MINTFORMER: A SPATIALLY AWARE CHANNEL EQUALIZER

Felicia Lim[1] *Mark R. P. Thomas*[2*] *Patrick A. Naylor*[1]

[1] Dept. of Electrical and Electronic Engineering, Imperial College London, UK,
{felicia.lim06, p.naylor}@imperial.ac.uk
[2] Microsoft Research, One Microsoft Way, Redmond WA 98052, USA, markth@microsoft.com

ABSTRACT

Reverberation is a process that distorts a wanted signal and impairs perceived speech quality. In the context of multichannel dereverberation, channel-based methods and beamforming are two common approaches. Channel-based methods such as the multiple input/output inverse theorem (MINT) can provide perfect dereverberation provided the exact acoustic impulse responses (AIRs) are known. However, they have been shown to be very sensitive to AIR estimation errors for which several modifications have consequently been proposed. Conversely, beamformers are significantly more robust but provide comparatively modest dereverberation. While the two approaches are conventionally considered independent, both can be formulated as a filter-and-sum operation with differing filter design criteria. We propose a unified framework, termed MINTForming, that exploits this similarity and introduces a mixing parameter to control the tradeoff between the potential performance of MINT and the robustness of beamforming. Empirical results show that the mixing parameter is a monotonic function of channel estimation error, whereby a MINT solution is preferred when channel estimation error is low.

Index Terms— channel equalization, dereverberation, beamformer, channel estimation errors

1. INTRODUCTION

A speech signal captured by a distant microphone is affected by reverberation due to the propagation of clean speech through a multipath acoustic channel. This contributes to the degradation of the signal, affecting the perceived quality of speech and reducing the performance of other speech devices such as speech recognizers, voice-controlled systems and hearing aids [1, 2].

One approach to dereverberation is acoustic channel equalization, where inverse systems are designed based on estimates of acoustic impulse responses (AIRs) between the speech source and the microphones to equalize the effect of the AIRs on the clean speech. In the multichannel case with M microphones, the multiple input/output inverse theorem (MINT) [3] provides an exact inverse system that achieves perfect dereverberation provided there are no common zeros between the AIRs and the length of the equalizing filter satisfies some criteria [3, 4]. However, its performance rapidly degrades in the presence of noise and AIR estimation errors, resulting in the addition of artificial reverberation in the equalized impulse response.

An alternative approach to dereverberation uses beamformers [5] that only require an estimation of the source signal's direction of arrival. Steering weights are designed to point the beamformer's look-direction towards the speech source while signals from all other directions are treated as noise and therefore attenuated. This approach is robust, particularly when additional constraints are placed upon the white noise gain, for example as in [6]. However, the performance of the beamformer is limited as reverberant signals will inevitably fall into the main lobe of the beamformer [1]. Further, the directivity factor averaged over all look directions is limited by the number of microphones [7].

Given a multichannel reverberant observation and the two different approaches described above, the motivation in this paper is to design a dereverberation algorithm which exploits both channel and spatial knowledge simultaneously. We demonstrate that the problems addressed by channel equalization and beamforming are very similar and establish a link between the solutions for MINT and a filter-and-sum beamformer (FSB). An independent variable is introduced as a mixer between the two solutions based on the expected level of AIR estimation errors to give the proposed MINTFormer solution.

The remainder of the paper is organized as follows. In Section 2, the problem is formulated. Section 3 describes the MINT solution, followed by the FSB solution in Section 4. The proposed MINTFormer framework is presented in Section 5 followed by simulation results in Section 6 and some conclusions in Section 7.

2. PROBLEM FORMULATION

An M-channel acoustic system can be modeled as M finite impulse responses, $\mathbf{h}_m = [h_m(0)\ h_m(1)\ \dots\ h_m(L-1)]^T$ for $m = 1, 2, \dots, M$. A speech signal $s(n)$ propagating through this acoustic system can be described at the m-th microphone as

$$x_m(n) = s(n) * h_m(n). \tag{1}$$

Given a set of equalizing filters $\mathbf{g}_m = [g_m(0)\ g_m(1)\ \dots\ g_m(L_i - 1)]^T$, the equalized impulse response is given as

$$r(n) = \sum_{m=1}^{M} h_m(n) * g_m(n), \tag{2}$$

and the equalized speech is given as

$$\hat{s}(n) = \sum_{m=1}^{M} x_m(n) * g_m(n). \tag{3}$$

The aim of dereverberation algorithms is to design $\mathbf{g}_m$ to recover the clean speech signal such that in the ideal scenario, $\hat{s}(n) = s(n)$ is obtained.

*This work is based on research started as a postdoctoral researcher at Imperial College London, UK.

3. MINT

The following reviews the MINT algorithm [3] in the time domain and reformulates it in the frequency domain with an iterative least squares solution.

In the time domain, channel equalization can achieve perfect dereverberation by designing a set of equalizing filters $\mathbf{g}_m$ such that

$$\sum_{m=1}^{M} \mathbf{H}_m \mathbf{g}_m = \mathbf{d}, \tag{4}$$

where $\mathbf{H}_m$ is the $(L + L_i - 1) \times L_i$ convolution matrix of $\mathbf{h}_m$, $\mathbf{d} = [\mathbf{0}_{1\times\tau'}, 1, \mathbf{0}_{1\times(L+L_i-\tau'-2)}]^T$ and τ' is the target delay of the equalized channel. Exact multichannel inverse filters $\mathbf{g} = [\mathbf{g}_1^T, \mathbf{g}_2^T, \ldots, \mathbf{g}_M^T]^T$ are provided by MINT [3], subject to there being no common zeros between the AIRs and $L_i \geq \lceil (L-1)/(M-1) \rceil$, where $\lceil \cdot \rceil$ is the ceiling operator [3, 4]. The solution is obtained as

$$\mathbf{g} = \mathbf{H}^{+}\mathbf{d}, \tag{5}$$

where $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_M]$ and $\{\cdot\}^+$ denotes the Moore-Penrose pseudo-inverse.

In the frequency domain, the K-point Discrete Fourier Transform (DFT) of $\mathbf{h}_m$ is given by

$$\tilde{H}_m(k) = \sum_{n=0}^{K-1} h_m(n) e^{-j2\pi nk/K}, \tag{6}$$

for $k = 0, \ldots, K-1$. The stacked multichannel DFT coefficients of the AIRs and equalizing filters can then be defined respectively as $\tilde{\mathbf{h}}(k) = [\tilde{H}_1(k), \ldots, \tilde{H}_M(k)]^T$ and $\tilde{\mathbf{g}}(k) = [\tilde{G}_1(k), \ldots, \tilde{G}_M(k)]^T$, where $\tilde{G}_m(k)$ is found in a similar manner to $\tilde{H}_m(k)$. To achieve linear convolution, $K \geq L + L_i - 1$ and therefore, $\mathbf{h}_m$ and $\mathbf{g}_m$ are zero-padded to length K before their DFTs are calculated. Equation (4) can therefore be reformulated in the frequency domain for the k-th DFT coefficient as

$$\sum_{m=1}^{M} \tilde{H}_m(k)\tilde{G}_m(k) = \tilde{\mathbf{h}}^T(k)\tilde{\mathbf{g}}(k) = 1. \tag{7}$$

Computation of $\tilde{\mathbf{g}}(k)$ by direct inverse is of limited practical use as AIRs are often non-minimum phase [8], resulting in non-causal inverse filters. These filters introduce distortion in the equalized signal usually referred to as pre-echo [1]. Instead, the least squares solution for $\tilde{\mathbf{g}}(k)$ can be found adaptively using the method of steepest descent [9] given the cost function for the i-th iteration and k-th DFT coefficient

$$\mathbf{J}_{\text{MINT}}(k, i) = \|\tilde{\mathbf{h}}^T(k)\tilde{\mathbf{g}}(k, i) - 1\|_2^2. \tag{8}$$

The partial derivative of $\mathbf{J}_{\text{MINT}}(i)$ is obtained as

$$\nabla \mathbf{J}_{\text{MINT}}(k, i) = \frac{\partial \mathbf{J}_{\text{MINT}}(k, i)}{\partial \tilde{\mathbf{g}}(k, i)} = \left(\tilde{\mathbf{h}}^H(k)\tilde{\mathbf{g}}^*(k, i) - 1\right)\tilde{\mathbf{h}}(k), \tag{9}$$

where $\{\cdot\}^H$ denotes the Hermitian transpose. The update equation for the complex conjugate of the equalizing filters is then given by

$$\tilde{\mathbf{g}}^*(k, i+1) = \tilde{\mathbf{g}}^*(k, i) - \mu \nabla \mathbf{J}_{\text{MINT}}(k, i), \tag{10}$$

where μ is the step-size and $\{\cdot\}^*$ denotes complex conjugate.

4. FILTER-AND-SUM BEAMFORMER (FSB)

A FSB consisting of M sensors can be characterized by its response to a complex plane wave with a given angle of arrival and frequency. For $k = 0, \ldots, K-1$ discrete frequencies $\omega(k) = 2\pi f_s k/K$, where f_s is the sampling frequency, and $q = 0, \ldots, Q-1$ discrete angles $\theta_q = 2\pi q/(Q-1)$, the beampattern is given as [10]

$$\tilde{Y}(k, \theta_q) = \sum_{m=1}^{M} \tilde{G}_m(k) e^{-j\omega(k)\tau_m(\theta_q)}, \tag{11}$$

where $\tilde{G}_m(k)$ is a beamshaping filter, $\tau_m(\theta_q) = d_m \cos\theta_q / c$, d_m is the distance between the m-th and first sensor for a uniform linear array (ULA), and c is the speed of sound.

For dereverberation, it is desirable to design $\tilde{G}_m(k)$ to approximate the desired beampattern for a given direction of source θ_d,

$$\tilde{Y}_d(k, \theta_q) = \begin{cases} 1 & \text{if } q = d; \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

This leaves the signal from the source direction undistorted while attenuating reverberant reflections from other directions. Using (11) in matrix notation, the minimization problem can be stated as

$$\arg\min_{\tilde{\mathbf{g}}(k)} \|\tilde{\mathbf{B}}(k)\tilde{\mathbf{g}}(k) - \tilde{\mathbf{y}}_d(k)\|_2^2 \tag{13}$$

where $[\tilde{\mathbf{B}}(k)]_{qm} = e^{-j\omega(k)\tau_m(\theta_q)}$, $\tilde{\mathbf{g}}(k) = [\tilde{G}_0(k), \ldots, \tilde{G}_{M-1}(k)]^T$, and $\tilde{\mathbf{y}}_d(k) = [\tilde{Y}_d(k, \theta_0), \ldots \tilde{Y}_d(k, \theta_{Q-1})]^T$. A larger weighting can be given to the desired response in the source direction θ_d with an additional constraint using the steering vector $\mathbf{a}(k) = [e^{-j\omega(k)\tau_1(\theta_d)}, \ldots, e^{-j\omega(k)\tau_M(\theta_d)}]^T$,

$$\tilde{\mathbf{g}}^T(k)\mathbf{a}(k) = 1. \tag{14}$$

As with MINT, the least squares solution can be found adaptively using the method of steepest descent. First, the Lagrangian cost function for (13) subject to (14) is defined for iteration i as

$$\mathbf{J}_{\text{FSB}}(k, i) = \|\tilde{\mathbf{B}}(k)\tilde{\mathbf{g}}(k, i) - \tilde{\mathbf{y}}_d(k)\|_2^2 + \lambda(\tilde{\mathbf{g}}^T(k, i)\mathbf{a}(k) - 1), \tag{15}$$

where λ is a Lagrange multiplier. The partial derivative of $\mathbf{J}_{\text{FSB}}(k, i)$ is obtained as

$$\nabla \mathbf{J}_{\text{FSB}}(k, i) = \frac{\partial \mathbf{J}_{\text{FSB}}(k, i)}{\partial \tilde{\mathbf{g}}(k, i)} = \nabla \mathbf{J}_{\text{FSB, unc}}(k, i) + \lambda \mathbf{a}(k), \tag{16}$$

where

$$\nabla \mathbf{J}_{\text{FSB, unc}}(k, i) = \tilde{\mathbf{B}}^H(k)\tilde{\mathbf{B}}(k)\tilde{\mathbf{g}}^*(k, i) - \tilde{\mathbf{B}}^H(k)\tilde{\mathbf{y}}_d(k). \tag{17}$$

The update equation for the complex conjugate of the beamshaping filters is then given as

$$\tilde{\mathbf{g}}^*(k, i+1) = \tilde{\mathbf{g}}^*(k, i) - \mu \nabla \mathbf{J}_{\text{FSB}}(k, i). \tag{18}$$

Recalling the constraint on $\tilde{\mathbf{g}}(k, i+1)$ from (14) gives

$$\begin{aligned} 1 &= \mathbf{a}^H(k)\tilde{\mathbf{g}}^*(k, i+1) \\ &= \mathbf{a}^H(k)\tilde{\mathbf{g}}^*(k, i) - \mu \mathbf{a}^H(k)\nabla \mathbf{J}_{\text{FSB}}(k, i). \end{aligned} \tag{19}$$

Solving for λ and rearranging the terms yields the iterative update

$$\tilde{\mathbf{g}}^*(k, i+1) = \mathbf{A}(k)\tilde{\mathbf{g}}^*(k, i) - \mu \mathbf{A}(k)\nabla \mathbf{J}_{\text{FSB,unc}}(k, i) + \mathbf{f}(k), \tag{20}$$

where

$$\mathbf{A}(k) \triangleq \mathbf{I} - \frac{\mathbf{a}(k)\mathbf{a}^H(k)}{\mathbf{a}^H(k)\mathbf{a}(k)} \quad \text{and} \quad \mathbf{f}(k) \triangleq \frac{\mathbf{a}(k)}{\mathbf{a}^H(k)\mathbf{a}(k)}. \tag{21}$$

Intuitively, $\mathbf{A}(k)$ is the projection matrix to the null space of $\mathbf{a}(k)$ and $\mathbf{f}(k)$ normalizes the signal energy contributed by $\mathbf{a}(k)$ in (14).

5. MINTFORMER

The MINT and FSB problems in (8) and (15) are clearly very similar, with adaptive least squares solutions (10) and (20) of a similar structure. The proposed MINTFormer aims to establish a link between both solutions so as to find a solution that exhibits the potential performance of MINT in achieving perfect dereverberation and the robustness of the beamformer. An independent new variable γ can be introduced to combine MINT and FSB as

$$\mathbf{J}_{\text{MF}}(k,i) = \gamma\mathbf{J}_{\text{FSB}}(k,i) + (1-\gamma)\mathbf{J}_{\text{MINT}}(k,i). \tag{22}$$

A unified framework for MINTFormer requires both MINT and FSB to have identical structures for the solution. To this end, the MINT solution given in (10) is first modified as follows.

$$\begin{aligned}\tilde{\mathbf{g}}^*(k,i+1) = \mathbf{A}_{\text{MINT}}\tilde{\mathbf{g}}^*(k,i) \\ - \mu\mathbf{A}_{\text{MINT}}\nabla\mathbf{J}_{\text{MINT}}(k,i) \\ + \mathbf{f}_{\text{MINT}},\end{aligned} \tag{23}$$

where $\mathbf{A}_{\text{MINT}} = \mathbf{I}_{[M\times M]}$ and $\mathbf{f}_{\text{MINT}} = \mathbf{0}_{[M\times 1]}$. The solutions (20) and (23) can now be combined using γ as

$$\begin{aligned}\tilde{\mathbf{g}}^*(k,i+1) = [\gamma\mathbf{A}(k) + (1-\gamma)\mathbf{A}_{\text{MINT}}]\,\tilde{\mathbf{g}}^*(k,i) \\ - \gamma\mu\mathbf{A}(k)\nabla\mathbf{J}_{\text{FSB,unc}}(k,i) \\ - (1-\gamma)\mu\mathbf{A}_{\text{MINT}}\nabla\mathbf{J}_{\text{MINT}}(k,i) \\ + \gamma\mathbf{f}(k) + (1-\gamma)\mathbf{f}_{\text{MINT}}.\end{aligned} \tag{24}$$

Setting $\gamma = 0$ yields the MINT solution while setting $\gamma = 1$ yields the filter-and-sum beamformer. The value of γ can then be adjusted given expected levels of AIR estimation errors to attain an optimum pre-defined dereverberation performance measure such as direct-to-reverberant ratio (DRR) or segmental signal-to-reverberant ratio [1].

Due to the complex conjugate property of DFT, only the first $\lfloor K/2 \rfloor + 1$ DFT coefficients of the equalizing filter need to be estimated, where $\lfloor\cdot\rfloor$ is the floor operator, since the remaining coefficients are simply mirroring complex conjugates of the first half. Similarly, since the beampattern is reflected around $\theta = \pi/2$, only the first $Q/2$ angles need to be processed.

6. SIMULATIONS

The performance of MINTFormer was evaluated over a range of γ in the presence of different levels of AIR estimation errors. Based on these results, experimentally optimum γ values are then determined for different levels of AIR estimation errors.

Two performance measures were used for evaluation. The first performance measure evaluates the level of reverberant tail suppression in the equalized impulse response, $r(n)$, using the direct-to-reverberant ratio given as [1]

$$\text{DRR} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_d} r^2(n)}{\sum_{n=n_d+1}^{\infty} r^2(n)}\right) \text{dB}, \tag{25}$$

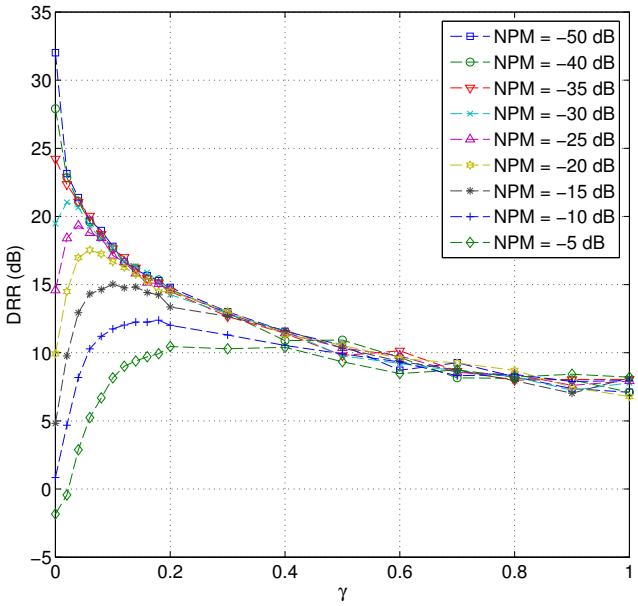

Figure 1: Averaged DRR over a range of γ for different levels of NPM.

where $r(n)$ for $n = 0,\ldots,n_d$ are assumed to represent only the direct-path propagation and $n_d = 0.005f_s$ is used in this work. The second performance measures the perceived quality of the equalized speech using ITU-T P.862 (PESQ) scores, which provides an estimate using a predicted mean opinion score (PMOS) ranging from 1 - 4.5 [11]. To allow comparison between the perceived quality of the reverberant microphone signals and the equalized speech signal, the difference in their PESQ scores is calculated as ΔP, where a positive ΔP indicates an improvement in the equalized signal over the microphone signals.

Speech signals from the TIMIT database were used as input after being resampled to $f_s = 8$ kHz and an $M = 8$ channel ULA with an inter-microphone distance of 0.04 m was used to receive the reverberant speech signals. The acoustic system was simulated using the image method [12, 13] with a distance of 1 m between the source and the microphone array centre, and a reverberation time $T_{60} = 250$ ms. The results were averaged over 100 Monte Carlo simulations with randomly varying room dimensions between the sizes of 3 x 4 x 2 m and 5 x 7 x 3 m and randomly varying locations of the source and microphone array. The fractional delay before the direct path in the AIRs was removed such that $\tau' = 0$ and the resulting length of the AIRs, $\mathbf{h}$, was taken as $L = T_{60}f_s$. AIR estimation errors were artificially introduced by the addition of Gaussian-distributed errors to $\mathbf{h}_m$ in the direction of the vector to achieve a desired level of normalized projection misalignment (NPM) over the range $[-50, -5]$ dB [14]. In this work, $Q = 37$ angles were used and the oracle case is assumed where the exact angle of arrival of the source signal, θ_d, is known and rounded to the nearest θ_q for practical beamformer design. For the iterative update of (24), the step-size $\mu = 0.01$ was used.

The DRR averaged over 100 Monte Carlo simulations are given in Figure 1. It can be seen that in the presence of small estimation errors where NPM ≤ -40 dB, optimal DRR is given by $\gamma = 0$, which gives the MINT solution. On the other hand, in the presence

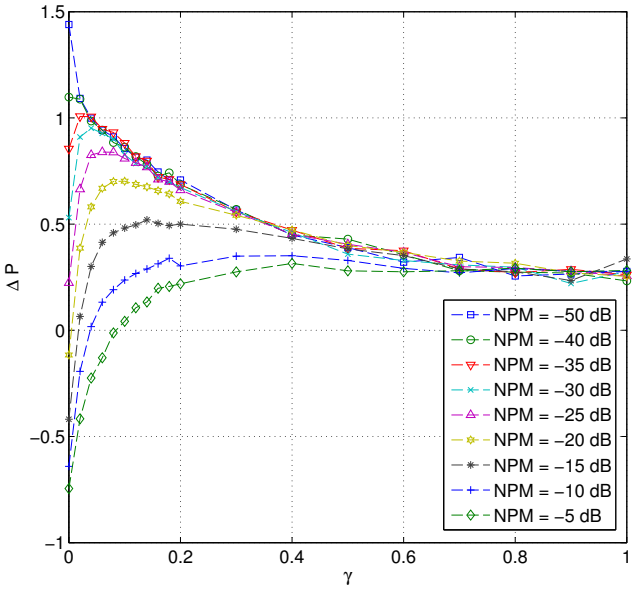

Figure 2: Averaged ΔP over a range of γ for different levels of NPM.

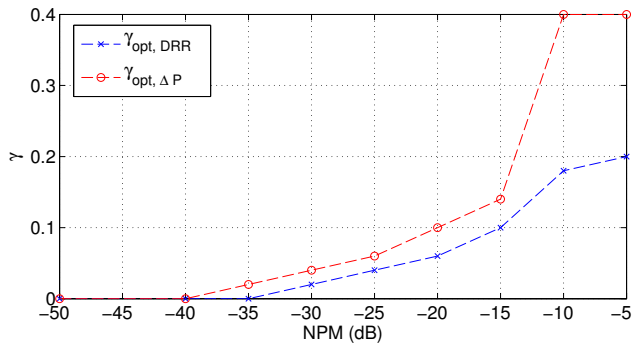

Figure 3: Experimentally optimal γ for different levels of NPM based on two performance measures, DRR and ΔP.

of larger estimation errors, a MINTFormer solution with $0 < \gamma < 1$ can give improved DRR. The ΔP scores averaged over 100 Monte Carlo simulations are given in Figure 2, where a similar trend to the DRR results can be observed.

From these results, two experimentally optimum γ values can be found based on the different performance measures. These are shown in Figure 3, where $\gamma_{\text{opt,DRR}}$ denotes the optimal γ based on the DRR results and $\gamma_{\text{opt},\Delta P}$ is based on the ΔP results.

7. CONCLUSION

Many multichannel dereverberation algorithms have been developed that can be broadly categorized as acoustic channel equalization or beamforming techniques. MINT is an example of the former and achieves perfect dereverberation in the multichannel case, but is very sensitive to AIR estimation errors. Beamforming provides moderate dereverberation performance but is more robust to angle of arrival estimations. We demonstrate that the problem formulations for both approaches are very similar and establish a link between the two solutions. The MINTFormer framework is proposed as a unified solution where both spatial and AIR knowledge are exploited to design a spatially aware channel equalizer. Experimental results demonstrate that in the presence of AIR estimation errors, MINTFormer provides improved dereverberation performance compared to a pure MINT or filter-and-sum beamformer solution.

8. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

[2] B. W. Gillespie, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 557–560.

[3] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[4] G. Harikumar and Y. Bresler, "FIR perfect signal reconstruction from multiple convolutions: minimum deconvolver orders," *IEEE Trans. Signal Process.*, vol. 46, pp. 215–218, 1998.

[5] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.

[6] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 77–80.

[7] E. Gilbert and S. Morgan, "Optimum design of directive antenna arrays subject to random variations," *Bell Syst. Tech. J.*, vol. 34, pp. 637–663, 1955.

[8] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, July 1979.

[9] S. Haykin, *Adaptive Filter Theory*, 4th ed. New Jersey, USA: Prentice-Hall, 2001.

[10] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[11] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunications Union (ITU-T), Recommendation P.862, Feb. 2001.

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

[13] E. A. P. Habets, "Room impulse response (RIR) generator," http://home.tiscali.nl/ehabets/rirgenerator.html, May 2008.

[14] W. Zhang and P. A. Naylor, "An algorithm to generate representations of system identification errors," *Research Letters in Signal Processing*, vol. 2008, pp. 13:1–13:4, Jan. 2008.