

6 Explicit Speech Modeling for Microphone Array Applications

Michael Brandstein and Scott Griebel

Harvard University, Cambridge MA, USA

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 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. We offer two examples of algorithms which incorporate this principle. The first is a frequency domain approach illustrating the utility of model-based microphone array processing for improved speech spectral analysis. The second is a time domain procedure which highlights the benefits of speech modeling/spatial filtering fusion for the purpose of improving speech quality.

6.1 Introduction

A major thrust of microphone array research has focused on improving the spatial filtering capability of the system in its operating environment. The previous chapters have directly addressed a number of the methods which have proven beneficial in this regard. The goal of this chapter is to present an alternative and complementary strategy which emphasizes the incorporation of explicit speech modeling into the microphone array processing. As will be shown, by combining specific knowledge of the desired time-series attributes with the advantages of multi-channel spatial filtering, this approach can significantly improve the quality of the speech signal and its derived quantities obtained in challenging acoustic environments.

Much of this work is motivated from current research in the field of single-channel speech systems where, due to necessity perhaps, improvements have been obtained by focusing on the underlying content of the signal of interest in addition to its environmental degradations. 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 [1–3]. 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, 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 [4–6]. 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, 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. Fixed and adaptive beamforming techniques generally assume the desired source is slowly varying and at a known location. While dynamic localization schemes, like those of [7], and robust 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.

Similarly, the class of post-filtering algorithms detailed in Chapter 3 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.

Another general 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 [8] 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 [9] 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-

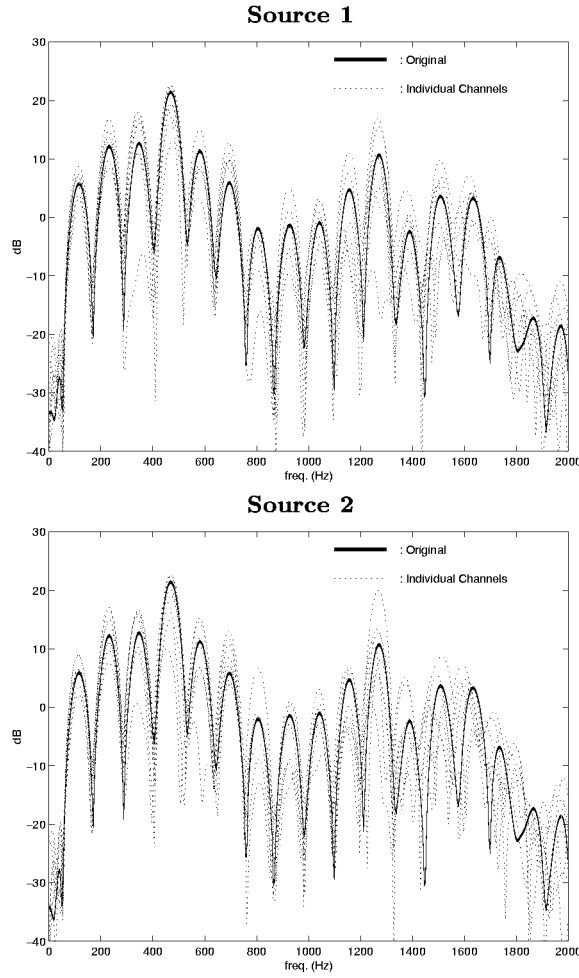


Fig. 6.1. Spectra of a voiced speech segment and its individual channels simulated at source locations spaced 10cm apart.

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 [10,11]. 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.

To illustrate the acoustic environment's sensitivity to source location variations, a simple example is presented. Two source locations spaced 10cm apart are simulated in the center of a noiseless 4m x 4m x 3m rectangular

room. The enclosure is assumed to have plane reflective surfaces and uniform, frequency-independent reflection coefficients equivalent to a 400ms reverberation time. Room impulse responses are generated for 8 microphones with 25cm spacing positioned along one wall of the enclosure using the Allen and Berkley image model technique [12] with intra-sample interpolation and up to sixth order reflections. Both the microphones and sources are assumed to have cardioid patterns and the sources are oriented toward the center of the array. Figure 6.1 plots the spectra of a voiced speech segment generated at each source location. The bold lines correspond to the spectrum of the original speech while the dotted lines plot the spectra of the data received at each of 8 microphones. The reverberation effects are multiplicative in the frequency domain and vary considerably from channel to channel. The results of this simulation show that there are significant variations in the spectra of the individual channels (up to 10dB for some frequencies) when the source is moved just a few inches. An implication of this example is 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.

6.2 Model-Based Strategies

Single-channel techniques exploit various features of the speech signal; multi-channel methods focus primarily on improving the quality of the spatial filtering process. In [11] we proposed an alternative to the traditional microphone array methods by explicitly incorporating the Dual Excitation Speech Model [5] into the beamforming process. Our work in [13] extended this idea using a multi-channel version of the Multi-Pulse Linear Predictive Coding (MPLPC) model [14,15] 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 concepts were then expanded upon in [16,17] by employing the wavelet domain for a multi-resolution signal analysis and reconstruction of the LPC residual. These works illustrated the ability of this approach to suppress the deleterious effects of both reverberations and additive noise without explicitly identifying the channel while being adaptive on a frame by frame basis. This allows the model-based processing paradigm to be applied for effective speech analysis and enhancement under general conditions.

Two examples of this prior work are summarized here. The first is a frequency domain approach illustrating the utility of model-based microphone array processing for improved speech spectral analysis. The second is a time domain procedure which highlights the benefits of the speech modeling/spatial filtering fusion advocated for the purpose of improving speech quality.

Example 1: A Frequency-Domain Model-Based Algorithm

In the single-channel, Dual Excitation (DE) Speech Model [5], a windowed segment of speech, $s[n]$, is represented as the sum of two components: a voiced signal, $v[n]$, and an unvoiced signal, $u[n]$. In the frequency-domain, the relationship may be expressed as:

$$S(\omega) = V(\omega) + U(\omega) \quad (6.1)$$

where $S(\omega)$, $V(\omega)$, and $U(\omega)$ correspond to the Fourier transforms of $s[n]$, $v[n]$, and $u[n]$, respectively. The voiced portion is assumed to be periodic over the time window and may be represented as the sum of the harmonics of a fundamental frequency, ω_0 :

$$V(\omega) = \sum_{m=-M}^M A_m W(\omega - m\omega_0) \quad (6.2)$$

where $W(\omega)$ is the Fourier Transform of the analysis window, A_m is the complex spectral amplitude of the m th harmonic, and M is the total number of harmonics ($M = \lfloor \pi/\omega_0 \rfloor$). Following [18], the fundamental frequency and harmonic amplitudes are estimated through minimization of the mean-squared error criterion:

$$\mathcal{E} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(\omega) - V(\omega)|^2 d\omega \quad (6.3)$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(\omega) - \sum_{m=-M}^M A_m W(\omega - m\omega_0)|^2 d\omega. \quad (6.4)$$

This non-linear optimization problem may be decoupled efficiently by noting that for a given fundamental frequency, the harmonic amplitudes which minimize the error are found through the solution of a set of linear equations. The optimal parameter set may then be calculated through global minimization of the error function in (6.4) versus all fundamental frequencies of interest. To effectively model the spectrum of the higher harmonics, a fundamental frequency resolution of less than 1 Hz is typically required. In practice, this exhaustive procedure may be computationally prohibitive. A more efficient approach is to evaluate a coarse, integer pitch estimate via a traditional time-domain pitch estimation procedure and then use the above frequency-domain analysis by synthesis procedure to refine the fundamental frequency estimate.

The estimated unvoiced signal plus noise spectrum, $\hat{U}(\omega)$, is then found from the difference spectrum:

$$\hat{U}(\omega) = S(\omega) - \hat{V}(\omega) \quad (6.5)$$

where $\hat{V}(\omega)$ is the estimated voiced spectrum derived from (6.2) using the estimated values of ω_0 and A_m .

The utility of the DE model for improving speech degraded by background noise lies in its independent enhancement of the voiced and unvoiced components of the speech. Assuming that the degrading noise is independent of the harmonic structure, the voiced spectrum is subjected to only a minor thresholding operation relative to the background noise power. The bulk of the enhancement is achieved by nulling out the unvoiced portions of strongly voiced harmonics and applying a modified Wiener filter to the remaining unvoiced spectral regions.

The Dual Excitation model is extended here for the multi-channel problem to improve its effectiveness for the additive noise case and to address the more general distant-talker scenario involving multipath channels and multiple sources. Consider first the extension of the DE error criterion in (6.4) to include data from N channels:

$$\mathcal{E}_N = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sum_{i=1}^N G_i(\omega) S_i(\omega) - \sum_{m=-M}^M A_m W(\omega - m\omega_0) \right|^2 d\omega \quad (6.6)$$

where $G_i(\omega)$ is the filter weighting associated with the i th channel and $S_i(\omega)$ is the short-term spectrum of the data received at the i th microphone. Alternatively, for environments where the dominant degradation effect is reverberant, it may be advantageous to recast the above error criterion as the L_2 norm in the log spectrum domain.

The voiced signal estimate, $\hat{V}_N(\omega)$, derived from the parameters minimizing (6.6) would then be used to produce the unvoiced signal plus noise spectrum from:

$$\hat{U}_N(\omega) = \frac{1}{N} \sum_{i=1}^N H_i(\omega) [G_i(\omega) S_i(\omega) - \hat{V}_N(\omega)] \quad (6.7)$$

The channel weightings, $G_i(\omega)$, could be designed to provide appropriate spatial filtering, addressing issues of noise-reduction, attenuation of interfering sources, and dereverberation. Additionally, the channel-dependent weighting filters, $H_i(\omega)$, could be incorporated as a multi-channel post-processor to exploit known signal characteristics.

In the simplest case of independent additive noise, the extension of the Dual Excitation model to a plurality of channels would stand to improve its enhancement performance by virtue of the data averaging alone. With the inclusion of the spatial filtering afforded through (6.6) and (6.7) it is possible to give the DE model a robustness to channel effects and interfering sources. With regard to multiple sources, the error criterion in (6.6) could be extended explicitly to include L sources and N channels by:

$$\mathcal{E}_{LN} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{1}{L} \sum_{j=1}^L \left(\sum_{i=1}^N G_{ij}(\omega) S_i(\omega) - \sum_{m=-M}^M A_{mj} W(\omega - m\omega_{0j}) \right) \right|^2 d\omega$$

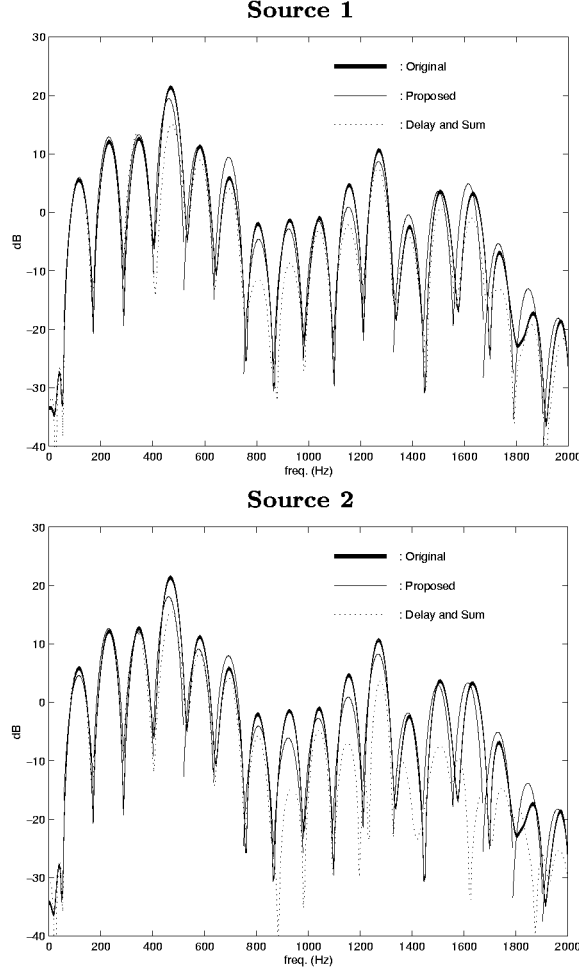


Fig. 6.2. Spectra of the Original, Delay and Sum Beamformer result, and Voiced Multi-Channel Dual Excitation result for the voiced segment in Figure 6.1.

where $G_{ij}(\omega)$ is the spatial filter associated with the i th channel and j th signal source, ω_{0j} is the fundamental frequency of the j th source, and A_{mj} is the amplitude of the m th harmonic associated with the j th source. Using this approach it would be possible to track individual sources through a combination of location and pitch data. Such a multi-channel DE model would have the ability to isolate and enhance a desired source signal by employing both spatial and signal-content information.

To illustrate the potential of such an approach, again consider the example of the voiced speech segment in Fig. 6.1. Figure 6.2 shows the relationship between the Delay and Sum Beamformer and the voiced signal estimate, $\hat{V}_N(\omega)$,

derived from the proposed multi-channel scheme for the two closely-spaced source locations. The pair of results was generated using delays appropriate for the source 1 location. This would correspond to a 10cm mis-aim in the source 2 case. The Delay and Sum method, like any beamforming technique, lacks any signal-dependent constraints on the output produced. As the plots suggest, by exploiting the periodic structure of the desired signal, the Multi-Channel Dual Excitation Model is significantly more robust to the local spectral variations produced by channel reverberations. Unlike the Delay and Sum method, the approach is relatively insensitive to imperfect knowledge of the source location suggesting a robustness to the small, but nominal, variations encountered in a practical operating environment. This result is confirmed by more quantitative methods, such as SNR and log spectral distortion scores.

Example 2: A Time-Domain Model-Based Algorithm

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

$$x_i[n] = h_i[n] * s[n] + u_i[n] \quad (6.8)$$

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 environment's acoustical and geometrical properties. 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 [2]. 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 correlated after performing time-delay compensation. Essentially, the approach will be to identify the clean speech excitation signal from a set of corrupted excitation signals. 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 proposed algorithm offers an effective method for estimating and then reconstructing the excitation signal by employing a class of wavelets to decompose the LPC residual signals. In [19], quadratic spline wavelets

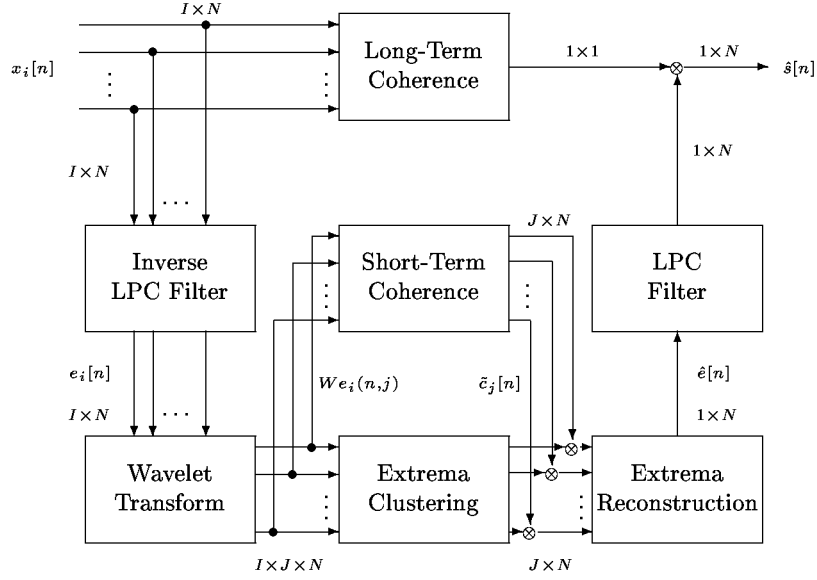


Fig. 6.3. Outline of the proposed algorithm for an N -point frame of speech.

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 microphone 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 [19].

The algorithm is outlined in Fig. 6.3. 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. For the following algorithm discussion, the analysis frame is 32ms ($N = 256$ samples at 8kHz sampling rate) and the number of channels is $I = 8$. Details of the simulation procedure are provided below.

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:

$$s[n] = \sum_{p=1}^P a_p s[n-p] + e[n] \quad (6.9)$$

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:

$$\frac{1}{I} \sum_{i=1}^I \sum_{n=1}^N e_i^2[n] = \frac{1}{I} \sum_{i=1}^I \sum_{n=1}^N \left\{ s[n] - \sum_{p=1}^P a_p s[n-p] \right\}^2. \quad (6.10)$$

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:

$$r[k] = \frac{1}{I} \sum_{i=1}^I \sum_{n=k}^N x_i[n] x_i[n-k]. \quad (6.11)$$

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 is 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 previously estimated LPC filter to generate an enhanced speech signal, $\hat{s}[n]$. For the current simulations, the LPC model order is $P = 13$.

The first J scales of the 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 $We_i(n, j)$ 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 [19]. 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 [20].

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.

Based upon the observation that residual impulses due to the original speech tend to be clustered across channels and scales (after time-delay compensation) while those due to reverberation effects are relatively uncorrelated

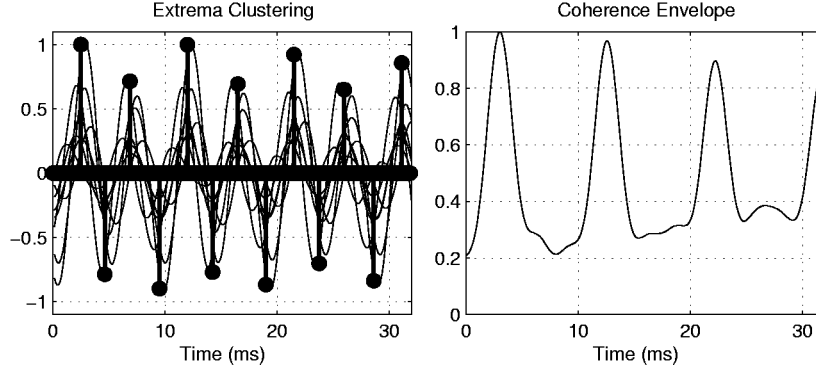


Fig. 6.4. Clustering Results and Coherence Envelope.

in both amplitude and time, a clustering technique is now applied to estimate the ‘clean’ extrema. In [13] we presented a nonlinear clustering criterion for this application, while in [16] we employed a Gaussian smoothing approach. However, the local extrema of WVT coefficients of the delay-and-sum beam-formed 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 6.4 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 later.

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 motivation is that wavelet extrema due to room effects will be less coherent from channel to channel than those due to the vocal excitation signal.

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

$$c_j[n] = \frac{2}{I-1} \frac{\sum_{i=1}^I \sum_{k=i+1}^I \sum_{l=-L}^L W e_i(n+l, j) W e_k(n+l, j)}{\sum_{i=1}^I \sum_{l=-L}^L W e_i(n+l, j)^2} \quad (6.12)$$

where $2L + 1 \ll N$ is the length of a small subwindow within the analysis frame, on the order of 3 to 5ms. With this window, any clustered extrema in areas of low coherence will be de-emphasized, while highly coherent extrema

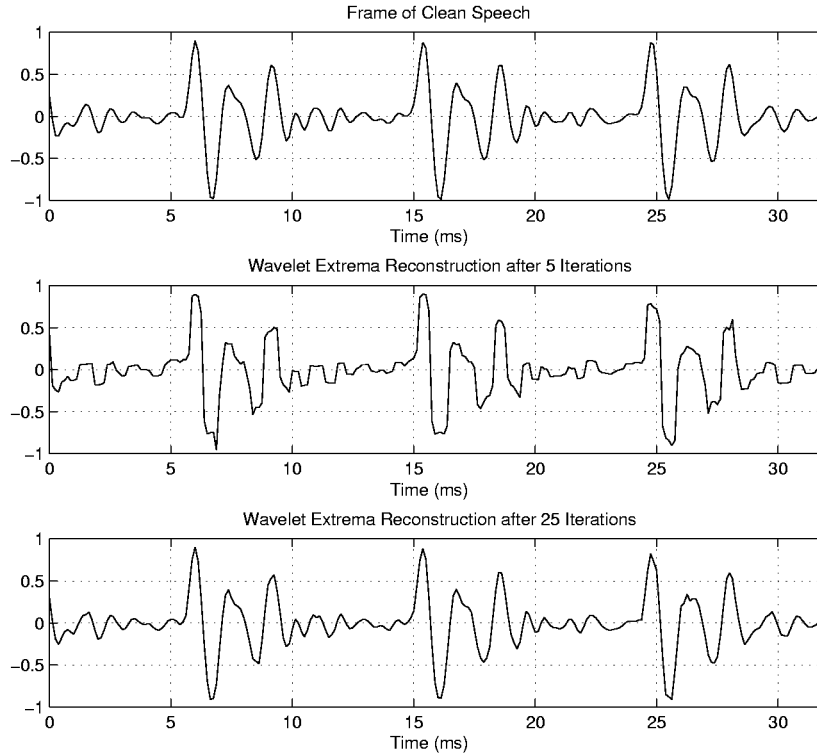


Fig. 6.5. Clean Speech and Wavelet Extrema Reconstructions After 5 and 25 Iterations.

will be relatively unaffected. The coherence function as given above is limited to values between -1 and 1 . A thresholding scheme is incorporated to maintain nonnegative scaling terms and to normalize the coherence values within the subwindow. The right-hand plot in Fig. 6.4 shows a typical coherence envelope used for weighting the clustered WVT extrema.

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 [21]. 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 reconstruct $\hat{f}[n]$ via a conjugate gradient approach [21]. Figure 6.5 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 studied, was inaudible.

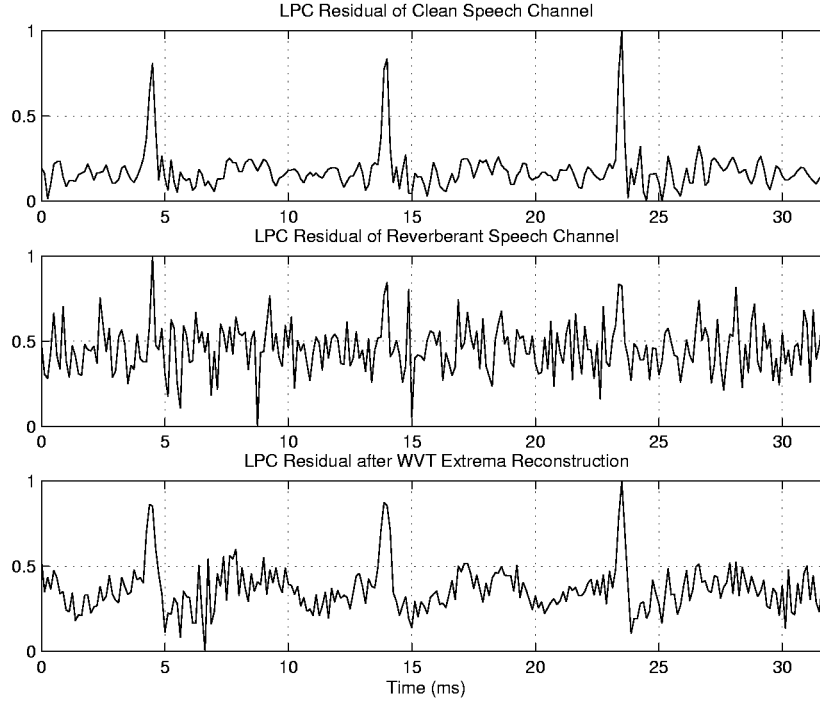


Fig. 6.6. LPC Residual of Clean Speech, After Beamforming, and After Wavelet Clustering Technique.

After multiplying the clustered peaks for each wavelet scale by the appropriate scale-dependent coherence envelope, the enhanced LPC residual is reconstructed using the wavelet extrema reconstruction algorithm just discussed. 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 [22] 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 6.6 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.

For each analysis frame, the reconstructed residual signal, $\hat{e}[n]$, is applied to the estimated LPC filter to generate an enhanced speech signal. Figures 6.7 and 6.8 offer the results for two different frames of speech. Figure 6.7 displays

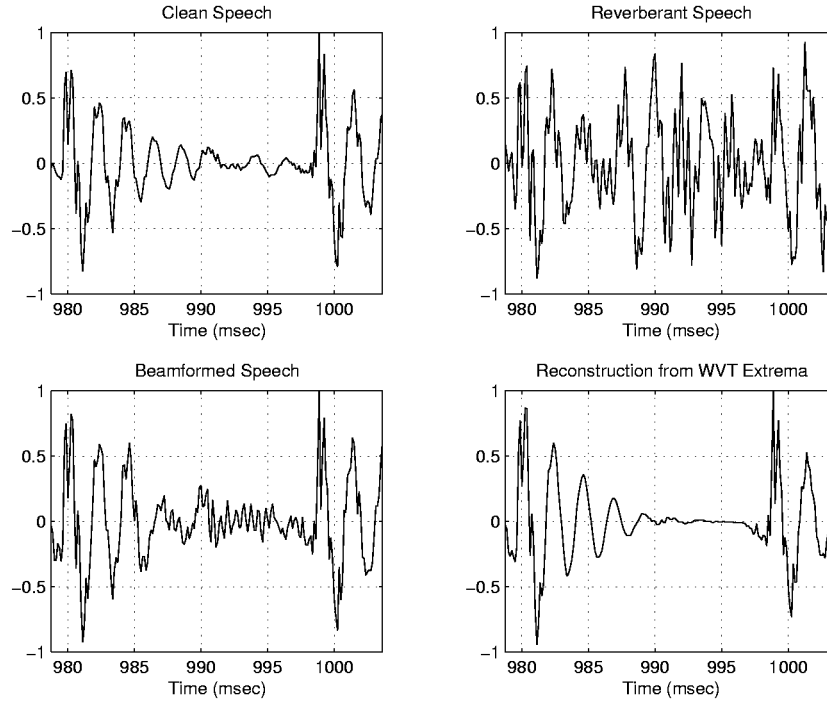


Fig. 6.7. Comparison 1: Clean, Reverberant, Beamformed, and WVT Extrema Reconstructed Speech.

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 original. Figure 6.8 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.

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

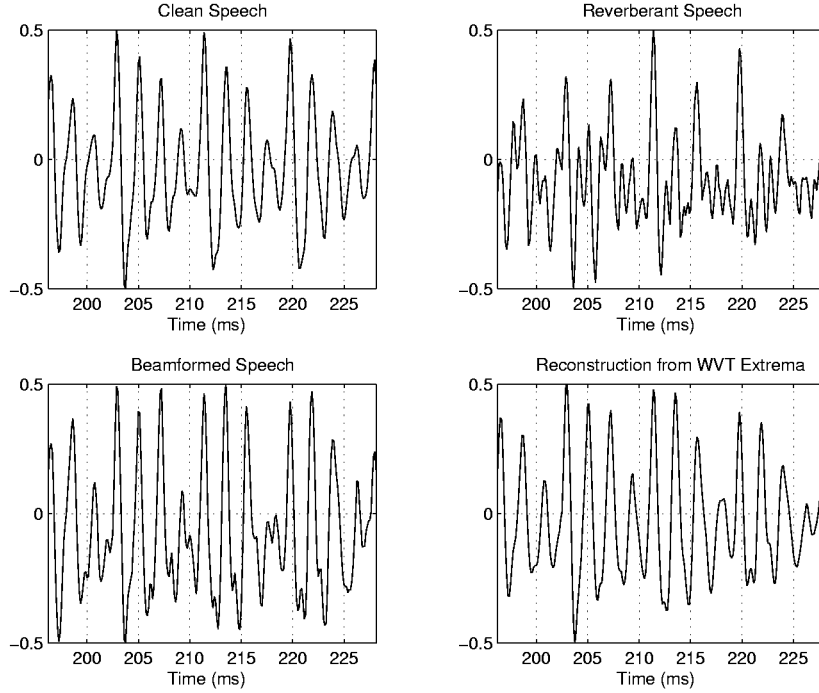


Fig. 6.8. Comparison 2: Clean, Reverberant, Beamformed, and WVT Extrema Reconstructed Speech.

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 useful at the end of voiced sounds.

Simulations have been performed to evaluate the quality of the enhanced speech derived from this procedure. Again, 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 was employed. A total of eight microphones were used: one pair on 3 of the walls and another pair on the ceiling, as shown in Figure 6.9. The source was located near the wall containing no microphones. The microphones and sources were all assumed to have cardioid patterns. Room reverberation times, T , ranged from 100ms to 1s. Uncorrelated Gaussian noise was added to each observed signal with SNR of 10dB.

Figures 6.10 and 6.11 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 6.10 shows the results derived for 400ms reverberant speech while Fig. 6.11 shows the case for 400ms re-

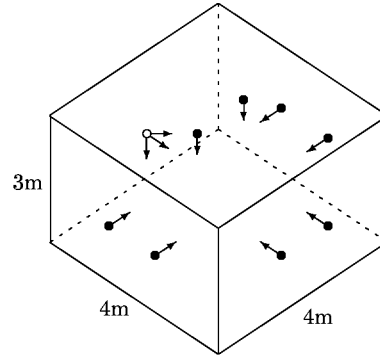


Fig. 6.9. Room Setup - • Represents Microphones, ○ Represents the Speech Source.

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 10dB. 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)[23] 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 6.12 shows the Bark spectral distance results for reverberation times from 100ms to 1s with additive white noise (SNR = 10dB). The proposed algorithm shows an improvement over delay-and-sum beamforming at all reverberation levels. These figures are consistent with informal listening tests.

6.3 Conclusion

This work has sought to show how the analysis and enhancement of speech acquired in multi-channel environments benefits through the exploitation of known features in the desired speech signal. Rather than focusing solely on the spatial filtering aspects of the problem, we have advocated strategies which incorporate an explicit speech model into the microphone array processing.

Two specific samples of this philosophy were presented. The first was a frequency-domain procedure which employed the Dual Excitation Speech

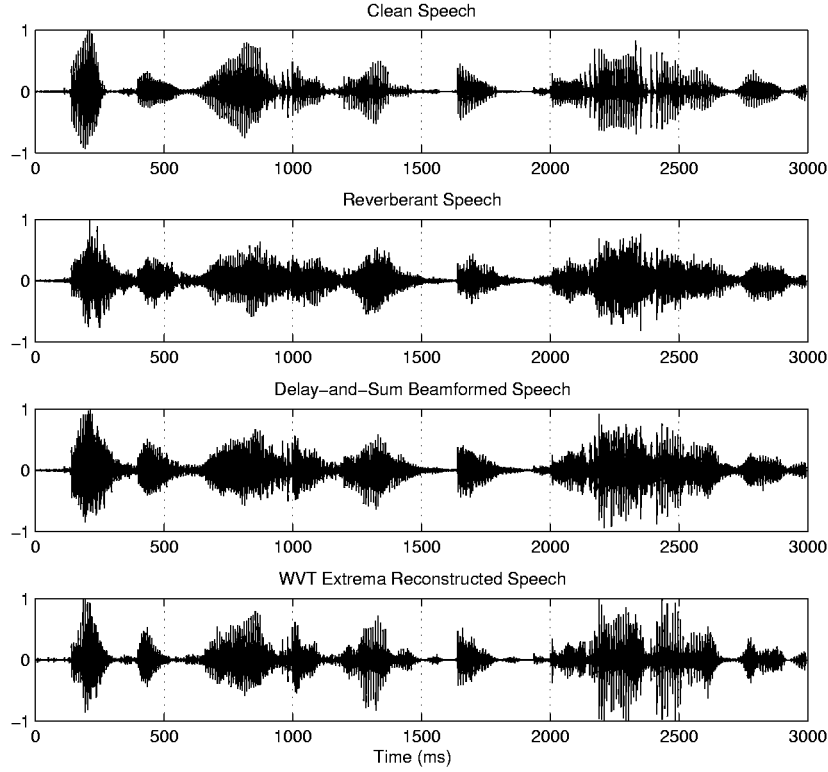


Fig. 6.10. Comparison of Clean, Reverberant, Beamformed, and the Proposed Algorithm (Reverberation-Only Case).

Model to utilize the periodicity of voiced speech. In the context of spectral analysis, we illustrated the improved performance obtained relative to Delay-and-Sum Beamforming in reverberant environments. We then outlined methods to address multi-source scenarios and for fusing this approach with traditional array processing techniques. While not presented here, we have applied this frequency-domain model to the problem of speech time-delay estimation in noisy and reverberant environments [24,25]. By emphasizing these spectral regions exhibiting a periodic structure, it is possible to achieve a time-delay estimator that outperforms standard methods based upon generalized cross-correlation.

The second example focused on time-domain residual onset data to illustrate the effectiveness and practicality of model-based processing for microphone array applications. 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 echoes and uncorrelated noise. The process achieves a significant attenuation of environmental rever-

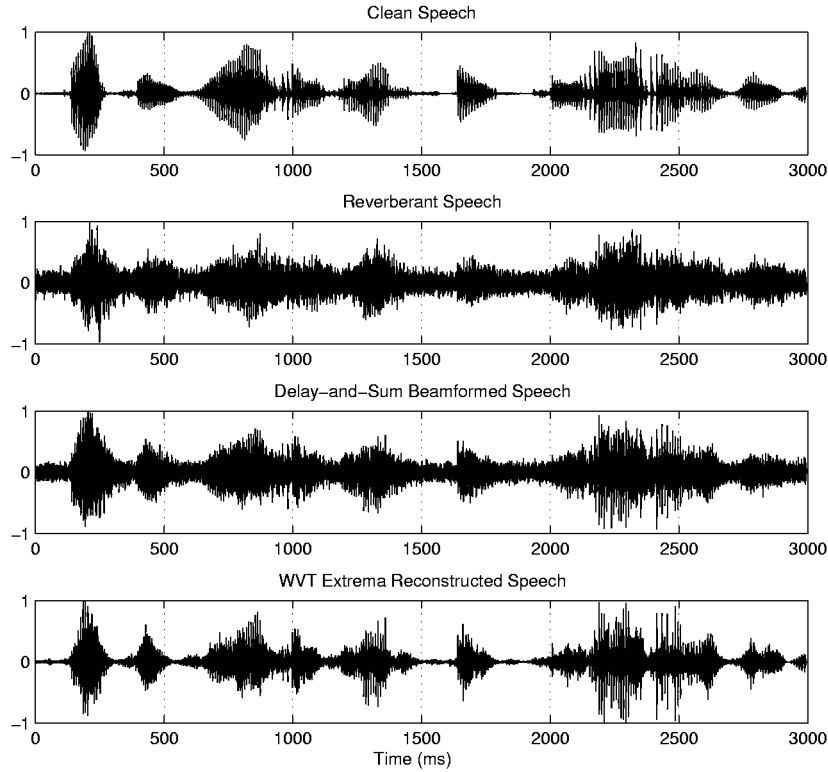


Fig. 6.11. Comparison of Clean, Reverberant, Beamformed, and the Proposed Algorithm (Reverberation plus Noise Case).

berations and additive noise without explicitly requiring estimation of the channel or noise characteristics. 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 outlined in this chapter, 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).

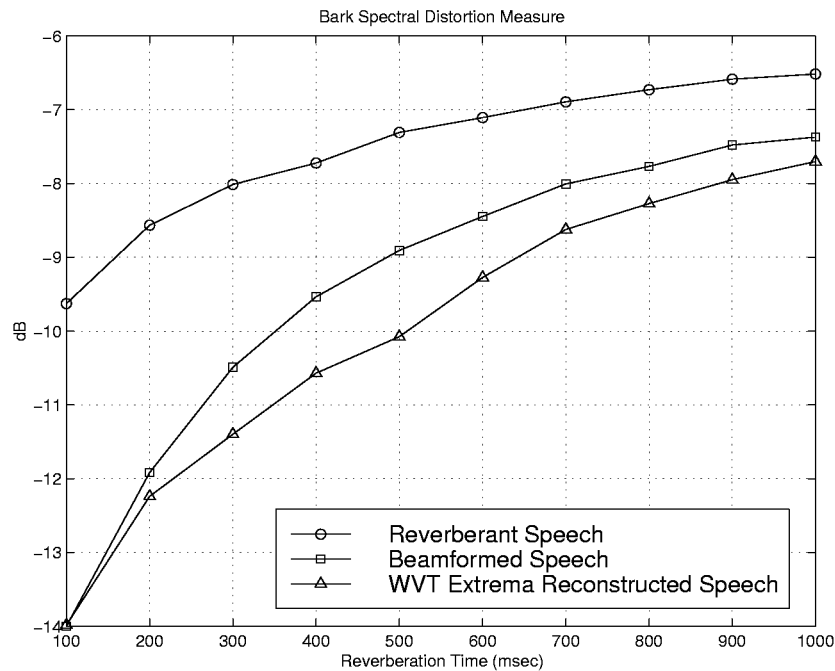


Fig. 6.12. Bark Spectral Distortion Results

There are a variety of speech models available for this application. This work has employed a pair of specific examples to illustrate the utility of the overall process. Similarly, the multi-channel methods based upon these models represent only two of the many possible analysis and enhancement schemes. Future work along these lines will investigate strategies to fuse proven methods in the single and multiple channel enhancement fields and to develop novel and effective algorithms.

References

1. J. Lim, ed., *Speech Enhancement*. Prentice Hall, 1983.
2. J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*. Prentice Hall, first ed., 1987.
3. S. Furui and M. Sondhi, eds., *Advances in Speech Signal Processing*. Marcel Dekker, first ed., 1992.
4. R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744-754, August 1986.
5. J. Hardwick, *The Dual Excitation Speech Model*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, June 1992.

6. J. Laroche, Y. Stylianou, and E. Moulines, "Hns: Speech modification based on a harmonic + noise model," in *Proceedings of ICASSP93*, pp. II-550-II-553, IEEE, 1993.
7. M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91-126, April 1997.
8. 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.
9. S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 425-437, September 1997.
10. B. Radlovic, R. Williamson, and R. Kennedy, "On the poor robustness of sound equalization in reverberant environments," in *Proc. ICASSP-99*, (Phoenix, AZ), pp. 881-884, March 15-19 1999.
11. M. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. ICASSP-98*, (Seattle, WA), pp. 3613-3616, May 12-15 1998.
12. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943-950, April 1979.
13. M. Brandstein, "An event-based method for microphone array speech enhancement," in *Proc. ICASSP-99*, (Phoenix, AZ), pp. 953-956, IEEE, March 15-19 1999.
14. B. S. Atal and J. R. Remde, "A new model of lpc excitation for producing natural-sounding speech at low bit rates," in *Proc. ICASSP-82*, pp. 614-617, 1982.
15. S. Singhal and B. S. Atal, "Improving performance of multi-pulse lpc coders at low bit rates," in *Proc. ICASSP-84*, pp. I-131-I-134, 1984.
16. 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), September 27-30 1999.
17. 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.
18. D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1223-1235, August 1988.
19. S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 710-732, July 1992.
20. S. Kadambe and G. F. Boudreaux-Bartels, "Applications of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Information Theory*, vol. 38, pp. 917-924, March 1992.
21. S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
22. S. M. Griebel, "Multi-channel wavelet techniques for reverberant speech analysis and enhancement," Tech. Rep. 5, HIMMEL, Harvard University, Cambridge, MA, February 1999.
23. S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Selected Areas in Communications*, vol. 10, pp. 819-829, June 1992.

24. M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2914–2919, 1999.
25. M. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, New York), October 19-22 1997.