MICROPHONE ARRAY SPEECH DEREVERBERATION USING COARSE CHANNEL MODELING

Scott M. Griebel and Michael S. Brandstein

Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
griebel@fas.harvard.edu, msb@hrl.harvard.edu

ABSTRACT

This paper presents a model-based method for the enhancement of multi-channel speech acquired under reverberant conditions. A very coarse estimate of the channel responses associated with each source-microphone pair is derived directly from the received data on a short-term basis. These estimates are employed to modify the LPC residuals of the channel data in an effort to deemphasize the effects of reverberant energy in the resulting synthesized signal. The approach is robust to conditions of partial and approximate channel information. Specifically, the incorporated channel model requires only approximate times and amplitudes of the initial multi-path reflections. In practice these impulses are responsible for the bulk of reverberant energy in the received speech signal and can be estimated to a sufficient degree on a time-varying basis.

1. INTRODUCTION

In recent years, microphone arrays have seen increasing application for the acquisition of speech in hands-free, distant-talker scenarios. A major distinction between these situations and the typical close-talker environment is the presence of significant room reverberation effects, the removal of which has proven to be a very difficult problem.

Reverberant distortion effects are convolutional and highly non-stationary. Multi-channel processing methods which are primarily oriented toward reducing the contributions of uncorrelated interfering sources and additive noise are of limited utility. To some degree, the signal averaging associated with beamforming is effective at attenuating long term echoes which tend to be uncorrelated across channels, but does little to reduce short term effects.

A number of approaches have been developed to identify the reverberant channel effects in some form and compensate for them. These include cepstral processing [1], matched filtering [2], and adaptive sub-space filtering [3]. However, the channel responses in even the simplest practical enclosure are very sophisticated and quickly time-varying. Motion as little as a few centimeters or a talker turning his or her head is frequently sufficient to compromise the behavior of these schemes [4].

In our prior work [5, 6, 7, 8], we have utilized a general strategy which emphasizes the incorporation of explicit speech modeling into the microphone array processing. By exploiting knowledge of the desired signal's attributes, this approach is capable of suppressing the deleterious effects of both reverberations and additive noise without explicitly identifying the channel and is adaptive on a frame by frame basis. In this work, we go a step further and include a very coarse model of the reverberant channels. While methods which perform some form of inverse filtering are very sensitive to the precision of the channel estimates, the approach taken here is robust to conditions of partial and approximate channel information. Specifically, the incorporated channel model requires only approximate times and amplitudes of the initial multi-path reflections. In practice these impulses are responsible for the bulk of reverberant energy in the received speech signal and can be estimated to a sufficient degree on a time-varying basis.

The next section outlines the model-based approach for multi-channel speech dereverberation. Methods for estimating the channel responses and using this information to enhance the speech are detailed. Section 3 presents some illustrations of the procedure and its results while Section 4 offers some conclusions.

This work was funded by National Science Foundation CAREER grant CCR-9983839.

2. SPEECH DEREVERBERATION ALGORITHM

A general model for speech production involves an impulse or noiselike signal exciting an all-pole filter. The proposed algorithm relies on the assumption that the detrimental effects of additive noise and reverberations introduce only zeros into the overall system and will primarily affect only the nature of the speech excitation sequence, not the all-pole filter. It is also assumed that the noise and errant impulses contributed to the excitation sequences are relatively uncorrelated across the individual channels, while the excitation impulses due to the original speech are invariant to the environmental effects.

Using this time-domain model for speech production, the approach is to identify the clean speech excitation sequence from a set of corrupted excitation signals and then reconstruct the speech with only the enhanced sequence. This method was applied in [6] using the LPC residual derived from I microphone channels. The excitation signals of the clean speech were identified through a pitch-synchronous clustering criterion. In [7, 8] the estimation of the residual impulses was carried out more effectively by employing a class of wavelets to decompose the LPC residuals. By locating the extrema which are well clustered across all channels, it was possible to capture the underlying impulsive structure of the original non-reverberant speech. We now show how a coarse estimate of the individual channel responses may be used to derive the clean speech excitation signal.

The reverberant speech signal, $x_i[n]$, observed at the i^{th} mi-

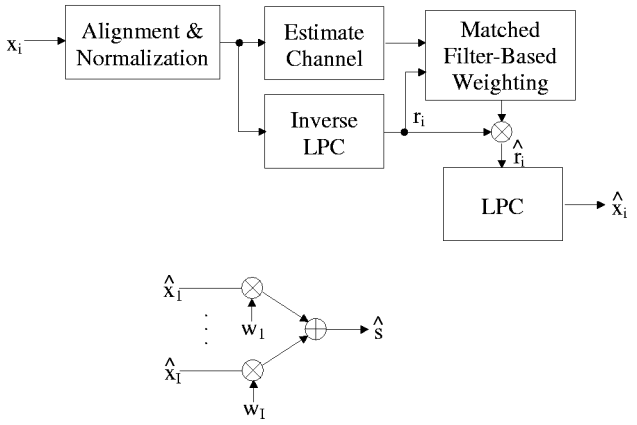

Fig. 1. Speech dereverberation algorithm for each channel (top) and ensuing weighted beamforming technique (bottom).

crophone ($i = 1, 2, \ldots, I$) can be modeled as:

$$x_i[n] = h_i[n] * s[n] + u_i[n] \tag{1}$$

where $s[n]$ is the clean speech signal, $u_i[n]$ represents ambient noise uncorrelated with the desired speech, and $h_i[n]$ is the room impulse response for the speech source relative to the i^{th} microphone. Our main objectives are to first estimate the individual channel responses, $h_i[n]$, and then to use these estimates to obtain an enhanced speech signal estimate, $\hat{s}[n]$.

The dereverberation algorithm is outlined in Figure 1. Using a 30ms half-overlapping analysis window, the time-aligned and power-equalized channel signals are individually whitened with a 13^{th} order inverse LPC filter. This produces a set of I residual signals, $r_i[n]$, which are then modulated by the results of a matched filtering based operation in an effort to deemphasize those portions which are due to noise and reverberation. The enhanced residual signals, $\hat{r}_i[n]$, are then passed through the appropriate LPC filter to produce an enhanced signal for the given channel, $\hat{x}_i[n]$. Finally, these enhanced channel signals are combined via a beamforming procedure, as shown in Figure 1, to produce a final estimate of the desired speech, $\hat{s}[n]$. Details of the channel estimation and matched filtering based weighting procedures are given below.

2.1. Channel Estimation

For a given source/microphone combination, the impulse response of the acoustic channel generally resembles a noise-like set of impulses with amplitude modulated by a decaying exponential. The rate of decay is roughly determined by the reverberation time of the environment. The direct path of the source corresponds to the initial impulse. It is followed by a set of increasingly delayed and attenuated impulses which are the result of one or more reflections from the enclosure surfaces. The Allen and Berkley image model technique [9] provides a means of simulating these impulse responses for simple room geometries and surfaces. We now illustrate a means for deriving a short-term estimate of the channel impulse responses using only the available microphone array data.

Assuming the image model is appropriate, the impulse response of channel i after time-delay compensation and normalization consists of an initial impulse representing the direct path and a set of delayed and attenuated impulses corresponding to the multipath reflections, i.e.:

$$h_i[n] = \delta[n] + \sum_l \alpha_{il}\delta[n - n_{il}] \tag{2}$$

and has the Fourier Transform:

$$H_i(\omega) \;=\; 1 + \sum_l \alpha_{il} e^{-j\omega n_{il}} \triangleq 1 + R_i(\omega)$$

where $R_i(\omega)$ is the response due to room reverberations.

We now make use of the Phase Transform (PHAT) version of the generalized cross-correlation function [10] to estimate the initial multipath components of the individual channel responses. The PHAT produces the cross-correlation of two input signals using only phase information derived from their respective power spectra. This phase-only procedure has the effect of whitening the data and emphasizing primarily the channel effects. The PHAT has a peak at the relative time delay of the two signals. This is typically used for estimating time-delay of arrival information. However, in practice, the function also includes a number of other peaks resulting from interactions between the reverberant impulses of the two channels. Using these additional local maxima judiciously, it is possible to evaluate information for the channel itself.

The phase transform of two observed signals $x_i[n]$ and $x_j[n]$ is computed from:

$$\phi\{x_i, x_j, n\} \;=\; \mathcal{F}^{-1}\left\{\frac{X_i(\omega)X_j^*(\omega)}{|X_i(\omega)||X_j(\omega)|}\right\}.$$

Using $X_i(\omega) = H_i(\omega)S(\omega) + U_i(\omega)$, ignoring the effects of the additive noise, and assuming that the channel responses are spectrally flat leads to:

$$\begin{aligned}\phi\{x_i, x_j, n\} &= \mathcal{F}^{-1}\left\{H_i(\omega)H_j^*(\omega)\right\} \\ &= \mathcal{F}^{-1}\{1 + R_i(\omega) + R_j^*(\omega) + R_i(\omega)R_j^*(\omega)\}.\end{aligned}$$

We can now average this result over all channels $j \neq i$ for a fixed channel i to obtain:

$$\begin{aligned}\bar{\phi}\{x_i, n\} &= \frac{1}{I-1}\sum_{j\neq i}\phi\{x_i, x_j, n\} \\ &= h_i[n] + \frac{1}{I-1}\mathcal{F}^{-1}\{\sum_{j\neq i}\left(R_j^*(\omega) + R_i(\omega)R_j^*(\omega)\right)\}.\end{aligned}$$

which is composed of the desired impulse response, $h_i[n]$, and attenuated versions of the time-reversed responses of the remaining channels as well as a number of attenuated cross terms. This suggests that the predominant peaks of $\bar{\phi}\{x_i, n\}$ will correspond to the multipath impulses of channel i. In practice, we approximate $h_i[n]$ from a clipped version of $\bar{\phi}\{x_i, n\}$.

The above procedure generates an independent estimate of the channel responses for each analysis frame. Assuming a very stable source and acoustic environment, the accuracy of the estimate can be increased by averaging over multiple analysis frames.

2.2. Matched Filtering-Based Weighting

We now show how the coarse estimate of the individual channel responses detailed above may be used to derive the clean speech

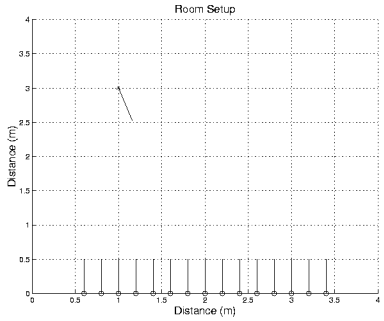

Fig. 2. Top view of the simulated room with microphone and source locations/orientations.

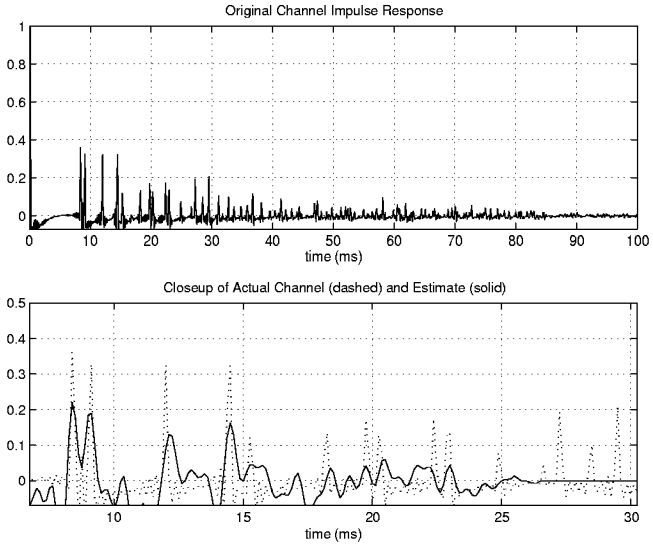

Fig. 3. Results of channel estimation procedure.

excitation signal. The channel residuals, $r_i[n]$, are noiselike in nature, roughly consisting of a set of seemingly random impulses. A matched filtering process is employed to identify which of these impulses are likely to have been present as part of the clean speech's excitation sequence.

The procedure is motivated by the observation that each impulse in the clean speech excitation is manifested in the residuals as a shifted and scaled version of the appropriate channel impulse response. The individual channel residuals are filtered with $m_i[n]$, a sequence constructed from a time reversed and normalized version of the channel estimate with the initial impulse at time $n = 0$ removed, i.e.

$$m_i[n] = \frac{h[-n] - h[0]\delta[n]}{\sum_n |h[n]| - |h[0]|}$$

Accordingly, any impulse in the residual corresponding to a relatively large value in this matched filter result is deemed to have been present in the clean speech excitation. A small value is indicative of impulsive energy due to reverberation effects. The modified residual, $\hat{r}_i[n]$, is then generated by multiplying the reverberant residual on a sample-by-sample basis by a weighting

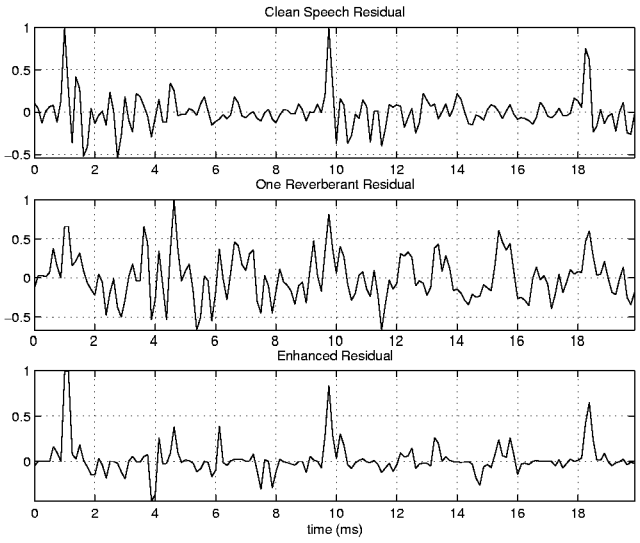

Fig. 4. Comparison of the clean speech residual (top), reverberant channel residual (middle), and the residual resulting from proposed algorithm (bottom).

function set proportional to the matched filter result. Specifically,

$$\hat{r}_i[n] = (|m_i[n] * r_i[n]|)^{\alpha} r_i[n]$$

where $\alpha = 1$ is typical.

3. RESULTS

The results in this section are based on a room with a reverberation time of 100ms simulated using the image model technique [9]. Figure 2 shows a top view of the $4 \times 4 \times 3$ meters enclosure. There are 15 microphones uniformly spaced in a linear array along one wall at a height of 1.5m. The source was located 3m from the array and displaced 1m from the center of the room at a location of $(x, y, z) = (1, 3, 1.5)$. Both the microphones and source are modeled with cardioid reception/radiation patterns. Their orientation angles are indicated in the figure.

Figure 3 shows the result of the channel estimation procedure for a single channel. The top plot shows the entire simulated channel response for one microphone. Because all channels were time-aligned before their estimation, each will have an impulse at time $n = 0$ corresponding to the direct path impulse. The remaining impulses are entirely due to multipath reflections. The bottom plot compares the channel estimate (solid line) to the actual channel impulse response (dotted line). The estimate provides a reasonable approximation for the initial reverberant impulses.

Figure 4 compares the residual signal associated with a 20ms segment of clean speech to that of the reverberant and enhanced residuals for a single channel. This is a portion of voiced speech. The clean residual is dominated by the glottal onset impulses at roughly 1ms, 10ms, and 18ms. The reverberant residual contains significant energy throughout the segment. The enhanced residual has deemphasized those portions of the segment dominated by reverberation effects. Figure 5 shows the corresponding original 20ms speech segment, the single channel of reverberant speech,

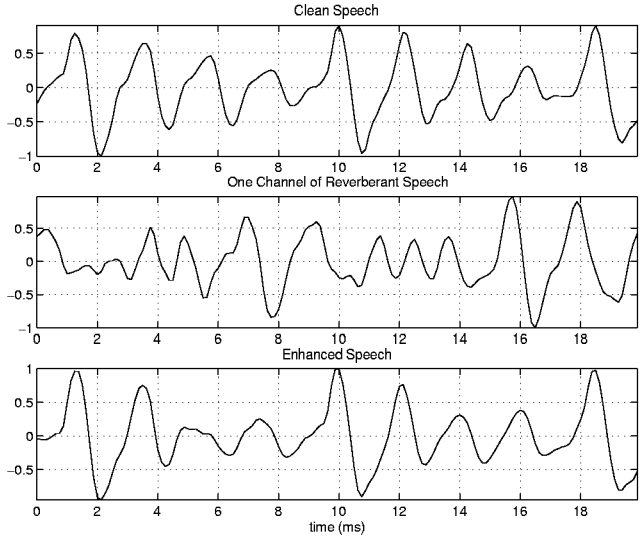

Fig. 5. Comparison of the clean speech (top), single channel of reverberant speech (middle), and speech resulting from proposed algorithm (bottom).

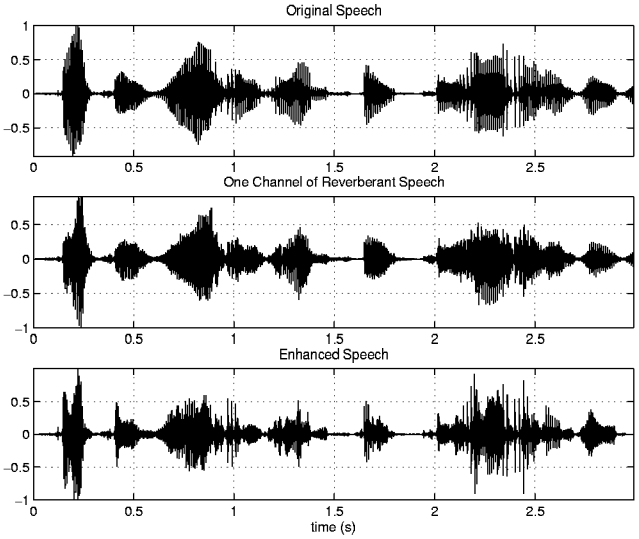

Fig. 6. Original, reverberant, and enhanced speech result for a 3s utterance.

and the enhanced speech signal generated by the proposed algorithm. The enhanced speech clearly represents a marked improvement over its reverberant counterpart over this short-term interval.

Finally, Figure 6 compares the original, reverberant, and enhanced speech for a 3s utterance. The visible effect of convolving the clean speech with the reverberant channel is to introduce a "smearing" in time. As is apparent from these plots, the enhancement of the speech undoes much of this degradation.

4. CONCLUSION

This paper has presented a method for multi-channel speech dereverberation which incorporates a specific model for the reverberant channels. This represents an extension of our earlier work which suppressed environmental effects by focusing on the known properties of the desired speech and without the necessity of any channel information. While this prior approach is advantageous in that it obviates the need for any channel estimation, it does not exploit what limited knowledge of the room impulse responses may be determined. In this work, we have detailed a means to produce a coarse estimate of the channel responses on a short-term basis and demonstrated their utility for speech enhancement. Ideally, this procedure will be used in conjunction with the speech model-based enhancement schemes detailed in [7, 8] to effectively combine the knowledge of both the channel and desired speech into a microphone array context.

5. REFERENCES

[1] S. Subramaniam, A. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Proc.*, vol. 4, no. 5, pp. 392–396, September 1996.

[2] J. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, no. 1-2, pp. 207–222, 1993.

[3] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Proc.*, vol. 5, no. 5, pp. 425–437, September 1997.

[4] B. Radlovic, R. Williamson, and R. Kennedy, "On the poor robustness of sound equalization in reverberant environments," in *ICASSP99*, Phoenix, AZ, March 15-19 1999, IEEE, pp. 881–884.

[5] M. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *ICASSP98*, Seattle, WA, May 12-15 1998, IEEE, pp. 3613–3616.

[6] M. Brandstein, "An event-based method for microphone array speech enhancement," in *ICASSP99*, Phoenix, AZ, March 15-19 1999, IEEE, pp. 953–956.

[7] S. Griebel and M. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverbearation," in *IEEE Workshop on Acoustic Echo and Noise Control*, Pocono Manor, Pennsylvania, September 27-30 1999.

[8] M. Brandstein and S. Griebel, "Nonlinear, model-based microphone array speech enhancement," in *Acoustic Signal Processing for Telecommunication*, S. Gay and J. Benesty, Eds., pp. 261–279. Kluwer, 2000.

[9] 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, April 1979.

[10] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, August 1976.