# AN EVENT-BASED METHOD FOR MICROPHONE ARRAY SPEECH ENHANCEMENT

*Michael S. Brandstein*

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

## ABSTRACT

This paper presents the Multi-Channel Multi-Pulse (MCMP) algorithm for the enhancement of speech degraded by reverberations and additive noise. The enhanced speech is synthesized from a sequence of impulses exciting a linear predictive filter. The excitation signal is computed from a nonlinear process which uses impulse clustering of the multi-channel speech data to discriminate portions of the linear prediction residual produced by the desired speech signal from those due to multipath effects and uncorrelated noise. The MCMP algorithm is shown to be capable of identifying and attenuating reverberant portions of the speech signal as well as reducing the effects of additive noise.

## 1. INTRODUCTION

There is a direct relationship between the effectiveness, functionality, and ease of use of a human-computer interface and the capabilities of the speech acquisition system upon which it depends. Currently, most speech acquisition systems require the user to be close to a microphone if reasonable sound quality is to be achieved. This close-talker condition significantly simplifies the acquisition problem by emphasizing the desired signal relative to background noise and other sources as well as by greatly reducing physical channel effects on the signal of interest. However, there are many situations in which it is neither possible nor desirable to have a talker or talkers physically linked to the acquisition device. This distant-talker condition introduces a number of signal degradations not present in the close-talker scenario; principle among these are reverberations and additive noise.

The dereverberation of speech has proven to be a very difficult problem. Since the distortion effects are convolutional and highly nonstationary, traditional speech enhancement methods designed for additive noise uncorrelated with the signal of interest are not applicable. While single-channel dereverberation techniques are available, [1] for example, multi-channel methods or microphone arrays offer an advantageous approach to the problem by virtue of their spatial filtering potential. Beamforming methods are effective at attenuating long term echoes [2] which tend to be uncorrelated across channels, but do little to reduce short term effects. More sophisticated approaches attempt to identify the channel effects in some form and compensate for them. These include cepstral processing [3], matched filtering [4], and adaptive sub-space filtering [5].

The channel effects in even a simple enclosure are very sophisticated and quickly time-varying. In [6] we address this point and argue that any system which attempts to estimate the reverberation effects and apply some means of inverse filtering would have to be adaptable on almost a frame-by-frame basis to be effective. However, the temporal averaging required by these processes prohibits adaptation at such a high rate. This imposes a fundamental limit on the effectiveness of this class of dereverberation approaches.

As an alternative to these filtering and least-squares methods, in [6] we proposed the incorporation of speech modeling into the beamforming process and illustrated the benefits of this paradigm with a frequency-domain method. Such an approach was shown to mitigate the reverberation effects without explicitly identifying the channel. Here we take this approach a step further utilizing a time-domain model based on LPC processing. As will be shown, it is possible to utilize some highly nonlinear, but still intuitive, techniques to suppress the deleterious effects of both reverberations and additive noise.

## 2. THE MULTI-CHANNEL MULTI-PULSE ALGORITHM

The Multi-Pulse Linear Predictive Coding (MPLPC) model [7, 8] represents an extension upon the traditional LPC speech synthesis method. Unlike the traditional LPC approach which employs a binary (voiced/unvoiced) excitation signal as the input to an all-pole filter, the MPLPC model represents the excitation sequence as a sum of weighted and delayed impulses. The impulse parameters are estimated through an analysis-by-synthesis technique which minimizes a perceptually motivated error criterion. With the inclusion of a long-term predictor to exploit pitch structure and truncating the number of impulses per frame to a handful, the MPLPC model is capable of producing very good quality synthesized speech for coding rates as low as 9.6kbps [9].

While MPLPC was developed as an efficient means of parameterizing and reproducing close-talker speech signals, it is adapted here for the multi-channel speech enhancement problem. The algorithm, termed Multi-Channel Multi-Pulse (MCMP), 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 excitation sequence, not the all-pole

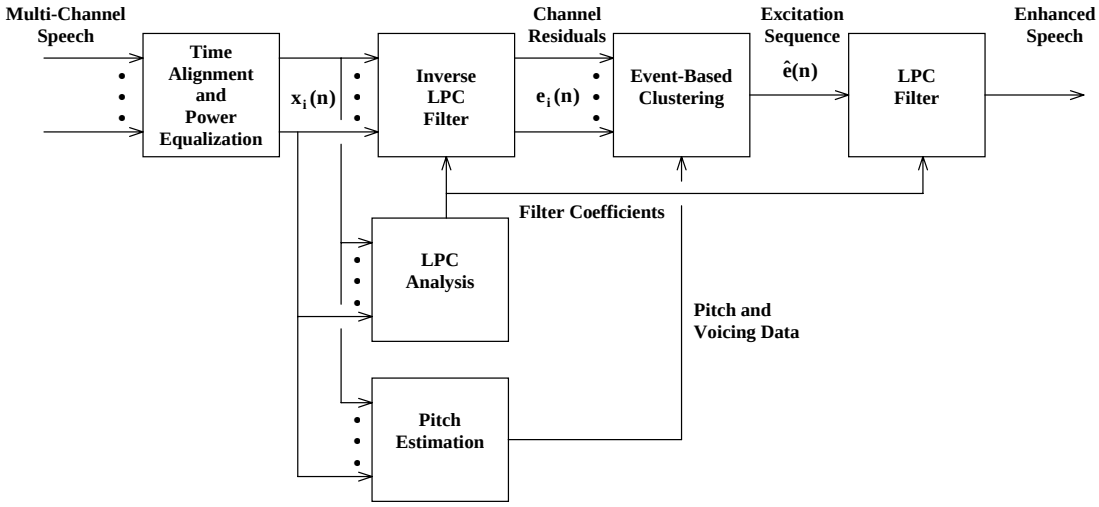


**Figure 1. Outline of the Multi-Channel Multi-Pulse (MCMP) Algorithm**

LPC filter. Furthermore, it is 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. As will be seen, these assumptions are appropriate in practice. Essentially, the approach will be to identify the MPLPC clean speech excitation sequence from a set of corrupted excitation signals.

The MCMP algorithm is outlined in Figure 1. The time-aligned and power-equalized channel signals, $x_i(n)$, from $N$ microphone channels are applied to LPC analysis and pitch estimation. Using a 30ms analysis window the $12^{\text{th}}$ order LPC filter is estimated for the frame of data using a multi-channel version of the autocorrelation method. The autocorrelation function associated with the frame is found from:

$$r(k) = \frac{1}{N}\sum_{i=1}^{N}\sum_{n=k}^{L} x_i(n)x_i(n-k)$$

where $L$ is the analysis window length. Once estimated, this joint LPC filter is used to generate $N$ residual signals, $e_i(n)$ for $1 \le i \le N$. The pitch period associated with the analysis frame is computed using an autocorrelation-based estimation scheme [10] employing $r(k)$ as given above.

The analysis-by-synthesis approach to excitation estimation employed by the traditional MPLPC algorithm is inappropriate in this context given the lack of a known desired signal. Instead, impulses in the enhanced excitation sequence are estimated from the minima of a clustering error criterion developed from the set of residual signals. The $N$ residuals are first lowpass filtered to 1kHz. For each channel the set of local minima and maxima, $\{n_{ij}, p_{ij}\}$, are then identified. Here $n_{ij}$ is the time index of the $j^{th}$ extrema for the $i^{th}$ channel and $p_{ij}$ is the corresponding extrema value.

For each time sample $n$ of the frame, the following clustering criterion is computed:

$$E(n) = \frac{1}{NK}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\left((kP+n) - \min_{j}\{kP+n-n_{ij}\}\right)^2$$

This error criterion finds the average Euclidean distance from the time index $n$ to the closest extremum in each of the residual sequences. $E(n)$ is small for time indices where the local extrema are aligned across the $N$ channels (corresponding to the impulses associated with the original speech) and is large for indices with non-aligned impulses. For additional robustness the criterion is estimated over $K$ pitch periods, where $P$ is the pitch estimated from the multi-channel data. For voiced segments, a value of $K$ equals 2 or 3 is effective, while for unvoiced segments $K$ is set to 1, effectively disabling the pitch feature.

The time indices of the synthesis impulses, $\hat{n}_j$, are found from the local minima of $E(n)$. The corresponding impulse values, $\hat{p}_j$, are calculated from the average of the $p_{ij}$ associated with the closest extrema to $\hat{n}_j$ and then inversely weighted using the tanh function of the value $E(\hat{n}_j)$ relative to the set of minima values of $E(n)$. Specifically,

$$\hat{p}_j = \frac{w(j)}{NK}\sum_{k=0}^{K-1}\sum_{i=1}^{N} p_{i\hat{l}}$$

where

$$\hat{l} = \arg\min_{l}\{kP + \hat{n}_j - n_{il}\}$$

$$w(j) = \frac{1}{2}\left[1 - \tanh\left(\frac{\frac{E(\hat{n}_j)-E_{min}}{E_{max}} - b}{a}\right)\right]$$

Here $E_{min}$ and $E_{max}$ are the extreme values of the local minima of $E(n)$. The parameters $a$ and $b$ are dependent on the environmental conditions. For the simulations presented in the next section, values of $a = .1$ and $b = .3$ were incorporated.

Finally, the resulting multi-pulse excitation sequence:

$$\hat{e}(n) = \sum_{j}\hat{p}_j\delta(n-\hat{n}_j)$$

excites the estimated LPC filter to produce the enhanced speech segment.

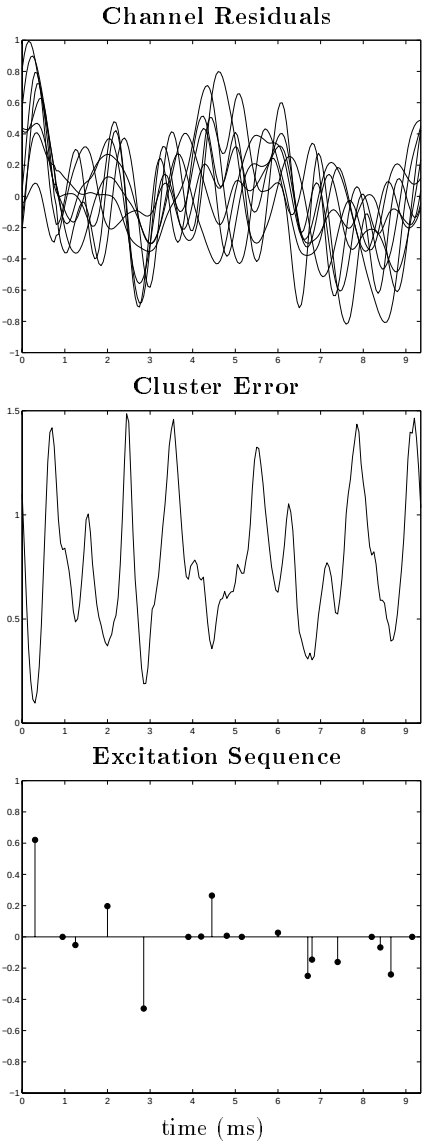


**Figure 2. Illustration of the Clustering Procedure**

Figure 2 illustrates the clustering procedure. The plots show the channel residuals $e_i(n)$ for a reverberant speech segment, the computed clustering criterion $E(n)$, and the corresponding multi-pulse excitation sequence derived from this process. For reference in the simulations to follow, this 9.5ms segment of speech corresponds to one period of the reverberated voiced sequence in Figure 3.

## 3. SIMULATION

A source is simulated in the center of a 4m x 4m x 3m rectangular room. The enclosure is assumed to have plane reflective surfaces and uniform, frequency-independent reflection coefficients. Room impulse responses are generated for 8 microphones with 25cm spacing positioned along one wall of the enclosure using the image model technique [11] with intra-sample interpolation and up to sixth order reflections. Both the microphones and sources are assumed to have cardioid patterns. The sources are oriented toward the center of the array. Additive white Gaussian noise is included in each channel response. The received signals' SNR are 20dB measured relative to the noiseless reverberated channels.

Figure 3 illustrates the results. A 30ms segment of 20kHz sampled voiced speech shown in plot (A) is subjected to a 200ms reverberation time multipath condition in the simulated enclosure detailed above. The reverberated and noise-added signal associated with a single channel is shown in plot (B) and the result of delay and sum beamforming with all 8 channels is given in plot (C). Plot (D) shows the synthetic speech derived from the MCMP algorithm. Observing the first 9.5ms period of this segment, it is apparent that the primary effect of the multipath distortion is to inflate the signal energy between 6ms and 9.5ms. The beamformer is ineffective at attenuating these distortions. The MCMP algorithm is capable of detecting the excitation energy due to the clean speech and produces a synthesized signal very similar to the original. In addition to reducing the effects of reverberation, this model-based synthesis procedure significantly diminishes the additive noise.

## 4. DISCUSSION

By focusing on event-based data and using a highly nonlinear filtering process, the MCMP algorithm is capable of discriminating impulses of the LP residual generated by the desired speech signal from those brought about by multipath echos and uncorrelated noise. The enhanced speech derived from an LP synthesis with the clean excitation sequence demonstrates a robustness to environmental reverberations and additive noise.

These principles may be extended to address the case of interfering sources. Essentially, the impulse events may be associated with one or more source locations by evaluating their relative delays across the channels given knowledge of the microphone placements. Once this association is performed, the individual speech signals may be independently reconstructed through the synthesis procedure outlined above. This approach to speaker isolation represents a distinct contrast to the spatial filtering paradigm which relies on very specific knowledge and assumptions regarding talkers' locations and radiation patterns to generate appropriate channel weightings. This event-based procedure, by virtue of its underlying speech model and exploitation of impulse data alone, has the potential to be much less sensitive to environmental uncertainties (e.g. imperfect source localization, non-ideal radiator effects, unknown channels).

## REFERENCES

[1] B. Yegnanarayana, P. Murthy, C. Avendano, and H. Hermansky. Enhancement of reverberant speech using lp residual. In *ICASSP98*, pages 405–408, Seattle, WA, May 12-15 1998. IEEE.

[2] J. Allen, D. Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *J. Acoust. Soc. Am.*, Vol.62(4):912–915, 1977.

[3] S. Subramaniam, A. Petropulu, and C. Wendt. Cepstrum-based deconvolution for speech dereverber-

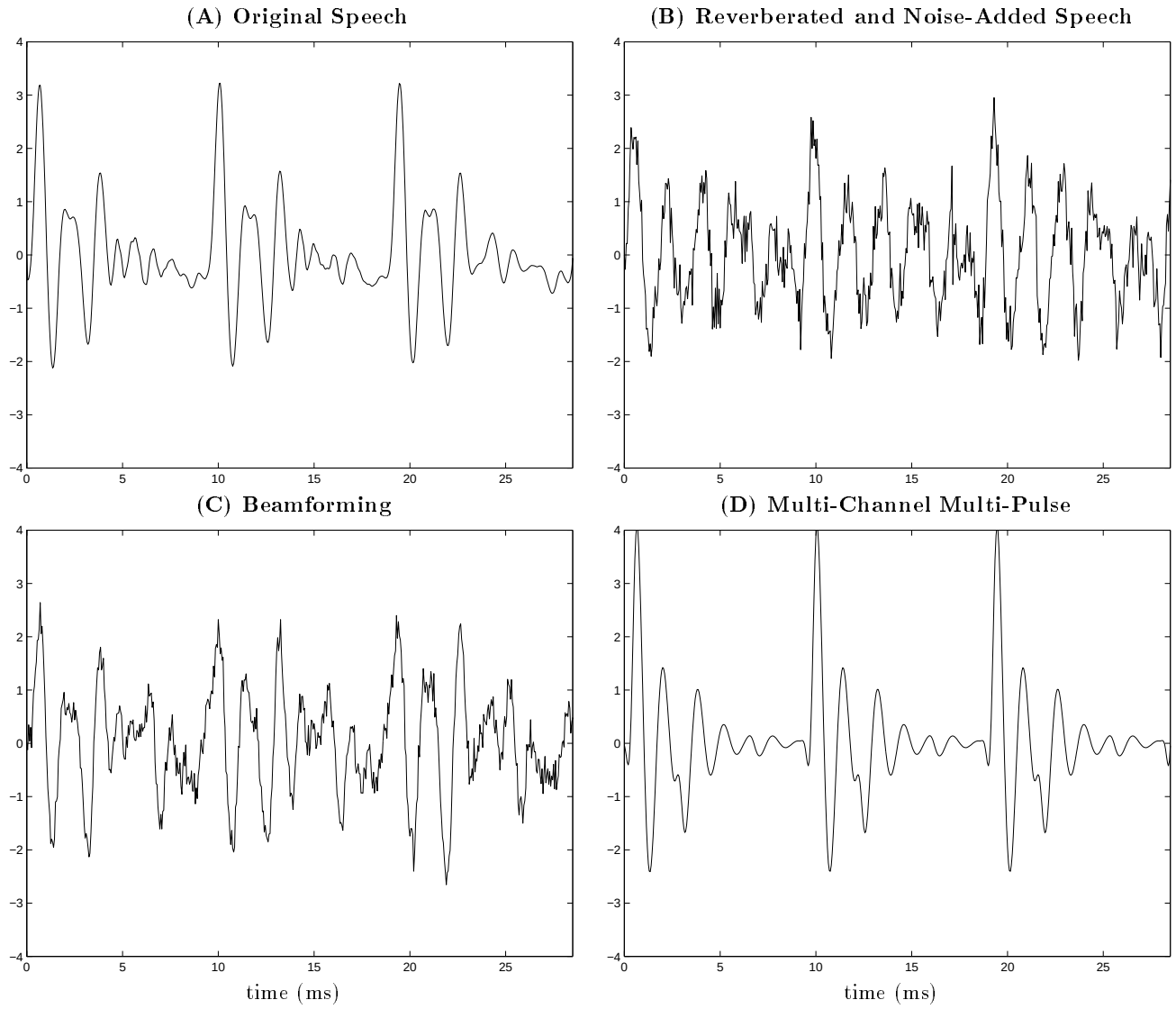


**Figure 3. Simulation Results: A reverberated speech segment enhanced via beamforming and the proposed multi-channel multi-pulse LPC method.**

ation. *IEEE Trans. Speech Audio Proc.*, 4(5):392–396, September 1996.

[4] J. Flanagan, A. Surendran, and E. Jan. Spatially selective sound capture for speech and audio processing. *Speech Communication*, 13(1-2):207–222, 1993.

[5] S. Affes and Y. Grenier. A signal subspace tracking algorithm for microphone array processing of speech. *IEEE Trans. Speech Audio Proc.*, 5(5):425–437, September 1997.

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

[7] B. S. Atal and J. R. Remde. A new model of lpc excitation for producing natural-sounding speech at low bit rates. In *Proceedings of ICASSP82*, pages 614–617. IEEE, 1982.

[8] S. Singhal and B. S. Atal. Improving performance of multi-pulse lpc coders at low bit rates. In *Proceedings of ICASSP84*, pages I–131–I–134. IEEE, 1984.

[9] J. Deller, J. Proakis, and J. Hansen. *Discrete-Time Processing of Speech Signals*. Prentice Hall, first edition, 1987.

[10] J. Wise, J. Capiro, and T. Parks. Maximum likelihood pitch estimation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24(5):418–423, October 1976.

[11] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, April 1979.