# SPEECH DEREVERBERATION VIA MAXIMUM-KURTOSIS SUBBAND ADAPTIVE FILTERING

*Bradford W. Gillespie* [1], *Henrique S. Malvar* [2] *and Dinei A. F. Florêncio* [2]

[1]Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA
[2]Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

## ABSTRACT

This paper presents an efficient algorithm for high-quality speech capture in applications such as hands-free teleconferencing or voice recording by personal computers. We process the microphone signals by a subband adaptive filtering structure using a modulated complex lapped transform (MCLT), in which the subband filters are adapted to maximize the kurtosis of the linear prediction (LP) residual of the reconstructed speech. In this way, we attain good solutions to the problem of blind speech dereverberation. Experimental results with actual data, as well as with artificially difficult reverberant situations, show very good performance, both in terms of a significant reduction of the perceived reverberation, as well as improvement in spectral fidelity.

## 1. INTRODUCTION

The quality of speech captured by personal computers in business offices is usually degraded by environment noise and by reverberation (caused by the sound waves reflecting off walls and other surfaces). Quasi-stationary noise produced by computer fans and air conditioning can be significantly reduced by spectral subtraction or similar techniques [1]. Reducing the distortion caused by reverberation is a difficult blind deconvolution problem, due to broadband nature of speech and the high order of the equivalent impulse response from the speaker's mouth to the microphone. The problem is, of course, alleviated by the use of microphone headsets, but those are usually inconvenient to the user.

In this paper we present an efficient algorithm for speech dereverberation using subband adaptive filtering for fast convergence. The key new concept is to control the adaptive subband filters not by a mean-square error criterion, but by a kurtosis metric on LP residuals. In this way, we make efficient use of the *a priori* knowledge that the signal to be recovered is speech. The algorithm is capable of reducing reverberation even when a single microphone signal is available, but better results are obtained with arrays containing several microphones.

We can model the signal received by the $c$th microphone as

$$x_c(n) = \mathbf{s}^T(n)\mathbf{g}_c(n) + w_c(n) \qquad (1)$$

where $\mathbf{s}(n) = \left[s(n-N+1)\ldots s(n)\right]^T$, with $s(n)$ the "clean" speech signal to be recovered, $w_c(n)$ are additive noises, and $\mathbf{g}_c(n)$ are the $N$-tap acoustic impulse responses. For a typical "wideband telephony" sampling rate of 16 kHz, $N$ can vary from 1,000 to over 4,000.

A simple multi-microphone speech enhancement system is the delay-and-sum beamformer [2], in which an estimate of $s(n)$ is formed by simply averaging $x_c(n-L_c)$. The delays, $L_c$, are computed to best enhance the desired speech signal. More efficient approaches have been reported, such as the use of subband envelope estimation [3], and decomposition of the received microphone signals into minimum-phase and allpass components [4]. Such techniques have shown only modest improvement over the delay-and-sum approach, in terms of reverberation reduction. The use of speech models to improve performance has been discussed in many reports, *e.g.* [5], [6]. In this paper, we extend use of explicit speech models by optimizing a metric of time-domain signal concentration to control the adaptation of the dereverberation filters. We achieve significant improvement in performance over the delay-and-sum beamformer, both in subjective signal quality and in spectral definition.

## 2. SPEECH ENHANCEMENT

For clean voiced speech, LP residuals have strong peaks corresponding to glottal pulses, whereas for reverberated speech such peaks are spread in time [6]. A measure of amplitude spread of LP residuals can serve as a reverberation metric. To test this concept, we performed the following experiment: in a standard 11'×11' office, we collected speech signals played back through a mouth simulator (Brüel &Kjær 4227) with a sampling frequency of 16 kHz, at fourteen locations, 6" to 84" (6" spacing) from a single omnidirectional electret microphone. We computed 10-th order LP residuals over 32 ms (512 samples) frames, and then the final kurtosis as the average of the frame kurtosis. A typical result is shown in Figure 1, for a female speaker in the presence of interfering office noise. We conclude that LP residual kurtosis is a reasonable measure of reverberation.

Our goal is to develop an online adaptive gradient-descent algorithm that maximizes LP residual kurtosis. In other words, we seek to find blind deconvolution filters that make the LP residuals as far as possible from being Gaussian – an idea that has been applied to blind deconvolution problems in underwater acoustics and geophysics [7],[8]. The following sections present our implementation of such an adaptive algorithm. We begin by developing an online single channel time-domain system. This is readily extended to handle multiple channels. While the approach is easier to describe in the time-domain, a frequency-domain implementation leads to better results, and thus we present the details of the frequency-domain mutichannel system.

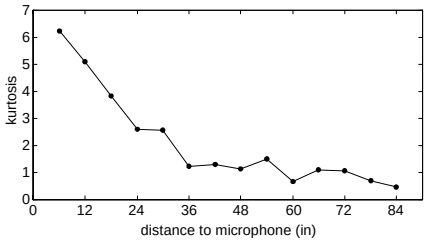



Figure 1. Kurtosis of LP residuals as a reverberation metric.

### 2.1 Single Channel Time-Domain Adaptation

This system is shown in Figure 2 (a). The received noisy reverberated speech signal is $x(n)$ and its corresponding LP residual is $\tilde{x}(n)$. $\mathbf{h}(n)$ is the $L$-tap adaptive filter at time $n$. The output is $\tilde{y}(n)=\mathbf{h}^T(n)\tilde{\mathbf{x}}(n)$, where $\tilde{\mathbf{x}}(n)=\left[\tilde{x}(n-L+1)\ldots\tilde{x}(n-1)\tilde{x}(n)\right]^T$. An LP synthesis filter yields $y(n)$, the final processed signal. Adaptation of $\mathbf{h}(n)$ is similar to the traditional LMS adaptive filter [9], except that instead of a desired signal we use a feedback function, $f(n)$, described below.

A problem with the system in Figure 2 (a) is LP reconstruction artifacts. This can be avoided in a simple manner. For small adaptation rates, the system in Figure 2 (a) is linear. $\mathbf{h}(n)$ can be computed from $\tilde{x}(n)$ but applied directly to $x(n)$, as shown in Figure 2 (b). LP reconstruction artifacts are avoided at the small price of running two filters.

To derive the adaptation equations, recall that we desire a filter that maximizes the kurtosis of $\tilde{y}(n)$, given by

$$J(n)=\mathrm{E}\left\{\tilde{y}^4(n)\right\}\Big/\mathrm{E}^2\left\{\tilde{y}^2(n)\right\}-3 \tag{2}$$

where the expectations E{} can be estimated from sample averages. The gradient of $J(n)$ with respect to the current filter is

$$\frac{\partial J}{\partial \mathbf{h}}=\frac{4\left(\mathrm{E}\left\{\tilde{y}^2\right\}\mathrm{E}\left\{\tilde{y}^3\tilde{\mathbf{x}}\right\}-\mathrm{E}\left\{\tilde{y}^4\right\}\mathrm{E}\left\{\tilde{y}\tilde{\mathbf{x}}\right\}\right)}{\mathrm{E}^3\left\{\tilde{y}^2\right\}} \tag{3}$$

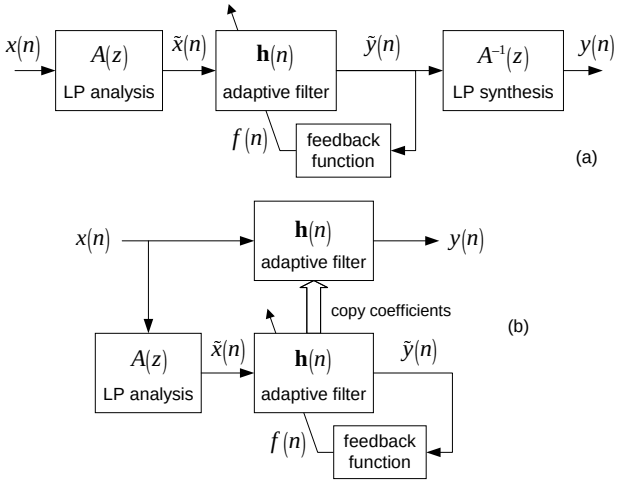



Figure 2. (a) A single channel online time-domain adaptive algorithm for maximizing kurtosis of the LP residual. (b) Equivalent system, which avoids LP reconstruction artifacts.

where the dependence on the time $n$ is not written for simplicity. In a manner similar to [10], we can approximate the gradient by

$$\frac{\partial J}{\partial \mathbf{h}}\approx\left(\frac{4\left(\left(\mathrm{E}\left\{\tilde{y}^2\right\}\tilde{y}^2-\mathrm{E}\left\{\tilde{y}^4\right\}\right)\tilde{y}\right)}{\mathrm{E}^3\left\{\tilde{y}^2\right\}}\right)\tilde{\mathbf{x}}=f(n)\tilde{\mathbf{x}}(n) \tag{4}$$

We refer to $f(n)$ as the feedback function. This function is used to control the filter updates. For continuous adaptation, $\mathrm{E}\left\{\tilde{y}^2(n)\right\}$ and $\mathrm{E}\left\{\tilde{y}^4(n)\right\}$ are estimated recursively. The final structure of the update equations for a filter that maximizes the kurtosis of the LP residual of the input waveform is then given by

$$\mathbf{h}(n+1)=\mathbf{h}(n)+\mu f(n)\tilde{\mathbf{x}}(n) \tag{5}$$

where

$$\begin{aligned} f(n)&=\frac{4\left[\mathrm{E}\left\{\tilde{y}^2(n)\right\}\tilde{y}^2(n)-\mathrm{E}\left\{\tilde{y}^4(n)\right\}\right]\tilde{y}(n)}{\mathrm{E}^3\left\{\tilde{y}^2(n)\right\}},\\ \mathrm{E}\left\{\tilde{y}^2(n)\right\}&=\beta\,\mathrm{E}\left\{\tilde{y}^2(n-1)\right\}+(1-\beta)\tilde{y}^2(n),\quad\text{and}\\ \mathrm{E}\left\{\tilde{y}^4(n)\right\}&=\beta\,\mathrm{E}\left\{\tilde{y}^4(n-1)\right\}+(1-\beta)\tilde{y}^4(n). \end{aligned} \tag{6}$$

Parameter $\mu$ controls the speed of adaptation, and $\beta$ controls the smoothness of the moment estimates.

### 2.2 Multichannel Time-Domain Adaptation

A multichannel time-domain implementation extends directly from this single-channel system just described. As before, our objective is to maximize the kurtosis of $\tilde{y}(n)$, the LP residual of $y(n)$. In this case, $y(n)=\sum_{c=1}^{C}\mathbf{h}_c^T(n)\mathbf{x}_c(n)$, where $C$ is the number of channels. Extending the analysis of the previous subsection, it is easy to see that the multichannel update equations become

$$\mathbf{h}_c(n+1)=\mathbf{h}_c(n)+\mu f(n)\tilde{\mathbf{x}}_c(n) \tag{7}$$

where the feedback function $f(n)$ is computed as in (6) using the multichannel $y(n)$. To jointly optimize the filters, each channel is independently adapted, using the same feedback function.

### 2.3 Frequency-Domain Implementation

Direct use of the time-domain LMS-like adaptation equations in (5) and (7) is not recommended, because the large variations in the eigenvectors of the autocorrelation matrices of the input signals may lead to very slow convergence, or no convergence at all under noisy situations [9]. We use subband adaptive filtering structure based on the modulated complex lapped transform (MCLT), as proposed in [11]. Since each subband signal has an approximately flat spectrum, we expect not only faster convergence but reduced sensitivity to noise [11]. A multichannel MCLT-based subband version of the structure of Figure 2 (b) is shown in Figure 3. Even though that figure shows only two channels, generalization to more channels is easy. Also, although two inverse MCLT blocks per channel are shown in Figure 3, it is clear that we can add the channels in the MCLT domain, so that only one IMCLT is needed for $y(n)$ and only one for $\tilde{y}(n)$.

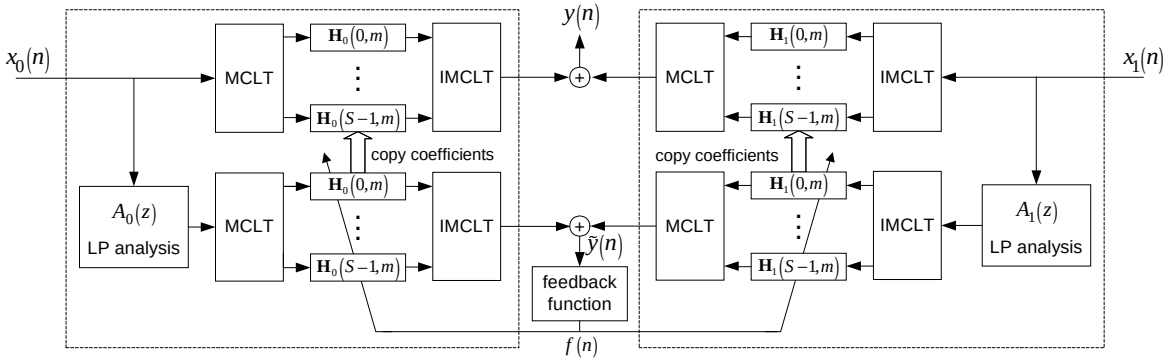



Figure 3. A two-channel online frequency-domain adaptive algorithm for speech dereverberation. A system with more than two channels extends directly from this system.

We assume that the microphone signals are decomposed via MCLTs into *M* complex subbands. To determine *M*, we consider the tradeoff that larger *M* are desired to whiten the subband spectra, whereas smaller *M* are desired to reduce processing delay. A good compromise is to set *M* such that the frame length is about 20–40 ms. Each subband *s* of each channel *c* is processed by a complex FIR adaptive filter with *L* taps, $\mathbf{H}_c(s,m)$, where *m* is the MCLT frame index. By considering that the MCLT approximately satisfies the convolution properties of the FFT [11], we can easily map the update equations in (7) to the frequency domain, generating the new update equation

$$\mathbf{H}_c(s,m+1) = \mathbf{H}_c(s,m) + \mu F(s,m)\tilde{\mathbf{X}}_c^*(s,m) \qquad (8)$$

where the superscript * denotes complex conjugation.

Unlike in an LMS formulation, the appropriate feedback function *F*(*s*, *m*) cannot be computed in the frequency domain. To compute the MCLT-domain feedback function *F*(*s*, *m*), we generate the reconstructed signal $\tilde{y}(n)$ and compute *f*(*n*) from (6). We then compute *F*(*s*, *m*) from *f*(*n*) using the MCLT. The overlapping nature of the MCLT introduces a one-frame delay in the computation of *F*(*s*, *m*). Thus, to maintain an appropriate approximation of the gradient, we use the previous input block in the update equation (8), generating our final update equation

$$\mathbf{H}_c(s,m+1) = \mathbf{H}_c(s,m) + \mu F(s,m-1)\tilde{\mathbf{X}}_c^*(s,m-1). \qquad (9)$$

Assuming the learning gain $\mu$ is small enough, the extra delay in the update equation above will introduce a very small error in the final convergence of the filter.

### 2.4 Implementation Issues

Kurtosis is insensitive to the total energy of the waveform. Therefore, like in most blind deconvolution problems, there is a gain uncertainty. As usual, we can solve that by maintaining a constant norm within the filter coefficients at each update cycle.

It is interesting to note that, although our optimization criterion of maximizing kurtosis of the LP residual makes more sense for voiced speech, we have not found a need to restrict this algorithm to adapt only during voiced segments. Continuously adapting the filters, even during unvoiced or silent periods, provides satisfactory results. This is because during these periods the input energy in $\tilde{\mathbf{x}}$ is generally small, reducing the adaptation rate.

For our dereverberation experiments, we obtained good results with the following parameters: $\beta = 0.99$, $\mu = 0.0004$, and $\mathbf{H}_c(s,0) = [1\ 0\ 0\ \cdots\ 0]^T$.

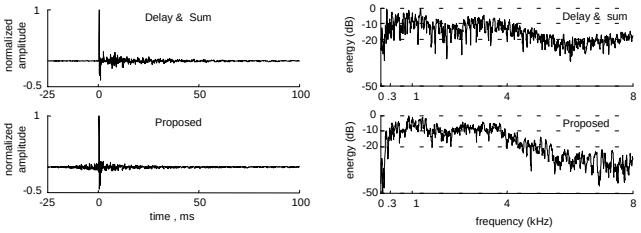
## 3. EXPERIMENTAL RESULTS

We present several experimental results from our proposed algorithm, comparing them to a delay-and-sum beamformer. As performance metrics, we use *equalized* room impulse responses and spectrograms. We refrained from computing mean-square error (MSE) between the original and reconstructed signals, because our system is not driven to minimize MSE, and minimum MSE does not necessarily correspond to better sounding speech.

### 3.1 Experiment 1

We collected data using a linear microphone array with 3” spacing between elements, at a distance of 7’ from the mouth simulator. To understand the performance of the algorithm we computed the impulse responses from the mouth simulator to each of the four microphone elements, by playing two minutes of white noise through the mouth simulator and correlating the received waveform with the transmitted white noise sequence. Without changing the room, ambient noise was collected (by turning on fans and computers) on the same array using the same system configuration. For reference, reverberated noisy speech was also collected by playing “clean” female speech signal through the mouth simulator. Finally, synthesized noisy speech was obtained by convolving the clean female speech signal with the impulse responses and adding the real room noise. Simulations were run using both the noisy speech and the synthetic speech, with no difference observed when the SNR was the same. Therefore, by acquiring real ambient noise from the same setup and room, we can realistically simulate (1) while being able to control the signal-to-noise (SNR) ratio and monitor the equalized room impulse response.

We used a 4-channel, 256-subband structure with only one tap in each subband adaptive filter. The results are shown in Figure 4. The equalized impulse response from our proposed approach is more impulsive than the equalized response from the delay and sum beamformer. Potentially more significant is the number of zeros in the spectrum of the equalized delay-sum impulse response that have been removed by the processing presented here. The spectrum of the equalized impulse response



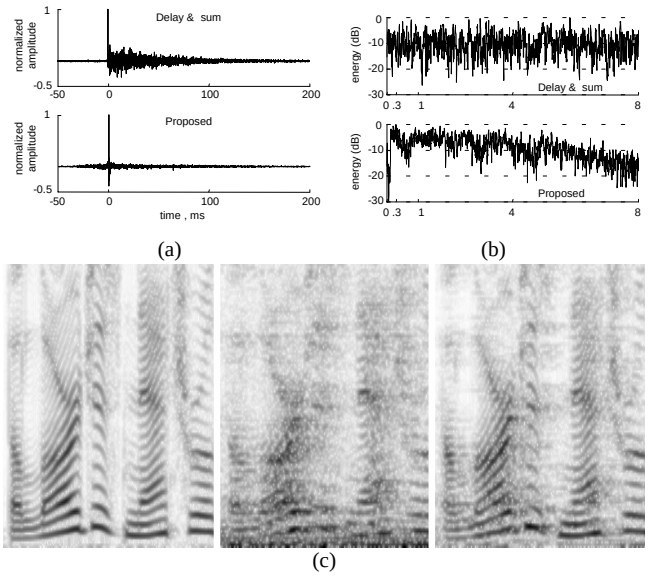

(a) (b)

Figure 4. Results for Experiment 1. Compare the equalized impulse response for a delay & sum beamformer to our proposed approach in (a) the time-domain (ideal result would be an impulse), (b) the frequency-domain.

from our proposed approach is considerably flatter in the important 0.5kHz to 4kHz region, compared to that of delay-and-sum.

### 3.2 Experiment 2

To test the ability of our proposed algorithm to equalize longer reverberation we simulate four impulse responses as white noise under a decaying exponential. A 4-channel 512-subband filter with one tap per band was used. Using these impulse responses we generate a received signal using the same female speaker and noise segments from Experiment #1. The result of this processing is shown in Figure 5. Listening to the processed waveform it is possible to hear a dramatic reduction in reverberation after about 5 seconds of adaptation. Figure 5 also shows that most of the spectral details of the original signal are recovered with our algorithm.




(a) (b)

(c)

Figure 5. Results for Experiment 2. Compare the equalized impulse response for a delay-and-sum beamformer to our proposed approach in (a) the time-domain (ideal result would be an impulse), (b) the frequency-domain. The three voiced-speech spectrograms (darker is more intense) in (c) are: original (left), delay-and-sum (center), and our proposed approach (right); the horizontal time window is 1 sec, and the vertical range is 0 – 4 kHz. Note the better spectral definition obtained with the proposed algorithm.

## 4. SUMMARY

In this paper we presented a new approach to dereverberate speech. Our approach is based on the principle that LP residual of reverberated speech (specifically voiced speech) is a time-spread version of the impulse-like LP residual of clean speech. We have shown that a kurtosis metric is effective in measuring reverberation. Computing the gradient of this metric with respect to the deconvolution filters is relatively easy. This yields a final form for adaptive filters that is simple and LMS-like.

For improved performance, we used a dual-filter structure to avoid LP reconstruction artifacts, and a subband filtering structure based on the MCLT. In that way, convergence is achieved within a few seconds, and computational complexity is not much higher than that of a standard LMS adaptive filter.

We validated the performance of our system on real world data, from both stationary and moving (not presented here) sources. It has also been validated on artificially difficult reverberation, with significantly better results than delay-and-sum beamforming.

## 5. REFERENCES

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

[2] J.L. Flanagan *et. al.*, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, **78**(11), pp. 1508–1518, Nov. 1985.

[3] H. Wang and F. Itakura, "An approach to dereverberation using multi-microphone sub-band envelope estimation," *Proc. ICASSP*, pp. 953–956. 1991.

[4] J. Gonzalez-Rodrigues, J. L. Sanchez-Bote, and J. Ortega-Garcia, "Speech dereverberation and noise reduction with a combined microphone array approach," *Proc. ICASSP*, pp. 953–956. 2000.

[5] M. Brandstein, "On the use of explicit speech modeling in microphone array applications," *Proc. ICASSP*, pp. 3613–3616, 1998.

[6] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, **8**(3), pp. 267–281, May 2000.

[7] R. A. Wiggins, "Minimum entropy deconvolution," *Geoexploration*, **16**, pp. 21–35, 1978.

[8] M. K. Broadhead and L. A. Pflug, "Performance of some sparseness criterion blind deconvolution methods in the presence of noise," *J. Acoust. Soc. Am.*, **107**(2), pp.885–893, Feb. 2000.

[9] S. Haykin, *Adaptive Filter Theory*. New Jersey: Prentice-Hall, 1996.

[10] O. Tanrikulu and A.G. Constantinides, "Least-mean kurtosis: a novel higher-order statistics based adaptive filtering algorithm," *Electronics Letters*, **30**(3), pp. 189–190, February 3, 1994.

[11] H. Malvar, "A modulated complex lapped transform and its application to audio processing," *Proc. ICASSP*, pp. 1421–1424, 1999.