

WAVELET TRANSFORM EXTREMA CLUSTERING FOR MULTI-CHANNEL SPEECH DEREVERBERATION

Scott Griebel and Michael Brandstein

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

ABSTRACT

This paper presents a method for enhancing multi-channel reverberant speech using event-based processing of wavelet transform coefficients. Clustering of the wavelet extrema across multiple channels is employed to obtain a single multi-scale extrema representation from which the enhanced signal is synthesized. Processing is done in the LPC residual domain, with the entire analysis being preceded by a multi-channel LPC inverse filter and followed by the corresponding forward LPC filter. The algorithm is compared to traditional delay-and-sum beamforming with results presented for reverberant, noisy conditions.

1. INTRODUCTION

In many real-world situations, speech signals are degraded by the reverberant effects of the environment in which they are produced. Traditional speech processing algorithms are designed to operate primarily under conditions where the noise is additive and uncorrelated with the desired speech signal. Their effectiveness is dramatically reduced in the presence of the non-stationary, convolutional noise inherent in these reverberant settings. A number of methods have been developed specifically to address the dereverberation problem. Beamforming methods are effective at attenuating long term echoes [1] 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 [2], matched filtering [3], and adaptive sub-space filtering [4]. However, the environmental degradations in even a simple enclosure are very sophisticated and quickly time-varying, and the temporal averaging required by these processes prohibits adaptation at a rate sufficient to track and undo these effects. This imposes a fundamental limit on the effectiveness of this class of dereverberation techniques.

In [5] we proposed an alternative to these methods by explicitly incorporating a speech model into the beamforming process. Our work in [6] extended this idea with an event-based processing method to discriminate impulses in the received signals due to channel effects from those present in the desired speech. These approaches were shown to suppress the deleterious effects of both reverberations and additive noise without explicitly identifying the channel and to be adaptable on a frame by frame basis. However, this earlier work was restricted to a small class of speech signals and limited to individual analysis frames. Here we take these concepts a step further by employing the wavelet domain for a multi-resolution signal analysis and reconstruction. This allows the model-based, event-based processing paradigm to be applied for effective speech enhancement under general conditions.

2. WAVELET TRANSFORM EXTREMA CLUSTERING

The reverberant speech signal, $x_i[n]$, observed at the i^{th} microphone ($i = 1, 2, \dots, I$) can be modeled as:

$$x_i[n] = h_i[n] * s[n] + u_i[n]$$

where $s[n]$ is the clean speech utterance, $u_i[n]$ is noise, and $h_i[n]$ is the room impulse response between the speech source and the i^{th} microphone.

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. Essentially, the approach will be to identify the clean speech excitation sequence from a set of corrupted excitation signals and then reconstruct the enhanced speech with only the desired sequence.

This method was applied in [6] using the LPC residual derived from I microphone channels. In this work estimation of the residual impulses is carried out more effectively by employing a class of wavelets to decompose the LPC residuals. In [7], quadratic spline wavelets were shown to be effective at detecting discontinuities in a signal. Because the quadratic spline wavelets are the derivatives of smoothing functions, significant wavelet transform (WVT) coefficients, or *wavelet extrema*, prove to be appropriate indicators of discontinuities in a signal. The extrema at each wavelet scale are an effective representation of the impulses in the received speech. By locating the extrema which are well clustered across all channels, we attempt to capture the underlying impulsive structure of the original non-reverberant speech. The estimated ‘clean’ multi-scale extrema can be used with an alternating projection algorithm to reconstruct the desired signal excitation with minimal approximation error.

The overall procedure is outlined in Figure 1. The received signals, $x_i[n]$, are first time-aligned using *a priori* knowledge of the source and microphone geometry. A 13th order joint LPC filter is then derived from independent N -point windows of the I channels of speech. This multi-channel LPC filter is estimated using the joint autocorrelation function:

$$r[k] = \frac{1}{I} \sum_{i=1}^I \sum_{n=k}^N x_i[n] x_i[n-k]$$

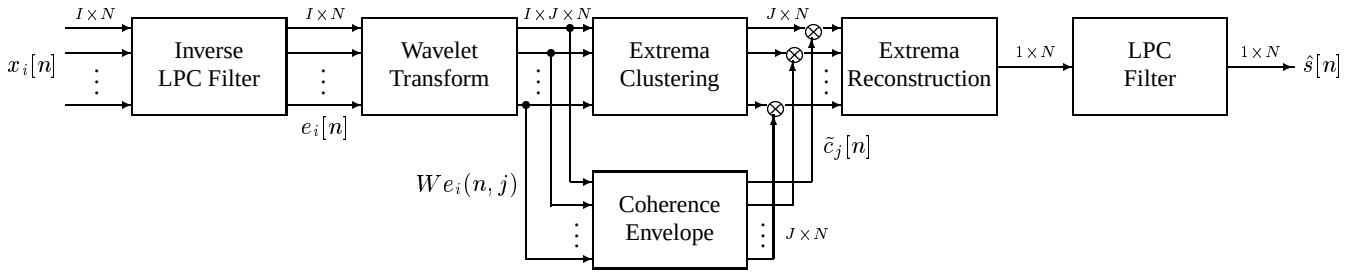


Figure 1: Outline of Wavelet Transform Extrema Clustering Algorithm.

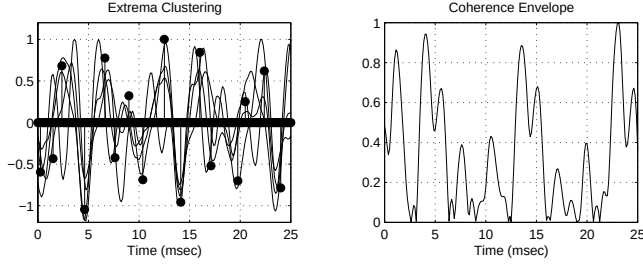


Figure 2: Clustering Results and Coherence Envelope.

and then applied to generate a set of residual signals, $e_i[n]$. For the remainder of this paper, the analysis frame length is $N = 200$ samples at 8kHz sampling (25 msec) and the number of channels is $I = 8$. The first J scales of the WVT, $We_i(n, j)$, for each of the I residual signals are then calculated. The WVT of a signal $e_i[n]$ is denoted as $We_i(n, j)$ where j is the scale parameter and n is the temporal index within that scale.

Based upon the observation that residual impulses due to the original speech tend to be clustered across channels and scales while those due to reverberation effects are relatively uncorrelated in both amplitude and time, a clustering technique is now applied to find the time indices associated with the ‘clean’ extrema. The clustering method for a given multi-channel wavelet scale, j , consists of the following steps:

1. For each channel, i , locate the extrema indices, $n_{i,k}$.
2. For each channel, form signals $y_i[n]$ such that

$$y_i[n] = \begin{cases} We_i(n, j) & n = n_{i,k} \\ 0 & \text{otherwise} \end{cases}$$

3. For each channel, form filtered extrema signals given by $\tilde{y}_i[n] = y_i[n] * g_j[n]$ where $g_j[n]$ is a Gaussian window with width dependent on the WVT scale.
4. Average the filtered extrema signals across channels to form $\bar{y}[n]$.
5. The extrema of $\bar{y}[n]$ are the ‘clustered extrema’.

Figure 2 shows the results of a typical clustering operation which has been performed for scales 1 through J ($J = 5$).

To further deemphasize clustered peaks due to room reverberations, the following short-term coherence window is applied at each scale j :

$$c_j[n] = \frac{2}{I-1} \frac{\sum_{i=1}^I \sum_{k=i+1}^I \sum_{l=-L}^L We_i(n+l, j) We_k(n+l, j)}{\sum_{i=1}^I \sum_{l=-L}^L We_i(n+l, j)^2}$$

where $2L + 1 \ll N$ is the length of a small subwindow within the analysis frame. With this window, any clustered extrema in areas of low coherence will be deemphasized, while highly coherent extrema will be relatively unaffected. Although the coherence window can assume values between -1 and 1 , the following thresholding scheme is used to form a new coherence window:

$$\tilde{c}_j[n] = \begin{cases} 1 & c_j[n] > T \\ 0 & c_j[n] < 0 \\ c_j[n] & \text{otherwise.} \end{cases}$$

This new coherence window is then lowpass filtered to remove any sharp discontinuities the threshold T might have caused. The value of T was determined experimentally to be between 0.3 and 0.5. Figure 2 shows a typical coherence envelope used for weighting the clustered WVT extrema.

After multiplying the clustered peaks for each wavelet scale by the coherence envelope for that scale, the enhanced LPC residual is reconstructed using the alternating projection algorithm in [7]. Figure 3 shows the LPC residual of clean and reverberant speech frames. Also shown is the enhanced LPC residual signal resulting from the WVT reconstruction algorithm. For each frame, this enhanced residual signal is reconstructed and passed through the forward LPC filter to generate the enhanced speech signal, $\hat{s}[n]$. The frames are then multiplied by a Hanning window and combined using an overlap-add method.

3. SIMULATIONS

The reverberant speech segments were generated using the image model of [8] with intra-sample interpolation and up to sixth order reflections. The room was a $4\text{m} \times 4\text{m} \times 3\text{m}$ rectangular enclosure assumed to have plane reflective surfaces and uniform, frequency-independent reflection coefficients. A total of eight microphones were used: one pair on 3 of the walls and another pair on the ceiling. The source was located near the wall containing no microphones. Both the microphones and sources were assumed to have cardioid patterns. A reverberation time of 200 msec was used. This corresponds to a wall reflection coefficient of 0.78. Uncorrelated, white Gaussian noise, $u_i[n]$, was added to each channel so that the SNR relative to the original *non-reverberant* utterance was -10dB.

Figures 4 and 5 show the results of using the proposed algorithm on the reverberant channels. Figure 4 displays a single analysis frame (25 ms) corresponding to a little over one period of a voiced speech segment. The four plots compare the reconstructed speech signal to the original speech utterance, a single reverberant channel, and a simple delay-and-sum beamforming of the I reverberant speech channels. It is apparent that the primary effect of the

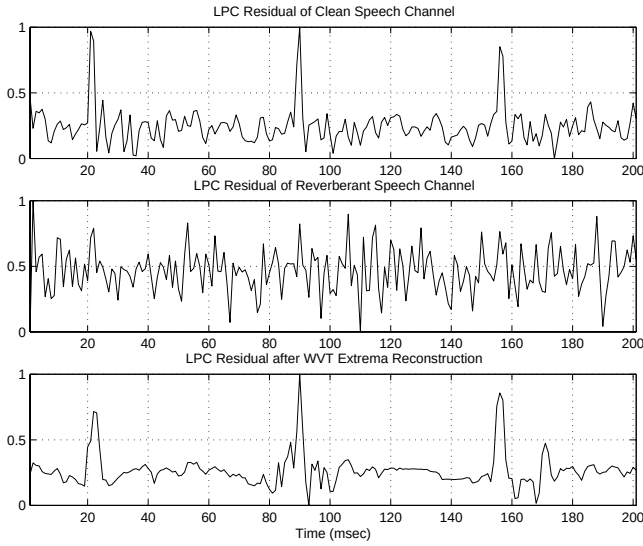


Figure 3: *LPC Residual After Beamforming and After Wavelet Clustering Technique.*

multipath distortion is to inflate the signal energy between 987 ms and 997 ms. The beamformer is ineffective at attenuating these distortions. The proposed 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 wavelet extrema-based synthesis procedure significantly diminishes the additive noise.

To illustrate the applicability of the proposed algorithm to continuous speech Figure 5 offers a similar comparison for a three second utterance. In this figure, the ability of the proposed algorithm to remove both long and short term reverberation effects as well as additive noise is apparent. Listening tests confirm the efficacy of the procedure. The synthesized speech lacks the audible reverberation and noise effects which are clearly present in both the degraded and beamformed speech.

4. DISCUSSION

By focusing on event-based data and a wavelet-domain analysis, the proposed algorithm is capable of discriminating impulses of the excitation residual generated by the desired speech signal from those brought about by multipath echos and uncorrelated noise. The enhanced speech derived from an all-pole 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 and unknown source locations. 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

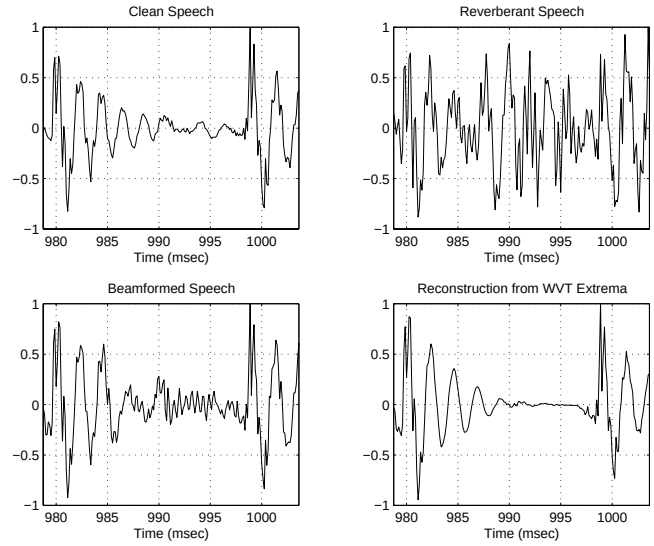


Figure 4: *Comparison of Clean, Reverberant, Beamformed, and WVT Extrema Reconstructed Speech*

to environmental uncertainties (e.g. imperfect source localization, non-ideal radiator effects, unknown channels).

5. REFERENCES

- [1] J. Allen, D. Berkeley, and J. Blauert, "Multimicrophone signal-processing techniques to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.
- [2] 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.
- [3] 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.
- [4] 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.
- [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. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 710–732, July 1992.
- [8] J. Allen and D. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 64, no. 4, pp. 943–950, 1979.

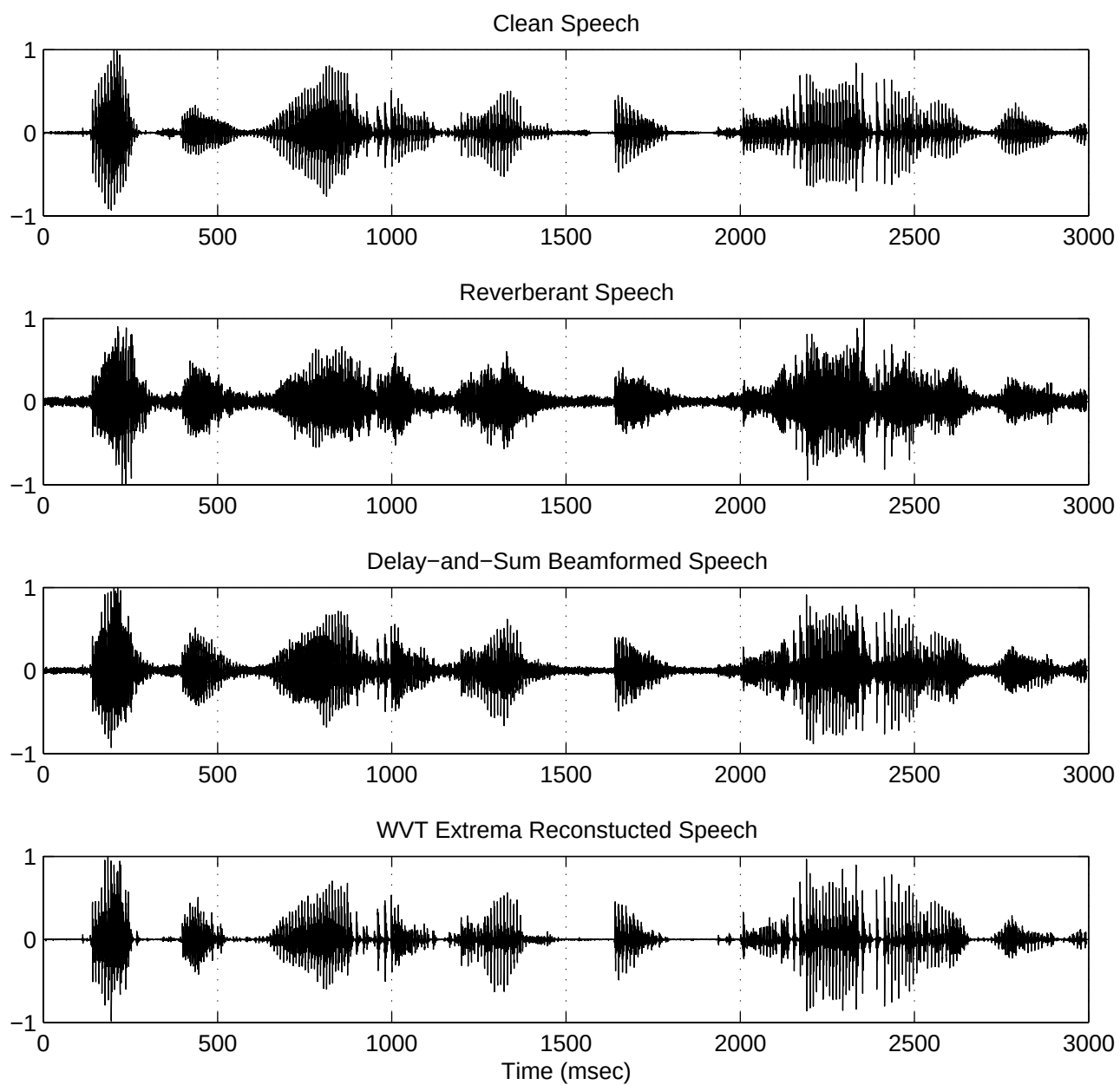


Figure 5: Comparison of Clean, Reverberant, Beamformed, and WVT Extrema Reconstructed Speech