Conversão de Contorno de *Pitch* para Aplicação em Sistemas de Conversão de Voz

Marcos Odebrecht Júnior e Rui Seara

***Resumo*— Este trabalho propõe uma nova técnica de conversão do contorno de *pitch* para aplicação em sistemas de conversão de voz. Na abordagem proposta, o contorno de *pitch* é modelado pela superposição de dois componentes: o macroprosódico e o microprosódico. Os componentes são estimados com auxílio do algoritmo MOMEL (*modelling melody*) e diferentes estratégias de conversão são utilizadas. O desempenho do método proposto é comparado com os dois métodos mais utilizados na literatura: conversão utilizando normalização gaussiana (GN) e modelo de misturas gaussianas (GMM). O desempenho subjetivo dos métodos considerados é avaliado através de dois testes subjetivos: de preferência e de similaridade. Os resultados experimentais ratificam a medida adotada, indicando uma preferência pelo método proposto.**

***Palavras-Chave*— Algoritmo MOMEL, conversão de voz, conversão do contorno de *pitch*, INTSINT, prosódia.**

***Abstract*— This work proposes a new technique of pitch contour conversion for application in voice conversion systems. In the proposed approach, the pitch contour is modeled as the superposition of two components: macroprosodic and microprosodic ones. The estimation of such components are performed by using the modelling melody (MOMEL) algorithm and different conversion strategies are considered. The performance of the proposed method is compared against two other commonly used methods: Gaussian normalization (GN) and Gaussian mixture model (GMM) based conversions. Two different subjective tests are used to assess the perceptual performance of the considered methods: preference and similarity tests. Experimental results ratify the adopted measure, pointing towards a preference of the proposed method.**

***Keywords*— MOMEL algorithm, voice conversion, pitch contour conversion, INTSINT, prosody.**

I. Introdução e Definição do Problema

Sistemas de conversão de voz possuem ampla gama de aplicações, tais como personalização da conversão texto-fala (TTS - *text-to-speech*), tradução automática, ensino de idiomas, auxílio no tratamento de doenças degenerativas do sistema fonador e, principalmente, entretenimento [1]–[7]. 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 dos avanços para refinar os resultados desses sistemas, a conversão do contorno de *pitch* tem atualmente experimentado um grande esforço de pesquisa.

Marcos Odebrecht Júnior e Rui Seara, LINSE - Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, SC, E-mails: {marcosode, seara}@linse.ufsc.br.

Este trabalho foi parcialmente financiado pela Coordenação de Aperfeiçoamento Pessoal de Nível Superior (CAPES) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Dentre as técnicas mais utilizadas para conversão do contorno de *pitch* discutidas na literatura, podemos destacar a utilização da normalização gaussiana (GN) e do modelo de misturas gaussianas (GMM - *Gaussian mixture model*), tanto pela flexibilidade quanto pela facilidade de integração com as demais etapas do processo de conversão de voz.

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 à média e ao desvio-padrão do contorno de *pitch* do locutor alvo. Exemplos de trabalhos de conversão de voz que utilizam a normalização gaussiana para a conversão do contorno de *pitch* podem ser encontrados em [6]–[19].

Considerando que as freqüências de *pitch* de ambos locutores possuem função distribuição de probabilidade normal, o contorno de *pitch* do locutor fonte é modificado por

$$\hat{\mathbf{y}} = \frac{(\mathbf{x} - \mu_x)}{\sigma_x}\sigma_y + \mu_y \tag{1}$$

onde $\mathbf{x} = [x(1)x(2)\dots x(N)]^{\mathrm{T}}$, $\mathbf{y} = [y(1)y(2)\dots y(N)]^{\mathrm{T}}$, e $\hat{\mathbf{y}} = [\hat{y}(1)\hat{y}(2)\dots \hat{y}(N)]^{\mathrm{T}}$ são, respectivamente, os vetores de contorno de *pitch* do locutor fonte, do locutor alvo e do contorno convertido. As variáveis μ_x, σ_x, μ_y, e σ_y representam a média e o desvio-padrão de $\mathbf{x}$ e $\mathbf{y}$, respectivamente [10].

A principal vantagem da conversão utilizando GN é sua simplicidade e facilidade na obtenção dos dados de treinamento, uma vez que se necessita apenas de valores médios e desvios-padrão dos contornos de *pitch* de ambos os locutores. Todavia, por alterar apenas a média e o desvio-padrão, a conversão por GN não realiza modificações dos detalhes do contorno de *pitch*, isto é, não é capaz de alterar o padrão de entonação da elocução [7], [20].

Como extensão da conversão por GN, o GMM considera a distribuição de probabilidade conjunta $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$ para estimar a função de conversão [12]. A distribuição de probabilidade de $\mathbf{z}$ é considerada como a soma de m funções gaussianas dadas por

$$p(\mathbf{z}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{z}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \sum_i \alpha_i = 1, \quad \alpha_i \geq 1 \tag{2}$$

onde $N(\mathbf{z}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ representa a distribuição normal bidimensional com vetor média $\boldsymbol{\mu}_i$ e matriz de covariância $\boldsymbol{\Sigma}_i$, e α_i representa o peso de cada mistura do modelo. A matriz de covariância é representada por

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix} \tag{3}$$

e vetor $\boldsymbol{\mu}$ dado por

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}. \quad (4)$$

Os parâmetros do modelo GMM podem ser estimados com o algoritmo de maximização do valor esperado (EM - *expectation maximization*) [21]. Em [12], é apresentada uma função de conversão que busca minimizar o erro quadrático médio (MSE - *mean-square error*) entre os contornos de *pitch* convertido e alvo. Assim,

$$\begin{aligned} \mathcal{F}(\mathbf{x}) = E[\mathbf{y}|\mathbf{x}] &= \int_{-\infty}^{+\infty} \mathbf{y} p(\mathbf{y}|\mathbf{x})\, dy \\ &= \sum_{i=1}^{m} P(\mathcal{C}_i|\mathbf{x})[\mu_i^y + \boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}(\mathbf{x} - \mu_i^x)] \end{aligned} \quad (5)$$

onde $P(\mathcal{C}_i|\mathbf{x})$ é obtido pela aplicação da regra de Bayes

$$P(\mathcal{C}_i|\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}, \mu_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}, \mu_j^x, \boldsymbol{\Sigma}_j^{xx})}. \quad (6)$$

A expressão (5) não faz qualquer consideração a respeito da função distribuição de probabilidade das freqüências de *pitch* do locutor alvo, visto que a segmentação do espaço de parâmetros é realizada a partir de observações de ambos os locutores [12].

Apesar do refinamento dos resultados da conversão utilizando GMM, frente aos resultados obtidos com GN, o processo de manipulação do contorno de *pitch* do locutor fonte visando conversão de voz pode ser facilitado através da seleção de contornos originalmente produzidos pelo locutor alvo. Como vantagem frente às técnicas apresentadas anteriormente, os métodos baseados na seleção de contorno são capazes de utilizar um contorno de *pitch* 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 [5].

Dentre as possíveis abordagens, podemos citar: (i) seleção de contorno da elocução em sua totalidade [10]; (ii) seleção de segmentos vozeados [5] e [20]; (iii) seleção de segmentos que representam sílabas [22]; e (iv) seleção de segmentos que representam fonemas, como proposto 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 treinamento. Após a seleção de um segmento, obtém-se, de um banco de fala paralelo e devidamente alinhado, o segmento produzido pelo locutor alvo. Por banco de fala paralelo entende-se que as elocuções são as mesmas para todos os locutores.

Com o objetivo de refinar os resultados dos atuais métodos de conversão do contorno de *pitch* para conversão de voz, propomos uma abordagem na qual o contorno de *pitch* é decomposto em dois componentes: macroprosódico e microprosódico, possibilitando assim diferentes estratégias de conversão para cada componente. Adicionalmente, é adotada uma medida de distância possibilitando uma avaliação objetiva dos resultados.

Este artigo está organizado como segue. Na Seção II, são detalhadas as técnicas utilizadas na conversão dos componentes macroprosódico e microprosódico, incluindo o índice de desempenho considerado bem como os resultados objetivos obtidos. A Seção III discute os testes subjetivos utilizados para avaliar o desempenho do método proposto. As conclusões e sugestões de trabalhos futuros são apresentadas na Seção IV.

II. Método Proposto

A abordagem proposta realiza a conversão do contorno de *pitch* em duas etapas distintas. Na primeira, 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* [23]. Na segunda etapa, são convertidos os detalhes do contorno de *pitch* não modelados pela curva macroprosódica, isto é, o componente microprosódico. Antes de apresentar a descrição detalhada do método proposto, importantes considerações sobre o banco de sinais de fala utilizado como também a medida objetiva adotada são elencadas.

A. Banco de Sinais de Fala

O banco de sinais de fala considerado no desenvolvimento do método proposto é baseado em 50 elocuções afirmativas pronunciadas por dois locutores do sexo feminino F_1 e F_2, totalizando aproximadamente $2,5$ minutos de material gravado por ambos os locutores. Além do sinal de fala propriamente dito, as marcas de *pitch* e a segmentação fonética fazem parte do banco de dados considerado.

As 40 primeiras elocuções são utilizadas para a etapa de treinamento, enquanto as 10 restantes, para a conversão e posterior avaliação dos resultados. Como é possível realizar a conversão de voz entre dois locutores em duas direções, isto é, alterando os locutores fonte e alvo, temos um total de 20 elocuções com o contorno de *pitch* convertido. Assim, 20 elocuções são convertidas sem que nenhum dado dessas elocuções, pronunciadas pelo locutor alvo, tenha sido considerado na etapa de treinamento. O contorno de *pitch* do locutor alvo das elocuções de teste somente é utilizado no cálculo do índice de desempenho (ver Seção II-B).

B. Avaliação Objetiva

Para possibilitar a comparação objetiva dos resultados, faz-se necessária a realização do alinhamento do contorno de *pitch* da elocução de teste em relação à elocução do locutor fonte. Esse alinhamento é realizado através de interpolação e/ou dizimação e tem por objetivo equalizar o número de marcas dos contornos de *pitch* alvo e convertido com o número de marcas do contorno de *pitch* do locutor fonte.

De posse dos contornos de *pitch* alinhados, é possível formular uma comparação objetiva dos resultados. Para tanto, é considerado o índice de desempenho (PI - *performance index*). O PI é uma medida objetiva proposta em [3] para avaliar o desempenho de diferentes sistemas de conversão

de voz. Originalmente proposto para avaliar a 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 um processo de conversão independentemente da técnica de conversão dos locutores e, até mesmo, do idioma considerado [3], [7] e [24]. O PI adaptado para conversão do contorno de *pitch* é dado por

$$\mathrm{PI} = 1 - \frac{D(\hat{\mathbf{y}}, \mathbf{y})}{D(\mathbf{x}, \mathbf{y})} \tag{7}$$

onde $D(\mathbf{x}, \mathbf{y})$ é a distância existente entre os contornos de *pitch* fonte e alvo, calculada como

$$D(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| \tag{8}$$

onde $||\cdot||$ representa a norma euclidiana e a distância $D(\hat{\mathbf{y}}, \mathbf{y})$ é obtida de forma semelhante à expressão (8), alterando apenas os contornos de *pitch* envolvidos.

É importante notar que o índice de desempenho 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.

Os índices de desempenho mínimo, médio e máximo obtidos com aplicação da conversão do contorno de *pitch*, considerando os métodos GN e GMM, são apresentados na Tabela I. Os valores mínimo e máximo de PI correspondem, respectivamente, ao pior e ao melhor desempenho de uma única elocução enquanto o valor médio representa a média do desempenho obtido com as 20 elocuções consideradas.

C. 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 convertido separadamente da estrutura detalhada do contorno de *pitch*. Tal componente é estimado através do algoritmo MOMEL [25] e armazenado no banco de treinamento após ser codificado pela transcrição INTSINT [23]. A transcrição INTSINT [considerada uma equivalente prosódica do alfabeto fonético internacional (IPA - *international phonetic alphabet*)] é constituída por oito símbolos: T, M, B, H, S, L, U, e D que significam, respectivamente, *top*, *mid*, *bottom*, *higher*, *same*, *lower*, *upstepped* e *downstepped*. Na Fig. 1, são ilustrados o contorno de *pitch*, o componente macroprosódico bem como a correspondente codificação INTSINT.

A primeira etapa de conversão do contorno de *pitch* é realizada através da conversão do componente macroprosódico pela aplicação das técnicas GN e GMM (de três misturas), denominados $\mathrm{MACRO_{GN}}$ e $\mathrm{MACRO_{GMM}}$, respectivamente. Os resultados obtidos nessa etapa são mostrados na Tabela I, enquanto um exemplo dos resultados é ilustrado na Fig. 2.

D. Conversão do Componente Microprosódico

Na segunda etapa de conversão, o componente microprosódico, que condicionado à natureza dos fonemas representa a variação local do contorno de *pitch*, é convertido através da seleção de segmentos de um banco de dados baseado em fonemas.

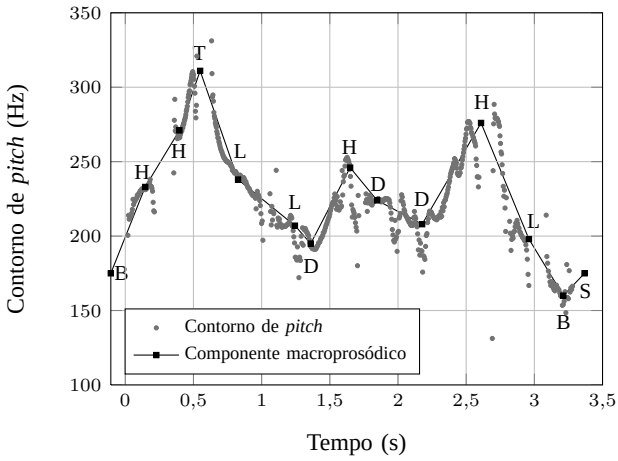


Fig. 1. Exemplo de contorno de *pitch*, componente macroprosódico e codificação INTSINT.

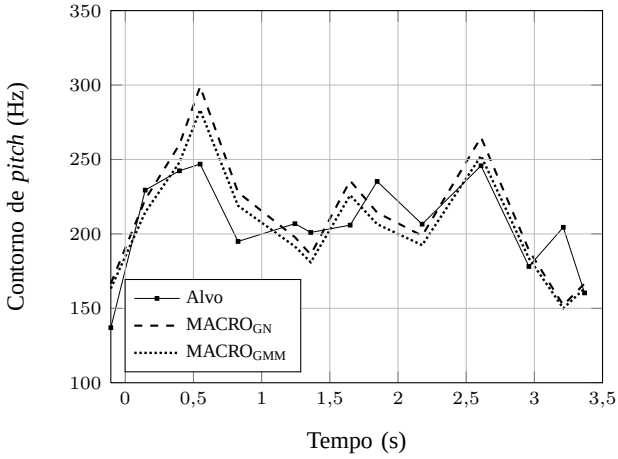


Fig. 2. Exemplo da conversão do componente macroprosódico do contorno de *pitch*.

Para possibilitar a seleção de segmentos, é necessário alinhar os dados de treinamento. Tal alinhamento é inicializado 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 tenha apenas dados que representem os segmentos de contorno de *pitch* que ambos os locutores pronunciaram. Dessa forma, são eliminadas possíveis inserções, apagamentos ou substituições fonéticas causadas por flutuações ou variantes. Uma vez que a conversão do

TABELA I
DESEMPENHO DOS DIFERENTES MÉTODOS DE CONVERSÃO

PI	GN	GMM	$\mathrm{MACRO_{GN}}$	$\mathrm{MACRO_{GMM}}$
Mínimo	$-0,055$	$-0,099$	$-0,071$	$0,004$
Médio	$0,153$	$0,081$	$0,101$	$0,145$
Máximo	$0,362$	$0,258$	$0,300$	$0,328$

componente microprosódico é realizada através da comparação do contorno de *pitch* de um fonema com todos os segmentos 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 no segundo passo, de forma similar à adotada em [22].

O processo de ajuste do tamanho dos segmentos de contorno de *pitch* é representado na Fig. 3, onde $f(n)$ representa o segmento de contorno de *pitch*, $F(k)$, sua transformada DCT, $\tilde{N}$, o número de coeficientes da transformada IDCT e $\tilde{f}(\tilde{n})$, o segmento $f(n)$, porém agora, de tamanho igual a $\tilde{N}$.

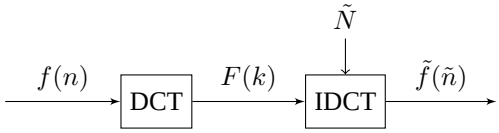


Fig. 3. Diagrama de blocos do processo de ajuste do tamanho dos segmentos.

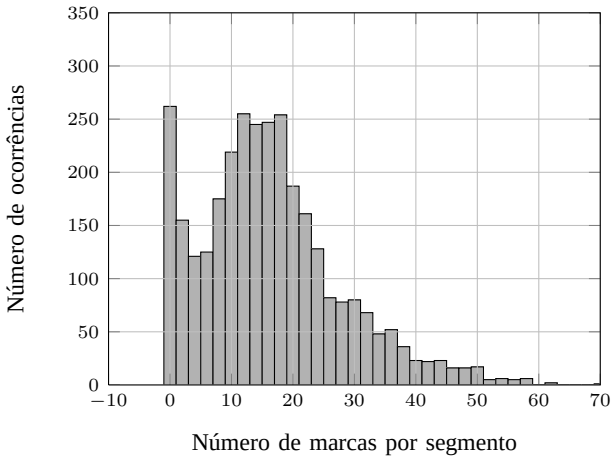
Observando o histograma mostrado na Fig. 4(a), constata-se a importância da escolha do valor de $\tilde{N}$, tendo em vista que níveis elevados de distorção podem ser inseridos no banco de dados quando o número de marcas de um determinado fonema for distante do valor de $\tilde{N}$ considerado. Buscando reduzir os efeitos negativos provenientes da escolha de valores inadequados de $\tilde{N}$, a abordagem proposta utiliza não apenas um, mas três valores de $\tilde{N}$. O valor dos coeficientes utilizados é estimado através do treinamento de um modelo GMM de três misturas (utilizando o algoritmo EM). Assim, cada um dos segmentos do contorno de *pitch*, extraídos das elocuções de treinamento, é codificado através de uma DCT seguida pela IDCT, utilizando $\tilde{N}$ igual ao valor médio das misturas do modelo estimado, especificamente, $6, 18$ e 33. A função distribuição de probabilidade estimada é ilustrada na Fig. 4(b).

Em contraste com a abordagem usada em [22], em que apenas um coeficiente $\tilde{N}$ é considerado, o método proposto busca modelar os segmentos de forma a reduzir os efeitos negativos provenientes da escolha do valor de $\tilde{N}$.

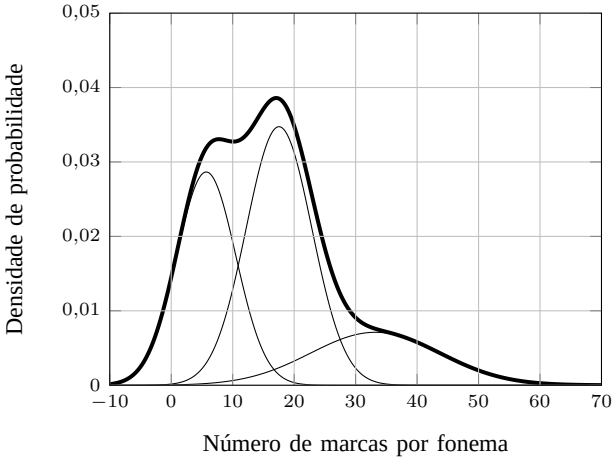
O valor médio de cada segmento de tamanho igual a $\tilde{N}$ é considerado como sendo nulo. Com isso, evita-se que um segmento de contorno de *pitch* seja selecionado mesmo que exista outro segmento similar ao desejado, porém, de maior distância euclidiana. Para tanto, basta remover o nível DC da DCT do sinal, ou seja, igualar a zero o primeiro coeficiente da DCT. Abordagens semelhantes são adotadas em [5], [22], [26].

A comparação do desempenho da seleção de segmentos, usando apenas um dos três valores de $\tilde{N}$ bem como para a utilização dos três valores de $\tilde{N}$, conforme proposto, é mostrada na Tabela II. Para todos os métodos considerados na Tabela II, a conversão do componente macroprosódico é realizada utilizando $\text{MACRO}_{\text{GMM}}$.

Mesmo que o PI máximo obtido com os três valores de $\tilde{N}$ seja inferior aos obtidos com $\tilde{N} = 18$ ou 33, os desempenhos mínimo e médio superiores (ver Tabela II) justificam seu uso.



(a)



(b)

Fig. 4. Modelagem do banco de treinamento por GMM. (a) Histograma do número de marcas por fonema. (b) Função distribuição de probabilidade estimada do GMM.

Um exemplo do resultado da superposição do componente macroprosódico (estimado utilizando GMM) e do componente microprosódico (estimado por seleção de segmentos de contorno de *pitch*), denominado $\text{SELEÇÃO}_{\text{GMM}}$, é ilustrado na Fig. 5, destacando os trechos bem modelados pelas regiões em cinza.

TABELA II
DESEMPENHO 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	$0,226$	$0,237$	$0,243$	$0,231$

III. AVALIAÇÃO SUBJETIVA

Duas abordagens podem ser consideradas nos testes subjetivos para a avaliação dos resultados das técnicas de conversão do contorno de *pitch* para a conversão de voz. Na primeira, as diferentes técnicas são aplicadas a sinais de voz obtidos de um sistema completo de conversão de voz. Essa abordagem é justificada pelas possíveis diferenças sutis obtidas quando

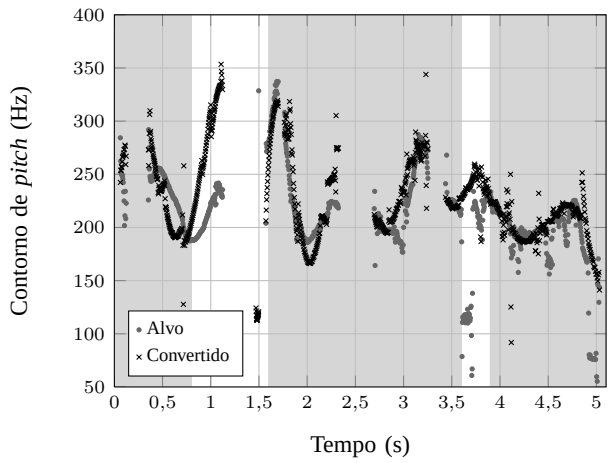


Fig. 5. Resultados obtidos com a conversão do contorno de *pitch* utilizando a abordagem proposta. Trechos bem modelados estão destacados em cinza.

apenas o contorno de *pitch* é convertido. Entretanto, a qualidade dos atuais sistemas de conversão de voz é considerada ainda um tanto *robótica* [22]. 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 [20].

Com o objetivo de isolar os efeitos da conversão do contorno de *pitch*, os testes subjetivos implementados utilizam sinais de voz alterando apenas o contorno de *pitch*. Dessa forma, as elocuções pronunciadas pelo locutor alvo tiveram o contorno de *pitch* substituído pelo contorno de *pitch* convertido a partir do contorno de *pitch* da elocução do locutor fonte.

Para evitar a fadiga dos avaliadores, apenas três técnicas de conversão do contorno de *pitch* são utilizadas nos testes subjetivos: GN, $\text{MACRO}_{\text{GMM}}$ e $\text{SELEÇÃO}_{\text{GMM}}$. Os principais objtivos da avaliação subjetiva são: confrontar o método proposto $\text{SELEÇÃO}_{\text{GMM}}$ com o GN e avaliar a necesidade de refinar os resultados da conversão $\text{MACRO}_{\text{GMM}}$ através da conversão do componente microprosódico por seleção de segmentos.

Os testes foram realizados em um estúdio provido de isolamento acústico com a colaboração de vinte e cinco avaliadores.

A. Teste de Preferência

O primeiro teste subjetivo aplicado para avaliar os resultados obtidos com os diferentes métodos é o teste de preferência. Nesse teste, dez elocuções (cinco de cada locutor) com o contorno de *pitch* convertido por três diferentes técnicas, totalizando trinta elocuções, foram submetidas à avaliação. Para cada elocução, o resultado das três técnicas de conversão foi apresentado de forma aleatória. Os avaliadores foram instruídos a escolher a elocução com entonação mais *natural*. Os resultados da primeira etapa dos testes subjetivos são mostrados na Tabela III.

B. Teste de Similaridade

O segundo teste subjetivo utilizado é o teste de similaridade. Nessa etapa, um novo conjunto de dez elocuções é apresentado aos avaliadores, sendo cinco elocuções de cada locutor e as elocuções são diferentes daquelas utilizadas no teste de preferência. Dessa vez, antes de cada alternativa foi apresentada a elocução original pronunciada pelo locutor alvo, isto é, a elocução de referência. Os avaliadores foram instruídos a escolher a elocução mais próxima à elocução de referência, considerando os aspectos prosódicos. Os resultados do teste de similaridade são apresentados na Tabela IV.

TABELA III
RESULTADOS DO TESTE DE PREFERÊNCIA

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

TABELA IV
RESULTADOS DO TESTE DE SIMILARIDADE

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

A primeira observação que pode ser feita, comparando a Tabela III com os valores médios do PI (Tabela I), é a coerência entre a medida objetiva e os resultados do teste subjetivo. Em segundo lugar, o melhor desempenho do método proposto (em suas duas etapas) frente ao método GN, o mais referenciado na literatura [6]–[19].

Ao final da primeira etapa de testes, a maioria dos avaliadores salientou a dificuldade de escolher entre duas das três elocuções apresentadas. Em testes subjetivos informais, verifica-se que os resultados mais semelhantes entre si correspondem à conversão por $\text{MACRO}_{\text{GMM}}$ e $\text{SELEÇÃO}_{\text{GMM}}$. Essa semelhança deve-se ao fato de a técnica $\text{SELEÇÃO}_{\text{GMM}}$ representar uma extensão da técnica $\text{MACRO}_{\text{GMM}}$.

A diminuição do índice de desempenho (quando incluídos os resultados da conversão do componente microprosódico aos resultados da conversão do componente macroprosódico) deve-se, principalmente, ao tamanho do banco de treinamento, que é de aproximadamente $2,5$ minutos de fala para cada locutor. Com a utilização de um banco de treinamento de maior duração, espera-se que o resultado da conversão do componente microprosódico possa ser refinado, tendo em vista o maior número de segmentos disponíveis para o processo de seleção de segmentos de contorno de *pitch*.

Os resultados obtidos com o método GN no teste de similaridade estão próximos aos obtidos com o teste de preferência. Por outro lado, os resultados oriundos do método $\text{SELEÇÃO}_{\text{GMM}}$ melhoraram significativamente quando comparados com aqueles do teste de preferência. Essa melhora de desempenho subjetivo deve-se ao refinamento introduzido com a conversão do componente microprosódico através da seleção de segmentos, mesmo que isso implique uma redução do desempenho objetivo como observado nas Tabelas I e II.

A melhora de desempenho proporcionada 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 conver-

são para os componentes macroprosódico e microprosódico. Ainda que o método proposto tenha levado a um desempenho superior, quando comparado com os métodos da literatura (GN e GMM), a distância $D(\tilde{\mathbf{y}}, \mathbf{y})$ remanescente após o processo de conversão do contorno de *pitch* ainda deixa margem para melhorar o desempenho do método proposto [ver Tabela V, considerando as distâncias iniciais $D(\mathbf{x}, \mathbf{y})$ mínima, média e máxima iguais a $90\,\mathrm{Hz}$, $115\,\mathrm{Hz}$ e $135\,\mathrm{Hz}$, respectivamente].

TABELA V
DISTÂNCIAS CONVERTIDO-ALVO OBTIDAS COM DIFERENTES MÉTODOS

Locutores	$D(\tilde{\mathbf{y}}, \mathbf{y})$	GN	$\text{MACRO}_{\text{GMM}}$	$\text{SELEÇÃO}_{\text{GMM}}$
$F_1 \rightarrow F_2$	Mínima	$77,68$	$70,88$	$81,11$
	Média	$86,35$	$89,46$	$97,03$
	Máxima	$95,75$	$107,05$	$109,31$
$F_2 \rightarrow F_1$	Mínima	$93,37$	$90,05$	$87,46$
	Média	$99,29$	$111,65$	$110,13$
	Máxima	$116,74$	$131,74$	$128,57$

IV. CONCLUSÕES

A flexibilidade do método proposto, frente aos métodos desenvolvidos exclusivamente para aplicação em sistemas TTS ou que dependem de parâmetros de simulação que precisam ser testados e reconfigurados, possibilita sua aplicação aos mais diferentes tipos de sistemas de conversão de voz. Com a separação do contorno de *pitch* em componentes macroprosódico e microprosódico, é possível implementar diferentes estratégias para cada etapa de conversão do contorno de *pitch*. Esse procedimento sugere a viabilidade de utilização do método proposto para a conversão do contorno de *pitch* em sistemas de conversão de voz com o objetivo de alterar também o padrão de entonação das elocuções.

Através do índice de desempenho PI, adaptado ao problema da conversão do contorno de *pitch*, é possível avaliar o desempenho das diferentes técnicas discutidas. Ainda, como o PI é normalizado pela distância inicial $D(\mathbf{x}, \mathbf{y})$, os resultados alcançados poderão ser facilmente confrontados com futuras técnicas desenvolvidas, mesmo que utilizem diferentes bancos de sinais de fala.

Como trabalho futuro, verifica-se a necessidade de testar o método proposto na conversão do contorno de *pitch* entre elocuções com diferentes padrões de entonação. 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 duração de fonemas.

AGRADECIMENTOS

Os autores gostariam de agradecer aos avaliadores pela inestimável colaboração no processo de avaliação subjetiva requerido para o desenvolvimento deste trabalho de pesquisa.

REFERÊNCIAS

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process. (ICASSP)*, New York, USA, Apr. 1988, pp. 655–658.

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

[3] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Sci. & Eng. at Oregon Health & Sci. Univ., Portland, USA, 2001.

[4] O. Türk, "Cross-lingual voice conversion," Ph.D. dissertation, Dept. Elec. Electron. Eng., Bogaziçi Univ., Istambul, Turkey, Oct. 2007.

[5] ——, "New methods for voice conversion," Master's thesis, Dept. Elec. Electron. Eng., Bogaziçi Univ., Istambul, Turkey, 2003.

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

[7] H. Duxans, "Voice conversion applied to text-to-speech systems," Ph.D. dissertation, Dept. Signal Theory Comm. Univ. Politècnica de Catalunya, Barcelona, Spain, 2006.

[8] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process. (ICASSP)*, Tampa, USA, 1985, pp. 748–751.

[9] M. Abe, "A segment-based approach to voice conversion," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process. (ICASSP)*, vol. 2, Toronto, Canada, May 1991, pp. 765–768.

[10] D. T. Chappell and J. H. Hansen, "Speaker-specific pitch contour modelling and modification," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process. (ICASSP)*, Seattle, USA, May 1998, pp. 885–888.

[11] L. M. Arslan, "Speaker transformation using sentence HMM based alignments and detailed prosody modification," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process. (ICASSP)*, vol. 1, Seattle, USA, 1998, pp. 289–292.

[12] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process. (ICASSP)*, vol. 2, Seattle, USA, 1998, pp. 285–288.

[13] H. Ye and S. Young, "Perceptually weighted linear transformations for voice conversion," in *Proc. European Conf. Speech Comm. Technol. (EUROSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 2409–2412.

[14] 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.

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

[16] K.-S. Lee, "Statistical approach for voice personality transformation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 641–651, Feb. 2007.

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

[18] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," in *Proc. European Conf. Speech Comm. Technol. (EUROSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 2413–2416.

[19] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process. (ICASSP)*, Salt Lake City, USA, 2001, pp. 841–844.

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

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

[22] E. E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process. (ICASSP)*, vol. 4, Honolulu, USA, 2007, pp. 509–512.

[23] 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.

[24] M. M. Wilde, "Controlling performance in voice conversion with probabilistic principal component analysis," Master's thesis, Dept. Elec. Eng. Comp. Sci., Tulane Univ., Nova Orleans, USA.

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

[26] O. Türk and L. Arslan, "Voice conversion methods for vocal tract and pitch contour modification," in *Proc. European Conf. Speech Comm. Technol. (EUROSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 2845–2848.