Dereverberação de voz baseada em dois estágio de predição linear utilizando múltiplos microfones

Alberto Y. Nakano e Phillip M. S. Burt

***Resumo*— Neste trabalho apresentamos um metodo de dereverberação de voz utilizando multiplos microfones baseado em predição linear (PL). Em um primeiro estagio de PL, poucos coeficientes são suficientes para branquear o envelope espectral do sinal de voz. Em um segundo estagio de PL, um grande numero de coeficientes e utilizado para branquear o envelope espectral devido a reverberação. O envelope espectral da voz e reintroduzido para produzir a estimativa do sinal dereverberado. Os sinais reverberados obtidos por meio de multiplos microfones são diferentes entre si, o que contribui para a melhora da analise de PL em cada estagio. O processo e adequado para filtragem adaptativa. Resultados baseados em respostas impulsivas de reverberação e filtragem adaptativa FTF apontam para um bom potencial de desempenho.**

***Palavras-Chave*— Dereverberação, multiplos microfones, predição linear, filtragem adaptativa**

***Abstract*— In this paper we present a multimicrophone speech dereverberation method based on linear prediction (LP). In a first multiple input LP stage, with a small number of coefficients, the spectral envelope of speech is approximately whitened. In a second multiple input LP stage, with a large number of coefficients, the spectral envelope due to reverberation is approximately whitened. The spectral envelope due to speech is reintroduced in order to produce the output. The reverberated signals obtained with a microphone array are not identical, which contributes to improve the LP analysis in each stage. The procedure is well suited for adaptive filtering. Results based on measured reverberation responses and FTF (Fast Transversal Filter) adaptive filtering point to a considerable performance potential.**

***Keywords*— Dereverberation, multimicrophone, linear prediction, adaptive filtering**

I. INTRODUÇÃO

Reverberação é a combinação de versões atenuadas e atrasadas de um sinal devido às múltiplas reflexões que ocorrem de forma aleatória em um espaço fechado. É um problema em reconhecimento de voz e sistemas de comunicação viva-voz. Alguns estudos recentes podem ser encontrados em [1], [2]. Porém, são métodos que empregam apenas um microfone, ou seja, apenas um sinal reverberado. Métodos com múltiplos microfones, nos quais se utiliza mais de um sinal reverberado, também são propostos [3], [4]. Recentemente, técnicas de dereverberação baseadas no processamento e melhoramento do resíduo de predição linear vêm sendo estudadas [5], [6]. O resíduo é processado de modo a destacar as características da voz e empregado no filtro responsável pela reconstrução do sinal de voz. O uso de múltiplos microfones e métodos baseados em PL são combinados em [7].

Neste trabalho propomos uma nova abordagem para obter coeficientes de PL de um conjunto de microfones e apresentamos um modo simples para processar o resíduo de PL do sinal utilizando outro estágio de PL para reduzir os efeitos da reverberação. Este artigo é dividido nas seguintes seções: na seção II aspectos da análise de PL sobre voz são revistos. Na seção III é apresentado o método de dereverberação proposto. Simulações e resultados são apresentados na seção IV e conclusões na seção V.

Alberto Y. Nakano e Phillip M. S. Burt¸ Departamento de Telecomunicações e Controle, Escola Politécnica, Universidade de São Paulo, São Paulo, Brasil, E-mails: {nakano,phillip}@lcs.poli.usp.br. Este trabalho foi financiado pela CAPES.

II. TEORIA DE PREDIÇÃO LINEAR

O modelo autoregressivo (AR) de produção de voz visto na Figura 1, consiste em um filtro

$$G(z) = \frac{1}{A(z)} \tag{1}$$

que modela o trato vocal cuja entrada varia entre um trem de impulsos de período de pitch T e ruído branco dependendo do tipo de segmento de voz desejado, vocálico ou não-vocálico. Esta relação pode ser expressa pela equação

$$s(n) = g(n) * u(n), \tag{2}$$

em que $s(n)$, $g(n)$, "$*$" e $u(n)$ são, respectivamente, o sinal de voz, a resposta impulsiva associada a $G(z)$, o operador convolução e a excitação de entrada.

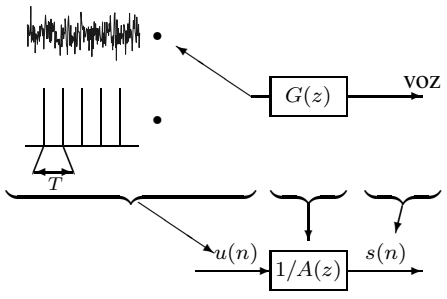

Fig. 1. Modelo básico de produção de voz. $s(n)$, $G(z)$ e $u(n)$ são, respectivamente, a voz, o modelo do trato vocal e a excitação de entrada.

Os coeficientes do filtro $A(z)$ podem ser estimados por PL. Isto é possível assumindo que as características do trato vocal variam lentamente no tempo e, que portanto, podemos considerar o sinal de voz estacionário em tempo-curto. Nestas

condições as K últimas amostras do sinal são usadas para estimar o valor presente do sinal

$$\hat{s}(n) = \sum_{k=1}^{K} w_k s(n-k). \tag{3}$$

Os coeficientes w_k são determinados minimizando-se o erro quadrático médio $E\{e^2(n)\}$, em que $e(n) \doteq s(n) - \hat{s}(n)$. O que nos conduz à equação de Wiener-Hopf

$$\mathbf{Rw} = \mathbf{p}, \tag{4}$$

em que $\mathbf{s}(n) = [s(n-1) \quad s(n-2) \quad \cdots \quad s(n-K)]^T$, $\mathbf{R} = E\{\mathbf{s}(n)\mathbf{s}(n)^T\}$, $\mathbf{p} = E\{\mathbf{s}(n)s(n)\}$ e $\mathbf{w} = [w_1 \quad w_2 \quad \cdots \quad w_K]^T$ são, respectivamente, vetor contendo amostras do sinal de voz, a matriz de autocorrelação, o vetor de correlação cruzada e os coeficientes de PL. O filtro $A(z) = \sum_{k=0}^{K} a_k z^{-k}$ é obtido diretamente da relação

$$\begin{cases} a_0 = 1, & k = 0 \\ a_k = -w_k, & k = 1, 2, \ldots, K \end{cases}. \tag{5}$$

III. Dereverberação por meio de múltiplos microfones em dois estágio

Considere inicialmente o caso em que usamos um microfone. O sinal reverberado $x(n)$ pode ser expresso como

$$x(n) = h(n) * s(n), \tag{6}$$

em que $h(n)$ é a resposta impulsiva entre a fonte de sinal e o microfone. Usando (2), temos que

$$x(n) = h(n) * g(n) * u(n). \tag{7}$$

Para obter uma estimativa de $s(n)$, passamos $x(n)$ por um estágio de PL com um número pequeno de coeficientes e posteriormente passamos sua saída por um segundo estágio de PL com um número elevado de coeficientes. A motivação para isto é que $g(n)$ e $h(n)$ possuem respostas em freqüência distintas: $|G(e^{j\omega})|$ possui variações em freqüência mais suaves que $|H(e^{j\omega})|$, como vistas na Figura 2, acima e abaixo, respectivamente. Devido ao efeito de branqueamento do processo de PL, o primeiro estágio de PL fornecerá $\widetilde{A}(z) \approx A(z) = 1/G(z)$, enquanto o segundo estágio fornecerá uma resposta em freqüência com o módulo aproximadamente dado por $1/|H(e^{j\omega})|$. Aplicando então $1/\widetilde{A}(z)$ ao sinal à saída do segundo estágio produz-se um sinal que, idealmente, possui apenas uma distorção de fase em relação a $s(n)$. Este procedimento é adequado ao uso de filtros adaptativos: para acompanhar as variações do sinal de voz, o primeiro estágio de PL deve adaptar-se rapidamente, o que é facilitado por ser de comprimento curto; o segundo estágio, por outro lado, tem um número grande de coeficientes, mas sua adaptação pode ser mais lenta, pois seu objetivo é seguir as variações da resposta acústica de $h(n)$. Como será visto em detalhes a seguir, este procedimento pode ser melhorado usando mais de um sinal de entrada

$$x_i(n) = h_i(n) * s(n) = h_i(n) * g(n) * u(n), \tag{8}$$

em que $h_i(n)$ é a resposta impulsiva entre a fonte de sinal e o microfone i, sendo $i = 1, 2, \ldots N$. Com isso, a estimação de

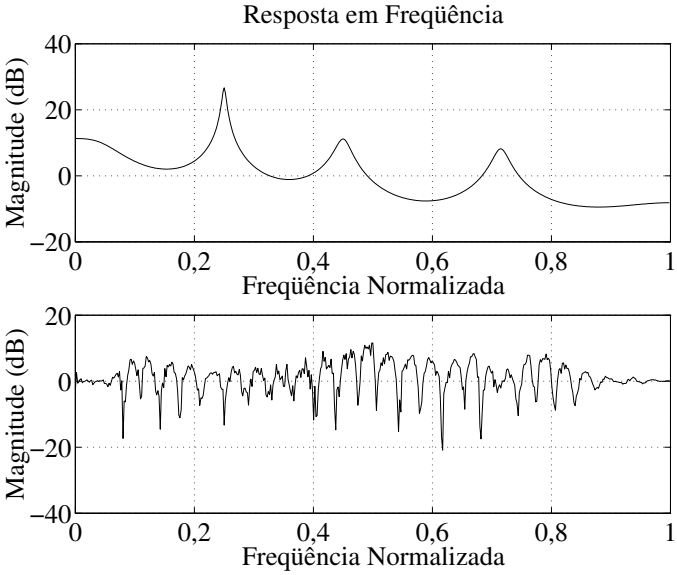

Fig. 2. Respostas em freqüência associadas a $g(n)$ (acima) e $h(n)$ (abaixo)

$A(z)$ no primeiro estágio é menos afetado pela reverberação e os filtros obtidos por PL no segundo estágio podem ser mais curtos.

A. Primeiro estágio

Gaubitch et al.[7] propuseram um método para estimar $A(z)$ aproveitando a diversidade espacial. É baseado na suposição de que tomando a média dos coeficientes de PL $\mathbf{w}_i$, de cada sinal $x_i(n)$, o efeito da reverberação na estimação de $A(z)$ é reduzido. Há dois inconvenientes neste procedimento: primeiro é necessário calcular os coeficientes de PL separadamente para cada sinal e segundo, a média dos coeficientes $\mathbf{w}_i$ pode produzir um filtro instável $1/\widetilde{A}(z)$.

Considerando inicialmente um cenário offline, propomos a minimização da soma dos erros quadráticos médios

$$\sum_{i=1}^{N} E\{e_i^2(n)\}, \tag{9}$$

em função dos coeficientes w_k, em que $e_i(n) = x_i(n) - \sum_{k=1}^{K} w_k x_i(n-k)$. O que nos conduz

$$\mathbf{w} = \left(\sum_{i=1}^{N} \mathbf{R}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{p}_i\right), \tag{10}$$

sendo $\mathbf{R}_i = E\{\mathbf{x}_i(n)\mathbf{x}_i(n)^T\}$, $\mathbf{p}_i = E\{\mathbf{x}_i(n)x_i(n)\}$ e $\mathbf{x}_i(n) = [x_i(n-1) \quad x_i(n-2) \quad \cdots \quad x_i(n-K)]^T$. Esta equação possui a mesma forma que a equação de Wiener-Hopf para entrada simples, que produz um $1/\widetilde{A}(z)$ estável. Portanto, este procedimento também produz um $1/\widetilde{A}(z)$ estável.

Em um cenário *online*, os coeficientes $\mathbf{w}$ são continuamente adaptados minimizando uma função custo relacionada a $\sum_{i=1}^{N} E\{e_i^2(n)\}$, por exemplo, $\sum_{m=0}^{n} \sum_{i=1}^{N} e_i^2(m)$. Além disto, as saídas $e_i(n)$ do primeiro estágio são as entradas do segundo estágio.

B. Segundo Estágio

No segundo estágio, um processo similar à equalização multicanal é realizado. Como é baseado em predição linear o processo apenas equaliza o módulo da resposta em freqüência. Porém, isto não é um grande problema já que estamos considerando sinais de voz.

Sem perda de generalidade, consideramos que as amostras passadas de todos os sinais $e_i(n)$ são usadas na predição do valor presente do sinal $e_1(n)$. Ou seja,

$$\varepsilon(n) = e_1(n) - \sum_{i=1}^{N}\sum_{k=1}^{M} f_{ik} e_i(n-k). \qquad (11)$$

Novamente, em uma situação offline, coeficientes $\mathbf{f}_i = [f_{i1} \;\; f_{i2} \;\; \cdots \;\; f_{iM}]^T$ são obtidos minimizando-se $E\{\varepsilon^2(n)\}$. O que conduz a

$$\mathbf{f} = \mathbf{R}_e^{-1}\mathbf{p}_e,$$

em que $\mathbf{f} = [\mathbf{f}_1^T \; \mathbf{f}_2^T \; \ldots \mathbf{f}_N^T]^T$, $\mathbf{R}_e = E\{\mathbf{e}(n)\mathbf{e}(n)^T\}$, $\mathbf{p}_e = E\{\mathbf{e}(n)e_1(n)\}$, com $\mathbf{e}(n) = [\mathbf{e}_1^T(n) \; \mathbf{e}_2^T(n) \; \ldots \mathbf{e}_N^T(n)]^T$ e $\mathbf{e}_i(n) = [e_i(n-1) \; e_i(n-2) \; \ldots e_i(n-M)]^T$. O processo pode ser visto na Figura 3, em que $\hat{e}_i(n)$ é a estimativa de $e_i(n)$. No procedimento *online* os filtros $F_1(z)$ e $F_2(z)$ são

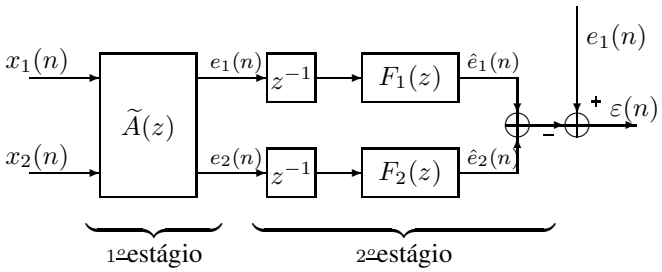

Fig. 3. Sistema de dereverberação em dois estágio de predição linear utilizando dois sinais reverberados $x_1(n)$ e $x_2(n)$

continuamente adaptados. A saída final é produzida filtrando-se $\varepsilon(n)$ por $1/\widetilde{A}(z)$, reinserindo as características da voz removidas no primeiro estágio.

IV. Simulações e Resultados

Na Figura 4 apresentamos uma comparação entre as respostas em freqüência associadas aos coeficientes de PL determinados pelos métodos discutidos na seção III-A, o método de Gaubitch e o proposto neste trabalho. Nas simulações usamos um conjunto de 6 microfones separados 15 cm um do outro para obter respostas impulsivas reais $h_i(n)$. Estas respostas são filtradas por um filtro de ordem 12 que modela o trato vocal, criando assim um conjunto de diferentes sinais reverberados. Ambos os métodos apresentam resultados similares, mas notamos a influência das respostas impulsivas da reverberação sobre as mesmas. Disto se conclui que embora $H_i(z)$ possua características de variações abruptas também possui algumas características de variações suaves influenciando desta forma a determinação de $\widetilde{A}(z)$. Isto depende do atraso entre a componente direta e as reflexões de cada $h_i(n)$: quanto maior

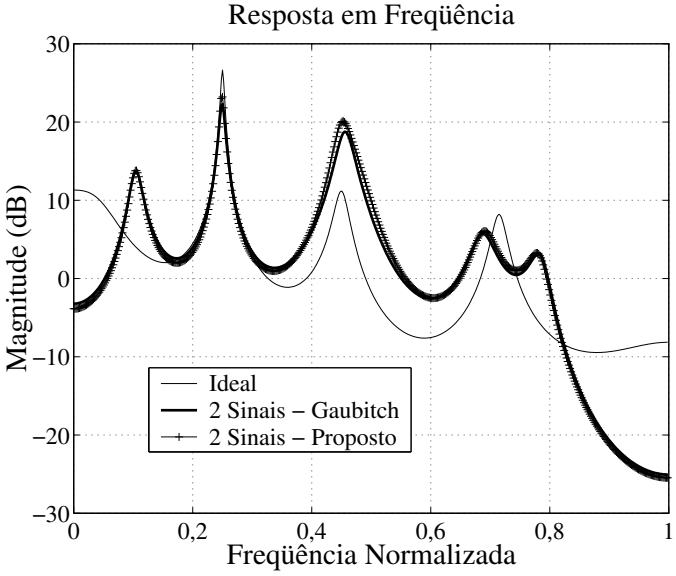

Fig. 4. Comparando as respostas em freqüência obtidas pelos método de Gaubitch e o proposto neste trabalho com o ideal ($G(z)$)

é este atraso, menor o efeito, pois desta forma reduz-se a influência entre o sinal e as reflexões.

Em uma segunda simulação, sinais de voz reverberados são produzidos filtrando-se uma amostra de voz por duas respostas impulsivas diferentes $h_1(n)$ e $h_2(n)$. Estas respostas foram medidas em um ambiente real e modificadas introduzindo-se o efeito de um longo atraso. Os microfones estavam separados de 15 cm e as respostas medidas possuiam um comprimento de 2000 amostras considerando uma freqüência de amostragem de 16 kHz. O algoritmo FTF [8] foi usado para a adaptação dos filtros devido a sua convergência rápida para sinais de voz. Os espectrogramas de um segmento de dois segundos do sinal de voz antes e depois da reverberação são mostrados na Figura 5. Como podemos ver a reverberação altera o espectrograma original. A Figura 6 apresenta os espectrogramas do sinal dereverberado: 1)usando 1 microfone e 7000 coeficientes na filtragem do segundo estágio; 2) usando 2 microfones e dois filtros com 2000 coeficientes cada no segundo estágio. Podemos ver que com 2 microfones, o número total de coeficientes é menor que 60% do número empregando 1 microfone, e a reverberação residual é ainda menor que o caso usando 1 microfone. Isto é verificado por uma análise subjetiva do sinal resultante e aponta vantagens no método proposto.

V. Conclusões

Neste trabalho apresentamos um método de dereverberação em estágios de PL empregando múltiplos microfones. O método, que é adequado para filtragem adaptativa, é baseado na suposição de que as respostas em freqüência associadas ao trato vocal e a reverberação possuem características diferentes. Aplicando o método proposto, o sinal de voz é melhorado analisando-o subjetivamente quando comparado com o caso utilizando um microfone.

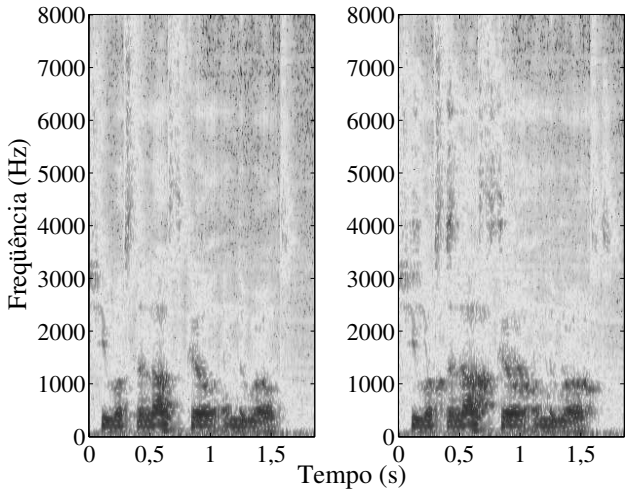

Fig. 5. Espectrograma do sinal de voz (esquerda) e o mesmo reverberado (direita).

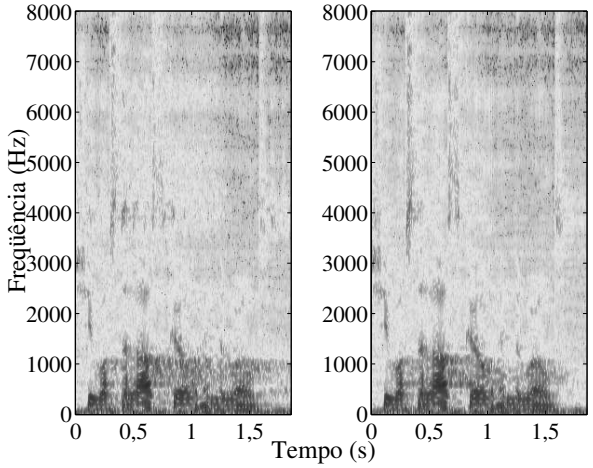

Fig. 6. Espectrograma do sinal de voz dereverberado por meio de 1 microfone (esquerda) e 2 microfones (direita).

AGRADECIMENTOS

Nossos agradecimentos a CAPES pelo apoio ao nosso trabalho.

REFERÊNCIAS

[1] M. Wu and D. Wang, "A one-microphone algorithm for reverberant speech enhancement," *Proceedings of ICASSP'03*, vol. 1, pp. 892–895, Abr. 2003.

[2] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proceedings of ICASSP'03*, vol. 1, pp. 92–95, Abr. 2003.

[3] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, Fev. 1988.

[4] J. G. Rodriguez, J. S. Bote, and J. O. Garcia, "Speech dereverberation and noise reduction with a combined microphone array approach," *Proceedings of ICASSP'00*, vol. 2, pp. 1037 – 1040, Jun. 2000.

[5] S. M. Griebel and M. Brandstein, "Microphone array speech dereverberation using coarse channel modeling," *Proceedings of ICASSP'01*, vol. 1, pp. 201–204, Mai. 2001.

[6] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proceedings of ICASSP'01*, vol. 6, pp. 3701 – 3704, Mai. 2001.

[7] N. D. Gaubitch, P. A. Nayloy, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," *Proceedings of IWAENC2003*, pp. 99–102, Set. 2003.

[8] S. Haykin, *Adaptive Filtering Theory*, 2nd ed. Prentice Hall, 1991.