Aperfeiçoamento de Algoritmo de Desreverberação Utilizando Medidas Perceptuais de Qualidade

Thiago de M. Prego, Amaro A. de Lima e Sergio L. Netto

***Resumo*— Este artigo apresenta um projeto otimizado para um algoritmo de desreverberação de sinais de fala. A otimização é calcada em aspectos perceptuais do efeito de reverberação utilizando a métrica de Allen. Resultados experimentais mostram que as propostas de modificações, quando comparadas com a estrutura original do algoritmo, alcançam um aumento de 7% considerando a métrica de Allen.**

***Palavras-Chave*— Reverberação, Avaliação de qualidade, Desreverberação, Subtração espectral, Filtragem inversa.**

***Abstract*— This paper describes an optimization procedure for designing a dereverberation algorithm. The algorithm tuning is performed based on perceptual concepts developed for the reverberation effect. Results indicate that the improved algorithm outperforms its original counterpart in about 7% considering the so-called Allen's reverberation score.**

***Keywords*— Reverberation, Quality assessment, Dereverberation, Spectral subtraction, Inverse filtering.**

I. INTRODUÇÃO

A reverberação pode afetar decisivamente o desempenho dos atuais sistemas de reconhecimento de fala/locutor ou mesmo dos sistemas de auxílio à reabilitação de deficientes auditivos. Isto motiva a utilização de técnicas apropriadas para reduzir os seus efeitos. Embora a reverberação possa ser bastante prejudicial, em pequena quantidade/intensidade ela pode até tornar a fala mais agradável para o ouvinte médio [1]. A utilização de arranjo de microfones é a configuração mais comumente usada em técnicas de desreverberação, porém para as aplicações anteriormente mencionadas a utilização de um único microfone é mais apropriada.

Neste artigo são sugeridas análises e propostas de modificações para o algoritmo de desreverberação de sinais de fala usando um único microfone como proposto em [2]. Esse algoritmo é dividido em dois blocos: o primeiro lida com os efeitos das primeiras reflexões e o segundo lida com os efeitos da reverberação tardia. A influência das primeiras reflexões é reduzida através de um processo adaptativo de filtragem inversa, que tem o objetivo de reconstruir o sinal de fala desejado. Já o efeito das componentes tardias de reverberação é mitigado usando subtração espectral, onde um modelo baseado na distribuição de Rayleigh é utilizado para emular o comportamento das componentes tardias.

No intuito de apresentar as modificações propostas, este trabalho é organizado da seguinte forma: Na Seção II, uma visão geral do algoritmo original de desreverberação de dois estágios é apresentada. A Seção III apresenta os detalhes do bloco de subtração espectral original, foco deste trabalho. A Seção IV descreve as modificações propostas. Na Seção V são descritas as bases de dados de 100 sinais cada, uma utilizada para treinamento e outra para teste, além dos resultados práticos considerando as alterações propostas na Seção IV. Finalmente, as conclusões referentes ao desempenho das modificações propostas estão na Seção VI.

Os autores estão com o Programa de Engenharia Elétrica, COPPE, Universidade Federal do Rio de Janeiro, Brasil. E-mails: {thprego, amaro, sergioln}@lps.ufrj.br

Amaro A. de Lima também está com o Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET-RJ), Brasil.

II. ALGORITMO DE 2 ESTÁGIOS

O algoritmo de dois estágios descrito em [2] consiste na utilização de blocos isolados de processamento de sinais com o intuito de reduzir o nível de reverberação do sinal de saída quando comparado com o sinal reverberante originalmente aplicado à entrada. Os dois blocos do algoritmo são chamados de filtragem inversa e subtração espectral, como mostrado na Fig. 1, onde $y(n)$, $z(n)$ e $x(n)$ são, respectivamente, os sinais de fala reverberante, inversamente filtrado e subtraído espectralmente (chamado neste trabalho de desreverberado).

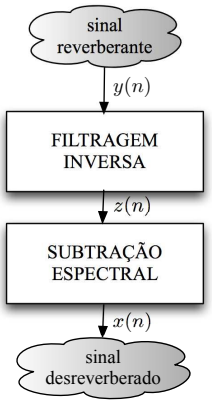

Fig. 1. *Diagrama de blocos do algoritmo de dois estágios de [2].*

Este algoritmo de Wu e Wang [2] lida com dois tipos de efeitos de reverberação presentes em sinais de fala reverberantes: coloração e reverberação de longa duração (*long-term reverberation*). A resposta ao impulso do ambiente (RIA) de um sinal reverberante pode ser modelada como a combinação dos seguintes componentes: do sinal de caminho direto entre a fonte e o ouvinte; das reflexões iniciais (*early reflections*), que possuem uma resposta em frequência não plana capaz de distorcer o espectro da fala; e da reverberação tardia (*late reverberation*), que causa distorção no espectro da fala, reduzindo desta forma a inteligibilidade e qualidade do sinal [2], [3].

O algoritmo em análise foi projetado para mitigar os efeitos devido às reflexões inicias e tardias, que estão respectivamente ligadas aos efeitos de coloração e reverberação de longa duração. Devido às propostas de modificações deste trabalho serem exclusivamente referentes ao bloco de subtração espectral, esta etapa será descrita em detalhes no próxima seção.

III. SUBTRAÇÃO ESPECTRAL

A Fig. 2 detalha os componentes do bloco de subtração espectral, onde se considera que as reflexões inicias e tardias são aproximadamente descorrelacionadas [2]. Este bloco utiliza como entrada o sinal de fala inversamente filtrado $z(n)$ e tem como saída o sinal de fala desreverberado $x(n)$. É importante ressaltar que a fase do sinal inversamente filtrado é usada para gerar o sinal desreverberado.

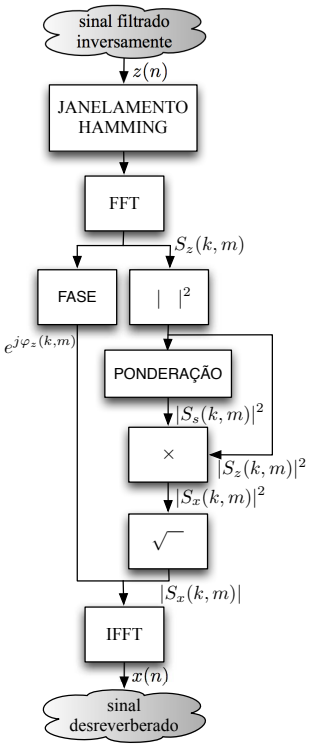

Fig. 2. *Diagrama de blocos da etapa de subtração espectral.*

A subtração espectral tem por objetivo reduzir o efeito da reverberação de longa duração causado pela componente de reverberação tardia da RIA. Inicialmente o sinal inversamente filtrado $z(n)$ é divido em blocos usando uma janela de Hamming de 32 ms com 24 ms de superposição entre blocos consecutivos.

Seja $S_z(k,m) = |S_z(k,m)|e^{j\varphi_z(k,m)}$ a transformada de Fourier em tempo curto (STFT - *Short Time Fourier Transform*) do sinal de fala inversamente filtrado $z(n)$, onde $k \in \mathbb{N}$ é o índice do *bin* da STFT e $m \in \mathbb{N}$ é o índice do bloco. Seja ainda $w(m)$ uma janela de atenuação que atende a distribuição de Rayleigh, dada por

$$\begin{cases} w(m) = \left(\frac{m+a}{a^2}\right) e^{\left(\frac{-(m+a)^2}{2a^2}\right)}, & \text{se } m > -a \\ w(m) = 0, & \text{caso contrário} \end{cases}, \quad (1)$$

onde a controla o espalhamento total da função.

Se ρ é o comprimento das primeiras reflexões e γ o fator de escala que estabelece a energia relativa da componente tardia depois da filtragem inversa, o espectro de potência da reverberação tardia é modelado por

$$|S_l(k,m)|^2 = \gamma w(i-\rho) * |S_z(k,m)|^2, \quad (2)$$

onde, o símbolo "$*$" representa a operação de convolução linear. Este modelo, apresentado na eq. (2), foi baseado no efeito de distorção das componentes tardias, que causam atenuação no espectro de sinal, levando a um modelo do espectro de potência das componentes tardias que é uma versão atenuada e deslocada no tempo do espectro de potência do sinal de fala inversamente filtrado.

Considerando que as componentes primeiras e tardias são descorrelacionadas, o espectro de potência das primeiras componentes pode ser estimado pela subtração do espectro de potência das componentes tardias do sinal inversamente filtrado. O módulo de subtração espectral faz uma espécie de ponderação no espectro de potência de $z(n)$, onde o bloco de PONDERAÇÃO da Fig. 2 é dado por

$$|S_s(k,m)|^2 = \max\left[1 - \frac{\gamma w(i-\rho) * |S_z(k,m)|^2}{|S_z(k,m)|^2}, \epsilon\right], \quad (3)$$

$\epsilon = 0.001$ correspondendo à máxima atenuação de 30 dB e finalmente o espectro de potência de $x(n)$ sendo

$$|S_x(k,m)|^2 = |S_z(k,m)|^2 \times |S_s(k,m)|^2. \quad (4)$$

No intuito de se calcular $x(n)$, a informação de fase obtida de $S_z(k,m)$ é combinada com o módulo de $S_x(k,m)$:

$$S_x(k,m) = |S_x(k,m)|e^{j\varphi_z(k,m)} \quad (5)$$

Os valores originais das constantes utilizadas em [2] são $\rho = 7$, $\gamma = 0{,}35$ e $a = 5$.

IV. MODIFICAÇÕES PROPOSTAS

Esta seção trata das propostas de modificações para gerar o sinal de fala desreverberado com maior qualidade perceptual. No intuito de se avaliar a qualidade do sinal de fala reverberante, a medida de qualidade conhecida como métrica de Allen [4], [5] é utilizada:

$$P = P_{max} - \sigma^2 T_{60}, \quad (6)$$

onde P_{max} é a nota máxima, σ^2 é a variância espectral da sala definida em [6] e T_{60} é o tempo de reverberação que é definido como sendo o período de tempo onde a pressão sonora cai em 60 dB. Na prática, um maior T_{60} indica um efeito de reverberação mais duradouro e a técnica usada para estimá-lo é o algoritmo de Karjalainen et al. [7]. Essas duas variáveis são obtidas diretamente da RIA, $h(n)$, que é estimada através da deconvolução entre o sinal de fala limpo e o degradado.

Este trabalho analisa três possibilidades de modificações do algoritmo de [2], todas no bloco de subtração espectral:

1) Otimização do fator de atenuação γ;
2) Otimização do limiar ϵ; e
3) Otimização conjunta dos parâmetros ρ e a usados na janela de atenuação.

Todas as propostas de modificações estão ligadas à Eq. (3), sendo que a última está também relacionada com a distribuição de Rayleigh definida em (1).

A otimização do fator de atenuação γ pode proporcionar uma melhor relação entre a energia das primeiras e das componentes tardias, proporcionando uma maior redução perceptual dos efeitos da reverberação tardia.

A otimização do limiar ϵ deve afetar a representação espectral do sinal desreverberado, fazendo com que uma representação não apropriada influencie fortemente na qualidade perceptual do sinal processado.

A otimização de ρ e a, que são o tamanho em blocos das primeiras reflexões e o fator de espalhamento da distribuição de Rayleigh, respectivamente, também afeta a Eq. (3). Os valores de a são dependentes de ρ uma vez que a não pode ser maior que ρ. A escolha apropriada desses parâmetros vai estabelecer o instante de tempo certo para a aplicação da atenuação das componentes tardias e o formato da janela de atenuação dessas componentes.

Uma análise detalhada e resultados experimentais das modificações propostas será apresentada na Seção V, onde será utilizada uma base de treinamento para a etapa de otimização e uma base de teste para a validação dos resultados. Estas bases são originárias de um única base que será descrita na seção seguinte.

V. Resultados experimentais

No intuito de avaliar as alterações propostas no contexto de avaliação de qualidade dos sinais de fala reverberantes é recomendado o uso de diferentes efeitos e níveis de reverberação. Consequentemente as bases de dados empregadas tanto no teste quanto no treinamento são essenciais para encaminhar a pesquisa para conclusões apropriadas. A base de dados utilizada neste trabalho foi criada pelos próprios autores.

A. Base de dados

A base de dados usada neste trabalho é chamada de Nova base de dados para o Português-Brasileiro (NPB) e inclui 200 sinais com $F_s = 48$ kHz, e diferentes tipos e intensidades de reverberação. A base completa foi gerada de 4 sinais de fala anecóicos (2 de locutores masculinos e 2 femininos) usando três enfoques distintos de reverberação:

- Efeito de reverberação artificial: Ele corresponde a 6 RIAs artificialmente geradas, onde as primeiras reflexões foram modeladas pelo método das imagens [8], com uma distância fonte-microfone fixa de $d = 1{,}8$ m numa sala virtual de dimensões comprimento×largura×altura = 4 m × 3 m × 3 m, e no que diz respeito a reverberação tardia, o método das redes de realimentação de atraso [9] (*feedback delay networks*) foi usado para emular tempos de reverberação na faixa de $T_{60} = \{200, 300, 400\}$ ms e uma versão modificada do método de Gardner [10], [11], que originariamente foi projetado para emular tempos de reverberação acima de 400 ms, foi usado para $T_{60} = \{500, 600, 700\}$ ms. O tempo de reverberação médio medido foi $\{196, 292, 387, 469, 574, 664\}$ ms.
- Efeito de reverberação natural: Suas RIAs foram obtidas através da gravação direta de 4 tipos de salas (cabine, sala de reuniões, escritório e sala de aula) com várias distâncias fonte-microfone para cada sala como detalhado em [12]. As 4 salas possuem diferentes dimensões e distâncias fonte-microfone fazendo um total de 17 diferentes RIAs. O tempo de reverberação médio medido para as 4 salas foram $\{120, 230, 430, 780\}$ ms.
- Efeito de reverberação real: É o único efeito onde os sinais degradados foram diretamente tocados/gravados nas salas, sem a utilização da operação de convolução entre a RIA e os sinais anecóicos. As 7 salas (cabine, escritório1, sala de aula1, sala de reuniões1, escritório2, sala de aula2 e sala de reuniões2) usadas nas gravações possuíam diferentes dimensões e empregaram 4 diferentes distância fonte-microfone, $\{1, 2, 3, 4\}$ m, exceto para a menor sala (cabine), onde somente foram empregadas 3 distâncias, $\{0{,}5, 1, 1{,}5\}$ m, emulando um total de 27 RIAs com tempo de reverberação médio na faixa de $\{140, 390, 570, 650, 700, 890, 920\}$ ms.

A base NPB foi dividida em duas bases menores de 100 sinais cada, sendo uma utilizada na otimização dos parâmetros, ou seja, na etapa de treinamento, e a outra utilizada na validação dos resultados, ou seja, na etapa de teste. A qualidade de cada sinal da base foi avaliada por 30 ouvintes utilizando uma escala de notas de 1 a 5, onde 1 representa a pior e 5 a melhor nota/qualidade. Devido a isso, a média das notas de cada sinal foi utilizada para ordená-los de forma que a base total fosse dividida em dois grupo com níveis equivalentes de reverberação. É importante ressaltar que neste trabalho os 4 sinais anecóicos **não** estão incluídos nos 200 sinais da NBP.

A métrica de Allen apresentada em Eq. (6) é usada para avaliar a qualidade do sinal desreverberado utilizando $P_{max} = 0$, que limita a máxima nota de Allen em $P = 0$, ou seja, quanto mais próximo de 0 menor é a reverberação percebida e consequentemente maior é a qualidade do sinal em análise.

B. Otimização conjunta do fator de atenuação γ e do limiar ϵ

Inicialmente foram realizadas buscas individuais dos valores ótimos de γ e ϵ para, depois, realizar uma otimização conjunta ao redor dos valores ótimos inicialmente encontrados.

A otimização inicial de γ da Eq. (3) foi realizada fixando todos os outros parâmetros em seus valores originais. No caso dos parâmetros em análise neste trabalho os valores foram $\epsilon = 10^{-3}$, $\rho = 7$ e $a = 5$. O parâmetro γ foi variado no intervalo $[0{,}1; 1]$ e a média P_m das notas P para os 100 sinais da base de treinamento foi computada para cada valor de γ. O melhor desempenho obtido foi $P_m = -2{,}17$ para $\gamma = 0{,}2$.

Na otimização inicial de ϵ, seus valores foram variados no interval $[10^{-9}; 10^{0}]$, fixando os parâmetros $\gamma = 0{,}35$, $\rho = 7$ e $a = 5$. Neste caso, o melhor resultado $P_m = -2{,}25$ foi alcançado com $\epsilon = 10^{-2}$.

Uma vez obtidos os melhores valores para os parâmetros γ e ϵ, foi realizada uma otimização conjunta numa faixa ao redor de $\gamma = 0{,}2$ e $\epsilon = 10^{-2}$, fixando $\rho = 7$ e $a = 5$. O resultado desta otimização conjunta é apresentado na Fig. 3, onde a

melhor combinação de parâmetros resultou em $\epsilon = 10^{-2}$ e $\gamma = 0{,}15$, acarretando $P_m = -2{,}15$.

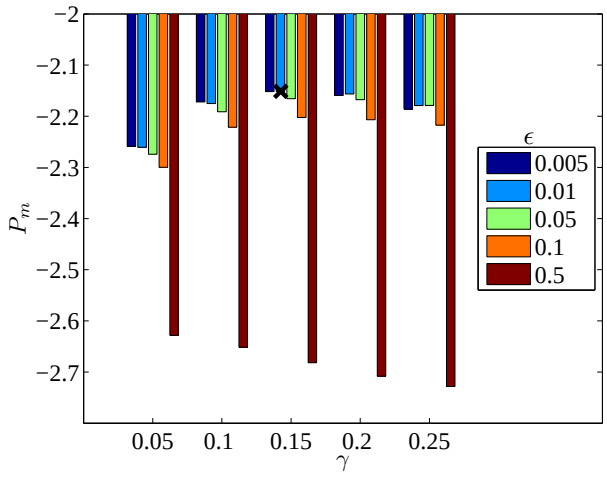

Fig. 3. *Performance de qualidade, P_m, da base de treinamento em volta dos valores ótimos γ e ϵ obtidos na Seção V-B.*

C. Otimização conjunta do atraso ρ e do fator de espalhamento a

A otimização do atraso ρ, que representa o tamanho em blocos das primeiras reflexões e do fator de espalhamento a da distribuição de Rayleigh é realizada de forma conjunta devido ao fato de existir uma certa dependência entre elas, como mencionado anteriormente na Seção III.

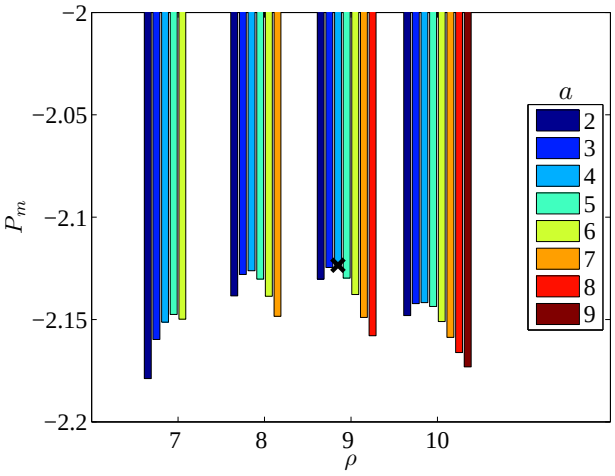

Fig. 4. *Performance de qualidade, P_m, da base de treinamento para o tamanho em blocos das primeiras reflexões, ρ, e para o fator de espalhamento, a, da distribuição de Rayleigh usados na Eq. (1).*

A Fig. 4 mostra a qualidade média dos sinais da base treinamento fixando os parâmetros $\gamma = 0{,}15$ e $\epsilon = 10^{-2}$ e variando $7 \leq \rho \leq 10$ e $2 \leq a \leq (\rho - 1)$. Embora a análise tenha sido realizada variando $5 \leq \rho \leq 14$ e $2 \leq a \leq (\rho - 1)$, a figura só representa uma pequena faixa desses valores para facilitar a visualização dos resultados. O melhor desempenho obtido foi $P_m = -2{,}12$, utilizando $\rho = 9$ e $a = 4$, o que gerou uma melhoria de 7% na qualidade percebida dos sinais quando comparada com a configuração original.

D. Validação dos resultados

Nesta seção a análise é realizada usando a base de teste para avaliar a qualidade percebida das versões original ($\gamma = 0{,}35$, $\epsilon = 10^{-3}$, $\rho = 7$ e $a = 5$) e modificada ($\gamma = 0{,}15$, $\epsilon = 10^{-2}$, $\rho = 9$ e $a = 4$) do algoritmo de desreverberação de dois estágios. O objetivo é utilizar a base de teste para validar os valores ótimos dos parâmetros obtidos utilizando a base de treinamento. Os resultados usando a base de treinamento também são apresentados no intuito de se confirmar a consistência dos resultados de teste.

TABELA I

Desempenho médio das medidas de avaliação de qualidade (métrica de Allen (P) e T_{60}) para a base de treinamento usando as versões original e modificada do algoritmo de [2].

Base de Treinamento			
Medidas de qualidade	Base de dados não processada	Algoritmo de 2 estágios	
		Original	Modificado
P_m	-2.93	-2.28	**-2.12**
T_{60}^m [ms]	509	327	**301**

As Tabelas I e II mostram os valores médios da métrica de Allen P_m e do tempo de reverberação T_{60}^m para as bases de treinamento e teste, respectivamente, usando as duas versões do algoritmo e os próprios sinais das bases não processados, ou seja, sem a realização de qualquer procedimento de desreverberação.

TABELA II

Desempenho médio das medidas de avaliação de qualidade (métrica de Allen (P) e T_{60}) para a base de teste usando as versões original e modificada do algoritmo de [2].

Base de Teste			
Medidas de qualidade	Base de dados não processada	Algoritmo de 2 estágios	
		Original	Modificado
P_m	-2.96	-2.34	**-2.17**
T_{60}^m [ms]	525	343	**312**

Comparando as melhorias na qualidade percebida para as bases de treinamento e teste podemos notar que os resultados são consistentes. Observando as duas versões do algoritmo de dois estágios, original e modificada, e comparando-as com os sinais não processados, as melhorias são de 22% e 28% para a métrica de Allen e de 36% e 41% para o T_{60}^m considerando somente a base de treinamento. Já observando somente a base de teste obtemos as melhorias de 21% e 27% para a medida de Allen e 35% e 41% para o T_{60}^m, respectivamente, para as versões original e modificada do algoritmo. Os percentuais de melhoria estão muito próximos mostrando uma adequação e consistência nos resultados. Tanto para o treinamento quanto para o teste a melhoria entre as versões originais e modificadas ficaram na faixa de 7% para P_m. Quanto ao T_{60}^m as melhorias foram de 8% e 9%, respectivamente, para o treinamento e

teste. Aplicando o t-teste nesses resultados do treinamento concluímos que as médias são diferentes com aproximadamente 97% e 99% de confiança para P e T_{60}. Já na base de teste as conclusões e valores são os mesmos, exceto pelo valor da confiança de 96% para P.

VI. Conclusões

Este trabalho analisa a influência de diversos parâmetros do algoritmo de desreverberação de dois estágios para sinais de fala com relação a qualidade percebida dos sinais desreverberados. Três propostas de modificações nos ajustes dos parâmetros usados no algoritmo foram analisadas gerando uma maior qualidade percebida do sinal desreverberado. As propostas consistiram na análise do fator de atenuação γ usado na subtração espectral, na escolha apropriada do limiar ϵ que evita o processamento de amostras com baixa energia espectral e na otimização conjunta do tamanho da primeiras reflexões ρ e do fator de espalhamento a da distribuição de Rayleigh. O treinamento e teste foram realizados utilizando-se duas bases de dados cada uma com 100 sinais de fala reverberantes, que levaram a melhorias na qualidade percebida na ordem de 7% e 8% na média da nota perceptual de Allen P_m e do T_{60}^m, respectivamente.

Agradecimentos

Os autores gostariam de agradecer ao Prof. M. Karjalainen, por fornecer o algoritmo de estimação do T_{60}; e ao Prof. D. Wang, por fornecer o algoritmo de dois estágios para sinais de fala reverberantes.

Referências

[1] R. Appel and J. Beerends, "On the Quality of Hearing One's Own Voice," *J. Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, April 2002.

[2] M. Wu and D. Wang, "A Two-Stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, May 2006.

[3] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," *Proc. IEEE Int. Workshop Acoustic Echo and Noise Control*, Paris, France, Sept. 2006.

[4] J. B. Allen, "Effects of small room reverberation on subjective preference," *J. Acoustic. Soc. Am.*, vol. 71, Apr. 1982.

[5] D. A. Berkley and J. B. Allen, "Normal Listening in Typical Rooms: The Physical and Psychophysical Correlates of Reverberation," *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., G.A. Studebaker and I. Hochberg eds., Allyn and Bacon, 1993.

[6] J. J. Jetz, "Critical distance measurement of rooms from the sound energy spectral response," *J. Acoustic. Soc. Am.*, vol. 65, pp. 1204–1211, May 1979.

[7] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy reponse measurements," *Proc. Conv. Audio Engineering Society*, Amsterdam, Netherlands, pp. 867–878, May 2001.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustic. Soc. Am.* vol. 65, no. 4, pp. 943–950, Apr. 1979.

[9] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," *Proc. 90th Conv. Am. Engineering Soc.*, Preprint 3030, Feb. 1991.

[10] W. G. Gardner, *Reverberation Algorithms*, in *Applications of Digital Signal Processing*, Ed. Mark Kahrs and Karl-Heinz Brandenburg, Kluwer, New York:NY, pp. 85–131, Mar. 1998.

[11] A. A. de Lima, F. P. Freeland, P. A. A. Esquef, L. W. P. Biscainho, B. C. Bispo, R. A. de Jesus, S. L. Netto, R. Schafer, A. Said, B. Lee, and A. Kalker, "Reverberation assessment in audioband speech signals for telepresence systems," *Proc. Int. Conf. Signal Processing in Multimedia Applications*, Porto, Portugal, pp. 257–262, July 2008.

[12] M. Jeub, M. Schäfer, and P. Vary, "A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms," *Proc. 16th Int. Conf. on Digital Signal Processing*, Santorini, Greece, 2009.