dois métodos de conversão do contorno de *pitch* mais difundidos na literatura, a saber: conversão utilizando a normalização gaussiana (GN) e utilizando o modelo de misturas gaussianas (GMM), introduzidos, respectivamente, nas Seções 1.2.1 e 1.2.3. Por fim, são apresentadas as conclusões acerca do método proposto.

3.1 Contorno de *Pitch*

Conforme introduzido no Capítulo 1, o contorno de *pitch* é a evolução temporal dos instantes de fechamento da glote durante a produção de sinais vozeados. Preferencialmente, o contorno de *pitch* é estimado com auxílio de um equipamento denominado laringógrafo. O laringógrafo utiliza um conjunto de eletrodos conectados em ambos os lados do pescoço do locutor, vislumbrando determinar os instantes de abertura e fechamento da glote. Todavia, para a maioria das aplicações de conversão de voz, é de grande interesse que o contorno de *pitch* seja determinado de forma rápida e sem a necessidade de equipamentos adicionais. Especialmente, se for necessário inserir novos locutores no banco de sinais de fala. Portanto, é desejável obter uma estimativa do contorno de *pitch* a partir da forma de onda do sinal de fala. O algoritmo de determinação do contorno de *pitch* utilizado neste trabalho é parte

integrante do programa Praat [38] detalhado em [39].

A Figura 3.1 apresenta um segmento de fala vozeada (fonema /i/) bem como as respectivas marcas de *pitch*, representadas por linhas verticais. Através desta ilustração pode ser observada a periodicidade de um sinal de fala vozeado. Um exemplo do contorno de *pitch* para a elocução "Existem muitos camundongos diferentes naquela ilha." realizada por dois locutores é mostrado na Figura 3.2(a).

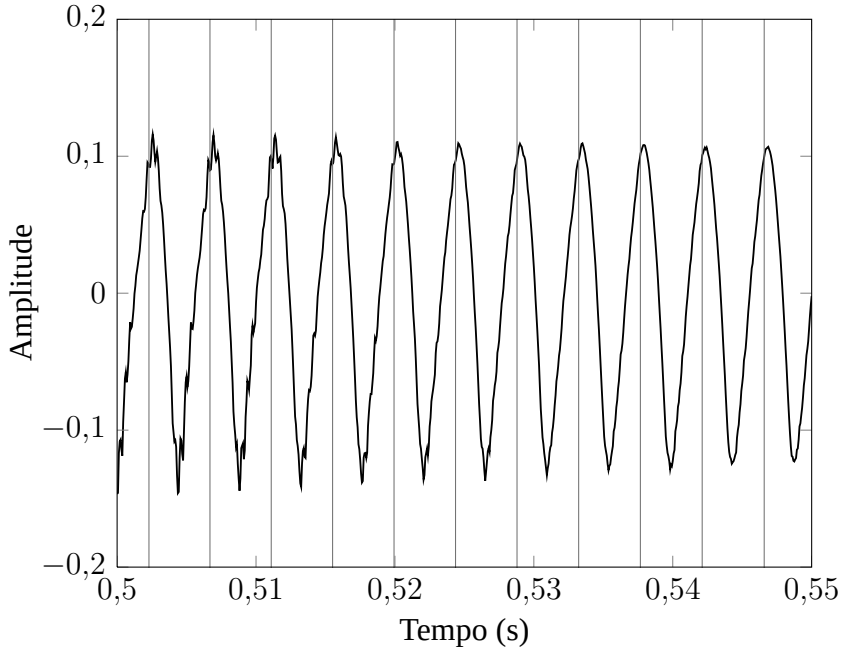
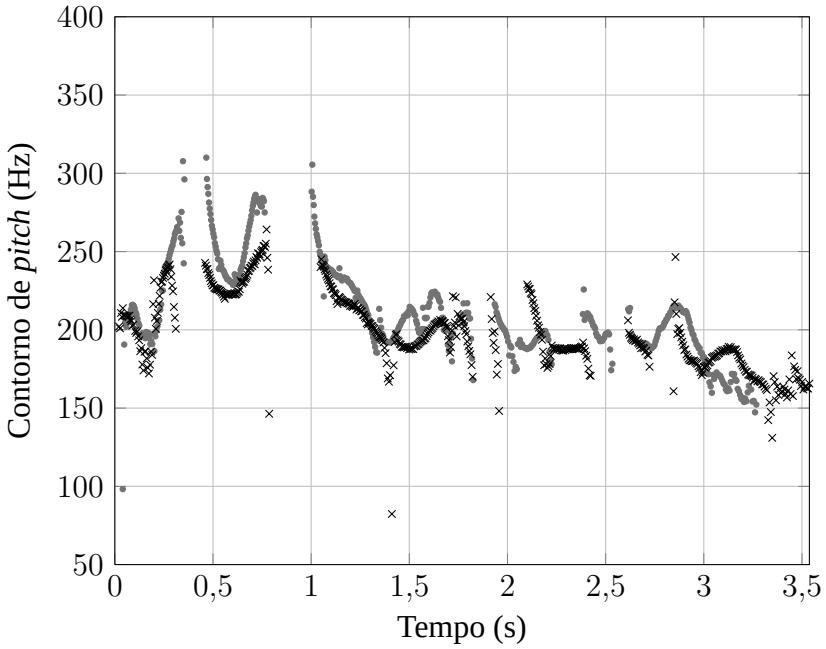


Figura 3.1: Segmento de fala vozeada com marcação de *pitch*, fonema /i/.

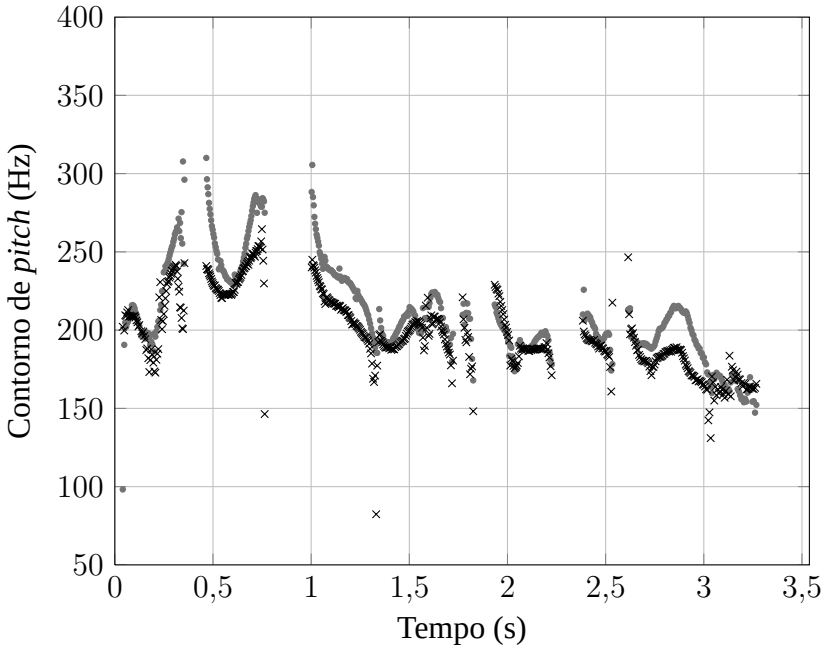
3.2 Alinhamento

Após determinar o contorno de *pitch* de todas as elocuções disponíveis na banco de sinais de fala é necessário alinhar tais elocuções. Elas devem estar alinhadas para que seja possível obter uma equivalência entre dados de diferentes locutores.

Conforme comentado na Seção 1.2.4, os sistemas de conversão do contorno de *pitch* discutidos na literatura convertem seja o contorno de *pitch* da elocução inteira [15], [24] ou através da seleção de segmentos. Para a seleção de segmentos, é possível defini-los como segmentos vozeados [4], [24] ou pela marcação silábica da elocução [28]. Entretanto, devido as flutuações ou a variantes causadas por harmonia vocálica (ver Seção 2.2.1) presentes nas elocuções produzidas por diferentes locutores ou mesmo em elocuções produzidas repetidas vezes pelo mesmo locutor, optou-se por trabalhar com o contorno de *pitch* de acordo com a segmentação fonética. Assim, buscamos evitar que formas particulares de pronúncia de



(a)



(b)

Figura 3.2: Exemplo do contorno de *pitch* da mesma elocução realizada por dois locutores. (a) Original. (b) Alinhado.

determinadas sílabas influenciem o resultado da conversão do contorno de *pitch*. Outra vantagem de se operar sobre segmentos que representam os fonemas vem do fato de os fonemas representarem a menor unidade sonora de uma língua (ver Seção 2.2).

De posse da marcação fonética das elocuções do banco de sinais de fala, são determinados o início, fim e duração de cada fonema[1]. Em seguida, o contorno de *pitch* (extraído conforme Seção 3.1) é segmentado de acordo com a marcação fonética. O alinhamento é então iniciado comparando a seqüência fonética da elocução dos locutores fonte e alvo. Após tal comparação, alguns fonemas podem ser eliminados. A eliminação de fonemas não coincidentes permite que o banco de treinamento possua apenas dados que representem os segmentos de contorno de *pitch* que ambos os locutores pronunciaram, isto é, são apenas mantidos os segmentos de contorno de *pitch* equivalentes entre os dois locutores e são eliminadas as possíveis inserções, apagamentos ou substituições fonéticas causadas por flutuações ou variantes.

Após o apagamento fonético, os segmentos de contorno de *pitch* dos fonemas coincidentes são armazenados no banco de treinamento na forma de pares de segmentos, garantindo assim que seja mantida uma equivalência entre os segmentos fonte e alvo. Para ilustrar o resultado do alinhamento em uma elocução inteira, a Figura 3.2(b) apresenta os mesmos contornos de *pitch* da Figura 3.2(a), porém agora alinhados.

3.3 Avaliação Objetiva

Uma vez que um mesmo locutor pode pronunciar determinada elocução de diferentes maneiras não existe um único contorno de *pitch* (fonte ou alvo). Entretanto, alguma medida de distância pode ser adotada para que seja possível comparar os resultados. Para tanto, é considerado que os contornos de *pitch* extraídos do banco de sinais de fala representem a forma *mais usual* do locutor pronunciar uma dada elocução e, portanto, são os contornos de *pitch* fonte e alvo utilizados na avaliação de desempenho.

Na alteração de parâmetros visando à conversão de voz, três medidas de distância são de maior interesse:

i) A distância inicial existente entre os contornos de *pitch* fonte e alvo, denominada distância fonte-alvo $D(\mathbf{x}, \mathbf{y})$.

ii) A distância final medida entre os contornos de *pitch* convertido e fonte, denominada de distância convertido-fonte $D(\hat{\mathbf{y}}, \mathbf{x})$.

[1] Segmentação automática, através de alinhamento forçado.

iii) A distância remanescente após o processo de conversão, medida entre os contornos de *pitch* convertido e alvo, denominada distância convertido-alvo $D(\hat{\mathbf{y}}, \mathbf{y})$. Essa medida avalia o quão perto do alvo o contorno de *pitch* convertido está.

A distância $D(\mathbf{x}, \mathbf{y})$ é obtida a partir dos contorno de *pitch* fonte $\mathbf{x}$ e alvo $\mathbf{y}$

$$D(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| \tag{3.1}$$

onde $|| \cdot ||$ representa a norma euclidiana. As distâncias $D(\hat{\mathbf{y}}, \mathbf{x})$ e $D(\hat{\mathbf{y}}, \mathbf{y})$ são obtidas de forma semelhante, alterando apenas o(s) contorno(s) de *pitch* em questão.

3.3.1 Índice de Desempenho

O índice de desempenho (PI - *performance index*) é uma formulação apresentada em [2] para avaliar o desempenho de sistemas de conversão de voz. O principal objetivo da utilização do PI é possibilitar a comparação objetiva entre diferentes técnicas de conversão. Originalmente proposto para avaliar o desempenho da conversão das características do trato vocal através do IHMD (*inverse harmonic mean distortion*), o PI é tido como capaz de representar adequadamente o desempenho de diferentes sistemas de conversão de voz, independentemente da utilização de diferentes métodos, locutores e até mesmo idiomas [2], [5], [40].

Adaptado para a conversão do contorno de *pitch*, o PI tem como objetivo fornecer uma medida que represente de forma satisfatória o desempenho geral de cada método implementado. O PI adaptado para conversão do contorno de *pitch* é dado por

$$\mathrm{PI} = 1 - \frac{D(\hat{\mathbf{y}}, \mathbf{y})}{D(\mathbf{x}, \mathbf{y})} \tag{3.2}$$

este índice possuirá valor $\mathrm{PI} = 0$ caso a função de conversão não altere o contorno de *pitch* do locutor fonte. No outro extremo, $\mathrm{PI} = 1$ se a conversão transformar idealmente o contorno de *pitch* do locutor fonte. Ainda, é possível obter valores negativos do PI quando a distância final $D(\hat{\mathbf{y}}, \mathbf{y})$ é maior do que a distância inicial $D(\mathbf{x}, \mathbf{y})$. Ao contrário do valor positivo máximo $\mathrm{PI} = 1$, valores negativos não têm limite.

3.3.2 Elocuções Convertidas

Todos os testes objetivos foram implementados baseados no índice de desempenho. Para a obtenção de uma estimativa do índice de desempenho médio de cada etapa desenvolvida, o banco de sinais de fala (ver Apêndice A) foi dividido em duas partes. A primeira,

constituída das primeiras quarenta elocuções de ambos locutores, foi utilizada para o treinamento. As elocuções restantes, de 41 a 50, formam a parte utilizada nos testes objetivos e subjetivos (ver Capítulo 4). Foram considerados dois locutores do sexo feminino, denominados de F_1 e F_2.

Como é possível realizar a conversão de voz entre dois locutores nas duas direções, isto é, alterando os locutores fonte e alvo, temos um total de vinte elocuções com o contorno de *pitch* convertido. Assim, vinte elocuções foram convertidas sem que nenhum dado dessas elocuções pronunciadas pelo locutor alvo tenha sido extraído na etapa de treinamento. Somente é utilizado o contorno de *pitch* do locutor alvo das elocuções de teste no cálculo do índice de desempenho. Os índices de desempenho mínimo, médio e máximo obtidos com aplicação da conversão do contorno de *pitch* utilizando GN e GMM são apresentados na Tabela 3.1.

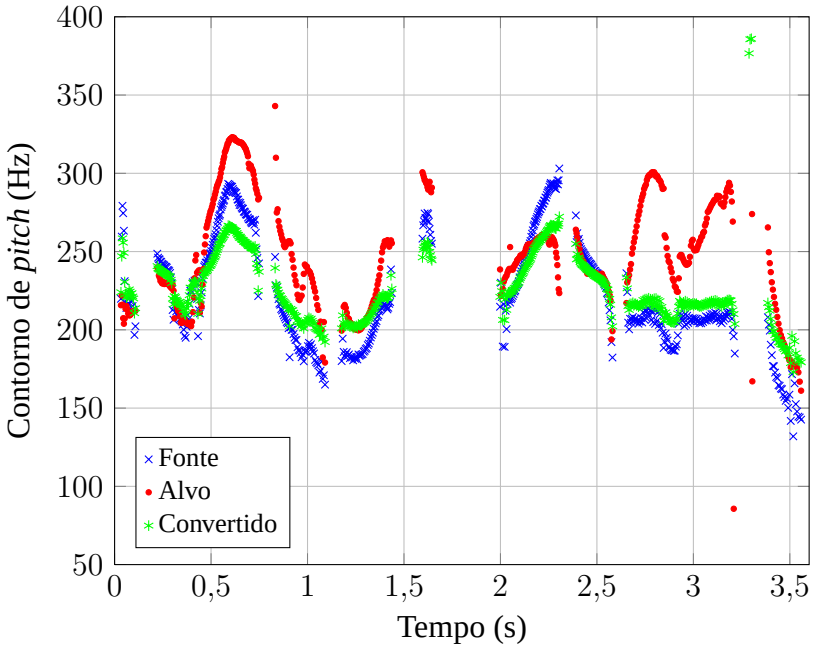
Tabela 3.1: PI mínimo, médio e máximo da conversão utilizando GN e GMM

Método	GN	GMM
PI Mínimo	$-0{,}055$	$-0{,}099$
PI Médio	$0{,}153$	$0{,}081$
PI Máximo	$0{,}362$	$0{,}258$

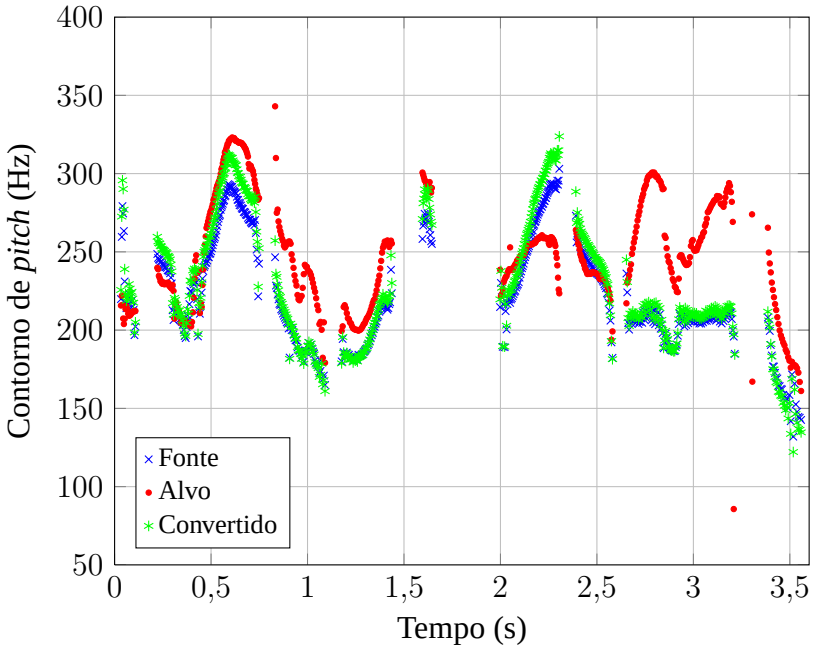
Apesar de a conversão do contorno de *pitch* utilizando GN resultar em um contorno de *pitch* semelhante ao contorno de *pitch* fonte (alterando apenas a média e o desvio padrão) o desempenho geral do método, para o banco de dados considerado, é superior ao método de conversão utilizando GMM. Para ilustrar o resultado da conversão utilizando GN e GMM, a Figura 3.3 mostra os contornos de *pitch* fonte, alvo e convertido utilizando GN na Figura 3.3(a) e GMM na Figura 3.3(b).

3.4 Método Proposto

Na abordagem proposta, a conversão do contorno de *pitch* é realizada através da superposição dos resultados da conversão de dois componentes: o componente macroprosódico que reflete o padrão de entonação da elocução e o componente microprosódico o qual caracteriza os segmentos fonemáticos da elocução. Na primeira etapa da conversão é estimada uma curva que representa o componente macroprosódico do contorno de *pitch* da elocução em questão. Por componente macroprosódico entende-se uma curva contínua e suave que modela o contorno de *pitch* [41]. Na segunda etapa são convertidos os detalhes do contorno de *pitch* não modelados pela curva macroprosódica, ou seja, o componente microprosódico.



(a)



(b)

Figura 3.3: Resultado da conversão do contorno de *pitch*. (a) Utilizando GN. (b) Utilizando GMM.

O processo de conversão do contorno de *pitch* é iniciado com a análise de uma elocução pronunciada pelo locutor fonte diferente das elocuções consideradas na fase de treinamento. Em linhas gerais, o componente macroprosódico do contorno de *pitch* é convertido com aplicação das técnicas GN e GMM, conforme detalhado na Seção 3.4.1. O componente microprosódico é convertido por seleção de segmentos (fonemas) do contorno de *pitch*. Em suma, cada segmento de contorno de *pitch* a ser convertido é comparado com todos os segmentos dos contornos de *pitch* das elocuções do locutor fonte utilizadas na fase de treinamento.

Como o banco de dados de treinamento possui apenas segmentos de contorno de *pitch* equivalentes entre os dois locutores, é possível obter o segmento da elocução pronunciada pelo locutor alvo que corresponde ao segmento extraído da elocução pronunciada pelo locutor fonte mais próximo ao segmento a ser convertido. O segmento de contorno de *pitch* do locutor fonte será posteriormente utilizado na avaliação objetiva dos resultados (ver Seção 3.3), enquanto o do locutor alvo será utilizado na conversão propriamente dita. Todo o processo de obtenção do banco de treinamento bem como a conversão propriamente dita são detalhados na Seção 3.4.2.

3.4.1 Conversão do Componente Macroprosódico

O padrão de entonação da elocução, modelado pelo componente macroprosódico do contorno de *pitch* e considerado independente da natureza dos fonemas é estimado com o algoritmo MOMEL (*modelling melody*) [42] e armazenado no banco de dados após ser codificado pela transcrição INTSINT (*international transcription system for intonation*) [41].

O INTSINT é considerado como sendo um equivalente prosódico do alfabeto fonético internacional (IPA - *international phonetic alphabet*), sendo constituído por um conjunto de 8 símbolos: T, M, B, H, S, L, U, e D que significam, respectivamente, *top, mid, bottom, higher, same, lower, upstepped* e *downstepped*. Os símbolos do INTSINT são divididos em absolutos (T, M e B) e relativos (H, S, L, U, e D). Os símbolos absolutos referem-se a faixa de possíveis valores das freqüências de *pitch* do locutor em questão. Por outro lado, os símbolos relativos são função dos símbolos precedentes. Os símbolos relativos são subdivididos em iterativos (U e D) ou não-iterativos (H, S e L) [41].

O mesmo contorno de *pitch* ilustrado na Figura 3.2(a) é apresentado na Figura 3.4 juntamente com o componente macroprosódico gerado pela interpolação quadrática dos pontos obtidos com o algoritmo MOMEL e a codificação INTSINT gerada pelo algoritmo descrito em [42].

Para cada elocução do banco de sinais de fala é obtida a codificação INTSINT e armazenada no banco de dados de treinamento. A conversão do contorno gerado pela in-

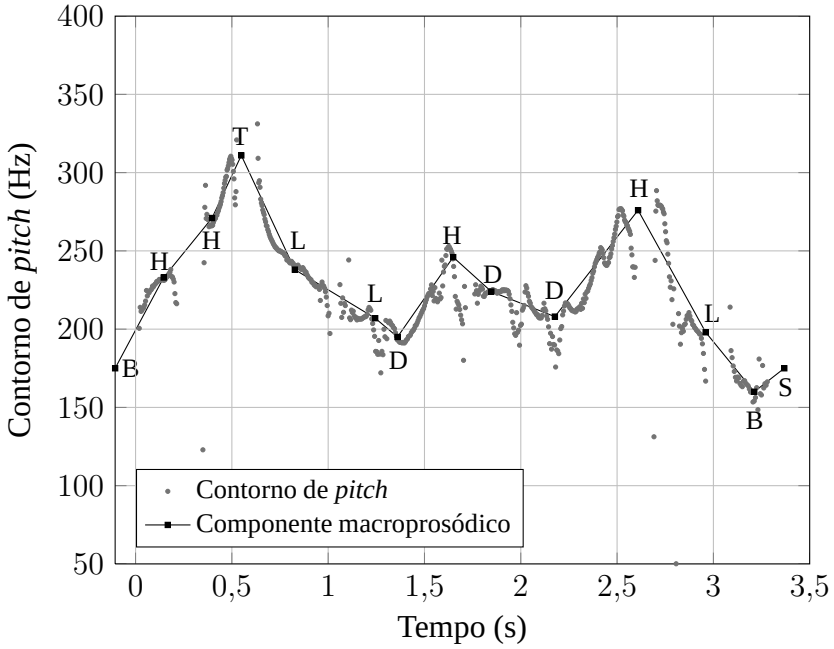


Figura 3.4: Exemplo de contorno de *pitch*, componente macroprosódico e codificação INTSINT.

terpolação dos pontos obtidos com o algoritmo MOMEL e codificados com o INTSINT é realizada de forma isolada da seleção de contorno (ver Seção 3.4.2).

Para realizar a conversão do contorno macroprosódico obtido com o algoritmo MOMEL foi implementada a conversão usando GN e GMM. Para conversão com o GMM, os testes iniciais apontaram para o problema de sobre-dimensionamento do modelo quando consideradas oito misturas (mesmo número de códigos INTSINT). Portanto, é considerado como número de misturas do modelo apenas os três códigos INTSINT absolutos, ou seja, T, M e B. A conversão é similar à apresentada na Seção 1.2.3 e os dados do modelo são obtidos com o algoritmo EM (ver Apêndice B). Os resultados obtidos na conversão do componente macroprosódico utilizando GN e GMM são apresentados na Tabela 3.2. Sendo que MACRO_{GN} e $\text{MACRO}_{\text{GMM}}$ correspondem, respectivamente, à conversão apenas do componente macroprosódico utilizando GN e GMM.

Tabela 3.2: PI mínimo, médio e máximo da conversão do componente macroprosódico utilizando GN e GMM

Método	MACRO_{GN}	$\text{MACRO}_{\text{GMM}}$
PI Mínimo	$-0{,}070$	$0{,}004$
PI Médio	$0{,}101$	$0{,}145$
PI Máximo	$0{,}300$	$0{,}328$

A Figura 3.5 ilustra o componente macroprosódico dos contornos de *pitch* fonte, alvo e convertidos, utilizando GN na Figura 3.5(a) e GMM na Figura 3.5(b).

3.4.2 Conversão do Componente Microprosódico

Uma vez que a conversão do componente microprosódico do contorno de *pitch* é realizada através da seleção de segmentos de um banco de dados constituído pelos fonemas observados na fase de treinamento, é necessário utilizar um procedimento para ajustar o tamanho dos segmentos. Tal procedimento consiste em aplicar a transformada discreta de cosseno (DCT - *discrete cosine transform*) seguida pela DCT inversa (IDCT - *inverse discrete cosine transform*), considerando um número fixo de coeficientes [28]. Inicialmente é obtida a transformada DCT $F(k)$ de todos os segmentos do contorno de *pitch* $f(n)$ mantendo o tamanho original N de cada segmento. A DCT utilizada é definida como [43]

$$F(k) = w(k) \sum_{n=1}^{N} f(n) \cos\left[\frac{\pi(2n-1)(k-1)}{2N}\right], \quad k = 1, \ldots, N \tag{3.3}$$

onde $w(k)$ é dado por

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 \leq k \leq N\,. \end{cases} \tag{3.4}$$

Em seguida, de posse do resultado da transformada DCT $F(k)$ (de tamanho K igual a N) de cada segmento de contorno de *pitch* $f(n)$ (de tamanho N), é aplicada a transformada IDCT resultando em um segmento aproximado $\tilde{f}(\tilde{n})$[2] do contorno de *pitch* $f(n)$, agora com número fixo de pontos $\tilde{N}$. A IDCT utilizada é dada por [43]

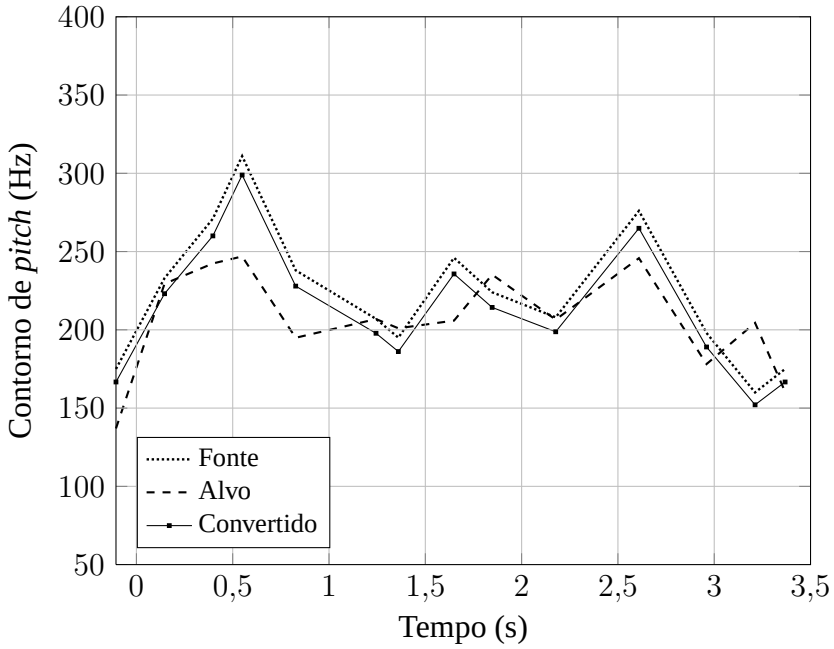
$$\tilde{f}(\tilde{n}) = \sum_{k=1}^{\tilde{N}} W(k) F(k) \cos\left[\frac{\pi(2\tilde{n}-1)(k-1)}{2\tilde{N}}\right], \quad \tilde{n} = 1, \ldots, \tilde{N} \tag{3.5}$$

onde $W(k)$ é

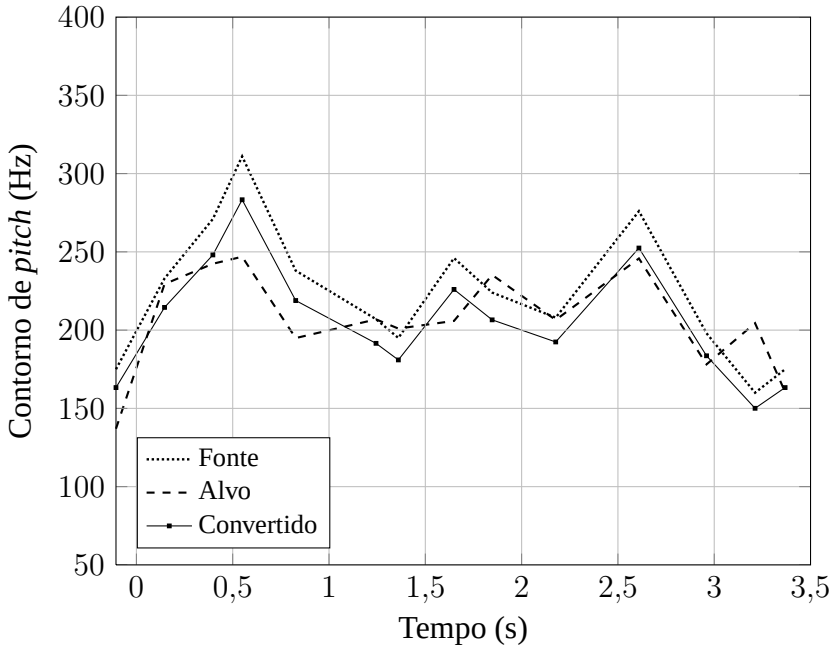
$$W(k) = \begin{cases} \frac{1}{\sqrt{\tilde{N}}}, & k = 1 \\ \sqrt{\frac{2}{\tilde{N}}}, & 2 \leq k \leq \tilde{N}\,. \end{cases} \tag{3.6}$$

Com o auxílio da DCT e IDCT é possível comparar o segmento de contorno de *pitch* a ser convertido com todos os segmentos existentes no banco de dados de treinamento independentemente do tamanho original do segmento de contorno de *pitch*. O processo de

[2]Para evitar ambigüidade é adotada a denominação $\tilde{n}$ para representar o domínio da transforma IDCT com número de coeficientes $\tilde{N} \neq N$.



(a)



(b)

Figura 3.5: Resultados da conversão do componente macroprosódico. (a) Utilizando GN. (b) Utilizando GMM.

normalização do segmento de contorno de *pitch* utilizando a transformada DCT seguida pela IDCT com número fixo de coeficientes $\tilde{N}$ é ilustrado pela Figura 3.6.

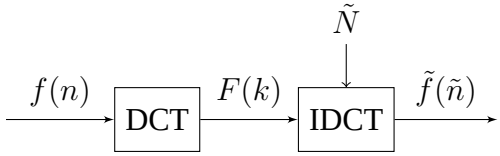
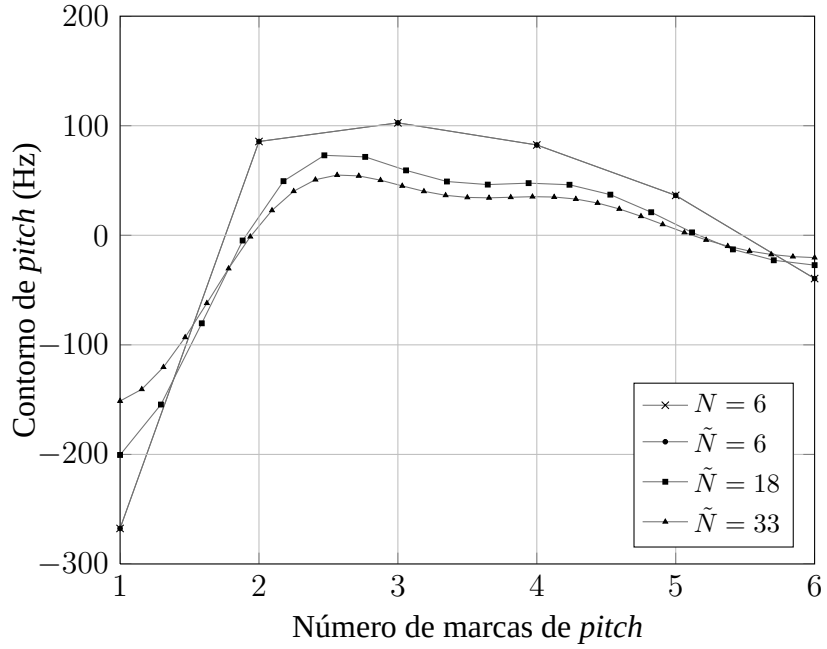


Figura 3.6: Diagrama do processo de normalização do segmento de contorno de *pitch*.

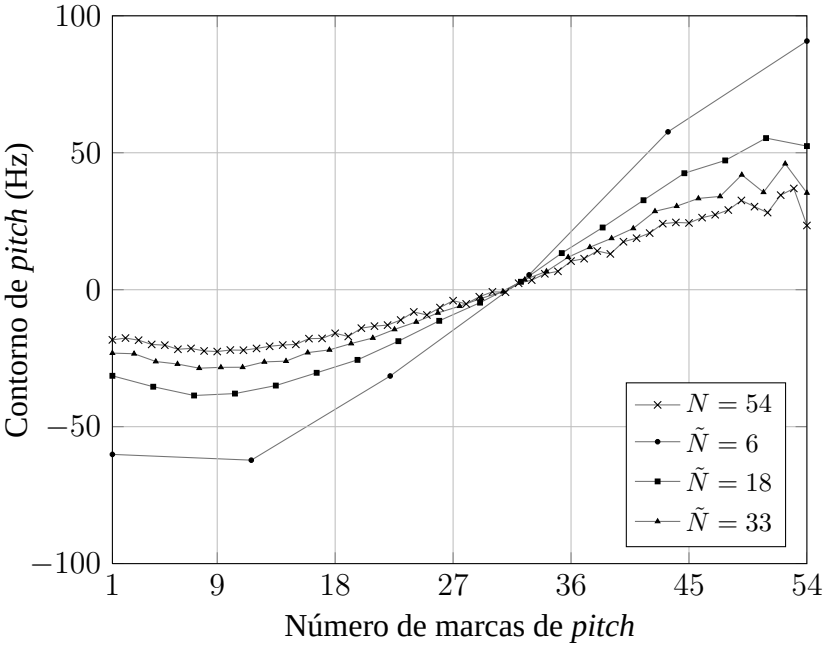
Em [28], todos os segmento de contorno de *pitch* (sílabas) são armazenados em um banco de dados de treinamento. O número de marcas de *pitch* de cada entrada do banco de dados é normalizado com auxílio da DCT de 8 coeficientes e a comparação é feita no domínio da transformada.

Durante a implementação da fase de treinamento do método proposto, foram observados tanto fonemas com apenas 2 marcas de *pitch* quanto fonemas com até 81 marcas. Devido à grande variabilidade do número de marcas de *pitch* existentes em cada fonema, a escolha do tamanho para qual todos os segmentos deveriam ser normalizados, ou seja, o número de coeficientes da transformada IDCT, torna-se de suma importância. Para ilustrar os possíveis efeitos da variação do número de coeficientes da IDCT na normalização do número de marcas de *pitch* em um segmento de contorno de *pitch*, a Figura 3.7 apresenta duas situações. Em ambos os casos um segmento de contorno de *pitch* tem seu tamanho normalizado através da DCT seguida pela IDCT de $6, 18$ e 33 coeficientes e com valores médios subtraídos. Na Figura 3.7(a), o segmento de contorno de *pitch* possui originalmente 6 marcas de *pitch* e é perfeitamente reconstruído pela IDCT de 6 coeficientes, enquanto o resultado da IDCT de 18 e 33 coeficientes introduz distorções. Já na Figura 3.7(b), o segmento de contorno de *pitch* possui originalmente 54 marcas e o melhor resultado é obtido com a IDCT de 33 coeficientes. Analisando estes exemplos, fica claro a necessidade de escolher adequadamente o número de coeficientes da IDCT para armazenar os segmentos do contorno de *pitch* no banco de dados.

A primeira alternativa considerada foi utilizar o número médio de marcas de *pitch* por fonema (igual a 18) como tamanho padrão para armazenar os segmentos do contorno de *pitch* no banco de treinamento. Entretanto, níveis elevados de distorção seriam inseridos no banco de dados quando efetuada a normalização do número de marcas de *pitch* para fonemas com um número de marcas distante do valor médio observado, conforme ilustrado na Figura 3.7. Com o objetivo de minimizar efeitos negativos provenientes do ajuste do tamanho dos segmentos do contorno de *pitch* e observando o histograma apresentado na Figura 3.8(a), optou-se por utilizar não apenas um, mas sim três diferentes valores de coeficientes para a transformada IDCT. O valor de cada coeficiente da IDCT foi estimado com o treinamento



(a)



(b)

Figura 3.7: Exemplos de segmentos de contorno de *pitch* e o resultado da normalização para 6, 18 e 33 marcas de *pitch*. (a) Segmento com 6 marcas de *pitch*. (b) Segmento com 54 marcas de *pitch*.

de um modelo GMM (utilizando o algoritmo EM, descrito no Apêndice B) de três misturas, conforme Figura 3.8(b). Assim, todos os segmentos do contorno de *pitch*, extraídos das sentenças de treinamento, foram codificado pela DCT seguida pela IDCT utilizando como número de coeficientes o valor médio de uma das três misturas do modelo GMM estimado, especificamente, $6, 18$ ou 33 coeficientes (os mesmos valores utilizados nos exemplos ilustrados na Figura 3.7). A classe $\mathcal{C}_i$ de maior probabilidade define para qual tamanho será normalizado o segmento de contorno de *pitch* em função do número de marcas do segmento em questão e da aplicação da regra de Bayes (ver Equação (1.3) na Seção 1.2.3) com os parâmetros estimados do GMM. Na fase de conversão propriamente dita, cada segmento de contorno de *pitch* a ser convertido é comparado com todos os segmento de contorno de *pitch* pertencentes à mesma classe $\mathcal{C}_i$, ou seja, normalizados com o coeficiente $\tilde{N}$ de mesmo valor.

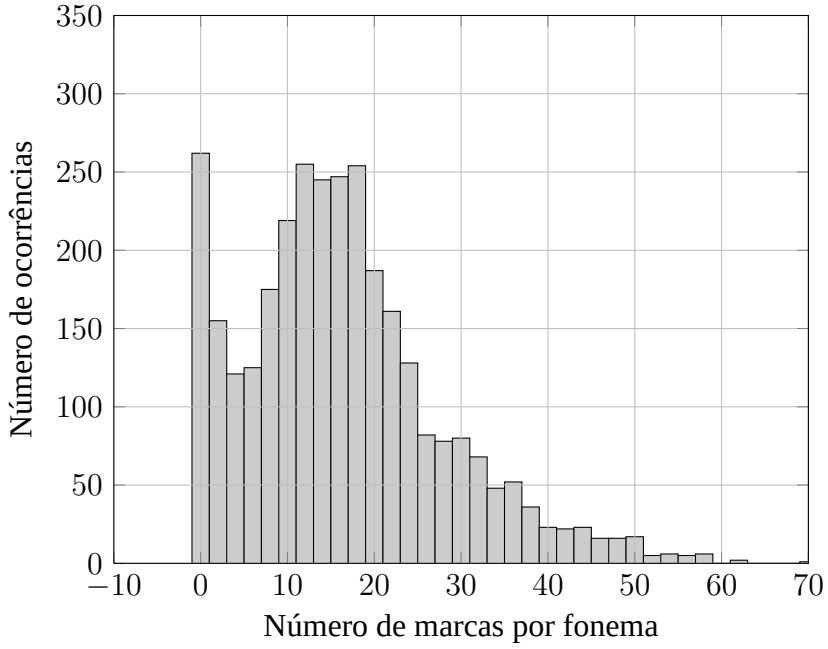
O valor médio de cada segmento de contorno de *pitch* de tamanho normalizado empregado no treinamento e na conversão é considerado como sendo nulo. Para tanto, basta igualar a zero o valor do primeiro coeficiente da transformada DCT. Assim, é evitado que o valor médio dos segmentos de contorno de *pitch* mascare o resultado, isto é, evita-se que um segmento de contorno de *pitch* do banco de treinamento seja selecionado mesmo que exista outro segmento com formato mais similar porém de maior distância euclidiana devido aos diferentes valores médios de cada segmento. Em [28], o primeiro coeficiente da DCT também é descartado, enquanto em [4] e [44], o valor médio de cada segmento de contorno de *pitch* armazenado no banco de dados de treinamento também é removido.

O desempenho do processo de seleção de segmentos de contorno de *pitch* usando a normalização do tamanho do segmento para apenas um coeficiente da transformada IDCT, para cada um dos três valores de $\tilde{N}$ considerados, bem como para a utilização do GMM de três misturas é apresentado na Tabela 3.3, sendo que para todos os métodos comparados na Tabela 3.3 a conversão do componente macroprosódico foi realizada com o emprego do GMM, conforme Seção 3.4.1.

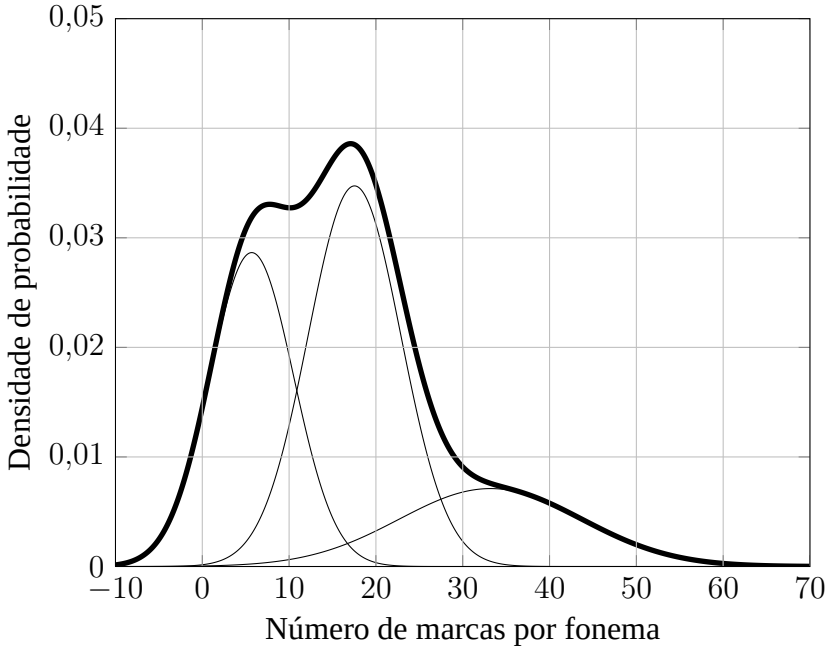
Tabela 3.3: PI mínimo, médio e máximo em função do número de coeficientes da IDCT

PI	6	18	33	6, 18 e 33
Mínimo	$-0{,}097$	$-0{,}121$	$-0{,}113$	$-0{,}093$
Médio	$0{,}085$	$0{,}085$	$0{,}087$	$0{,}091$
Máximo	$0{,}226$	$0{,}237$	$0{,}244$	$0{,}231$

Apesar do índice de desempenho máximo obtido com a utilização de $\tilde{N} = 33$ ou 18 ser superior ao obtido com os três valores de $\tilde{N}$, os desempenhos mínimo e médio superiores, quando considerados os três valores de $\tilde{N}$, justificam seu uso. Portanto, quando mencionada a conversão do componente microprosódico por seleção de segmentos do contorno de *pitch*,



(a)



(b)

Figura 3.8: Modelagem do banco de treinamento por GMM. (a) Histograma do número de marcas por fonema. (b) Função distribuição de probabilidade estimada do GMM.

entende-se a utilização dos três valores de $\tilde{N}$, modelados pelo GMM. Exemplos dos resultados obtidos com a combinação da conversão do componente macroprosódico utilizando GMM em conjunto com a conversão do componente microprosódico por seleção de segmentos do contorno, denominado SELEÇÃO$_{\text{GMM}}$, são ilustrados na Figura 3.9.

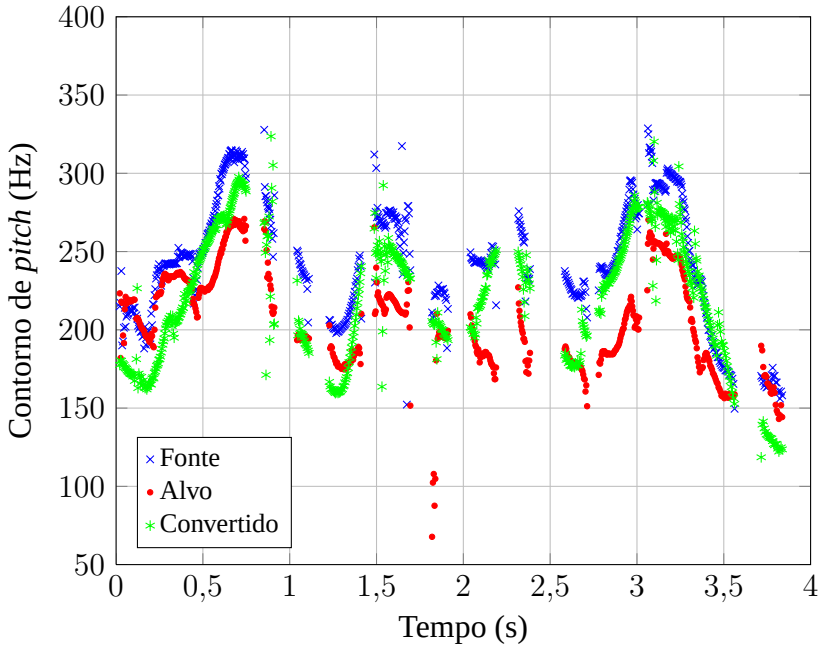
3.5 Conclusões

Mesmo que não exista uma métrica universalmente aceita para comparação objetiva dos resultados obtidos com diferentes métodos de conversão do contorno de *pitch*, a utilização do índice de desempenho se justifica, visto que tal índice permite a comparação dos métodos frente diferentes locutores e/ou elocuções. A simples comparação da distância final $D(\hat{\mathbf{y}}, \mathbf{y})$ entre os contornos de *pitch* convertido e alvo não é capaz de prover resultados significativos, uma vez que seu desempenho é altamente dependente dos locutores como também das elocuções em questão. O índice de desempenho é um parâmetro útil para a comparação objetiva dos resultados obtidos com os diferentes métodos, uma vez que seu valor é normalizado em função da distância inicial entre os contornos de *pitch* fonte e alvo $D(\mathbf{x}, \mathbf{y})$.

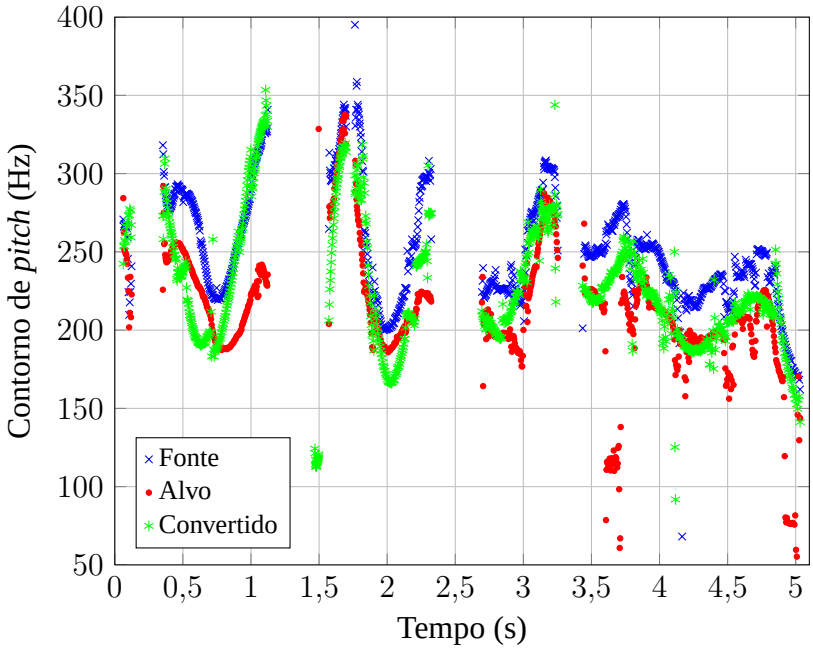
Dentre os métodos propostos, a marcação e a transcrição fonética fazem-se necessárias apenas para o método de seleção de segmentos do contorno de *pitch*. A conversão do componente macroprosódico requer apenas a codificação INTSINT, obtida com auxílio do algoritmo MOMEL, sendo que neste caso o alinhamento é necessário somente para o cálculo do índice de desempenho. Com objetivo de refinar os resultados obtidos através da conversão do componente macroprosódico do contorno de *pitch*, é proposta uma seleção de segmentos do contorno de *pitch*.

Para possibilitar a seleção de segmentos, deve ser utilizada também a normalização do tamanho dos segmentos. Para evitar os problemas inerentes à interpolação é avaliado o emprego da transformada DCT seguida pela IDCT de tamanho fixo, com três diferentes valores de coeficientes. Assim, é esperado que possíveis distorções provenientes do processo de normalização dos segmentos de contorno de *pitch* sejam minimizadas.

Com uso de GMM e normalização do tamanho dos segmentos, considerando três diferentes valores de coeficientes da transformada IDCT é possível reduzir consideravelmente as distorções introduzidas no banco de dados, quando comparados com a estratégia de utilização de apenas um valor de coeficiente da transformada IDCT para a normalização. Assim, sempre que o segmento de contorno de *pitch* possuir número de marcas de *pitch* igual a $6, 18$ ou 33, o segmento será armazenado em seu formato original. Distorções serão introduzidas sempre que o número de marcas do fonema for diferente daquelas dos coeficientes estimados. As diferenças no desempenho do procedimento de seleção de segmentos de contorno



(a)



(b)

Figura 3.9: Resultados da conversão dos componentes macroprosódico e microprosódico. (a) Elocução número 42. (b) Elocução número 45.

de *pitch*, quando considerados um ou três valores de coeficientes da transformada IDCT, são comparadas e os resultados permitem indicar uma melhoria no desempenho da conversão quando utilizado GMM de três misturas.

Os resultados obtidos com a seleção de segmentos do contorno de *pitch* indicam um melhor desempenho frente aos dois métodos mais utilizados: GN e GMM. Todavia, comparando os índices de desempenho mostrados nas Tabelas 3.3 e 3.2, observamos uma sensível redução desse índice após combinar os resultados da conversão dos componentes macroprosódico e microprosódico. Para melhor avaliar os efeitos dessa redução no desempenho geral dos métodos, no Capítulo 4 serão apresentados os testes subjetivos realizados.

Capítulo 4

Testes Subjetivos

Como não existe um único contorno de *pitch certo* ou *errado*, o objetivo da conversão do contorno de *pitch* para conversão de voz é obter um contorno que seja o mais *natural* possível para uma dada elocução e um locutor. Apesar de o PI possibilitar uma comparação objetiva dos resultados obtidos com os diferentes métodos, é necessário realizar testes subjetivos para verificar se tal avaliação é consistente com o desempenho dos diferentes métodos estudados após a síntese do sinal usando o contorno de *pitch* convertido.

Em suma, duas abordagens distintas podem ser utilizadas no desenvolvimento de testes subjetivos para avaliação dos resultados das técnicas de conversão do contorno de *pitch* para a conversão de voz. Na primeira, as diferentes técnicas de conversão do contorno de *pitch* são aplicadas a sinais de voz obtidos de um sistema completo de conversão de voz. Esta abordagem é justificada pela possível sutileza de diferenças quando apenas o contorno de *pitch* é convertido. Entretanto, a qualidade dos resultados dos atuais sistemas de conversão de voz é descrito como um sinal de voz *robótico*, dificultando assim a sua avaliação subjetiva [28]. A segunda abordagem procura isolar os efeitos da conversão do contorno de *pitch* dos efeitos da conversão de voz. Para tanto, apenas o contorno de *pitch* é manipulado, mantendo as demais características do sinal de fala inalteradas [24].

Em [28], a avaliação subjetiva é dividida em duas etapas, ambas utilizando noventa elocuções de treinamento, vinte e cinco elocuções de teste e dezenove avaliadores. Antes de iniciar a primeira etapa de testes, os avaliadores ouviram amostras de elocuções do banco de sinais de fala e foram instruídos a prestar atenção nos aspectos prosódicos das elocuções. Na primeira fase foram apresentadas elocuções obtidas por conversão de voz utilizando a conversão do contorno de *pitch* por GMM e pelo método em análise. As instruções solicitavam que os avaliadores desprezassem possíveis distorções do sinal e escolhessem a opção mais semelhante às amostras apresentadas antes do início do teste. Na segunda etapa dos testes, as elocuções foram repetidas e, segundo as instruções, os avaliadores deveriam escolher a

opção que soasse menos *robótica*. Em ambas etapas, foi dada ainda uma terceira opção para o caso de não haver diferenças perceptíveis entre as amostras apresentadas.

Outros dois testes são apresentados em [24]. No primeiro, de similaridade, para cada uma das oito elocuções avaliadas foram apresentadas aos avaliadores as elocuções alvo original e quatro elocuções com o contorno de *pitch* convertido por diferentes técnicas. As instruções fornecidas solicitavam a escolha da elocução mais semelhante à elocução de referência, ou seja, a elocução original pronunciada pelo locutor alvo. Já no segundo teste, de preferência, as quatro técnicas de conversão do contorno de *pitch* avaliadas foram confrontadas sem a apresentação da elocução de referência. Em ambos os testes, apenas o contorno de *pitch* das elocuções foi alterado.

Com o objetivo de isolar os efeitos da conversão do contorno de *pitch*, os testes subjetivos implementados utilizam sinais de voz com apenas o contorno de *pitch* manipulado. Para tanto, as elocuções pronunciadas pelo locutor alvo tiveram seu contorno substituído pelo contorno convertido a partir da elocução do locutor fonte, de acordo com a metodologia de testes subjetivos adotada em [24].

Conforme detalhado na Seção 3.3.2, vinte elocuções com o contorno de *pitch* convertido foram utilizadas, sendo dez em cada direção de conversão. A Tabela 4.1 resume os resultados apresentados no Capítulo 3, sendo que SELEÇÃO$_{\mathrm{GMM}}$ corresponde à conversão do componente macroprosódico por GMM refinada pela seleção de segmentos do contorno utilizando $\tilde{N}$ igual a $6, 18$ e 33. Para evitar a fadiga dos avaliadores dos testes subjetivos, apenas três técnicas de conversão do contorno de *pitch* estudadas foram utilizadas nos testes subjetivos: como método *base*, a normalização gaussiana e como método proposto a conversão do componente macroprosódico empregando GMM, denominado MACRO$_{\mathrm{GMM}}$, e a extensão desta técnica incluindo a conversão do componente microprosódico por seleção de segmentos de contorno de *pitch*, SELEÇÃO$_{\mathrm{GMM}}$. Mesmo que os índices de desempenho mínimo e médio da conversão MACRO$_{\mathrm{GN}}$ sejam melhores do que os obtidos com a utilização da GN, optou-se por não avaliar o desempenho do método MACRO$_{\mathrm{GN}}$ nos testes subjetivos pois o método MACRO$_{\mathrm{GMM}}$ representa uma abordagem semelhante porém com resultados superiores. Outro método não avaliado nos testes subjetivos é o GMM devido ao seu baixo desempenho apresentado, em especial os valores de PI mínimo e médio.

Dentre os objetivos da avaliação subjetiva podemos citar: confrontar o resultado dos dois métodos propostos frente ao método de normalização gaussiana, avaliar a medida de distância adotada e verificar a necessidade de refinar os resultados da conversão do componente macroprosódico por seleção de segmentos de contorno de *pitch*.

Tabela 4.1: PI mínimo, médio e máximo dos diferentes métodos de conversão do contorno de *pitch*

PI	GN	GMM	MACRO_{GN}	$\text{MACRO}_{\text{GMM}}$	$\text{SELEÇÃO}_{\text{GMM}}$
Mínimo	$-0{,}055$	$-0{,}099$	$-0{,}071$	$0{,}004$	$-0{,}093$
Médio	$0{,}153$	$0{,}081$	$0{,}101$	$0{,}145$	$0{,}091$
Máximo	$0{,}362$	$0{,}258$	$0{,}300$	$0{,}328$	$0{,}231$

4.1 Teste de Preferência

O primeiro teste subjetivo aplicado para avaliar os resultados obtidos com os diferentes métodos é o teste de preferência. Nessa primeira etapa, o contorno de *pitch* convertido foi estimado a partir do contorno de *pitch* do locutor fonte e transplantado para a elocução pronunciada pelo locutor alvo. Cinco elocuções de dois locutores com o contorno de *pitch* convertido por três diferentes técnicas, totalizando trinta elocuções, foram apresentadas aos avaliadores. Para cada elocução foram apresentadas as três opções de forma aleatória e as instruções solicitavam a escolha da elocução mais *natural* com vista às características prosódicas das elocuções. Os resultados da primeira etapa dos testes subjetivos, para um total de vinte e cinco avaliadores, são detalhados na Tabela 4.2.

Tabela 4.2: Resultados do teste de preferência

	GN	$\text{MACRO}_{\text{GMM}}$	$\text{SELEÇÃO}_{\text{GMM}}$
Número de Votos	19	198	33
Total (%)	7,6	79,2	13,2

A primeira observação que pode ser feita comparando a Tabela 4.2 com os valores médios de PI da Tabela 4.1 é a coerência entre a medida objetiva e os resultados desta etapa do teste subjetivo. Ainda, ambos os métodos propostos sobrepujam o método mais utilizado na literatura pesquisada, ou seja, a conversão do contorno de *pitch* utilizando GN.

Ao final da primeira etapa dos testes. a maioria dos avaliadores comentou a dificuldade de escolher entre duas das três opções apresentadas. Em testes subjetivos informais, verifica-se que as opções com resultados mais próximos correspondem aos dois métodos propostos. Esta semelhança deve-se ao fato de ambos os métodos compartilharem a mesma técnica de conversão do componente macroprosódico do contorno de *pitch*. A diminuição do índice de desempenho quando adicionada a conversão do componente microprosódico à conversão do componente macroprosódico utilizando GMM deve-se, em parte, a problemas na obtenção das marcas de *pitch* em regiões limítrofes entre segmentos vozeados e não-

vozeados. Outro fator que limita o desempenho da seleção de segmentos do contorno de *pitch* é o tamanho do banco de treinamento, uma vez que com maior número de elocuções de treinamento as chances de repetições de fonemas com pouca ocorrência é maior.

4.2 Teste de Similaridade

O segundo teste subjetivo aplicado é o teste de similaridade. Nessa etapa, um novo conjunto de dez elocuções foi apresentado aos avaliadores, sendo cinco elocuções de cada locutor e diferentes das elocuções utilizadas no teste de preferência. Novamente, o contorno de *pitch* convertido a partir do contorno de *pitch* de locutor fonte foi transplantado para a elocução pronunciada pelo locutor alvo. Desta vez, antes de cada alternativa foi apresentada a elocução original pronunciada pelo locutor alvo, ou seja, a elocução de referência. Os avaliadores foram instruídos a escolher a opção mais similar à elocução original, considerando os aspectos prosódicos. Os resultados do teste de similaridade, para um total de vinte e cinco avaliadores, são apresentados na Tabela 4.3.

Tabela 4.3: Resultados do teste de similaridade

	GN	$\text{MACRO}_{\text{GMM}}$	$\text{SELEÇÃO}_{\text{GMM}}$
Número de Votos	18	178	54
Total (%)	7,2	71,2	21,6

Os resultados obtidos com o método GN no teste de similaridade estão próximos aos obtidos no teste de preferência. Já os resultados obtidos pelo método $\text{SELEÇÃO}_{\text{GMM}}$ melhoraram significativamente quando comparados aos resultados dos testes de preferência. Essa melhoria deve-se ao refinamento introduzido com a seleção de segmentos do contorno de *pitch*, mesmo que para isto o desempenho objetivo seja sensivelmente reduzido.

4.3 Conclusões

Apesar da dificuldade de se definir uma metodologia adequada para avaliar a conversão do contorno de *pitch*, os resultados dos testes subjetivos sugerem que a medida objetiva adotada, o índice de desempenho, é adequada para comparar os resultados das diferentes técnicas de conversão. Os resultados apresentados indicam ainda a preferência dos avaliadores pelos métodos propostos de conversão do contorno de *pitch* frente ao método de normalização gaussiana.

Mesmo que o melhor índice de desempenho de uma única elocução seja obtido com o método GN (Tabela 4.1), no teste de similaridade esse método não foi escolhido nenhuma vez para a elocução que obteve tal desempenho. Isto se deve à característica intrínseca do método de não alterar o padrão de entonação da elocução, mantendo sempre padrão de entonação geral da elocução pronunciada pelo locutor fonte. O mesmo fenômeno pôde ser observado em outras elocuções que obtiveram resultados semelhantes.

Ainda que os dois métodos propostos tenham obtido desempenho superior quando comparados com os métodos usualmente utilizados na literatura pesquisada (GN e GMM), a distância $D(\hat{\mathbf{y}}, \mathbf{y})$ remanescente após o processo de conversão do contorno de *pitch* ainda deixa margem para melhorar o desempenho do método proposto (ver Tabela 4.4, considerando as distâncias iniciais $D(\mathbf{x}, \mathbf{y})$ mínima, média e máxima iguais a $89{,}48\,\mathrm{Hz}$, $116{,}01\,\mathrm{Hz}$ e $135{,}07\,\mathrm{Hz}$, respectivamente).

Tabela 4.4: Distâncias $D(\hat{\mathbf{y}}, \mathbf{y})$ (Hz) mínima, média e máxima obtidas com os diferentes métodos de conversão do contorno de *pitch*

$D(\hat{\mathbf{y}}, \mathbf{y})$	GN	$\text{MACRO}_{\text{GMM}}$	$\text{SELEÇÃO}_{\text{GMM}}$
Mínima	77,679	70,879	81,106
Média	92,819	100,556	103,581
Máxima	116,737	131,745	128,571

Todos os métodos confrontados na Tabela 4.4 apresentaram índice de desempenho negativo em, ao menos, uma elocução. Este fenômeno pode ser facilmente visualizado comparando as distâncias $D(\hat{\mathbf{y}}, \mathbf{y})$ máxima de cada método com a distância $D(\mathbf{x}, \mathbf{y})$ máxima ($135{,}07\,\mathrm{Hz}$). Independente do método de conversão utilizado, os resultados apresentados evidenciam a dependência do desempenho do método com a elocução em questão e com os locutores considerados. A dependência dos resultados objetivos em função dos locutores fica melhor evidenciada com os dados apresentados de forma separada para cada direção de conversão (ver Tabela 4.5).

Tabela 4.5: Distâncias $D(\hat{\mathbf{y}}, \mathbf{y})$ (Hz) mínima, média e máxima obtidas com os diferentes métodos de conversão do contorno de *pitch*, em função dos locutores considerados

Locutores	$D(\hat{\mathbf{y}},\mathbf{y})$	GN	$\text{MACRO}_{\text{GMM}}$	$\text{SELEÇÃO}_{\text{GMM}}$
$F_1 \rightarrow F_2$	Mínima	77,679	70,879	81,106
	Média	86,354	89,463	97,031
	Máxima	95,745	107,049	109,314
$F_2 \rightarrow F_1$	Mínima	83,373	90,045	87,457
	Média	99,285	111,649	110,129
	Máxima	116,737	131,745	128,571

Capítulo 5

Conclusões Finais

Sistemas de conversão de voz possuem ampla gama de aplicações, desde a personalização de TTS, tradução automática, ensino de idiomas, auxílio no tratamento de doenças degenerativas do sistema fonador e, principalmente, entretenimento. Apesar da vasta possibilidade de aplicações, os atuais sistemas de conversão de voz carecem de uma técnica robusta de conversão das características prosódicas do sinal de fala. Como parte do esforço para refinar os resultados, a conversão do contorno de *pitch* tem levado atualmente a um grande esforço de pesquisa.

A partir da discussão inicial das técnicas de conversão do contorno de *pitch* para aplicação em conversão de voz é possível levantar as principais características e limitações de cada abordagem considerada. Dentre as técnicas mais utilizadas para conversão do contorno de *pitch* destacam-se os métodos GN e GMM, tanto pela flexibilidade quanto pela facilidade de integração com as demais etapas do processo de conversão de voz. Entretanto tais métodos não são capazes de fornecer resultados satisfatórios.

Com a introdução das características fisiológicas do aparelho fonador humano e do sistema sonoro do português brasileiro foi possível avaliar diferentes fenômenos comumente observados em bancos de sinais de fala. Em especial, as flutuações e harmonias vocálicas devido as quais optou-se pela implementação da seleção de segmentos de contorno de *pitch* em conformidade com a segmentação fonética. Ao contrário de outras técnicas apresentadas na literatura, especificamente a seleção de segmentos de contorno de *pitch* definidos como segmentos vozeados, a segmentação fonética possibilita eliminar certas flutuações e harmonias vocálicas.

Como o principal objetivo é possibilitar a aplicação do método proposto aos mais diferentes tipos de sistemas de conversão de voz, não são empregados dados usualmente disponíveis apenas em banco de dados desenvolvidos para TTS.

Considerando que apenas a transcrição e a marcação fonética são necessários para

implementar a seleção de segmentos, é possível adaptar o método proposto aos mais diferentes tipos de sistemas de conversão de voz. Com a divisão do contorno de *pitch* em componentes macroprosódico e microprosódico é possível implementar diferentes estratégias para cada etapa da conversão. O fato de a conversão do contorno de *pitch* ser implementada em duas etapas distintas sugere a possibilidade de utilização do método proposto para aplicação em sistemas de conversão de voz com o objetivo de alterar também o padrão de entonação das elocuções.

Infelizmente não foi possível comparar os resultados do método proposto frente a alguns dos resultados publicados na literatura. Dentre os motivos que impossibilitam tal comparação podemos enumerar: a não adoção de uma medida objetiva comum entre os diferentes grupos de pesquisa e a escassez de detalhes dos resultados publicados. Apesar do método publicado em [28] ser mais refinado do que os métodos de conversão do contorno de *pitch* utilizando GN e GMM, não foi possível estabelecer uma comparação direta dos resultados, pois tal método utiliza dados provenientes de um banco de dados desenvolvido para TTS. Mesmo que a comparação objetiva não tenha sido possível entre o método proposto com o publicado em [28], a maior flexibilidade do método proposto deve ser considerada na determinação de qual método empregar. Ainda que em [24] sejam detalhadas as distâncias remanescentes $D(\hat{\mathbf{y}},\mathbf{y})$ após o processo de conversão do contorno de *pitch*, as distâncias iniciais $D(\mathbf{x},\mathbf{y})$ não são apresentadas, impedindo assim uma comparação mais justa e criteriosa.

Apesar dos problemas inerentes à avaliação de desempenho dos diferentes sistemas de conversão de voz, a utilização do índice de desempenho, adaptado ao problema da conversão do contorno de *pitch*, possibilita comparar o desempenho de diferentes técnicas discutidas. Neste trabalho, foram confrontados o desempenho das duas técnicas de conversão de contorno de *pitch* mais populares para aplicação em conversão de voz, conversão utilizando GN e GMM. Além do mais, são apresentados os resultados obtidos em cada etapa de conversão do contorno de *pitch* do método proposto. Como o índice de desempenho é normalizado pela distância inicial $D(\mathbf{x},\mathbf{y})$ é possível comparar os resultados do método proposto com os que poderão vir a ser obtidos.

A avaliação subjetiva aplicada busca isolar os efeitos da conversão do contorno de *pitch* dos efeitos provenientes da conversão de voz e, portanto, são avaliadas as elocuções pronunciadas pelos locutores alvo apenas com o contorno de *pitch* alterado. Os contornos de *pitch* convertidos são obtidos a partir do contorno de *pitch* da elocução pronunciada pelo locutor fonte. Os dados extraídos da elocução de teste, pronunciado pelo locutor alvo são empregados na comparação objetiva do desempenho da conversão do contorno de *pitch*, bem como no teste de similaridade como elocução de referência. Essa abordagem evita a degradação do sinal de voz após o processo de conversão de voz.

Os resultados obtidos com a avaliação subjetiva demonstram a preferência dos avaliadores pelo método de conversão do contorno de *pitch* $\text{MACRO}_{\text{GMM}}$. Mesmo com o decremento do índice de desempenho do método $\text{SELEÇÃO}_{\text{GMM}}$ frente ao método $\text{MACRO}_{\text{GMM}}$, os resultados obtidos com o teste de similaridade apontam para a possibilidade de incremento na similaridade do sinal com contorno de *pitch* convertido usando $\text{SELEÇÃO}_{\text{GMM}}$. Todavia, observando a distância $D(\hat{\mathbf{y}},\mathbf{y})$ é possível verificar que melhorias significativas ainda podem ser obtidas com refinamento do processo de conversão do componente microprosódico.

A primeira observação que pode ser feita comparando os resultados objetivos e subjetivos indica a coerência dos métodos de avaliação empregados. Entretanto, algumas discrepâncias podem ser observadas. Em especial, o fato de a elocução com o contorno de *pitch* convertido que obteve o melhor índice de desempenho isolado não ter sido escolhida nenhuma vez no teste de similaridade. Tal fenômeno pode ser explicado pela característica intrínseca do método de conversão utilizando GN de alterar apenas a média e o desvio padrão do contorno de *pitch*.

O incremento de desempenho proporcionado pelo método proposto frente aos métodos de conversão do contorno de *pitch* utilizando GN e GMM deve-se, em grande parte, à possibilidade de implementar diferentes estratégias de conversão para os componentes macroprosódico e microprosódico, unindo assim as vantagens das abordagens que empregam modelos estatísticos, como por exemplo o GMM, e aqueles que utilizam contornos (ou segmentos) naturais, isto é, produzidos pelo locutor alvo.

Os resultados obtidos indicam que o índice de desempenho, adaptado para a conversão do contorno de *pitch*, é adequado para avaliação objetiva dos diferentes métodos. Todavia, uma vez que a maior parte das aplicações da conversão de voz é direcionada para consumo humano, os testes subjetivos são de extrema importância na avaliação dos resultados.

Como trabalho futuro, verifica-se a necessidade de testar o método proposto com maior número de locutores bem como para diferentes bancos de sinais de fala. Em especial, a conversão do contorno de *pitch* entre elocuções com diferentes padrões de entonação. Como a distância remanescente $D(\hat{\mathbf{y}},\mathbf{y})$ do processo de conversão do contorno de *pitch* ainda é elevada, acredita-se que melhorias significativas possam ser alcançadas com o refinamento da conversão do componente microprosódico do contorno de *pitch*. Dentre as possibilidades existentes, o aumento do tamanho do banco de sinais de fala pode prover melhorias pontuais visto que a maior ocorrência de determinados fonemas pouco observados no banco de treinamento pode facilitar o processo de seleção de segmentos de contorno de *pitch*. Em detrimento à flexibilidade do método proposto, é possível adaptar o processo de seleção de segmentos de contorno de *pitch* para busca baseada na classificação fonética dos segmentos. É ainda possível estender a abordagem proposta para a conversão de outras características prosódicas do

sinal de fala, dentre elas a flutuação da energia e a duração dos fonemas. Por fim, apesar da dificuldade de isolar os resultados da conversão do contorno de *pitch*, um sistema completo de conversão de voz deve ser implementado empregando as diferentes técnicas de conversão do contorno de *pitch* para uma avaliação final do método proposto.

Apêndice A

Banco de Sinais de Fala

O banco de sinais de fala utilizado para o treinamento, conversão, testes subjetivos e demais testes informais é formado por um conjunto de cinqüenta elocuções, foneticamente ricas. As gravações foram realizadas em canal único, sendo os sinais amostrados a $16\,\mathrm{kHz}$ e quantizados com $16\,\mathrm{bits}$, resultando em um sinal de $256\,\mathrm{kbps}$. As sentenças estão listadas na Tabela A.1.

Tabela A.1: Sentenças do banco de dados

Nº	**Sentença**
1	Existem muitos camundongos diferentes naquela ilha.
2	Humberto ganhou um monte de tinta de presente da sua avó.
3	José Pinto não era contra o filho do bondoso velhinho húngaro.
4	Os potes de geléia de ameixa foram oferecidos à mãe no fim do jantar.
5	O dedo mutilado do rapaz deixava seu rosto alegre sombreado.
6	Sinto que aquele homem corcunda não tem certeza do resultado.
7	Dois abelhões se dirigiram ao chuchu dependurado na cerca.
8	Ontem, ninguém montou a fera na feira de montaria de Corumbá do Norte.
9	Pela manhã do dia onze, Júnior brincou na gangorra do bosque Boi Bumbá.
10	O índio pulava no bonde enquanto a chuva caía lá fora.
11	Deslumbrado, o rapaz cumprimentou seu ídolo.
12	Aquele símbolo suntuoso sempre esteve associado a coisas boas.
13	Ontem, Rafael e Rosa estavam irreconhecíveis antes da festa à fantasia.
14	Na tela do pintor, as cores vermelha e marrom representavam distintas emoções.
15	Independente se isso é bom ou ruim, eu acho que é errado.
16	As crianças também podem brincar na creche com a educadora.
17	Nada lhe impede de ser bondoso com seus semelhantes.

Continua na próxima página

Nº	Sentença
18	Vim caminhando pela avenida Trompowsky sob abundante chuva.
19	O bebê chorava, atrapalhando o show de Tom Jobim.
20	Na reunião do fórum era um xingamento só.
21	Umbelino ainda não teve como contactar seu tio pintor.
22	João ganhou só um presente de aniversário e ficou indignado.
23	Fiz um bolo no aniversário de Xuxa e degustei um bom vinho.
24	Um objeto estranho entupiu a pia da minha cozinha.
25	Faço psicanálise e acupuntura há muito tempo.
26	Um velho amigo foi merecidamente homenageado.
27	China e Índia não são países vizinhos.
28	Trabalho com síntese e conversão de fala inteligível.
29	Enfim, cada um seguiu seu rumo.
30	Umbigo e *umbrella* não são sinônimos.
31	O engenho foi confundido com o velho moinho de vento.
32	Faltou um membro suplente para a banca ficar completa.
33	Um garoto levou uma injeção contra tétano no bumbum.
34	Milagrosamente, no domingo, ressurgiu o velho cachorro Totó.
35	A homenagem às mães foi feita na quinta-feira pelos bombeiros.
36	Infelizmente, ele levou um choque quando mexia no chuveiro perto da garagem.
37	Sem sombra de dúvida, a roda foi fundamental para outras invenções.
38	Em Tupandi, o varandão e a ponte eram cobertos com muitas telhas de zinco.
39	Na antiga cidade, as casas não têm bombas entulhadas no porão.
40	A rainha suntuosa sambou com uma velha e simples roupa azul.
41	Ele é profundamente bondoso com um simpático cachorrinho.
42	Admirei esse espetáculo até a septuagésima volta.
43	Abneguei de meu feriado em função do diagnóstico.
44	O padrinho inteligente não entendia bem o assunto.
45	Fixei um ninho, cinco galhos e algumas folhas no umbral abandonado.
46	Lanchei com um empresário de invejável currículo.
47	Júlio é um velho indígena conhecido de minha afilhada.
48	A pressão atmosférica é medida com barômetros.
49	Mesmo cansado, fui longe só pra ver onde o jacaré estava.
50	Ganhei um embrulho do etnarca há vinte e cinco anos.

Apêndice B

Maximização do Valor Esperado

Este apêndice apresenta os detalhes do algoritmo de maximização do valor esperado (EM - *expectation maximization*) aplicado para estimar os parâmetros do modelo GMM, para conversão do contorno de *pitch*, para conversão do componente macroprosódico do contorno de *pitch* e para determinar o número de coeficientes da transformada IDCT para cada segmento de contorno de *pitch*.

O objetivo do algoritmo EM é maximizar a função de verossimilhança do conjunto de parâmetros que define o modelo de misturas gaussianas, sendo constituído de dois passos: cálculo da verossimilhança esperada e a maximização da função obtida no primeiro passo.

A função de verossimilhança (*likelihood*) $\mathcal{L}(\mathbf{X}|\theta)$ dos parâmetros $\theta = \{\mu_j, \mathbf{\Sigma}_j, \alpha_i : i = 1, \ldots, M\}$ do modelo GMM é estimada da seguinte forma:

i) Inicialização. Atribui valores iniciais aos parâmetros θ, utilizando o conjunto de dados $\mathbf{Z} = \{z_n : n = 1, \ldots, N\}$.

ii) Passo E (*Expectation*). Estima a probabilidade *a posteriori* $P(\mathcal{C}_i|z_n)$ para cada z_n e cada componente i do GMM, similar a (1.3). Assim,

$$P(\mathcal{C}_i|z_n) = \frac{\alpha_i \mathcal{N}(z_n; \mu_i, \mathbf{\Sigma}_i)}{\sum_{j=1}^{M} \alpha_j \mathcal{N}(z_n; \mu_j, \mathbf{\Sigma}_j)} \tag{B.1}$$

iii) Passo M (*Maximization*). Reestima os parâmetros de cada componente do GMM, utilizando os resultados do passo (ii). O conjunto de parâmetros $\hat{\theta} = \{\hat{\mu}_j, \hat{\mathbf{\Sigma}}_j, \hat{\alpha}_i\}$ é

estimado com as seguintes equações

$$\hat{\alpha}_i = \frac{1}{N}\sum_{n=1}^{N} P(\mathcal{C}_i|z_n) \tag{B.2}$$

$$\hat{\mu}_i = \frac{\sum_{n=1}^{N} P(\mathcal{C}_i|z_n) z_n}{\sum_{n=1}^{N} P(\mathcal{C}_i|z_n)} \tag{B.3}$$

$$\hat{\mathbf{\Sigma}}_i = \frac{\sum_{n=1}^{N} P(\mathcal{C}_i|z_n)[z_n - \hat{\mu}_i][z_n - \hat{\mu}_i]^T}{\sum_{n=1}^{N} P(\mathcal{C}_i|z_n)} \tag{B.4}$$

iv) Finalização. Repetir os passos (ii) e (iii) até se obter a convergência do algoritmo, ou seja, até quando a diferença da verossimilhança logarítmica $\ln(\mathcal{Z}|\theta)$ de duas iterações sucessivas atingir um valor pré-definido.

Mesmo que seja garantida a convergência para um mínimo local, especial atenção deve ser dada à inicialização do algoritmo, pois com parâmetros iniciais *adequados*, o algoritmo converge mais rapidamente e com maiores chances de convergir para a máxima verossimilhança global. Uma prática comum para evitar que as matrizes de covariância se tornem singulares é adicionar uma pequena constante positiva $\varepsilon\mathbf{I}$ à matriz de covariância após cada passo (iii).

Referências Bibliográficas

[1] E. Moulines and Y. Sagisaka, "Voice conversion: State of the art and perspectives," *Speech Communication*, vol. 16, no. 2, pp. 125–126, Feb. 1995.

[2] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Science & Engineering at Oregon Health & Science University, Portland, OR, 2001.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[4] O. Türk, "New methods for voice conversion," Master's thesis, Department Electrical Electronics Engineering, Bogaziçi University, Istanbul, Turkey, 2003.

[5] H. Duxans, "Voice conversion applied to text-to-speech systems," Ph.D. dissertation, Department Signal Theory Communication University Politècnica de Catalunya, Barcelona, Spain, 2006.

[6] O. Türk, "Cross-lingual voice conversion," Ph.D. dissertation, Department Electrical Electronics Engineering, Bogaziçi University, Istanbul, Turkey, Oct. 2007.

[7] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, vol. 2, Seattle, WA, May 1998, pp. 285–288.

[8] D. Sündermann, "Text-independet voice conversion," Ph.D. dissertation, Bundeswehr University, München, Germany, Jul. 2008.

[9] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on GMM and STRAIGHT," in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, Sep. 2001, pp. 361–364.

[10] M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion using bilingual and non-bilingual databases," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Sep. 2002, pp. 293–296.

[11] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "TC-star: Cross-language voice conversion revisited," in *Proceedings of TC-Star Workshop on Speech-to-Speech Translation*, Barcelona, Spain, Jun. 2006, pp. 231–236.

[12] O. Türk and L. M. Arslan, "Subband based voice conversion," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Sep. 2002, pp. 289–292.

[13] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," in *Proceedings of IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, Tampa, FL, Mar. 1985, pp. 748–751.

[14] M. Abe, "A segment-based approach to voice conversion," in *Proceedings of IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, vol. 2, Toronto, Canada, May 1991, pp. 765–768.

[15] D. T. Chappell and J. H. Hansen, "Speaker-specific pitch contour modelling and modification," in *Proceedings of IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, vol. 2, Seattle, WA, May 1998, pp. 885–888.

[16] L. M. Arslan, "Speaker transformation using sentence HMM based alignments and detailed prosody modification," in *Proceedings of IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, vol. 1, 1998, pp. 289–292.

[17] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proceedings of IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 841–844.

[18] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on radial basis function networks," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Sep. 2002, pp. 285–288.

[19] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 2413–2416.

[20] H. Ye and S. Young, "Perceptually weighted linear transformations for voice conversion," in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 2409–2412.

[21] Z. Shuang, R. Bakis, and Y. Qin, "Voice conversion based on mapping formants," in *TC-Star Workshop on Speech-to-Speech Translation*, Barcelona, Spain, Jun. 2006, pp. 219–223.

[22] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, 2006.

[23] K.-S. Lee, "Statistical approach for voice personality transformation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 641–651, Feb. 2007.

[24] Z. Inanoglu, "Transforming pitch in a voice conversion framework," Master's thesis, St. Edmund's College, University of Cambridge, Cambridge, England, 2003.

[25] L. C. Schwardt and J. A. D. Preez, "Voice conversion based on static speaker characteristics," in *Proc. of the 1998 South African Symp. on Communications and Signal Processing*, 1998, pp. 57–62.

[26] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons, 1985.

[27] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: John Wiley & Sons, 2000.

[28] E. E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *Proceedings of IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, vol. 4, Honolulu, HI, 2007, pp. 509–512.

[29] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs: Prentice Hall, 1993.

[30] J. L. Flanagan, "Some properties of the glottal sound source," *Journal of Speech and Hearing Research*, vol. 1, pp. 99–116, 1958.

[31] L. F. Brosnahan and B. Malmberg, *Introduction to Phonetics*. Cambridge: Cambridge University Press, 1976.

[32] F. N. Gregio, "Configuração do trato vocal supraglótico na produção das vogais do português brasileiro: Dados de imagens de resssonância magnética," Dissertação de Mestrado, Pontifícia Universidade Católica de São Paulo, São Paulo, 2006.

[33] W. F. Netto, *Introdução à Fonologia da Língua Portuguesa*, 1. ed. São Paulo: Hedra, 2001.

[34] J. M. Câmara Jr., *Estrutura da Língua Portuguesa*, 8. ed. Petrópolis: Vozes, 1977.

[35] L. Bisol, *Introdução a Estudos de Fonologia do Português Brasileiro*, 3. ed. Porto Alegre: EDIPUCRS, 2001.

[36] J. C. W. Ribeiro, "Estudo comparativo da estrutura silábica em espanhol e português," Dissertação de Mestrado, Universidade Federal de Santa Catarina, Florianópolis, 2003.

[37] T. C. Silva, *Fonética e Fonologia do Português: roteiro de estudos e guia de exercícios*, 8. ed. São Paulo: Contexto, 2005.

[38] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (versão 5.0.47)," Amsterdam, Jan. 2009, programa de computador. [Online]. Available: $http://www.praat.org$

[39] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences 17*, Institute of Phonetic Sciences. University of Amsterdam, 1993, pp. 97–110.

[40] M. M. Wilde, "Controlling performance in voice conversion with probabilistic principal component analysis," Master's thesis, Department of Electrical Engineering and Computer Science, Tulane University, Nova Orleans, LA, 2004.

[41] D. J. Hirst, A. D. Cristo, and R. Espesser, *Levels of representation and levels of analysis for the description of intonation systems, in Prosody: Theory and Experiment*. New York: Kluwer Academic Press, 2000, ch. 3, pp. 51–87.

[42] D. J. Hirst, "A PRAAT plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation," in *Proceedings of International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, Aug. 2007, pp. 1233–1236.

[43] E. K. R. Rao and P. Yip, *The Transform and Data Compression Handbook*. Boca Raton: CRC Press LLC, 2000.

[44] O. Türk and L. Arslan, "Voice conversion methods for vocal tract and pitch contour modification," in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 2845–2848.