
THEORY AND APPLICATIONS OF ACOUSTIC SIGNAL PROCESSING FOR TELECOMMUNICATION

Chapter 2

NONLINEAR, MODEL-BASED MICROPHONE ARRAY SPEECH ENHANCEMENT

Michael S. Brandstein

Division of Engineering and Applied Sciences

Harvard University

msb@hrl.harvard.edu

Scott M. Griebel

Division of Engineering and Applied Sciences

Harvard University

griebel@fas.harvard.edu

Abstract

In this chapter we address the limitations of current approaches to using microphone arrays for speech acquisition and advocate the development of multi-channel techniques which employ non-traditional processing and an explicit model of the speech signal. The goal is to combine the advantages of spatial filtering achieved through beamforming with knowledge of the desired time-series attributes and intuitive nonlinear processing. We then offer a multi-channel algorithm which incorporates these principles. The enhanced speech is synthesized using a linear predictive filter. The excitation signal is computed from a nonlinear wavelet-domain process. It uses extrema 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 algorithm is shown to be capable of identifying and attenuating reverberant portions of the speech signal and reducing the effects of additive noise.

Keywords: Microphone Arrays, Multi-Channel, Speech Enhancement, Noise Reduction, Dereverberation, Wavelet, Nonlinear, Model-Based

1. INTRODUCTION

Speech has become an increasingly vital component in modern human-machine interfaces. Background noise, reverberation effects, and competing signals produce aesthetically undesirable effects and diminish the system's ability to convey information across the interface. Signal quality also plays a key role in the performance of related analysis systems. For instance, speech recognition and speaker identification techniques exhibit dramatic performance degradations when unpristine data is employed.

There is a direct relationship between the ease of use, effectiveness, and functionality of these interfaces and the capabilities of the speech acquisition system upon which it depends. Currently, most speech acquisition systems rely on the user to be physically close to a device to achieve reasonable sound quality. 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 from the signal of interest. However, there are many situations in which it is not possible or desirable to have a talker or talkers physically linked to the acquisition device.

The ultimate goal of this work is to allow users the opportunity to roam unfettered in diverse environments while still providing a high quality speech signal and a robustness to background noise, interfering sources, and reverberation effects. Besides enhancing current human-machine interfaces (such as the speakerphone), advances in this technology are providing for a variety of novel approaches and new avenues for human-machine interaction. In recent years, the use of multiple sensors (microphone arrays) has received considerable attention as a means for dramatically improving the performance of traditional, single-microphone systems. This multi-sensor approach has opened up a wide range of additional topics of study [1, 2].

There is a great deal of potential for advancement in distant-talker speech acquisition research and a wealth of current and future technology dependent upon advances in this area. For single channel systems, recent research has focused on appropriate modeling of the speech signal, while multi-channel system research has emphasized improved spatial filtering. This work advocates an alternative approach to microphone array speech acquisition which incorporates nonlinear, model-based processing. By combining the advantages of multi-channel spatial filtering, non-traditional processing methods, and specific knowledge of the desired time-series attributes, this approach has the potential to sig-

nificantly improve the quality of speech obtained in these challenging environments.

The next section provides a short summary of the existing single- and multi-channel approaches to distant-talker speech enhancement and discusses their limitations. We then introduce the principles of the nonlinear, model-based paradigm and detail a specific example of its use along with some results illustrating its merit. Finally, some conclusions and future issues for investigation are given.

2. SPEECH ENHANCEMENT METHODS

There is a rich history of work addressing the use of single channel methods for speech enhancement. Summaries of these techniques may be found in texts on the subject, such as [3, 4, 5]. While capable of improving perceived quality in restrictive environments (additive noise, no multipath, high to moderate signal-to-noise ratio (SNR), single source), these approaches do not perform well in the face of reverberant distortions, competing sources, and severe noise conditions. In recent years, more sophisticated speech models have been applied to the enhancement problem. In addition to utilizing the periodic features of the speech, as in the case of comb filtering, these systems exploit the signal's mixture of harmonic and stochastic components [6, 7, 8]. Such model-based techniques offer an improved performance, both in speech quality and intelligibility. Additionally, these methods, by virtue of their specific parameterization of the speech signal, offer some applicability to the more general acquisition problem. Currently, however, these model-based estimation schemes have been limited to single channel applications.

By employing spatial filtering in addition to temporal processing, multi-channel systems (or microphone arrays) offer a distinct performance advantage over single-channel techniques in the presence of additive noise, interfering sources, and distortions due to multipath channel effects. (See [9] for a review of beamforming methods.) The simplest microphone array method is the Delay and Sum Beamformer which derives its output via an averaging of the time-synchronized microphone data. A variety of more sophisticated algorithms exists for adaptively 'steering' the array in the direction of the desired source and simultaneously adjusting the microphone 'weightings' to minimize the contributions of noise sources. These techniques usually assume the desired source is slowly varying and at a known location. While dynamic localization schemes, like those of [10], and weighting constraints may be incorporated into the adaptation procedure, these methods are very sensitive to steering errors which limit their noise source attenuation performance

and frequently distort or cancel the desired signal. Furthermore, these algorithms are oriented solely toward noise reduction and have limited effectiveness at enhancing a desired signal corrupted by reverberations.

Another class of algorithms attempts to enhance the speech signal by employing any of the beamforming methods described above followed by a post-filter. The post-processor may take the form of multi-channel Wiener filtering [11, 12, 13, 14], Spectral Subtraction [15, 16], or a combination of the two [17]. These algorithms are limited in their assumption that the additive noise is uncorrelated. Once again the emphasis here is additive noise reduction. The performance (and appropriateness) of these post-filters quickly degrades in the presence of multi-path distortions and coherent noise.

A variant approach is based upon attempting to undo the effects of multipath propagation. The channel responses themselves are in general not minimum phase and are thus non-invertible. By beamforming to the direct path and the major images, it is possible to use the multipath reflections constructively to increase SNR's well beyond those achieved with a single beamformer. The result is a matched filtering process [18] which is effective in enhancing the quality of reverberant speech and attenuating noise sources. Unfortunately, this technique has a number of practical shortcomings. The matched filter is derived from the source location-dependent room response and as such is difficult to estimate dynamically. The channel responses obtained in this manner do not address the issue of non-stationary or unknown source locations or changing acoustic environments. These problems are addressed by Affes and Grenier [19] in attempting to adaptively estimate the channel responses and incorporate the results into an adaptive beamforming process.

In general, beamforming research has dealt with algorithms to attenuate undesired sources and noise, track moving sources, and deconvolve channel effects. These approaches, while effective to some degree, are fundamentally limited by the nature of the distant talker environment. Array design methods are overly sensitive to variations in their assumptions regarding source locations and radiation patterns, and are inflexible to the complex and time-varying nature of the enclosure's acoustic field. Motion as little as a few centimeters or a talker turning his or her head is frequently sufficient to compromise the optimal behavior of these schemes in practical scenarios [21, 20]. Similarly, matched filter processing, while shown to be capable of tracking source motion to a limited degree, requires significant temporal averaging and is not adaptable at rates sufficient to effectively capture the motions of a realistic talker.

3. NONLINEAR, MODEL-BASED PROCESSING

Single-channel techniques exploit various features of the speech signal; multi-channel methods focus primarily on improving the quality of the spatial filtering process. While advances will continue to be made, these methods pursued independently have provided limited success. It is unlikely that such schemes in their present forms will achieve the stated goal of this research, namely to acquire a high-quality speech signal from an unconstrained talker in a hands-free environment surrounded by interfering sources (the “cocktail party” problem). In addition to the spatial filtering benefits of microphone arrays, an effective solution to this challenging problem will require the application of knowledge regarding the desired signal content. Furthermore, existing approaches to this problem have relied almost entirely on traditional linear filtering and least-squares methods. As will be shown, it is possible to utilize some highly nonlinear, but still intuitive, techniques to suppress the effects of reverberation.

In [20] we proposed an alternative to these methods by explicitly incorporating the Dual Excitation Speech Model [7] into the beam-forming process. Our work in [22] extended this idea using a multi-channel version of the Multi-Pulse Linear Predictive Coding (MPLPC) model [23, 24] and a nonlinear 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 adaptive 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 of the LPC residual. This allows the nonlinear, model-based processing paradigm to be applied for effective speech enhancement under general conditions. The proposed algorithm is now explained in detail.

4. A MULTI-CHANNEL SPEECH ENHANCEMENT ALGORITHM

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

$$x_i[n] = h_i[n] * s[n] + u_i[n] \quad (2.1)$$

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. Under all but ideal circumstances, the room impulse response is a complicated function of the enclosure environment. The noise term, $u_i[n]$, is assumed to be uncorrelated, both temporally and spatially.

A very general model for speech production approximates the vocal tract as a time-varying all-pole filter [4]. In the case of voiced speech, the filter excitation is modeled as a quasi-periodic impulse train where the average width between consecutive impulses is the pitch period. For unvoiced speech, the excitation signal is approximated by random noise. 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 signal from a set of corrupted excitations. The enhanced speech is then reconstructed by using the enhanced excitation signal as input to an estimate of the all-pole filter representing the vocal system.

The idea of enhancing the excitation signal was explored in [22] using the LPC residual derived from I microphone channels. The proposed algorithm represents a more effective method for estimating and then reconstructing the excitation signal by employing a class of wavelets to decompose the LPC residuals. In [25], quadratic spline wavelets were shown to be effective at detecting singularities in a signal. Because these 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 set of extrema at each wavelet scale is an effective representation of the impulses in the excitation signals. By locating the extrema which are well clustered across all channels, we attempt to capture the underlying impulsive structure (for voiced speech) of the original non-reverberant LPC residual signal. The estimated ‘clean’ multi-scale extrema can be used to reconstruct the desired excitation signal with minimal approximation error [25].

The overall procedure is outlined in Figure 2.1. The received signals, $x_i[n]$, are assumed to have been time-aligned either through *a priori* knowledge of the source and microphone geometry or as the result of applying a relative time-delay estimation procedure to the original channel data. The microphone signals are converted to LPC residuals and trans-

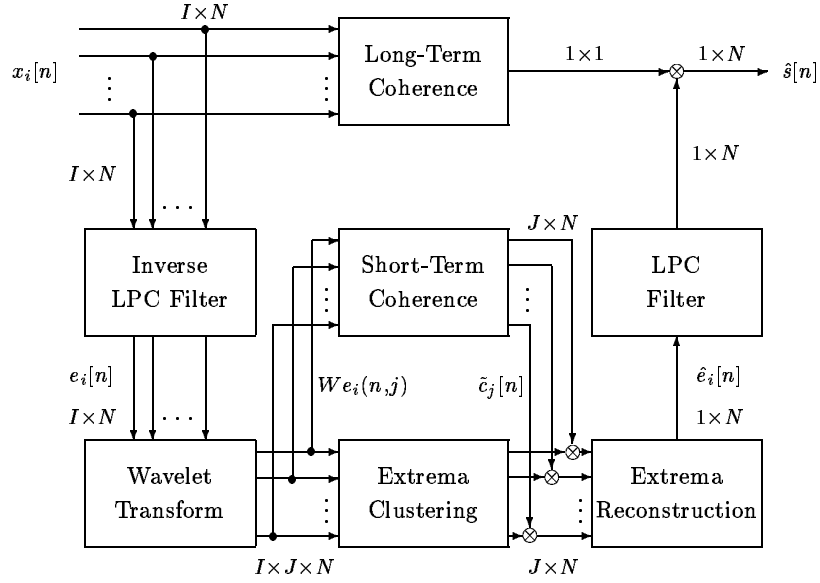


Figure 2.1 Outline of the Proposed Algorithm

formed to the wavelet domain. For each wavelet scale the multi-channel extrema are clustered and scaled by a local coherence measure. The resulting extrema at each scale are used to generate a set of synthetic wavelets which produce the enhanced excitation signal. This synthesized excitation (presumably with less of the detrimental reverberation and noise effects) is then fed into the estimated LPC filter. Finally, a global coherence measure is used to modulate the LPC filter output in an effort to reduce larger time-scale reverberation effects. The algorithm stages will now be discussed in detail. For the illustrative plots that follow in the remainder of this section we employ 400 msec reverberated speech with additive white Gaussian noise with SNR of 10 dB. Details of the simulation procedure are provided in the following section.

4.1 ALGORITHM DETAILS

Inverse LPC Filtering

A joint LPC filter is derived from N -point windows of the I channels of speech. The linear prediction formalism assumes the speech source signal can be modeled as:

$$s[n] = \sum_{p=1}^P a_p s[n-p] + e[n] \quad (2.2)$$

where $e[n]$ is the residual or excitation signal discussed previously. The LPC coefficients, $\{a_1, a_2, \dots, a_P\}$, are estimated by minimizing the average energy of the residual signal. For this multi-channel setting the criterion is expressed as:

$$\frac{1}{I} \sum_{i=1}^I \sum_{n=1}^N e_i^2[n] = \frac{1}{I} \sum_{i=1}^I \sum_{n=1}^N \left\{ s[n] - \sum_{p=1}^P a_p s[n-p] \right\}^2. \quad (2.3)$$

The minimum energy solution is achieved by solving a system of linear equations as in the single-channel case, but involves a multi-channel autocorrelation function, given by:

$$r[k] = \frac{1}{I} \sum_{i=1}^I \sum_{n=k}^N x_i[n] x_i[n-k]. \quad (2.4)$$

The resulting LPC coefficients can be used to generate a set of residual signals, $e_i[n]$, via inverse filtering. The goal of the following wavelet extrema reconstruction will be to use the multiple residual signals to create a single enhanced residual signal, $\hat{e}[n]$. This residual will then be used as input to the LPC filter to generate an enhanced speech signal, $\hat{s}[n]$.

For the work presented in this chapter, the analysis frame is 32 msec ($N = 256$ samples at 8kHz sampling rate), the number of channels is $I = 8$, and the LPC model order is $P = 13$.

Wavelet Transform

The first J scales of the wavelet transform (WVT) are calculated for each of the I residual signals. Here the WVT is calculated for the first $J = 5$ scales. The WVT of the residual signal of the i^{th} channel, $e_i[n]$, is denoted by $W e_i(n, j)$ where j is the scale parameter and n is the temporal index within that scale. We use the non-decimated quadratic spline wavelet discussed in [25]. Because the wavelet is the derivative of a smoothing function, the resulting WVT coefficients will contain local extrema at large changes in the signal. Therefore, impulses in the residual signal of voiced speech segments will be manifested as large WVT coefficients at different scales. This same wavelet has been used on non-reverberant speech for pitch detection [26].

Another advantage of working in the wavelet domain is the ability to approximately reconstruct a signal from the local extrema of its WVT coefficients at each scale [27]. To summarize this process, suppose the wavelet transform of a signal, $f[n]$, with local extrema at $n = n_{j,k}$ in scale j is to be approximated by $\hat{f}[n]$. By minimizing the norm of $\hat{f}[n]$ and imposing the constraint $Wf(n_{j,k}, j) = W\hat{f}(n_{j,k}, j)$, it is possible to construct $\hat{f}[n]$ via a conjugate gradient approach [27]. Figure 2.2 shows a clean voiced speech segment and its wavelet extrema reconstruction after 5 and 25 iterations of the conjugate gradient method. Note that this method tends to approximate a signal to within a very high frequency error term, which, for the speech signals we examined, was inaudible.

In the following stages, this algorithm is used to reconstruct an enhanced residual signal, where wavelet extrema at different scales correspond to impulses in the voiced excitation. While the significance of the wavelet extrema in the unvoiced case is not as clearly delineated, this approach is effective in reconstructing an appropriate noise-like residual signal.

Extrema Clustering

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 estimate the ‘clean’ extrema. In [22] we presented a nonlinear clustering criterion for this application, while in [28] we employed a Gaussian smoothing approach. However, the local extrema of WVT coefficients of the delay-and-sum beamformed speech have been found to perform sufficiently well in this context. Because of its computational advantages and near comparable performance, we will employ this delay and sum strategy for extrema clustering in what follows. The left-hand plot in Figure 2.3 illustrates a single scale’s WVT coefficients for 8 channels and the resulting clustered extrema. In general, for voiced speech, extrema due to glottal excitations tend to be more closely clustered than those due to additional impulses from the reverberant channel responses. This causes the less clustered peaks to have a relatively small amplitude. The clustering operation maps the multi-channel WVT extrema data to a single WVT extrema representation suitable for input to the WVT extrema reconstruction algorithm described previously.

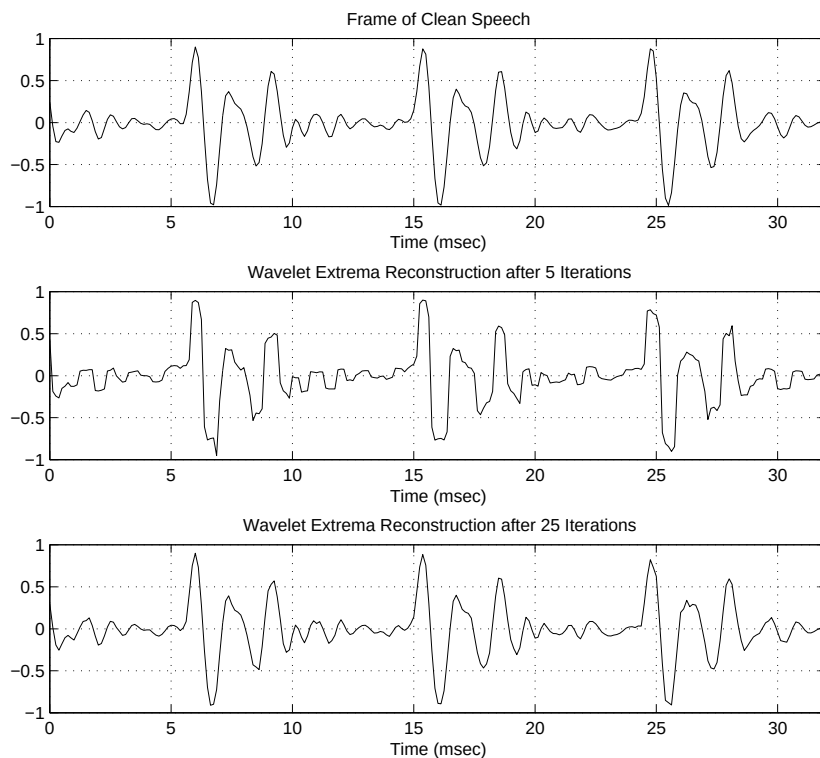


Figure 2.2 Clean Speech and Wavelet Extrema Reconstructions After 5 and 25 Iterations

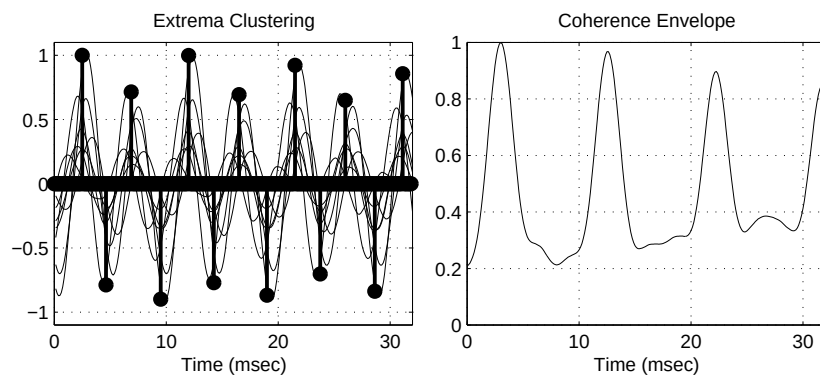


Figure 2.3 Clustering Results and Coherence Envelope

Short-Term Coherence Scaling

We now apply the first of two coherence-based scaling functions. The purpose of the amplitude modulation scheme detailed here is to further

de-emphasize the clustered peaks due to room reverberations beyond what is achieved by the averaging process associated with the extrema clustering.

The following short-term coherence window is multiplied by the wavelet extrema at each scale j :

$$c_j[n] = \frac{2}{I-1} \frac{\sum_{i=1}^I \sum_{k=i+1}^I \sum_{l=-L}^L W e_i(n+l, j) W e_k(n+l, j)}{\sum_{i=1}^I \sum_{l=-L}^L W e_i(n+l, j)^2} \quad (2.5)$$

where $2L+1 \ll N$ is the length of a small subwindow within the analysis frame, on the order of 3 to 5 msec. With this window, any clustered extrema in areas of low coherence will be de-emphasized, while highly coherent extrema will be relatively unaffected. The coherence function as given above can assume values between -1 and 1 . A thresholding scheme is incorporated to maintain nonnegative scaling terms and to normalize the coherence values within the subwindow. For each scale, the modified coherence window is calculated from:

$$\tilde{c}_j[n] = \begin{cases} c_j[n] / \max_n c_j[n] & c_j[n] \geq T \\ T / \max_n c_j[n] & c_j[n] < T \end{cases} \quad (2.6)$$

where $0 \leq T \leq 1$. A reasonable value of T was determined experimentally to be 0.1. With $T > 0$, the thresholding ensures that the coherence window does not completely eliminate any wavelet extrema, which could result in excessive smoothing of the resulting speech signal. The right-hand plot in Figure 2.3 shows a typical coherence envelope used for weighting the clustered WVT extrema.

Extrema Reconstruction

After multiplying the clustered peaks for each wavelet scale by the appropriate scale-dependent coherence envelope, the enhanced LPC residual is reconstructed using the conjugate gradient algorithm discussed previously. Listening tests show that the variations in amplitude produced by the extrema synthesis process across consecutive periods of voiced speech can produce audible distortions in the resulting speech signal. To lessen this effect, a quasi-pitch-synchronous smoothing approach is employed at the coarser wavelet scales to ensure that the peaks from consecutive voiced periods are slowly varying. This procedure involves averaging WVT extrema with their corresponding extrema in neighboring pitch periods and implicitly assumes a pitch estimate is available. In [29] a variety of multi-channel pitch estimators are evaluated in the environmental conditions considered here. Several are shown to be effective for reverberation levels of 0 to 1000ms.

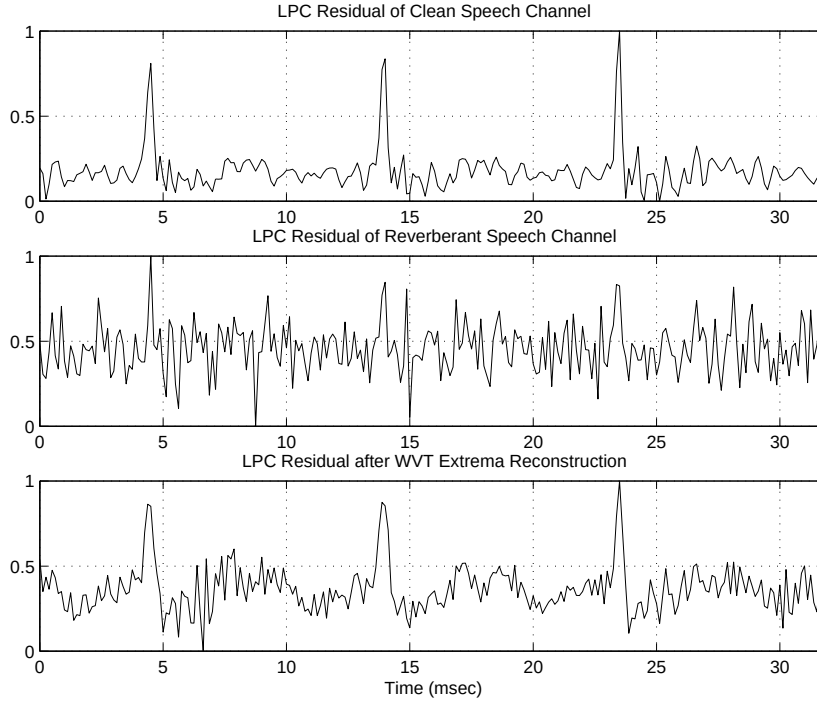


Figure 2.4 LPC Residual of Clean Speech, After Beamforming, and After Wavelet Clustering Technique

Figure 2.4 shows an example of the LPC residual for clean and reverberant speech frames followed by the residual signal resulting from the WVT extrema reconstruction algorithm.

LPC Filter

For each analysis frame, the reconstructed residual signal, $\hat{e}[n]$, is applied to the estimated LPC filter to generate the enhanced speech signal. Figures 2.5 and 2.6 offer the results for two different frames of speech. Figure 2.5 displays a single analysis frame 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 multipath distortion is to inflate the signal energy between 987ms and 997ms. 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

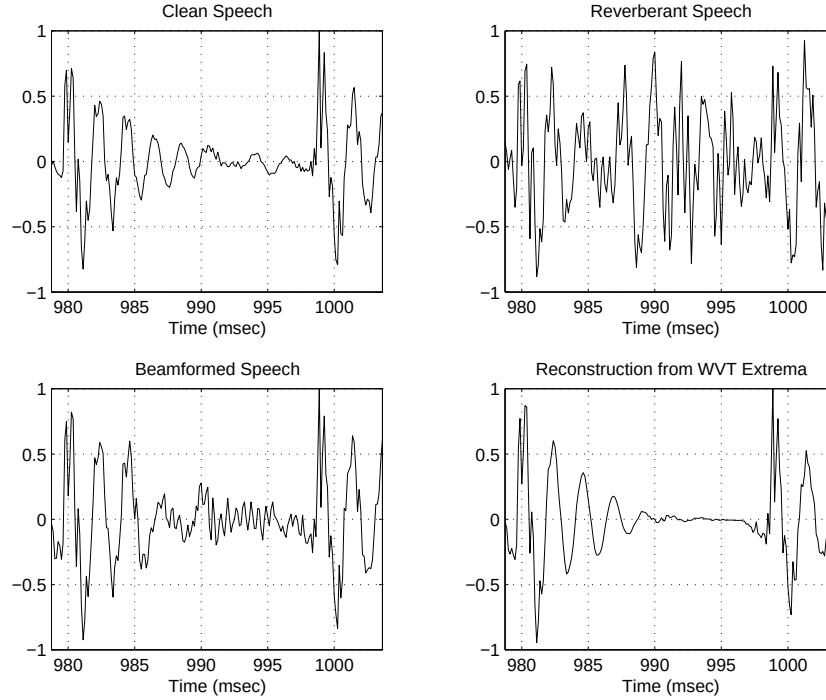


Figure 2.5 Comparison 1: Clean, Reverberant, Beamformed, and WVT Extrema Reconstructed Speech

original. Figure 2.6 demonstrates a voiced speech segment from early in a vowel segment. At this point, the speech is more affected by additive noise than reverberations, and the enhancement procedure diminishes its presence.

Long-Term Coherence Scaling

The final step in the enhancement procedure is to scale the reconstructed speech frames by a measure that is based on both the long-term coherence of the speech channels and the energy of the beamformed signal. This second weighting procedure serves two purposes. First, it adjusts the energy of each frame to provide a relatively smoothly varying output. The short-term coherence window can vary substantially from frame to frame. The energy of the resulting speech will vary accordingly, potentially resulting in audible distortions. The long-term coherence window incorporates an energy statistic to address this issue. Second, it allows for the de-emphasis of lengthy signal portions which are due primarily to reverberation effects. This is especially the case at

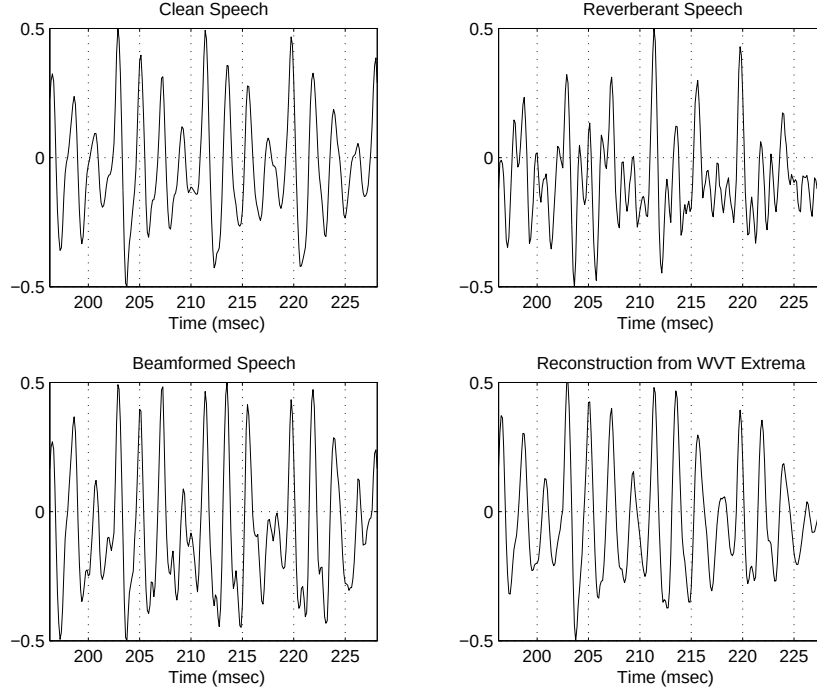


Figure 2.6 Comparison 2: Clean, Reverberant, Beamformed, and WVT Extrema Reconstructed Speech

the end of voiced sounds. The long-term coherence window is able to detect this condition and is used to adjust the corresponding regions of the reconstructed signal.

The calculation of this coarse scaling term, $g[n]$, is as follows:

1. Calculate the coherence of observed speech signals using:

$$c[n] = \frac{2}{I-1} \frac{\sum_{i=1}^I \sum_{k=i+1}^I \sum_{l=-L}^L x_i[n+l] x_k[n+l]}{\sum_{i=1}^I \sum_{l=-L}^L x_i[n+l]^2} \quad (2.7)$$

2. Calculate the short term energy, $b[n]$, of the beamformed signal.
3. If the energy of the beamformed signal decreases rapidly, then:

$$g[n] = b[n] \cdot c[n] / \max_{k \in [0, N]} c[n-k] \quad (2.8)$$

otherwise, $g[n] = b[n]$.

Figure 2.7 shows one channel of reverberant speech and the resulting $g[n]$. Each reconstructed speech frame is multiplied by the average

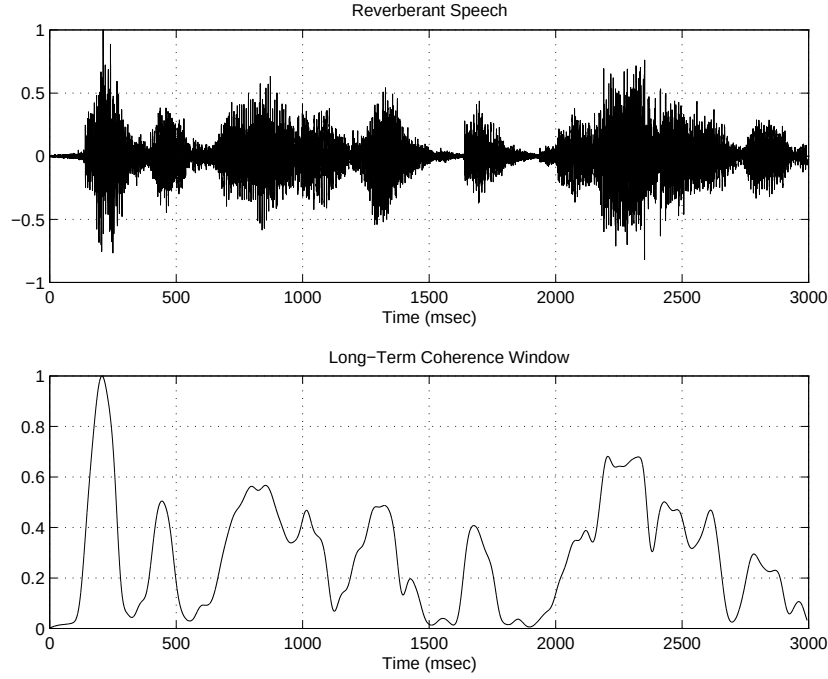


Figure 2.7 Long-Term Coherence Window

value of $\sqrt{g[n]}$ over the length of the frame. The individual frames are then combined using an overlap-add procedure. The resulting enhanced speech signal has an energy profile which follows $g[n]$.

4.2 SIMULATIONS

Reverberant speech segments were generated using the image model of [30] with intra-sample interpolation [31] and up to sixth order reflections. Room reverberation times, T , ranged from 100 msec to 1 sec. The corresponding reflection ratio, β , used by the image model was calculated via Eyring's formula [32]:

$$\beta = \exp\{-13.82/[c(L_x^{-1} + L_y^{-1} + L_z^{-1})T]\}$$

where L_x , L_y , and L_z are the room dimensions and c is the speed of sound in air (≈ 342 m/sec). 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, as shown in Figure 2.8. The source was located near the wall con-

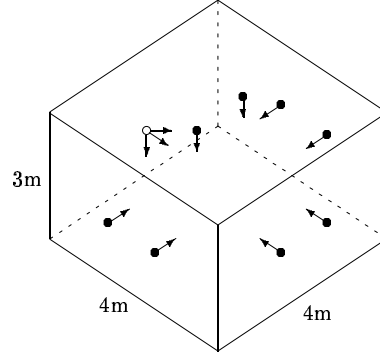


Figure 2.8 Room Setup - • Represents Microphones, ○ Represents the Speech Source.

taining no microphones. The microphones and sources were all assumed to have cardioid patterns.

Figures 2.9 and 2.10 illustrate a series of three second speech segments which have been simulated in the reverberant enclosure described above. Each figure plots the original speech, the signal received at a single microphone, the product of delay-and-sum beamforming, and the result achieved using the proposed enhancement algorithm. Figure 2.9 shows the results derived for 400 msec reverberant speech while Figure 2.10 shows the case for 400 msec reverberant speech plus additive noise. In this latter case, uncorrelated white Gaussian noise was added to each channel so that the SNR relative to the original non-reverberant utterance was 10 dB. The superior performance of the proposed algorithm relative to delay-and-sum beamforming at removing both long and short term reverberation effects and additive noise is apparent. Listening tests confirm the efficacy of the procedure. The synthesized speech reduces the audible reverberation and noise effects which are clearly present in both the degraded and delay-and-sum beamformed speech.

To quantify the improvements of the proposed algorithm, the average Bark Spectral Distortion (BSD)[33] was calculated over all frames of speech. The BSD is a speech quality assessment method which employs a series of perceptually motivated transformations to the signal spectra prior to computing a spectral distance. The result is an objective distortion measure which correlates well with perceptual quality ratings. Figure 2.11 shows the Bark spectral distance results for reverberation times from 100 msec to 1 sec with additive white noise (SNR = 10 dB). The proposed algorithm shows an improvement over

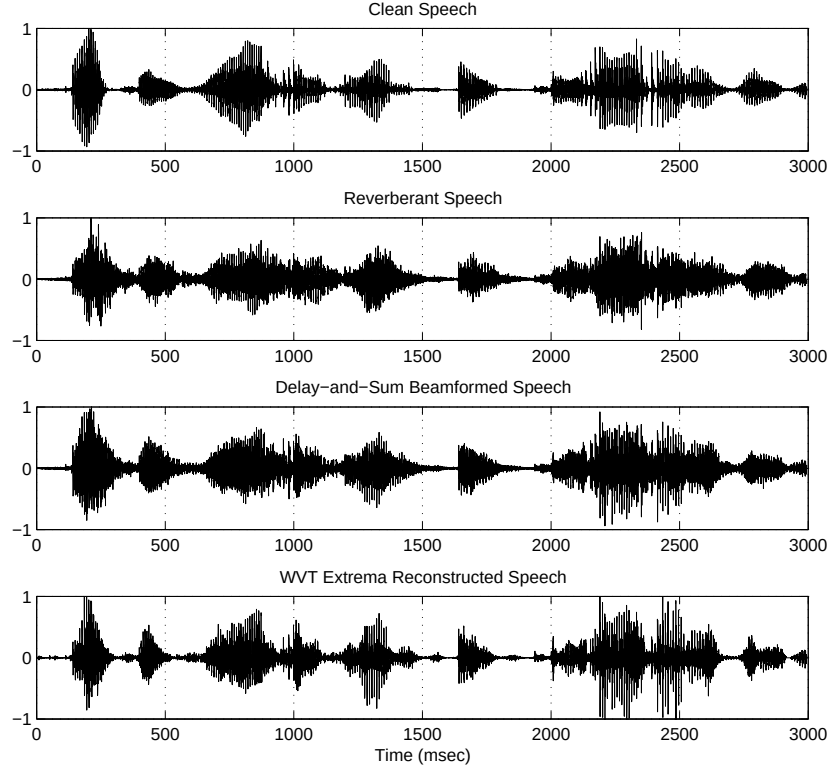


Figure 2.9 Comparison of Clean, Reverberant, Beamformed, and the Proposed Algorithm (Reverberation-Only Case)

delay-and-sum beamforming at all reverberation levels. These figures are consistent with informal listening tests.

5. CONCLUSION

By focusing on residual onset data and incorporating specific knowledge of the nature of the speech signal, the proposed wavelet-domain 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 process achieves a significant attenuation of environmental reverberations and additive noise without explicitly requiring estimation of the channel or noise characteristics. This illustrates the effectiveness and practicality of the advocated fusion of nonlinear and model-based processing for microphone array applications.

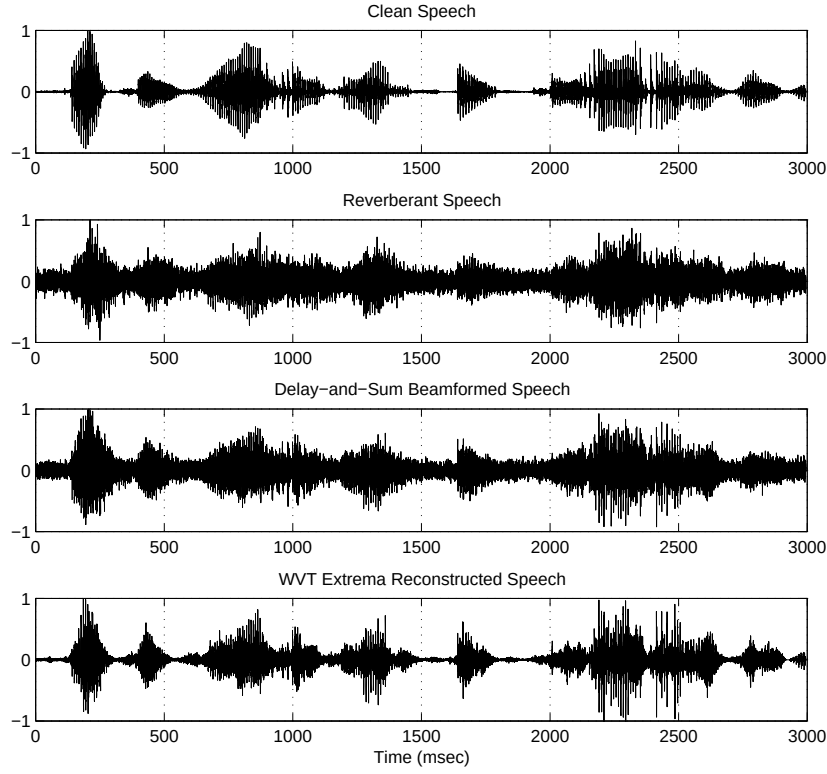


Figure 2.10 Comparison of Clean, Reverberant, Beamformed, and the Proposed Algorithm (Reverberation plus Noise Case)

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. The 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, and unknown channels).

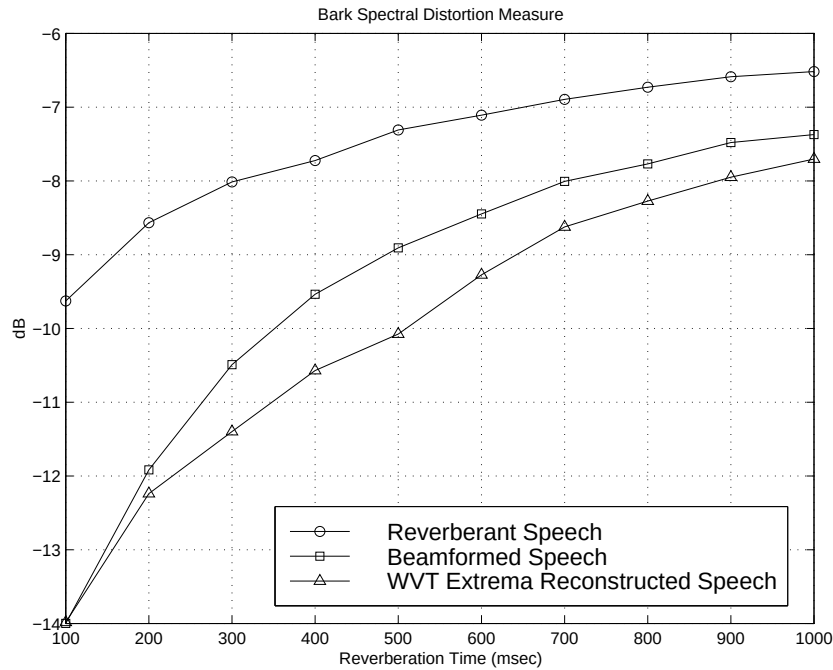


Figure 2.11 Bark Spectral Distortion Results

References

- [1] J. L. Flanagan and H. F. Silverman, "Material for international workshop on microphone-array systems: Theory and practice," LEMS Technical Report 113, LEMS, Division of Engineering, Brown University, Providence, RI 02912, Oct. 1992.
- [2] J. L. Flanagan and H. F. Silverman, "Material for international workshop on microphone-array systems: Theory and practice," Technical report, CAIP, Rutgers University, Piscataway, NJ 08855, Oct. 1994.
- [3] J. Lim, editor, *Speech Enhancement*, New Jersey: Prentice-Hall, 1983.
- [4] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*, New Jersey: Prentice Hall, first edition, 1987.
- [5] S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, New York: Marcel Dekker, first edition, 1992.
- [6] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744-754, Aug. 1986.

- [7] J. Hardwick, *The Dual Excitation Speech Model*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, June 1992.
- [8] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," in *Proc. IEEE ICASSP*, 1993, pp. II-550-II-553.
- [9] D. Johnson and D. Dudgeon, *Array Signal Processing- Concepts and Techniques*, New Jersey: Prentice Hall, first edition, 1993.
- [10] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, 11(2):91-126, April 1997.
- [11] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE ICASSP*, 1988, pp. 2578-2580.
- [12] K. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Second Cost 229 Workshop on Adaptive Algorithms in Communications*, pp. 185-194, Bordeaux, France, Sept. 1992.
- [13] Z. Yang, K. Simmer, and A. Wasiljeff, "Improved performance of multi-microphone speech enhancement systems," in *Proceedings of the 14th GRETSI Symposium*, pp. 479-482, 1993.
- [14] C. Marro, Y. Mahieux, and K. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Proc.*, 6(3):240-259, May 1998.
- [15] S. Gierl, "Noise reduction for speech input systems using an adaptive microphone array," in *Proceedings 22nd ISATA*, pp. 517-524, 1990.
- [16] M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," in *Proc. IEEE ICASSP*, 1997, pp. 239-242.
- [17] J. Meyer and K. Simmer, "Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction," in *Proc. IEEE ICASSP*, 1997, pp. 1167-1170.
- [18] J. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, 13(1-2):207-222, 1993.
- [19] 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, Sept. 1997.

- [20] M. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. IEEE ICASSP*, 1998, pp. 3613–3616.
- [21] B. Radlovic, R. Williamson, and R. Kennedy, "On the poor robustness of sound equalization in reverberant environments," in *Proc. IEEE ICASSP*, 1999, pp. 881–884.
- [22] M. Brandstein, "An event-based method for microphone array speech enhancement," in *Proc. IEEE ICASSP*, 1999, pp. 953–956.
- [23] B. S. Atal and J. R. Remde, "A new model of lpc excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE ICASSP*, 1982, pp. 614–617.
- [24] S. Singhal and B. S. Atal, "Improving performance of multi-pulse lpc coders at low bit rates," in *Proc. IEEE ICASSP*, 1984, pp. I–131–I–134.
- [25] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(7):710–732, July 1992.
- [26] S. Kadambe and G. Faye Boudreaux-Bartels, "Applications of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Information Theory*, 38(2):917–924, Mar. 1992.
- [27] S. Mallat, *A Wavelet Tour of Signal Processing*, Boston: Academic Press, 1998.
- [28] S. Griebel and M. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverberation," in *IEEE Workshop on Acoustic Echo and Noise Control*, Pocono Manor, Pennsylvania, pp. 52–55, Sept. 1999.
- [29] S. M. Griebel, "Multi-channel wavelet techniques for reverberant speech analysis and enhancement," Technical Report 5, HIMMEL, Harvard University, Cambridge, MA, Feb. 1999.
- [30] 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.
- [31] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, 80(5):1527–1529, Nov. 1986.
- [32] H. Kuttruff, *Room Acoustics*. London: Elsevier, third edition, 1991.
- [33] S. Wang and, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Selected Areas in Communications*, 10(5):819–829, June 1992.