STRATEGIES FOR IMPROVING AUDIBLE QUALITY AND SPEECH RECOGNITION ACCURACY OF REVERBERANT SPEECH

Bradford W. Gillespie[*] *and Les E. Atlas*
Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA

ABSTRACT

We showed in [1] that penalizing long-term reverberation energy is more effective than maximizing the signal-to-reverberation ratio (SRR) for improving audible quality and automatic speech recognition (ASR) accuracy. Using this knowledge we propose a blind approach to speech dereverberation that reduces the length of the equalized speaker-to-receiver impulse response. The approach reduces the long-term correlation in the linear prediction (LP) residual of reverberant speech. We show that this approach improves both the audible quality (measured with subjective listening tests) and ASR accuracy (measured with two commercial ASR systems) of reverberant speech.

1. BACKGROUND

The quality of speech captured by personal computers in real-world environments is invariably degraded by acoustic interference. This interference can be broadly classified into two distinct categories: additive and convolutional.

A received speech signal, $x(n)$, can be modeled as

$$x(n) = \mathbf{r}^T(n)\mathbf{s}(n) + w(n), \quad (1)$$

where the clean speech is $\mathbf{s}(n) = [s(n-M+1)\ldots s(n)]^T$. Corrupting this clean speech waveform is additive noise given by, $w(n)$, and time varying convolutional interference, $\mathbf{r}(n) = [r(0,n)\ldots r(M-1,n)]^T$.

Additive noise can be significantly reduced by spectral subtraction or similar techniques (see for example [2] and references therein). These approaches have been successful, and are used in commercial products.

The still existing convolutional interference problem can be modeled as $\mathbf{r}(n)$, an M-tap acoustic impulse response. Typical causes of this convolutional distortion are room and microphone effects. When $\mathbf{r}(n)$ is caused by room effects it is commonly referred to as reverberation. The focus of this work is developing a system for mitigating the effects of reverberation in captured speech to improve audible quality and ASR accuracy.

It appears easy enough to remove reverberation: estimate the speaker-to-receiver filter, $\mathbf{r}(n)$, and then design an L-tap inverse filter $\mathbf{g}(n)$ to undo its effects: $y(n) = \mathbf{g}^T(n)\mathbf{x}(n)$. From linearity, we can speak of an *equalized* source-to-receiver impulse response, $\mathbf{h}(n)$, that captures the combined effects of $\mathbf{r}(n)$ and $\mathbf{g}(n)$. If $\mathbf{g}(n)$ perfectly equalizes the speaker-to-receiver impulse response then $\mathbf{h}(n)$ is a delta function and $y(n) = s(n)$ (neglecting additive noise in the received signal). Although this sounds simple, it is difficult in practice; $\mathbf{r}(n)$ must be estimated blindly from only the received signal(s). Further complicating matters, $\mathbf{r}(n)$ is often plagued by zeros near the unit circle, making inversion difficult [3]. Finally, $\mathbf{r}(n)$ is well modeled as an finite impulse response (FIR) filter [4]. As such $\mathbf{r}(n)$ can only be inverted by a corresponding infinite impulse response (IIR) filter. To avoid using IIR inverse filters, most researchers use long FIR filters [5]. The requirement of an IIR inverse is only applicable in a single microphone input system, there is no such IIR constraint on multiple input single output systems.

[*] Currently affiliated with Microsoft Corporation.

It is possible to exploit multiple microphones to make the problem somewhat more tractable. We experimentally showed in [1] that multiple microphones are necessary for complete equalization of the speaker-to-receiver impulse response. Furthermore, if complete equalization is not possible, penalizing long-term reverberation energy was shown to be more effective than maximizing the signal-to-reverberation ratio (SRR) for improving audible quality and ASR accuracy. Using this knowledge we proposed a nonblind equalizing strategy for reverberant speech. The performance of this approach was shown to exceed traditional speech enhancement techniques.

In a real system, we do not have the luxury of direct access to the speaker-to-receiver impulse response. Without this, the approach proposed in [1] (WLLS equalization) has limited applicability. Here we propose a new blind approach to speech dereverberation. This approach reduces the length of the equalized speaker-to-receiver impulse response. We show that this approach can improve both the audible quality and ASR accuracy of reverberant speech.

2. INTRODUCTION

Given a reverberated speech waveform, estimating the speaker-to-receiver impulse response is exceedingly difficult. If accurate estimates were available, the speaker-to-receiver impulse response could be directly equalized with a linear least squares (LLS) equalizer (or, as we have shown [1], more preferably a weighted linear least squares (WLLS) equalizer). Unfortunately, in practice, current techniques are not particularly effective at estimating the speaker-to-receiver impulse responses from reverberant speech [6].

Our goal is to mitigate the effects of reverberation by shortening the equalized speaker-to-receiver impulse response. What then can be done with access only to the received reverberant waveform?

Compare in Figure 1, the equalized speaker-to-receiver impulse responses (obtained in our previous work [1]) and their corresponding autocorrelations for a variety of processing approaches.

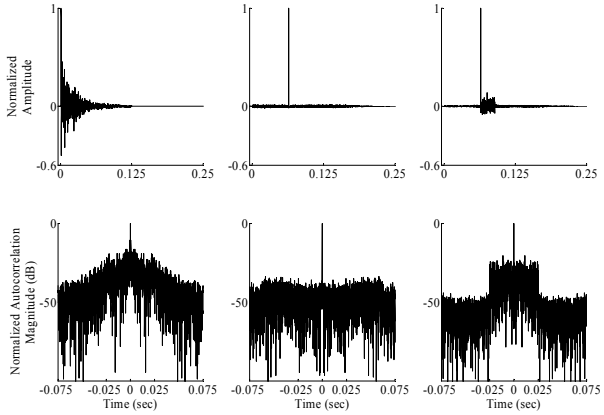

Figure 1: Comparison of the speaker-to-receiver impulse response and the autocorrelation of the speaker-to-receiver impulse response from (left) unprocessed single channel, (middle) 1-channel LLS equalized response, and (right) the 1-channel WLLS equalized response. The BWLLS equalizer reduced the energy at longer lags in the autocorrelation compared with either the unprocessed single channel or LLS equalized response.

Specifically compare that obtained using 1-channel binary weighted linear least squares (BWLLS) equalization to that obtained using either no processing or 1-channel LLS processing. As we have seen in [1], the WLLS processed speaker-to-receiver impulse response had less long-term reverberation energy than either the original impulse response or the impulse response corresponding to LLS processing. In addition, the corresponding autocorrelation sequence also had less energy at longer lags than either the original impulse response or the impulse response corresponding to LLS processing. This reduced energy in the autocorrelation sequence is a direct consequence of the reduced long-term reverberation energy in the equalized speaker-to-receiver impulse response. We showed in [1] that this shortening of the equalized speaker-to-receiver impulse response translates to improved audible quality and ASR accuracy.

Since direct estimation of the speaker-to-receiver impulse response is difficult, we instead propose to reduce the long-term correlation energy in the received reverberant waveform. The basic idea is this: WLLS processing explicitly reduces the long-term reverberation energy in the equalized speaker-to-receiver impulse response. Indirectly, the correlation energy at long lags is also reduced. In our proposed blind approach, we instead explicitly reduce the correlation energy at long lags. This indirectly shortens the equalized speaker-to-receiver impulse response.

Most of the long-term correlation that exists in reverberant speech is due to the speaker-to-receiver impulse response, very little is due to the correlation in speech. Furthermore, most of the correlation due to speech can be removed using linear prediction (LP).

To accomplish this shortening of the output autocorrelation function we develop a technique we call "correlation shaping". In effect, we shape the processed autocorrelation sequence to have a desired response. For blind speech dereverberation, the desired response is zero at long lags. This, we show, has the intended effect of reducing the length of the equalized speaker-to-receiver impulse response. The result is that our proposed processing simultaneously improves audible quality and ASR accuracy blindly.

3. CORRELATION SHAPING

Our proposed correlation shaping technique modifies the correlation structure of the processed waveform, y. One or more inputs, x_c, are modified using a set of adaptive linear filters, $\mathbf{g}_c$, to minimize the weighted mean square error between the actual output autocorrelation sequence, $\mathbf{R}_{yy}$, and the desired output autocorrelation sequence, $\mathbf{R}_{dd}$. A set of feedback functions continuously adjusts the individual equalizing filters to perform this minimization. The block diagram of a 2-channel implementation of this approach is shown in Figure 2. The actual system is not limited solely to 2 channels, but can be generalized to any number of inputs.

The adaptive filter we employ uses a gradient descent approach to accomplish this minimization. The gradient takes on a simple form, relying only on the autocorrelation of the output, $\mathbf{R}_{yy}$, the crosscorrelation between the output and input, $\mathbf{R}_{yx_c}$, and the desired output autocorrelation, $\mathbf{R}_{dd}$.

Update Equation

The multichannel input sequence $x_c(n)$ has a corresponding autocorrelation sequence $R_{x_c x_c}(\tau)$ given by

$$R_{x_c x_c}(\tau) = \sum_n x_c(n)\, x_c(n-\tau). \tag{2}$$

The correlation shaping approach will be implemented as a multi-input single-output linear filter expressed as follows

$$y(n) = \sum_{c=0}^{C-1} \mathbf{g}_c^T(n)\, \mathbf{x}_c(n). \tag{3}$$

The output signal, $y(n)$, has a corresponding autocorrelation sequence given by $R_{yy}(\tau)$

$$R_{yy}(\tau) = \sum_n y(n)\, y(n-\tau). \tag{4}$$

The goal of our proposed correlation shaping approach is to minimize the weighted mean square error between $R_{yy}(\tau)$ and the autocorrelation shape we desire, $R_{dd}(\tau)$. This error is given by

$$e(\tau) = W(\tau)\big(R_{yy}(\tau) - R_{dd}(\tau)\big)^2, \tag{5}$$

where $W(\tau)$ is a real valued weight. A large positive of $W(\tau)$ gives more importance to the error at a particular lag, τ. To compute the gradient that minimizes $e(\tau)$ with respect to the filter coefficients, evaluate the partial derivative of the error with respect to the filter,

$$\frac{\partial e(\tau)}{\partial g_c(l)} = 2W(\tau)\big(R_{yy}(\tau) - R_{dd}(\tau)\big)\frac{\partial R_{yy}(\tau)}{\partial g_c(l)}. \tag{6}$$

we obtain the gradient for each filter coefficient. The gradient is given by

$$\nabla(l) = \sum_\tau W(\tau)\big(R_{yy}(\tau) - R_{dd}(\tau)\big)\big(R_{x_c y}(l-\tau) + R_{x_c y}(l+\tau)\big) \tag{7}$$

This gradient will be used in the following update equation to perform correlation shaping,

$$g_c(l, n+1) = g_c(l,n) + \mu \nabla_c(l). \tag{8}$$

This update equation is used in the system shown in Figure 2. This system can be readily generalized to any number of inputs.

Don't Care Region

A don't care region can be introduced in the correlation shaping technique similar to that used in BWLLS (from [1]) equalization. In this case, we use a don't care region for autocorrelation lags close to the zeroth lag. The lags from, say, 1 to Z and -1 to $-Z$ are given a weight of 0, and everywhere else is given a weight of 1. In addition $\mathbf{R}_{dd}$ is 0 for all lags except the zeroth, which is given a weight of 1.With these two constraints, we can modify the gradient given in Equation (7). We can neglect the term corresponding to $\tau = 0$ in Equation (7), as this is to control the value of the zeroth lag of the output autocorrelation function. Since the zeroth lag affects only the energy in the output waveform the exact value is unimportant. It is only important that it be nonzero. This can be addressed by normalizing the update equation. Furthermore, due to the symmetry in Equation (7), the sum only needs to be evaluated for positive values of τ. The final expression is given by

$$\nabla(l) = \sum_{\tau > Z} \big(R_{yy}(\tau)\big)\big(R_{x_c y}(l-\tau) + R_{x_c y}(l+\tau)\big). \tag{9}$$

The filter design equation becomes

$$g_c(l, n+1) = g_c(l,n) + \mu \frac{\nabla_c(l)}{\sqrt{\sum_c \sum_l \nabla_c(l)^2}}. \tag{10}$$

The term in the denominator serves two purposes. It provides the normalization we require to remove the summation due to the zeroth lag. It also normalizes the gradient, improving the convergence properties in a manner similar to Normalized LMS.

Correlation Shaping for Speech Dereverberation

Correlation shaping suppresses the long-term correlation energy in the received waveform. Since the reverberated speech wave-

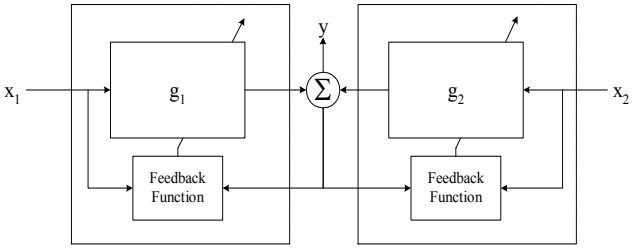

Figure 2: Block diagram for the 2-channel correlation shaping algorithm. This directly extends to any number of channels.

form contains correlation due to both speech production and room reverberation, applying our correlation shaping approach attempts to remove both the correlation due to speech and correlation due to the room.

To overcome this, we propose to remove the correlation due to speech prior to processing with our correlation shaping technique. We show that the LP residual of reverberant speech contains most of the correlation from the room but very little of the correlation from speech. Thus, processing the LP residual of reverberant speech using our correlation shaping technique does not remove the vital speech correlation. It instead removes only the correlation due to reverberation.

The top subplot in Figure 3 shows the autocorrelation of a single channel of the speaker-to-receiver impulse response from the small office environment described in [1]. This is the true autocorrelation of the speaker-to-receiver impulse response. The middle subplot of Figure 3 shows the autocorrelation of the LP residual of clean speech (FAKS0) convolved with the speaker-to-receiver impulse response. Not surprisingly, the autocorrelation of the LP residual of clean speech convolved with the speaker-to-receiver impulse response was very similar to the autocorrelation of the speaker-to-receiver impulse response alone. The correlation coefficient between the two is 0.98, demonstrating the high degree of similarity. This confirms that the LP residual of clean speech is approximately white. The lower subplot of Figure 3 shows the autocorrelation of the LP residual of reverberant speech (using the same speaker-to-receiver impulse response). The autocorrelation of the LP residual of reverberant speech and the autocorrelation of the speaker-to-receiver impulse response has significant structural similarity. The correlation coefficient between the two is 0.95.

Broadly speaking, the LP analysis of reverberant speech separates the correlation due to the speech and the correlation due to the vocal tract. The LP coefficients generally contain most of the information about the vocal tract, while the LP residual contains most of the information about reverberation. The residual could be processed to remove most of the long-term correlation. The cleaned speech waveform could be reconstructed using the modified LP residual and original LP coefficients. Unfortunately, structuring the algorithm in this manner has a significant drawback. As pointed out in [7] this structure causes LP reconstruction artifacts that result in an unnatural sounding degradation of the audible waveform.

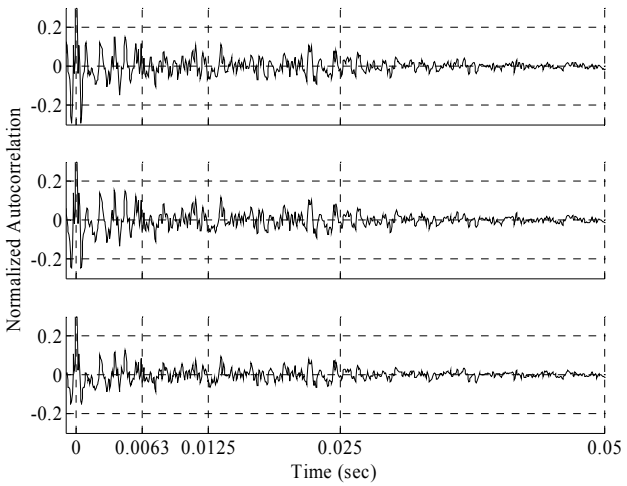

Figure 3: (Top) Normalized Autocorrelation of the speaker-to-receiver impulse response from the small office environment, (consider this the ground truth). (Middle) Normalized Autocorrelation of the LP residual of clean speech convolved with the speaker-to-receiver impulse response from the small office environment. (Bottom) Normalized Autocorrelation of the LP residual of reverberant speech (this is the realistic scenario).

A better approach is to use the double filtering technique proposed in [7]. Processing is broken into two separate, but connected paths. In one path, the LP residual is processed. In [7] processing was done to maximize kurtosis. Here, we process the LP residual to remove long-term correlation. The adaptive filter continuously tries to suppress the long-term correlation energy in the LP residual. At every time instant, the filter coefficients in this path are copied to corresponding filters in the second path. The filters in this second path are applied to the reverberated speech waveform. This avoids reconstruction artifacts, while at the same time removing the long-term correlation present in the LP residual of reverberant speech. A block diagram of the proposed system is shown in Figure 4. The update equations are the same as before, the only change is that the LP residual is processed instead of the original waveform.

4. EXPERIMENTS

We now evaluate the performance of our proposed correlation shaping approach. The full TIMIT testset corpus was used as our clean speech database. The same set of 4 speaker-to-receiver impulse responses described in [1] will be used to corrupt the clean testset. To simulate reverberant speech the full clean TIMIT testset will be convolved with each of the speaker-to-receiver impulse responses. This gives 4 separate reverberant waveforms for each speech example, simulating a 4-channel microphone array.

ASR Accuracy

To evaluate the ASR accuracy of reverberant speech processed with our correlation shaping technique we used both the Microsoft Speech SDK 5.0 and the IBM ViaVoice 8.0 recognizer. We provided no additional "out-of-box" training to the system, and all online adaptation was disabled. This ensured that the ASR systems remained as general as possible.

ASR accuracy (Table 1) for the clean testset was 58.8% with the Microsoft recognizer and 66.0% with the IBM recognizer. Since the ASR systems received no "out of the box" training, there is a mismatch between training and testing conditions. This was the primary reason for the low ASR accuracy.

An unprocessed single channel of reverberant speech reduced the ASR accuracy to 28.7% for the Microsoft recognizer and to 28.6% for the IBM recognizer (Table 1). To measure the improvement in ASR accuracy using a delay-sum (DS) preprocessor [8], the reverberant speech is processed by a 2, 3, and 4-element delay-sum array. Results are shown in Table 2.

Reverberant speech was processed with our 4-channel correlation shaping technique (*i.e.* the system shown in Figure 4) The equalizing filters, $\mathbf{g}_k$, had 1000 taps per channel and the learning rate was fixed at $3e^{-4}$. The processing was repeated for a variety of don't care region durations. Table 3 shows the SRR and RT_{60} for these various durations. Table 3 also shows the ASR accuracy as a

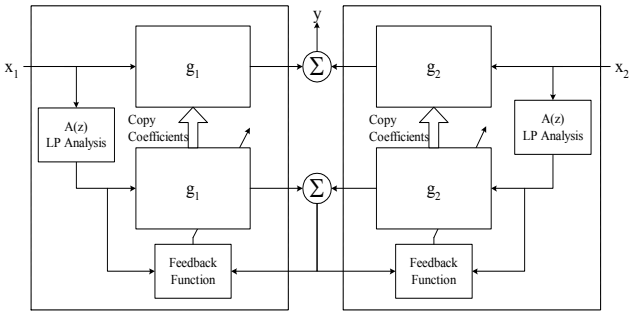

Figure 4: Block diagram for the 2-channel correlation shaping based speech enhancement algorithm. This directly extends to any variety of channels.

Table 1: ASR accuracy of the clean TIMIT dataset and a single unprocessed reverberant channel using the Microsoft and IBM ASR systems.

	Clean	Single Reverberant Channel
Accuracy (MS)	58.8%	28.7%
Accuracy (IBM)	66.0%	28.6%

Table 2: Accuracy of the TIMIT data using delay-sum processing.

	2-channels	3-channels	4-channels
Accuracy (MS)	30.6%	33.4%	34.5%
Accuracy (IBM)	31.3%	32.7%	33.8%

function of the don't care region. We do not show the ASR accuracy of the output as it was adapting, but rather show the accuracy of the reverberant speech processed with the final converged filter. Using the Microsoft recognizer, reverberant speech processed with a 4-channel correlation shaping preprocessor with a don't care region of 0.0187 s gave a 33% better accuracy than the same speech processed with the 4-channel DS beamformer. Using the IBM recognizer the gain in the same situation was 42%

Audible Quality

The audible quality of this speech dereverberation technique was evaluated by subjective testing. The experimental protocol was as follows

- 3-second sound clips from 8 speakers in the TIMIT dataset were used for testing. The speakers were the following: FADG0, FAKS0, FCMH0, FCMR0, MTLS0, MWBT0, MWJG0, and MWVW0.
- Six variations of each three second clip were generated, 1) the original clean clip, 2) the reverberated single channel, 3) the 4 reverberated channels processed by the DS array, 4) the 4 reverberated channels processed by the correlation shaping approach with a don't care region of 0.0 s 5) 0.0187 s, and 6) 0.0375 s.
- Listeners were randomly presented with 2 variations of the same clip and asked to indicate which they preferred. They were not told to select which variation was less reverberant, only which variation they preferred. They could listen to each variation as many times as they liked.

12 subjects participated in this test. The results are shown in Table 4. These results indicate that using our proposed correlation shaping technique with a don't care region of 0.0187 s provides a significant improvement over an unprocessed single channel, and 4-channel DS processing. 100% of the time the average listener preferred reverberant speech processed by the 4-channel correlation shaping equalizer with a don't care region of 0.0187 s to the unprocessed speech. The speech processed with a 4-channel correlation shaping equalizer with a don't care region of 0.0187 s was also superior to 4-channel DS equalization, 90% of the time (7.2 out of 8 speakers) the average listener preferred correlation shaping. In addition using the correlation shaping technique with a don't care region of 0.0187 s gave better audio quality than the same technique using either a don't care region of 0.0 s or 0.0375

5. SUMMARY

Here we developed a blind approach to speech dereverberation that reduces the length of the equalized speaker-to-receiver impulse response. The approach reduces the long-term correlation in the LP

Table 3: SRR, RT_{60}, and ASR accuracy using 4-channel correlation shaping.

	Don't Care Region					
	0.0 s	**0.0063 s**	**0.0125 s**	**0.0187 s**	**0.025 s**	**0.0313 s**
SRR (dB)	1.62	0.59	0.00	−0.05	−0.19	−0.30
RT_{60} (s)	0.31	0.29	0.24	0.22	0.20	0.19
Accuracy (MS)	40.7%	43.0%	45.6%	45.9%	44.3%	42.8%
Accuracy (IBM)	42.3%	43.8%	47.9%	48.0%	45.9%	44.6%

Table 4: Result for the average listener on the subjective test. Scores are reported as the percentage of the time the average listener preferred A to B over the eight different clips each listener is asked to compare.

Percentage of sound clips for which the average listener …	mean	sd
preferred Correlation Shaping (0.0187 s) to Unprocessed Single Channel	100%	0%
preferred Correlation Shaping (0.0187 s) to 4-channel DS	90%	12%
preferred Correlation Shaping (0.0187 s) to Clean	2%	5%
preferred Correlation Shaping (0.0187 s) to Correlation Shaping (0.0 s)	79%	15%
preferred Correlation Shaping (0.0187 s) to Correlation Shaping (0.0375 s)	81%	10%
preferred DS to Unprocessed Single Channel	98%	5%
preferred DS to Clean	0%	0%
preferred Clean to Unprocessed Single Channel	100%	0%

residual of reverberant speech. We have shown that this correlation shaping approach can simultaneously improve both the audible quality and ASR accuracy of reverberant speech.

- Using the Microsoft recognizer, reverberant speech processed with 4-channel correlation shaping (0.0187 s) provided a 60% relative gain in ASR accuracy over an unprocessed reverberant channel alone, while the same reverberant speech processed with a 4-channel DS array provided only a 20% improvement.
- Under the same conditions, but with the IBM recognizer, reverberant speech processed with 4-channel correlation shaping (0.0187 s) provided a 68% relative gain in ASR accuracy over an unprocessed reverberant channel alone, while the same reverberant speech processed with a 4-channel DS array provided only a 18% improvement.

In addition, reverberant speech processed with 4-channel correlation shaping (0.0187 s) was shown provided superior audio quality to either 4-channel DS processing or an unprocessed single channel.

ACKNOWLEDGEMENTS

Acoustic impulse response data used in this work were provided by Dr. Henrique Malvar and Microsoft Research (MSR). This work was supported through an MSR Graduate Fellowship.

REFERENCES

[1] B. W. Gillespie and L. E. Atlas, "Acoustic Diversity for Improved Speech Recognition in Reverberant Environments," presented at the 2002 International Conference on Acoustics, Speech, and Signal Processing, 2002.

[2] W. Jiang and H. Malvar, "Adaptive Noise Reduction of Speech Signals," Microsoft Research MSR-TR-2000-86, 2000.

[3] H. Wang and F. Itakura, "Dereverberation of speech signals based on sub-band envelope estimation," IEICE Transactions, vol. E74, pp. 3576-83, 1991.

[4] B. E. D. Kingsbury, "Perceptually Inspired Signal Processing Strategies for Robust Speech Recognition in Reverberant Environments," University of California, Berkeley, 1998.

[5] J. N. Mourjopoulos, "Digital equalization of room acoustics," Journal of the Audio Engineering Society, vol. 42, pp. 884-900, 1994.

[6] L. Juan and H. Malvar, "Blind deconvolution of reverberated speech signals via regularization," presented at 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Piscataway, NJ, 2001.

[7] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech Dereverberation via Maximum Kurtosis Subband Adaptive Filtering," presented at the 2001 International Conference on Acoustics, Speech, and Signal Processing, 2001.

[8] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko, "Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms," Journal of the Acoustical Society of America, vol. 78, pp. 1508-1518, 1985.