Análise Paramétrica de Sinais de Voz Baseada em Estimação Conjunta do Modelo Fonte-Filtro

Mário Uliani Neto, José Eduardo de C. Silva, Leandro de C. T. Gomes, Diego A. Silva, Thiago de A. M. Campolina, João Pedro H. Sansão, Hani C. Yehia e Maurílio N. Vieira

***Resumo*— Este trabalho apresenta um algoritmo de análise por síntese baseado em estimação conjunta para análise de sinais de voz. A principal vantagem do algoritmo proposto é a parametrização automática e simultânea do filtro do trato vocal e da fonte de excitação utilizados no processo de produção da voz. O modelo proposto para o trato vocal é capaz de identificar os formantes do sinal de fala. O modelo de excitação modela a forma de onda glotal e os ruídos de aspiração e fricção. Para um melhor modelamento dos ruídos de aspiração e fricção, é proposto o uso do algoritmo TMS (*Transient Modeling Synthesis*).**

***Palavras-Chave*— Análise por síntese, modelo fonte-filtro, TMS.**

***Abstract*— This paper presents an analysis-by-synthesis algorithm based on joint estimation of speech parameters. The main advantage of the proposed algorithm is that both vocal tract and glottal source parameters are estimated simultaneously in an automatic way. The proposed vocal tract model is able to identify the formants of the speech signal. The glottal source model models the glottal waveform and the aspiration and friction noises. To better model the aspiration and friction noises, the use of the TMS (Transient Modeling Synthesis) algorithm is proposed.**

***Keywords*— Analysis-by-synthesis, source filter model, TMS.**

I. Introdução

Uma das abordagens mais utilizadas para modelar o processo de produção da fala é o chamado modelo fonte-filtro [1]. Nesse modelo, o aparelho fonador humano é separado em dois componentes distintos: um filtro linear, cuja função de transferência está relacionada às frequências de ressonância das cavidades supra-glotais do trato vocal humano (boca, faringe, fossas nasais), e uma fonte geradora que excitará esse filtro com um sinal de entrada.

O tipo de sinal emitido pela fonte depende das características do sinal de fala a ser analisado. Nos trechos de fala vozeados, cujo exemplo típico são as vogais, o sinal da fonte é quase periódico, resultado da vibração das pregas vocais. Nos trechos não vozeados, como por exemplo os sons fricativos das consoantes /s/ e /f/, o sinal da fonte é tratado como um ruído gaussiano branco. Já nos trechos híbridos, o sinal da fonte é visto como uma soma dos dois componentes descritos anteriormente.

Mário Uliani Neto, José Eduardo De Carvalho Silva, Leandro de Campos Teixeira Gomes, Diego Augusto Silva e Thiago de Almeida Magalhães Campolina, CPqD, Campinas, SP, Brasil, Emails: {uliani, jcsilva, tgomes, diegoa, thiagomc}@cpqd.com.br. Hani Camille Yehia e Maurílio Nunes Vieira, Departamento de Engenharia Eletrônica, UFMG, MG, Brasil, Emails: {hani, maurilionunesv}@cpdee.ufmg.br. João Pedro Hallack Sansão, Departamento das Engenharias de Telecomunicações e Mecatrônica, UFSJ, MG, Brasil, Email: joao@ufsj.edu.br.

A análise de sinais de fala, tendo como base o paradigma do modelo fonte-filtro, consiste em definir modelos matemáticos para a fonte de excitação e para o filtro do trato vocal, estimando-se parâmetros para os modelos de forma a minimizar um critério de erro entre o sinal original e aquele produzido ao aplicar o sinal da fonte ao filtro.

Neste trabalho, a representação dos trechos vozeados do sinal de fala é feita por meio de um modelo fonte-filtro simplificado, propondo-se a utilização de algoritmos evolutivos para estimar conjuntamente os parâmetros da fonte e do filtro do trato vocal. O modelo proposto é baseado em características físicas do locutor: o modelo do trato vocal é capaz de identificar os formantes do sinal de voz, e o modelo de excitação baseia-se na forma de onda glotal e nos ruídos de aspiração e fricção. Para o melhor ajuste deste último, é proposto o uso do algoritmo TMS (*Transient Modeling Synthesis*).

O artigo está estruturado como descrito a seguir. A seção II apresenta o algoritmo de análise por síntese proposto. Na seção III, são apresentados resultados experimentais que ilustram o método apresentado. Finalmente, a seção IV traz algumas conclusões e perspectivas de trabalhos futuros.

II. Análise por Síntese baseada em Estimação Fonte-Filtro Conjunta

O modelo de produção de voz proposto neste trabalho está ilustrado na Figura 1. Os trechos vozeados e não vozeados do sinal de voz são modelados de formas diferentes. Para os trechos vozeados, a derivada da forma de onda glotal é modelada pelo chamado modelo *Liljencrants-Fant*, ou modelo LF [2]. Os ruídos de aspiração e fricção são modelados utilizando o TMS e um ruído gaussiano branco com amplitude modulada. O trato vocal é modelado por um filtro contendo apenas polos, composto por duas estruturas: uma baseada em frequência e largura de banda dos formantes, e outra complementar, representando a informação não contemplada pelo filtro de formantes. Para os quadros não vozeados, o ruído de turbulência é modelado por um ruído gaussiano branco, enquanto o trato vocal é modelado por um filtro contendo apenas polos. Os detalhes do sistema proposto são apresentados nas subseções seguintes.

A. Deconvolução Fonte-Filtro Conjunta Baseada em Algoritmos Evolutivos

O método desenvolvido para estimar os parâmetros da fonte glotal e do trato vocal é baseado em um paradigma

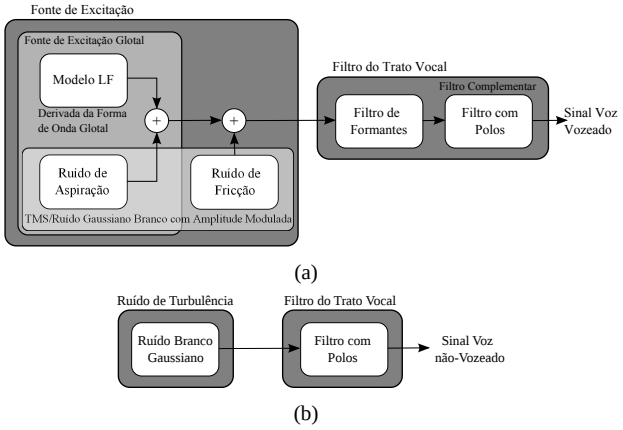


Fig. 1. Modelo fonte-filtro proposto. a) Modelo para sinais vozeados e híbridos. b) Modelo para sinais não vozeados.

conhecido como deconvolução fonte-filtro. O algoritmo de deconvolução proposto é baseado em computação evolutiva [3]. Ele estima de forma conjunta os parâmetros do filtro do trato vocal (modelado através de um conjunto de formantes em cascata), os parâmetros da fonte de excitação (nesta etapa, por simplicidade, a fonte é modelada através do modelo Rosenberg-Klatt (RK) [4]) e o instante de fechamento da glote (GCI, *Glottal Closure Instant*).

1) Modelo da Fonte Vozeada - Modelo RK: O modelo RK é um modelo parametrizado no domínio do tempo, capaz de modelar um período fonatório da derivada da forma de onda glotal, caracterizando os momentos em que a glote está aberta e fechada. O modelo RK é dado por:

$$\hat{g}_{RK}(n) = \begin{cases} 0 & 1 \leq n < n_c \\ 2a(n-n_c) - 3b(n-n_c)^2 & n_c \leq n < T_0 \end{cases}, \quad (1)$$

onde T_0 corresponde ao período fonatório, n_c à duração de fase fechada e os parâmetros a e b devem respeitar a relação $a = b.(T_0 - n_c).T_0$.

O modelo RK possui ainda um filtro passa-baixas com função de transferência $\frac{1}{1-\mu z^{-1}}$ com $\mu > 0$, para controlar o chamado *decaimento espectral* (*spectral tilt*).

2) Modelo do Trato Vocal - Filtro de Formantes: O filtro do trato vocal adotado é formado por um conjunto de ressoadores conectados em cascata [4]. Cada ressoador é especificado através de dois parâmetros: a frequência de ressonância (formante) F e a largura de banda de ressonância BW:

$$\begin{aligned} C &= -e^{-2\pi\, BW\, T} \\ B &= 2e^{-\pi\, BW\, T}\cos(2\pi\, F\, T) \\ A &= 1 - B - C \end{aligned} \quad (2)$$

onde T é o período de amostragem.

A função de transferência do trato vocal tem a seguinte resposta em frequência:

$$vt(z) = \prod_{i=1}^{\substack{quantidade \\ formantes}} \frac{A_i}{1 - B_i z^{-1} - C_i z^{-1}} \quad (3)$$

onde A_i, B_i e C_i são determinados, segundo a equação 2, pelos valores da frequência e largura de banda do i-ésimo formante.

3) Otimização Baseada em Algoritmos Evolutivos: O método utilizado para estimação conjunta dos parâmetros da fonte glotal e filtro de trato vocal é baseado em computação evolutiva. Para diminuir o número de parâmetros a serem otimizados e limitar o espaço de busca desses parâmetros, reduzindo a complexidade computacional, são utilizados modelos simplificados para a fonte de excitação (seção II-A.1) e para o filtro de trato vocal (seção II-A.2). Além do menor número de parâmetros, os modelos simplificados permitem que se defina um conjunto de restrições para que as configurações dos modelos representem uma estrutura física válida de um locutor durante o processo de produção de voz; essas restrições contribuem no processo de otimização baseado em algoritmos evolutivos, promovendo um ganho significativo no que diz respeito ao custo computacional e diminuindo a quantidade de mínimos locais presentes na superfície de *fitness*. O modelo RK é utilizado nesta etapa com as restrições presentes na tabela I.

Parâmetro	Restrição	Codificação
n_c	$0,4 \times T_0 < n_c \leq 0,7 \times T_0$	Inteiro
a	$a \leq 0$	Real
μ	$0 < \mu < 0,99$	Real
GCI	$0 \leq GCI \leq T_0$	Inteiro

TABELA I

PARÂMETROS DA FONTE E GCI: RESTRIÇÕES E CODIFICAÇÃO.

O filtro do trato vocal é modelado através de um filtro apresentando as restrições enumeradas na tabela II [4].

Formante	Frequência (Hz)		Largura de banda (Hz)	
	Mínimo	Máximo	Mínimo	Máximo
1	150	900	40	500
2	500	2500	40	500
3	1300	3500	40	500
4	2500	4500	100	500
5	3500	5500	150	700
6	4000	8000	200	2000
7	4000	8000	200	2000

TABELA II

RESTRIÇÕES DOS VALORES DE FREQUÊNCIA E LARGURA DE BANDA PARA O FILTRO DE FORMANTES EM CASCATA.

Para a otimização conjunta dos parâmetros dos modelos, foi usado uma estratégia evolutiva. Tanto os parâmetros da fonte (n_c, a, μ,GCI) quanto os do filtro (F_i, BW_i) fizeram parte do cromossomo, de forma que os modelos foram estimados conjuntamente. Os detalhes desta implementação foram apresentados em [5].

B. Ajuste do Trato Vocal Através de Filtragem Adaptativa

Para melhorar a modelagem do trato vocal, é proposto o uso de um filtro adaptativo capaz de estimar os parâmetros de um filtro complementar que modele a informação do trato vocal

não presente no filtro de formantes. O filtro adaptativo utilizado neste trabalho é baseado no filtro de *Wiener*, otimizado com o algoritmo RLS (*Recursive Least Squares*) [6]. A figura 2 detalha o modelo de filtragem utilizado.

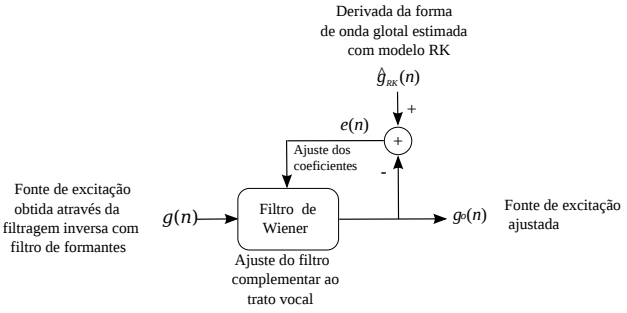


Fig. 2. Filtro adaptativo para ajuste do filtro complementar ao trato vocal.

Este método está baseado na comparação da saída $g_o(n)$ do filtro de *Wiener* com a estimativa da derivada do sinal glotal $\hat{g}_{RK}(n)$, obtido através dos parâmetros otimizados no processo de estimação conjunta. Quando os coeficientes do filtro de *Wiener* são adequadamente ajustados, o sinal de erro $e(n)$ é minimizado, o que implica em um sinal na saída do filtro $g_o(n)$ o mais próximo possível do sinal $\hat{g}_{RK}(n)$. O sinal $g_o(n)$ é a derivada da forma de onda glotal, obtida através da deconvolução do quadro de voz original com o filtro do trato vocal formado pelos filtros de formantes e complementar.

O processo de estimação conjunta dos parâmetros do filtro de formantes e do modelo RK não é capaz de estimar de forma satisfatória a duração em que a glote permanece fechada (parâmetro n_c), devido a erros provocados por simplificações do modelo de formantes utilizado no trato vocal.

Para o ajuste do n_c, é proposto um processo de busca linear, com variação do n_c durante a etapa de filtragem adaptativa. Assumindo que esse parâmetro pode variar entre 40% e 70% do período fonatório [7], o algoritmo de filtragem adaptativa é executado para valores de n_c entre estes limites, com passo de 10%. A solução escolhida como ótima é aquela que apresenta o menor erro quadrático médio entre o quadro original temporal e a forma de onda sintetizada a partir do sinal $\hat{g}_{RK}(n)$ e do filtro do trato vocal.

C. Ajuste da Fonte Glotal - Modelo LF

O modelo LF [2] é capaz de descrever a derivada da forma de onda glotal com uma precisão maior do que o modelo RK. Esse modelo pode ser descrito como:

$$\hat{g}_{LF}(n) = \begin{cases} E_0 e^{\alpha n} \operatorname{sen}(\omega_g n), & 0 \le n < T_e \\ -\frac{E_e}{\epsilon T_a}\left[e^{-\epsilon(n-T_e)} - e^{-\epsilon(T_c-T_e)}\right], & T_e \le n < T_c \le T_0 \end{cases} \quad (4)$$

A forma de onda produzida pelo modelo LF pode ser determinada através de quatro parâmetros temporais: $\{T_p, T_e, T_a, T_c\}$, além da magnitude $\{E_e\}$, conforme a figura 3.

O ajuste do modelo LF é realizado em duas etapas. Inicialmente, é realizada uma estimativa dos parâmetros temporais do modelo LF (T_p, T_e, T_a, T_c) e da excitação glotal E_e através de métodos de estimação direta [8]. Esta técnica mede os parâmetros diretamente da forma de onda temporal obtida com o modelo RK.

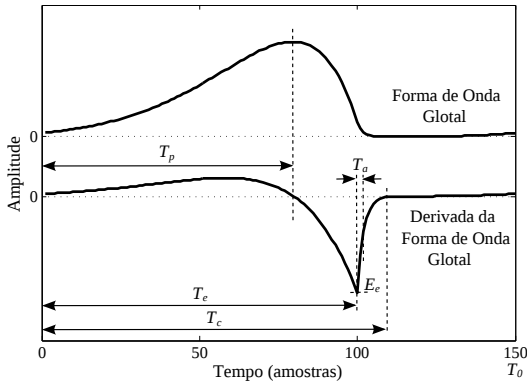


Fig. 3. Formas de onda glotal e sua derivada geradas pelo modelo LF.

Em seguida, os parâmetros do modelo LF são refinados através de uma estratégia evolutiva. As estimativas de T_c e T_e são consideradas confiáveis [8] e não sofrem alterações na otimização. O parâmetro T_a é confinado a variações entre 0 e $T_c - T_e$; o T_p pode variar entre $\pm 20\%$ da estimativa inicial; E_e pode variar entre $\pm 10\%$ da estimativa inicial. O *fitness* baseia-se no erro quadrático entre a derivada da forma de onda glotal do quadro original e forma de onda ajustada através do modelo LF.

D. Modelagem do Ruído Residual

As técnicas de modelagem da fonte glotal descritas nas seções anteriores não incorporam o ruído de aspiração e fricção no sinal de voz. Devido a esta limitação na modelagem, a diferença entre o sinal da fonte $\hat{g}_{LF}$ e o sinal obtido pela filtragem inversa do quadro original produz um ruído residual. Uma forma de interpretar este ruído é considerá-lo como um sinal gaussiano branco modulado em amplitude [9]. Mas nem sempre o ruído gaussiano branco modela satisfatoriamente o ruído residual. A fim de manipular o ruído residual com o objetivo de aproximá-lo de um ruído gaussiano branco, propõe-se utilizar o algoritmo TMS [10]. Essa técnica explora a dualidade tempo-frequência de senoides e impulsos, e como mostrado em [10] pode ser usada para separar trechos que apresentem característica impulsiva no domínio do tempo. O processo está ilustrado na figura 4.

E. Modelagem do Resíduo Final

O modelo do resíduo final usado neste trabalho baseia-se na técnica proposta em [9]. O método consiste na aproximação do resíduo por um ruído gaussiano síncrono com o período fonatório e modulado em amplitude pelo modelo LF.

O resíduo final é parametrizado da seguinte forma: primeiro, um ruído gaussiano de média zero e variância unitária é modulado pela forma de onda glotal obtida através do modelo LF. Em seguida, a energia deste ruído gaussiano é ajustada para um valor igual ao da energia do resíduo final.

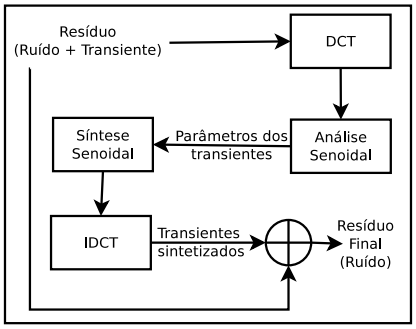


Fig. 4. Diagrama de blocos do funcionamento do TMS.

III. Análise de Resultados

Para avaliação do modelo proposto, são apresentados resultados com dois sinais: no primeiro caso, foi sintetizado um sinal amostrado a 8 kHz usando-se uma fonte LF com período de 120 Hz e um trato com 4 formantes com largura de banda de 100 Hz e posicionados em 500 Hz, 1500 Hz, 2500 Hz, 3500 Hz; no segundo caso, é apresentado a análise de um quadro da vogal "a" tônica (/a/) para uma voz masculina gravada em estúdio, livre de ruído, codificada em PCM linear com taxa de amostragem de 8 kHz e 16 bits por amostra.

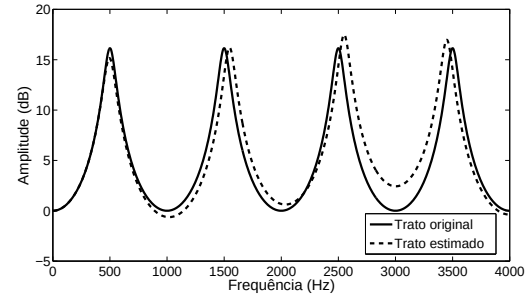
Na otimização, foi utilizado um algoritmo baseado na estratégia evolutiva ($\mu + \lambda$), configurado para iterar por 100 gerações, considerando-se um total de 400 indivíduos (μ) por geração. A cada geração foram gerados 200 filhos (λ), com taxa de *crossover* de 1 e probabilidade de mutação de 1.

A figura 5 apresenta o resultado para o sinal sintetizado. Observa-se em 5(a) que o trato estimado é muito próximo do original. Há pequenas diferenças que se concentram especialmente nos formantes de ordem mais elevada. O fato de o cálculo do *fitness* do algoritmo evolutivo (seção II-C) ser baseado apenas no erro quadrático do sinal da glote pode justificar este comportamento. A maior parte da energia do pulso glotal está concentrada em baixas frequências ($f < 2000$ Hz), de modo que um erro nos formantes de ordem mais alta não afeta de modo expressivo o sinal da glote, o que, por consequência, não altera significativamente o *fitness*. As figuras 5(b) e 5(c) mostram que a estimativa da derivada da forma de onda glotal e o sinal sintetizado estão próximos de suas respectivas referências. As diferenças existentes podem ser justificadas pelos erros cometidos na otimização do trato, que acabam refletindo diretamente em desvios no sinal da glote. Apesar disso, o sinal sintetizado aproxima-se do sinal original, como mostrado na figura 5(c).

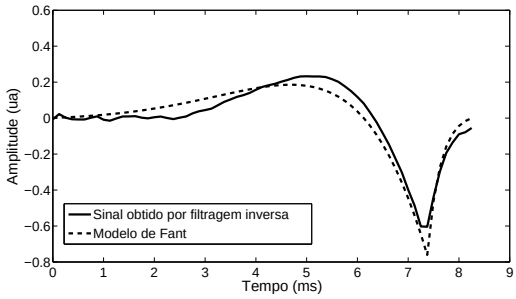
A figura 6 apresenta a otimização de um sinal com trato e fonte idênticos ao apresentado na figura 5, porém, com adição de ruído de aspiração à fonte glotal. O ruído de aspiração foi gerado através de um ruído gaussiano de média 0 e variância 1, modulado pelo sinal glotal. A figura 6(a) apresenta o sinal temporal original e o sinal recuperado com o algoritmo proposto (incluindo o uso do algoritmo TMS). Pode-se observar que os sinais apresentam diferenças, no entanto, seguem um mesmo contorno.

A figura 6(b) apresenta a função de autocorrelação do ruído do sinal (obtido através da diferença entre o modelo LF e o sinal obtido através da filtragem inversa do sinal original) com e sem o uso do algoritmo TMS. Diferentemente da função de autocorrelação do ruído produzido sem o uso do TMS, a autocorrelação do ruído gerado com o uso do algoritmo de TMS se mantém dentro do intervalo de confiança de 95% (definido pelas linhas horizontais pontilhadas presentes na figura) para todos os atrasos maiores que 2 amostras. Isso evidencia que esse é mais próximo de um ruído branco que o aquele gerado quando o TMS não é usado. É importante dizer que não é foco deste trabalho avaliar o impacto dos parâmetros do algoritmo TMS sobre os resultados apresentados.

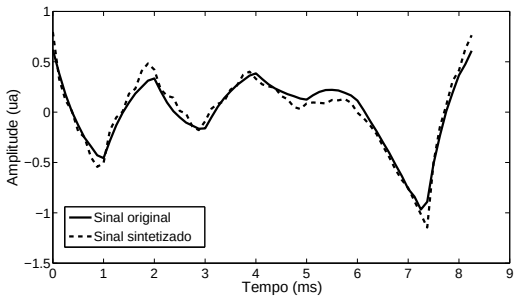
A figura 7 apresenta um resultado preliminar da otimização para um quadro de um sinal real da vogal /a/. A forma de onda do sinal sintetizado apresenta, marcadamente no início e final da figura, grandes desvios em relação ao sinal original. Isto provavelmente aconteceu por causa das limitações existentes nos modelos usados neste trabalho. A identificação precisa das razões para este descasamento será alvo de trabalhos futuros.



(a)

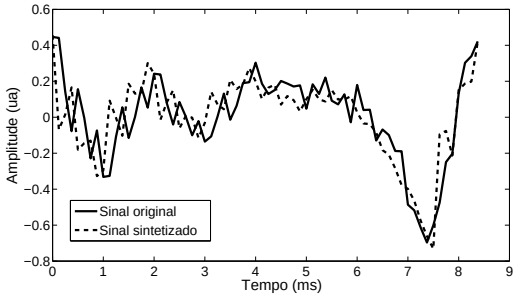


(b)

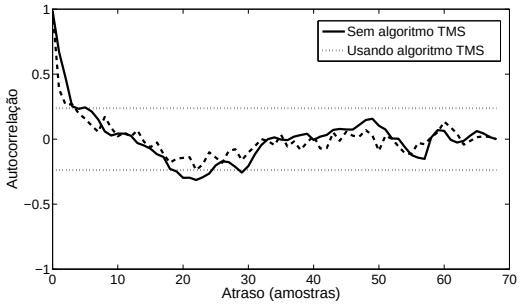


(c)

Fig. 5. Otimização de um sinal sintetizado. a) Trato vocal original e estimado. b) Derivada do sinal glotal original e otimizado. c) Forma de onda temporal do sinal original e sintetizado.



(a)



(b)

Fig. 6. a) Sinal original e sinal recuperado usando-se o algoritmo TMS. b) Função de autocorrelação do ruído do sinal com e sem o uso do algoritmo TMS.

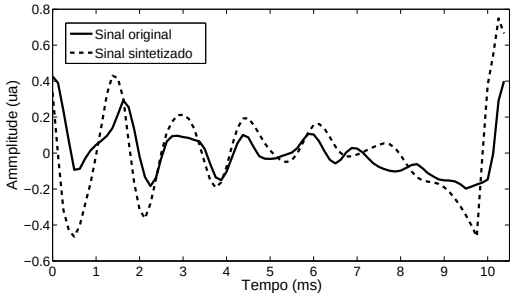


Fig. 7. Forma de onda temporal do sinal real e sintetizado (um quadro da vogal /a/).

IV. Conclusões

Neste trabalho, foi apresentada uma técnica de otimização conjunta dos parâmetros da fonte sonora e do filtro de trato vocal para a produção de trechos vozeados de sinais de fala. Foi utilizado o modelo LF para representação da fonte, e uma estratégia baseada no algoritmo TMS para modelamento das componentes de ruído. O filtro foi obtido através do cascateamento de quatro formantes, modelados através de uma frequência central e uma largura de banda. A otimização foi realizada por meio de uma estratégia evolutiva.

A computação evolutiva apresenta a vantagem de permitir a otimização conjunta dos parâmetros de fonte e filtro, mesmo que a função de *fitness* represente um problema multimodal. Em todas as simulações realizadas, os algoritmos encontraram uma solução factível e satisfatória para a otimização. A principal desvantagem está relacionada ao custo computacional, sendo necessário um tempo significativo para convergência dos algoritmos.

Uma vantagem do método de otimização apresentado é sua capacidade de determinar com eficiência o coeficiente ótimo de decaimento espectral, bem como o instante de fechamento glótico, GCI. Em outros trabalhos, como apresentado em [9], é comum o uso de algoritmos específicos para detecção do coeficiente de ajuste do decaimento espectral e detecção do GCI. Além disso, o uso do algoritmo TMS possibilita um branqueamento do ruído residual obtido através da diferença da derivada da forma de onda glotal original e do modelo LF, levando a um melhor ajuste do ruído.

A abordagem apresentada neste trabalho para modelamento de fonte e filtro permite que se faça uma interpretação física dos parâmetros obtidos na otimização, uma vez que o modelo LF expressa a derivada do pulso glotal e o filtro com formantes em cascata representa a envoltória espectral dos quadros de fala. Esta técnica mostra-se viável para aplicações como compressão de sinais de fala, transformação de voz (pois os parâmetros obtidos na otimização podem ser alterados), e suavização de parâmetros de quadros adjacentes que necessitem ser concatenados (permitindo a evolução de algoritmos de síntese concatenativa de fala).

Em trabalhos futuros, prevêem-se o estudo e a análise da viabilidade da aplicação da técnica proposta em sistemas de transformação de voz.

Agradecimentos

Os autores agradecem o apoio dado a este trabalho, desenvolvido com recursos do Fundo para o Desenvolvimento Tecnológico das Telecomunicações (Funttel) do Ministério das Comunicações administrados pela Finep.

Referências

[1] Fant, G., *Acoustic Theory of Speech Production.* Mouton, The Hague. 1970.

[2] G. Fant, J. Liljencrants, and Q. Lin, *A four-parameter model of glottal flow*. STL-QPSR, 26(4):1-13, 1985

[3] de Castro, L. N., *Fundamentals of Natural Computing: Basic Concepts, Algorithms and Applications*. Flórida, Estados Unidos: Chapman & Hall/CRC, 2006

[4] H. Klatt, *Software for a cascade/parallel formant synthesizer*. Journal of the Acoustical Society of America, 67(3):971-995, March 1980

[5] M. U. Neto, B. Costa, F. Simões, R. Violato, M. Leal *Estimação conjunta do processo de produção de sinais de fala utilizando computação evolutiva*. In IV Congresso Tecnológico InfoBrasil, 2011

[6] Simon Haykin, *Adaptive Filter Theory*. Prentice Hall, United States, fourth edition, 2001

[7] Vieira, M.N. *Automated measures of dysphonias and the phonatory effects of asymmetries in the posterior larynx*. Ph.D. dissertation, University of Edinburgh, 1997.

[8] J. Perez, A. Bonafonte, *Automatic voice-source parameterization of natural speech*. In Proceedings of Interspeech, pages 1065-1068, Lisboa, Portugal, Sep 2005

[9] A. Del Pozo and S. J. Young, *The linear transformation of LF glottal waveforms for voice conversion*. In Proceedins Interspeech, pages 1457-1460, Brisbane, Australia, Sep 2008

[10] Verma, T.S., and Meng, T.H.Y., *Extending spectral modeling synthesis with transient modeling synthesis*. Computer Music Journal 24(2), volume 24, MIT Press, 47-59, 2000