Microphone Array Processing

For Robust Speech Recognition

Michael L. Seltzer

Ph.D. Thesis Prospectus

Submitted to the Department of Electrical and Computer Engineering

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy at

Carnegie Mellon University

Pittsburgh, Pennsylvania

June 2001

1. Introduction

State-of-the-art speech recognition systems are known to perform reasonably well when the speech signals are captured in a noise-free environment using a close-talking microphone worn near the mouth of the speaker. The progress of such systems has reached the point where real speech recognition systems have been deployed in the marketplace for a variety of uses. As recognition performance continues to improve, it is expected that demand for such systems will further increase. However, many of the target applications for this technology do not take place in noise-free environments. To further compound the problem, it is often inconvenient for the speaker to wear a close-talking microphone. As the distance between the speaker and the microphone increases, the speech signal becomes increasingly susceptible to background noise and reverberation effects that significantly degrade speech recognition accuracy. This is especially problematic in situations where the locations of the microphones or the users are dictated by physical constraints of the operating environment, as in meeting rooms or automobiles.

This problem can be greatly alleviated by the use of multiple microphones to capture the speech signal [9]. Microphone arrays record the speech signal simultaneously over a number of spatially separated channels. Many array-signal-processing techniques have been developed to combine the signals in the array to achieve a substantial improvement in the signal-to-noise ratio (SNR) of the output signal.

Currently, microphone array-based speech recognition is performed in two independent stages: array processing and recognition. Array-processing algorithms, typically designed for speech enhancement, process the captured waveforms and the output waveforms are passed to the speech recognition system. These systems implicitly assume that the array processing methods which provide the best enhancement will result in the best recognition performance. However, recognition systems, unlike enhancement algorithms, do not operate on the speech waveform itself, but rather a set of features extracted from the waveform. As a result, improvements in the quality of the output waveform may not necessarily translate into improvements in the quality of the recognition features and, by extension, improvements in recognition performance.

The goal of this thesis is to improve the performance of microphone array-based speech recognition systems. We propose to design microphone array-processing strategies specifically for use with speech recognition systems, without regard to SNR, perceptual quality of the signal, or other speech enhancement

metrics. We will consider the array-processing front end and the speech recognition system as one complete system, not two independent entities cascaded together. This approach will enable us to integrate information from the recognition system into the design of the array processing strategy to achieve better recognition performance than conventional array processing methods. Specifically, the microphone-array/ speech recognition system will be treated as a single closed-loop system, with information from the statistical models of the recognition system used as feedback to tune the parameters of the array processing scheme. We believe this will enable us to achieve better recognition performance than conventional array processing methods in microphone-array speech recognition systems.

This document is organized as follows:

Chapter 2 discusses previous approaches to microphone array processing for speech and their use for speech recognition. Conventional speech recognition compensation techniques which have been applied to array-processed speech are also considered. In Chapter 3, a new approach to array processing motivated solely by speech recognition performance is presented, along with some preliminary results using this approach. Chapter 4 describes proposed work to be performed in this thesis to expand the ideas described in Chapter 3. Chapter 5 outlines the overall goals for the thesis and a preliminary timetable for the proposed research.

2. Review of Previous Microphone Array Processing Strategies for Speech Recognition

Array signal processing is a very mature field, with applications not just in speech processing, but in radar, sonar, and other areas. In fact, many of the most successful microphone array speech processing strategies are not specific to speech signals at all. These classic array-processing techniques utilize well-tested, well-understood array properties to enhance any distant noisy target signal. More recently, the demand for hands-free speech communication and recognition has increased and as a result, newer techniques have been developed to address the specific issues involved in the enhancement of speech signals captured by a microphone array. In addition, several speech recognition compensation algorithms, originally developed for degraded single-channel speech, have improved the recognition performance of speech processed by a microphone array. Some of the more successful array processing algorithms and speech recognition compensation methods will be presented in this chapter, along with their benefits and drawbacks.

2.1 Array Processing Methods

2.1.1 Fixed Beamforming

The most widely used array-processing method is called beamforming [13]. Beamforming refers to any method that algorithmically (rather than physically) steers the sensors in the array toward a target signal. The direction the array is steered is called the “look direction”. Beamforming algorithms can either be fixed, meaning that the array-processing parameters are “hardwired” and do not change over time, or adaptive, where parameters are time varying and adjusted to track changes in the target signal and environment. The most common form of fixed beamforming is the delay-and-sum method. In delay-and-sum, signals from the various microphones are first time-aligned to adjust for the delays caused by path length differences between the target source and each of the microphones, using a variety of methods (*e.g.* [7][24]). The aligned signals are then summed together. Any interfering noise sources that do not lie along the look direction remain misaligned and are attenuated by the averaging. It can be shown that if the noise signals corrupting each microphone channel are uncorrelated to each other and the target signal, delay-and-sum processing results in a 3 dB increase in the SNR of the output signal for every doubling of the number of

microphones in the array [13]. Many microphone array-based speech recognition systems have successfully used delay-and-sum processing to improve recognition performance, and because of its simplicity, it remains the “method of choice” for many array-based systems (*e.g.* [12]). Most other array-processing procedures are variations of this basic delay-and-sum scheme or its natural extension, filter-and-sum processing, where each microphone channel has an associated filter and the captured signals are first filtered before being combined.

2.1.2 Adaptive Beamforming

In adaptive beamforming, the array-processing parameters are dynamically adjusted according to some optimization criterion. The Frost algorithm [10] is a weighted delay-and-sum technique in which the weights applied to each signal in the array are adaptively adjusted, subject to a unity-gain constraint. In the Griffiths-Jim algorithm [11], a fixed beamformer and an adaptive beamformer are combined to obtain the desired target signal. In some cases, the filter parameters can be calibrated to a particular environment or user. In [23], such a calibration scheme is designed for a hands-free telephone environment in an automobile. A series of “typical” target signals from the speaker, as well as jammer signals from the hands-free loudspeaker, are captured in the car and used for initial calibration of the parameters of a filter-and-sum beamforming system. These parameters are then adapted during use based on the stored calibration signals and updated noise estimates.

These adaptive-filter methods assume that the target and jammer signals are uncorrelated. When this assumption is violated, as is the case for speech signals in a reverberant environment, the methods suffer from signal cancellation because reflected copies of the target signal appear as unwanted jammer signals. This seriously degrades the quality of the output signal and results in poor speech recognition performance. Van Compernolle [33] showed that signal cancellation in adaptive filtering methods can be reduced somewhat by adapting the parameters only during silence regions when no speech is present in the signals.

2.1.3 Dereverberation Techniques

Reverberation is a significant cause of poor speech recognition performance in microphone array-based speech recognition systems [29]. Because none of the traditional beamforming methods successfully compensate for the negative effects of reverberation on the speech signal, much recent research has focused in this area. Most of the research effort has focused on estimating and then inverting the impulse response of

the room which characterizes effect of the room on the target signal as it travels to the microphone. However, room impulse responses are generally non-minimum phase [22] which causes the inverted dereverberation filter to be unstable. As a result, approximations to the true inverse of the transfer functions have to be used. Miyoshi and Kaneda [19] show that if multiple channels are used and the room transfer functions of all channels are known, the exact inverse is possible to obtain if the transfer functions have no common zeros. However, concerns about the numerical stability and hence, practicality, of this method have been raised because of the large matrix inversions it requires [27][30]. Liu *et al.* [18] break up room transfer functions into minimum phase and all-pass components and process these components separately to remove the effects of reverberation. However, even in simulated enviroments, they report implementation difficulties in applying this method to continuous speech signals. Raghaven *et al.* [29] take a slightly different approach to the reverberation problem. They estimate the transfer function of the source-to-sensor room response for each microphone in the array using [5], and then use a truncated, time-reversed version of this estimate as a matched-filter for that source-sensor pair. The matched filters are used in a filter-and-sum manner to process the array signals. They show that this method is able to reduce the effects of reverberation significantly and obtain recognition improvements in highly reverberant environments.

These dereverberation methods, however, require that the room transfer functions, from the source to each microphone in the array, be static and known *a priori*. While the transfer functions can be measured [5], this is both inconvenient and unrealistic, as it requires the use of additional hardware to estimate the impulse responses and assumes that the transfer functions are fixed, which implies the location of the talker and the environmental conditions in the room will not change over time.

2.1.4 Blind Source Separation

Blind source separation (BSS) has also been applied to microphone array environments, *e.g.* [15]. In the general BSS framework, observed signals from multiple sensors are assumed to be the result of a combination of source signals and some unknown mixing matrix. In one family of BSS techniques, called independent component analysis (ICA), the inverse of this unknown mixing matrix is estimated in the frequency domain for each DFT bin independently using iterative optimization methods [6]. Using this estimated inverse matrix, the microphone signals are “separated” on a frequency-component basis, and then recombined to form the output signal. Informal listenings of the separation produced by this method applied to a

recording of two sources captured by two microphones are quite compelling [16]. However, these methods assume that the number of competing sound sources is both known and identical to the number of microphones present. Additionally, these methods assume that the sources are mutually independent point sources and are unable to process target signals in correlated or diffuse noise, both of which are common in microphone array recordings. Acero *et al.* [2] attempt to relieve some of these problems by removing some of "blindness" in the source separation. They consider the source mixtures to contain only one signal, the target speech signal of interest, and treat the other signal as unwanted noise. A probabilistic model of speech (a vector quantized codebook of linear prediction coefficient vectors representing clean speech) is then used to guide the source separation process to obtain the desired signal. However, no measurable results of the performance of this method were reported.

2.1.5 Auditory model-based Array Processing

The auditory system is an unbelievably good array processor, capable to isolating target signals in extremely difficult acoustic conditions. In auditory model-based methods, no output waveform is produced, but rather some representation of the combined signal that models processing believed to occur in the auditory system. Features can be extracted from this auditory representation and used directly in speech recognition. Sullivan [31][32] devised such a scheme in which the speech from each microphone was bandpass filtered and then the cross-correlations among all the microphones in each subband were computed. The peak values of the cross-correlation outputs were used to derive a set of speech recognition features. While the method was quite promising in pilot work, the speech recognition performance on real speech was only marginally better than conventional delay-and-sum techniques and was much more computationally expensive.

2.2 Speech Recognition Compensation Methods

Once the multiple array signals have been processed into a single output signal, there are several classical speech recognition compensation techniques that have been successfully applied to improve speech recognition performance. These techniques are not specific to microphone array-based speech recognition, and can be applied to any conventional compensation situation.

2.2.1 Maximum Likelihood Linear Regression

Maximum Likelihood Linear Regression (MLLR) assumes that the Gaussian means of the state distributions of the Hidden Markov Models (HMM) representing noisy speech are related to the corresponding clean speech Gaussian means by a linear regression [17]. The regression has the form

$$\mu_{\boldsymbol{n}} = \boldsymbol{A}\mu_{\boldsymbol{c}} + \boldsymbol{b} \tag{2.1}$$

where $\mu_{\boldsymbol{n}}$ is the Gaussian mean vector of the noisy speech, $\mu_{\boldsymbol{c}}$ is the Gaussian mean vector of the clean speech and $\boldsymbol{A}$ and $\boldsymbol{b}$ are regression factors that transform $\mu_{\boldsymbol{c}}$ to $\mu_{\boldsymbol{n}}$. These parameters are estimated from noisy adaptation data to maximize the likelihood of the data. MLLR adaptation can be either supervised or unsupervised. In the supervised adaptation scheme, MLLR requires a set of adaptation data to learn the noisy means. For the unsupervised adaptation scheme, the adaptation is performed on the data to be recognized itself. MLLR has been observed to work very well in many situations, including microphone array environments [14]. However, since the adapted models are assumed to be truly representative of the speech to be recognized, all of the adaptation data and the test data need to be acoustically similar. This amounts to requiring that the corrupting noise be quasi-stationary.

2.2.2 Codeword Dependent Cepstral Normalization and Vector Taylor Series

Codeword Dependent Cepstral Normalization (CDCN) [1][3] and Vector Taylor Series (VTS) [20][21] are model-based compensation methods that assume an analytical model of the environmental effects on speech. Noisy speech is assumed to be clean speech that has been passed through a linear filter and then corrupted by additive noise. This model is represented in the cepstral domain by a non-linear equation:

$$z = x + h + IDFT\{\ln(1 + e^{DFT(n-h-x)})\} \tag{2.2}$$

relating the cepstrum of the noisy speech z to the cepstrum of the clean speech x, the cepstrum of the unknown noise n, and the cepstrum of the impulse response of the unknown filter, h.

Both CDCN and VTS are algorithms that assume no prior knowledge of the filter or the noise. These methods estimate the parameters by maximizing the likelihood of the observed cepstra of noisy speech, given a Gaussian mixture distribution for the cepstra of clean speech. Since the transformation that relates the noisy cepstra to the clean cepstra is nonlinear, both CDCN and VTS approximate it as a truncated Tay-

lor series in order to estimate it.While CDCN uses a zeroth-order Taylor series approximation, VTS uses a first-order approximation. The estimated filter and noise parameters are then used to estimate the clean speech cepstra from the noisy cepstra or to adapt the HMMs to reflect the noisy conditions of the speech to be recognized. Both CDCN and VTS are highly efficient at medium levels of noise (*i.e.*at SNRs of 10 dB and above), but VTS performs slightly better. However, both algorithms assume that the noise is stationary, and thus, both perform poorly when this assumption is violated In [31], CDCN was applied to features derived from both delay-and-sum beamforming and cross-correlation-based auditory processing with improvements in recognition performance seen in both cases.

In this chapter, we have presented array-processing techniques that have been developed for multi-channel speech processing. Most of these techniques have been able to achieve some improvement in array-based speech recognition performance, but also make assumptions about the environment or speaker that are either unrealistic or highly restrictive. Furthermore, with the exception of the auditory model-based techniques, the algorithms are all speech enhancement algorithms designed to improve the SNR and perceived listenability of the target waveform, not speech recognition performance. We also presented some compensation algorithms originally developed for single-channel speech recognition, which have been successfully applied to microphone array speech recognition. It should be noted that our goal should be to make our front-end array processing methods and our speech recognition compensation methods as complementary as possible. That is, any array processing algorithm and speech recognition compensation algorithm applied in conjunction should result in better recognition performance than either of the methods applied in isolation.

In the next chapter, we present a framework for a new array processing methodology specifically designed for improved speech recognition performance, and some pilot experimental work demonstrating its preliminary implementation.

3. Preliminary Work in Recognizer-based Array Processing

As stated in the introduction, the goal of this work is to develop array-processing strategies specifically designed to improve speech recognition performance. As described earlier, most previous methods suffer from the drawback that they are inherently speech *enhancement* schemes, aimed at improving the quality of the speech waveform as judged perceptually by human listeners or quantitatively by SNR. While this is certainly appropriate if the speech signal is to be interpreted by a human listener, it may not be the right criterion if the signal is to be interpreted by a speech recognition system. Speech recognition systems do not interpret the waveform itself, but a set of *features* derived from the speech waveform. Furthermore, recognition systems are large statistical pattern classifiers which typically operate in a maximum likelihood framework [28]. By ignoring the manner by which the recognition system processes incoming signals, these speech enhancement algorithms are treating speech recognition systems as equivalent to human listeners, which is clearly not the case.

In this chapter, we describe some preliminary work in the development of an array-processing scheme that will be the foundation of the work proposed in this thesis. We propose a new filter-and-sum microphone array processing scheme that *integrates the speech recognition system directly into the filter design process*. We believe that incorporating the speech recognition system into the array processing design strategy ensures that the algorithm enhances those components of the output signal that are important for recognition, without undue emphasis on the unimportant components.

3.1 Filter-and-sum array-processing

We will employ traditional filter-and-sum processing to combine the signals captured by the array. In the first step, the speech source is localized and the relative channel delays caused by path length differences to the source are resolved so that all waveforms captured by the individual microphones are aligned with respect to each other. Several algorithms have been proposed in the literature to do this, *e.g.* [7][24]. In this work, we have used cross-correlation to determine the delays among the multiple channels.

Once the signals are time aligned, each signal is passed through an FIR filter whose parameters are determined by the calibration scheme described in the following section. The filtered signals are then added to obtain the final signal, as shown in Figure 3.1. This procedure can be represented as:

$$y[n] = \sum_{i=1}^{N} \sum_{k=0}^{K} h_i[k] x_i[n-k-\tau_i] \quad (1)$$

where $x_i[n]$ represents the n^{th} sample of the signal recorded by the i^{th} microphone, τ_i represents the delay introduced into the i^{th} channel to time align it with the other channels, $h_i[k]$ represents the k^{th} coefficient of the FIR filter applied to the signal from the i^{th} microphone, and $y[n]$ represents the n^{th} sample of the final output signal. K is the order of the FIR filter and N is the total number of microphones in the array. Once $y[n]$ is obtained, it can be parameterized to derive a sequence of feature vectors to be used for recognition.

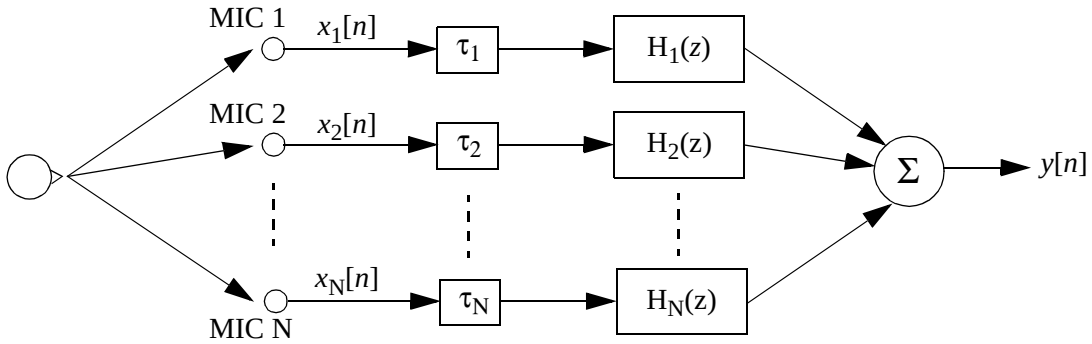

Figure 3.1 The filter-and-sum microphone-array-processing algorithm.

3.2 Speech Recognition-based Filter Calibration

As stated earlier, we propose designing a speech recognition-specific array-processing scheme. In the filter-and-sum approach, this means choosing the filter parameters $h_i[k]$ that will optimize speech recognition performance. One possible approach is to maximize the likelihood of the *correct* transcription for the utterance, thereby increasing the difference between its likelihood and that of other competing hypotheses. However, because the correct transcription of any utterance is unknown, we optimize the filters based on a single *calibration utterance* with a known transcription. Before using the speech recognition system, a user records a calibration utterance, and the filter parameters are optimized based on this. All subsequent utterances are processed using the derived filters in the filter-and-sum scheme described previously.

The sequence of recognition features derived from any utterance $y[n]$ is a function of the filter parameters $h_i[n]$ of all of the microphones, as in (1). In this work, recognition features are assumed to be mel-fre-

quency cepstra. The sequence of mel-frequency cepstral coefficients is computed by segmenting the utterance into overlapping frames of speech and deriving a mel-frequency cepstral vector for each frame. If we let $\boldsymbol{h}$ represent the vector of all filter parameters $h_i[k]$ for all microphones, and $\boldsymbol{y}_j(\boldsymbol{h})$ the vector of observations of the j^{th} frame expressed as a function of these filter parameters, the mel-frequency cepstral vector for a frame of speech can be expressed as

$$\boldsymbol{z}_j = DCT(\log(\boldsymbol{M}\left|DFT(\boldsymbol{y}_j(\boldsymbol{h}))\right|^2)) \tag{2}$$

where $\boldsymbol{z}_j$ represents the mel-frequency cepstral vector for the j^{th} frame of speech and $\boldsymbol{M}$ represents the matrix of the weighting coefficients of the triangular mel filters.This feature extraction process is shown in Figure 3.2.

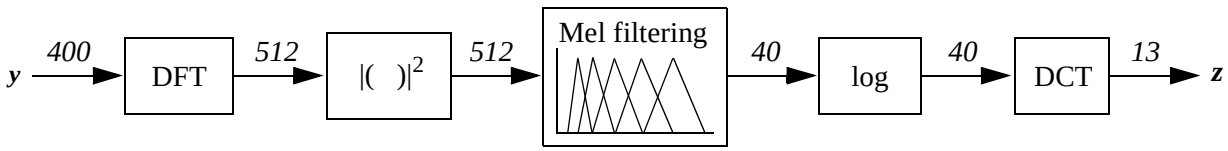

Figure 3.2 The derivation of mel-frequency cepstral coefficients (MFCC) for a frame of speech. The numbers on the arrows represent the number of terms generated by each block.These numbers may vary but are typical for most state-of-the-art speech recognizers.

The likelihood of the correct transcription must be computed using the statistical models employed by the recognition system. In this work, we use SPHINX-III, an HMM-based speech recognition system. For simplicity, we further assume that the likelihood of the utterance is largely represented by the likelihood of the most likely state sequence through the HMMs. Using this assumption, the log-likelihood of the utterance can be represented as

$$L(\boldsymbol{Z}) = \sum_{j=1}^{T} \log(P(\boldsymbol{z}_j|s_j)) + \log(P(s_1, s_2, s_3, \ldots, s_T)) \tag{3}$$

where $\boldsymbol{Z}$ represents the set of all feature vectors $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_T\}$ for the utterance, T is the total number of feature vectors (frames) in the utterance, s_j represents the j^{th} state in the most likely state sequence, and $\log(P(\boldsymbol{z}_j|s_j))$ is the log likelihood of the observation vector $\boldsymbol{z}_j$ computed on the state distribution of s_j. The *a priori* log probability of the most likely state sequence, $\log(P(s_1, s_2, s_3, \ldots, s_T))$, is determined by the transition probabilities of the HMMs. In order to maximize the likelihood of the correct transcription,

$L(\boldsymbol{Z})$ must be jointly optimized with respect to both the filter parameter vector $\boldsymbol{h}$ and the state sequence $s_1, s_2, s_3, \ldots, s_T$.

For a given $\boldsymbol{h}$, the most likely state sequence can be easily determined using the Viterbi algorithm. However, for a given state sequence, in the most general case, $L(\boldsymbol{Z})$ cannot be directly maximized with respect to $\boldsymbol{h}$ for two reasons. First, the state distributions used in most HMMs are complicated distributions, *i.e.* mixtures of Gaussians. Second, $L(\boldsymbol{Z})$ and $\boldsymbol{h}$ are related through many levels of indirection, as can be seen from (1), (2), and (3). As a result, iterative non-linear optimization methods must be used to solve for $\boldsymbol{h}$. Computationally, this can be expensive. A few additional approximations are made that reduce the complexity of the problem. We assume that the state distributions of the various states of the HMMs are modelled by single multivariate Gaussian, not mixtures of Gaussians. Furthermore, we assume that to maximize the likelihood of a vector on a Gaussian, it is sufficient to minimize the Euclidean distance between the observation vector and mean of the Gaussian. This assumption is equivalent to assuming that all Gaussians in all HMMs have independent components with equal variance. Thus, given the optimal state sequence, we can define an objective function to be minimized with respect to $\boldsymbol{h}$ as follows:

$$Q(\boldsymbol{Z}) = \sum_{j=1}^{T} \left\| \boldsymbol{z}_j - \mu_{s_j} \right\|^2 \tag{4}$$

where μ_{s_j} is the mean vector of the Gaussian distribution of the state s_j. Because the dynamic range of mel-frequency cepstra diminishes with increasing cepstral order, it is clear that our previous assumption regarding Gaussian components with equal variance is invalid. As a result, low-order cepstral terms will have a much more significant impact on the objective function (4) than higher ones. To avoid this potential problem, we redefine the objective function in the log mel-spectral domain, where the assumption of components with equal variance is more reasonable:

$$Q(\boldsymbol{Z}) = \sum_{j=1}^{T} \left\| IDCT(\boldsymbol{z}_j - \mu_{s_j}) \right\|^2 \tag{5}$$

Note that the IDCT operation in (5) transforms a thirteen-dimensional cepstral vector back to a forty-dimensional log mel-spectral vector. Using (1), (2), and (5), the gradient of the objective function with respect to $\boldsymbol{h}$, $\nabla_{\boldsymbol{h}} Q(\boldsymbol{Z})$, was determined. The gradient formulation is unwieldy and for brevity, is not included here. Using the objective function and its gradient, we can minimize (5) using gradient descent

1. Determine the array path length delays τ_i and time-align the signals from each the N microphones.
2. Initialize the filter parameters: $h_i[0] = 1/N$; $h_i[k]=0,\ (k \neq 0)$.
3. Process the signals using (1) and derive recognition features.
4. Determine the optimal state sequence from the obtained recognition features using Viterbi.
5. Use the optimal state sequence and (5) to estimate optimal filter parameters.
6. If the value of the objective function using the estimated filter parameters has not converged, go to Step 3.

Table 3.1 The calibration algorithm for filter-and-sum processing for speech recognition.

[26] to obtain locally optimal filter parameters $\boldsymbol{h}$. The entire algorithm for estimating the filter parameters for an array of N microphones using the calibration utterance is shown in Table 3.1.

An alternative to estimating the state sequence and filter parameters iteratively is to record the calibration utterance simultaneously through a close-talking microphone. The recognition features derived from this clean speech signal can either be used to determine the optimal state sequence, or used directly in (5) instead of the Gaussian mean vectors. However, even in the more realistic situation where no close-talking microphone is used, a single pass through Steps 1 through 6 seems to be sufficient to estimate the filter parameters. The estimated filter parameters are then used to process all subsequent signals in the filter-and-sum manner described in Section 3.1.

3.3 Experimental results

Experiments were performed using two databases to evaluate the proposed calibration algorithm, one using simulated microphone array speech data and one with actual microphone array data.

A simulated microphone array test set, “WSJ_SIM”, was designed using the test set of the Wall Street Journal (WSJ0) corpus [25]. Room simulation impulse response filters were designed using the well-known image method [4] for a room 4m x 5m x 3m with a reverberation time of 200ms. The microphone array configuration consisted of eight microphones placed around an imaginary 0.5m x 0.3m flat panel display on one of the 4m walls. The speech source was placed one meter from the array at the same height as

the center of the array, as if a user were addressing the display. A noise source was placed above, behind, and to the left of the speech source. A room impulse response filter was created for each source/microphone pair. To create a noise-corrupted microphone array test set, clean WSJ0 test data were passed through each of the eight speech source room impulse response filters and white noise was passed through each of the eight noise source filters. The filtered speech and noise signals for each microphone location were then added together. The test set consisted of eight speakers with 80 utterances per speaker. Test sets were created with SNRs from 0-25 dB. The original WSJ0 test data served as a close-talking control test set.

The real microphone array data set, “CMU_TMS”, was collected at CMU [31]. The array used in this data set was a horizontal linear array of eight microphones spaced 7cm apart placed on a desk in a noisy speech lab approximately 5m x 5m x 3m. The talkers were seated directly in front of the array at a distance of one meter. There are ten speakers each with fourteen unique utterances comprised of alphanumeric strings and strings of command words. Each array recording has a close-talking microphone control recording for reference.

All experiments were performed using a single pass through Steps 1-6 in the calibration algorithm described in the previous section. In all experiments, the first utterance of each data set was used as the calibration utterance. After the microphone array filters were calibrated, all test utterances were processing using the filter-and-sum method described in Section 3.1. Speech recognition was performed using the SPHINX-III speech recognition system with context-dependent continuous HMMs (eight Gaussian/state) trained on clean speech using 7000 utterances from the WSJ0 training set.

In the first series of experiments, the calibration procedure was performed on the WSJ_SIM test set with an SNR of 5 dB and the CMU_TMS test set. In the first experiment, the close-talking recording of the utterance was used in (5) for calibration. The stream of target feature vectors was derived from the close-talking recording and used in to estimate a 50-point filter for each of the microphone channels.

In the second experiment, the HMM state segmentation derived from the close-talking calibration recording was used to estimate the filter parameters. The calibration recording used in the previous experiment was force-aligned to the known transcription to generate an HMM state segmentation. The mean vectors of one Gaussian/state HMMs in the state sequence were used to estimate a 50-point filter for each

Array Processing Method	WSJ_SIM	CMU_TMS
Close-talking mic (CLSTK)	16.52	19.36
Single mic array channel	93.84	62.32
Delay and Sum (DS)	64.48	39.36
Calibrate Optimal Filters w/ CLSTK Cepstra	33.37	35.0
Calibrate Optimal Filters w/ CLSTK State Segmentations	36.5	37.07
Calibrate Optimal Filters w/ DS State Segmentations	40.2	34.95

Table 3.1 Word error rate for the two microphone array test corpora, WSJ_SIM at 5 dB SNR, and CMU_TMS, using conventional delay and sum processing and the optimal filter calibration methods described.

microphone channel.

Finally, we assumed that no close-talking recording of the calibration utterance was available. Delay-and-sum processing was performed on the time-aligned microphone channels and the resulting output was used with the known transcription to generate an estimated state segmentation. The Gaussian mean vectors of the HMMs in this estimated state sequence were extracted and used to estimate 50-point filters as in the previous experiment. The word error rates (WER) from all three experiments are shown in Table 3.1. The results using conventional delay-and-sum beamforming are shown for comparison. Large improvements over conventional beamforming schemes are seen in all cases. With the exception of the calibration using the close-talking-microphone-based state segmentation for the CMU_TMS test set (WER 37.07), all improvements in recognition accuracy between delay-and-sum beamforming and the calibration methods are significant with better than 95% confidence. Having a close-talking recording of the calibration utterance is clearly beneficial, yet substantial improvements in word error rate can be seen even when no close-talking recording is used.

Figure 3.3 shows WER as a function of SNR for the WSJ_SIM data set, using the proposed calibration scheme and, for comparison, conventional delay-and-sum processing. For all SNRs, no close-talking recordings were used. All target feature-vector sequences were estimated from state segmentations generated from the delay-and-sum output of the array.

Clearly, at low to moderate SNRs, there are significant gains over conventional delay-and-sum beam-

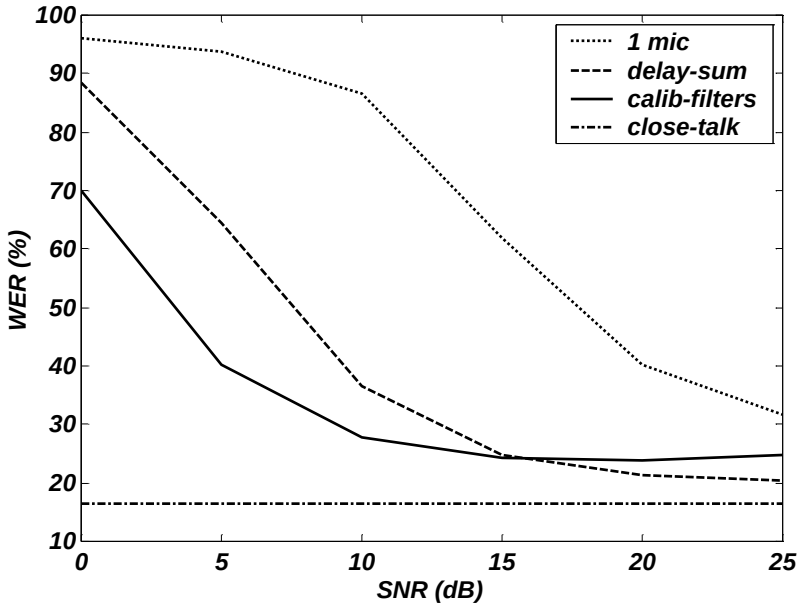

Figure 3.3 Word error rate vs. SNR for the WSJ_SIM test set using filters calibrated from delay-and-sum state segmentations.

forming. However, at high SNRs, the performance of the calibration technique drops below that of delay-and-sum processing. We believe that this is the result of using the mean vectors from one Gaussian/state HMMs as the target feature vectors. In doing so, we are effectively quantizing our feature space, and forcing the data to fit single Gaussian HMMs rather than the Gaussian mixtures which more accurately describe the data [28] and result in better recognition accuracy.

To demonstrate the advantage of estimating the filter parameters for each microphone channel jointly, rather than independently, a final experiment was conducted. The recognition performance using jointly optimized filters was compared to two other strategies: 1) performing delay-and-sum, then optimizing a single filter for the resulting output signal, and 2) optimizing the filters for each channel independently. These optimization variations were performed on the WSJ_SIM test set with an SNR of 10 dB. Again, 50-point filters were designed in all cases. The results are shown in Table 3.2 Joint optimization is significantly better than all other methods with better than 99% confidence.

It is clear from these experiments that significant gains in recognition accuracy can be achieved for microphone-array-based systems if the speech recognition system is incorporated into the design of the array processing strategy. We have empirically shown that by tuning the filter parameters to maximize the

Filter Optimization Method	WSJ_SIM
Delay and Sum	36.43
Optimize Single Filter for D & S output	36.29
Optimize Mic Array Filters Independently	48.19
Optimize Mic Array Filters Jointly	27.79

Table 3.2 Word error rate for the WSJ_SIM test set with an SNR of 10dB for delay-and-sum processing and three different filter optimization methods.

likelihood of the feature vectors derived from the resulting output signal, we are able to improve our speech recognition performance over conventional array processing techniques. In the next chapter, ideas for expanding this work are proposed in order to exploit this speech recognition design methodology.

4. Proposed Work

The results of the experiments described in the previous chapter confirm that further research into speech recognition-based array processing algorithms is merited. In this chapter, we describe the directions that will be pursued in this thesis research.

4.1 Reverberation Compensation

The experiments in the previous section have shown that the likelihood-based filter optimization strategy was effective at reducing the effects of noise on the recognition performance. However, reverberation is a significant source of performance loss in speech recognition systems, as noted in Section 2. Previous dereverberation methods require a complicated process involving external hardware to estimate the impulse response of the room and assume the reverberation levels do not change over time, which, in realistic environments, is not true. We propose to apply our technique to train longer filters in order to automatically compensate for the effect of room reverberation on the recognition features. It should be noted that we are not attempting to invert or “undo” the reverberation in the signal, just its effect on the derived recognition features. Compensating for the reverberation in this way allows the system to operate in environments where the reverberation levels change over time, or are unknown *a priori*.

4.2 Improved Objective Function

The current objective function employed in the filter parameter optimization is a Euclidean distance metric which compares the estimated log-mel spectra to the target log-mel spectra determined from the HMMs via an inverse DCT. Operating in the log-mel spectral domain is necessary because in the cepstral domain, the coefficients are not of equal dynamic range. Ideally, we would like to formulate and implement a true maximum likelihood objective function for filter optimization to match the criterion used in training and testing the speech recognizer. However, speech recognition systems operate on feature vectors consisting of not just the features themselves but their first and second derivatives, *e.g.* delta and delta-delta cepstra, as well. Formulating a maximum likelihood objective function and its gradient in terms of this full feature vector is difficult, if not impossible. As an approximation, we propose to incorporate the cepstral variances from the HMMs into an objective function defined over the features themselves and not

the derivatives. This is equivalent to using a separate speech recognizer trained only on feature vectors (and not their derivatives) for the filter optimization process.

In addition, it was seen in Figure 3.3 that in low-noise, reverberant conditions, approximating the target feature values of clean speech as the means of single Gaussians results in worse performance than conventional delay-and-sum processing. Therefore, we propose to refine the estimate of the target clean speech vectors by deriving them from mixtures of Gaussians.

4.3 Unsupervised Processing

The experiments performed in the previous section showed the potential for improved recognition in a calibration scenario. The resulting filter parameters, used in a conventional filter-and-sum array processing scheme could be applied in real time. However, if the real-time constraints are relaxed, as is often the case in transcription tasks, this algorithm could be applied in an unsupervised manner to each utterance to tune the array-processing filters to each utterance or group of utterances. This is expected to provide improved accuracy, especially in non-stationary noise situations.

4.4 Incorporation of Confidence

Confidence measures, such as [8], are used in various ways to quantify the reliability of the statistical hypotheses of the recognition system. In a degraded environment, there will be portions of the signal which will be less corrupted than others depending on the relative energies of the speech and noise at any given time. The less corrupted regions will typically result in better recognition accuracy. Therefore, estimating parameters using only the more reliable portions of the signal should allow better filter estimation. We can apply confidence scores to help decide which portions of the signal to use to tune the filter parameters.

4.5 Filter Adaptation

Performance of the algorithm could be improved by adapting the filter parameters over time. There are many possibilities for doing so. The filter can be adapted when a new speaker or a significant change in the environment is detected. This would not slow the algorithm, as the adaptation could be done in the background based on recent utterances.

4.6 Alternative Objective functions

Mel frequency cepstral coefficients (MFCC) are the most common speech recognition features used today. They are relatively simple to compute and provide good performance over a wide range of conditions relative to other feature sets. However, because of the non-linearity present in the formulation of MFCCs, minimization of objective function proposed in this work requires iterative optimization methods to find a solution. We can speed up the filter parameter computation tremendously if we can derive a linear feature formulation, such as Linear Prediction Coefficient-derived Cepstra (LPCC), whose objective function minimization would have a closed-form solution. We therefore plan to investigate the application of the ideas in this thesis to alternative feature sets which are linear in nature. It is believed that while the performance may not be as good as systems using MFCCs as a feature set, the improved speed of the algorithm would be a tremendous benefit.

4.7 Application to Single-Channel Speech

The algorithms presented have been applied in the context of a multi-channel input signal. Still, there is nothing inherent in the work that restricts its application to multiple channels. We propose to evaluate all algorithms in a single-channel context and compare it to other compensation methods, such as those presented in Section 2.

5. Thesis Goals and Timetable

5.1 Resources and Databases

Experiments to evaluate the effectiveness of the work in this thesis will be performed on actual multi-channel speech data collected by us and other researchers. We propose to test our methods on three tasks that represent a wide range of microphone array-based speech recognition environments, in terms of noise levels, reverberation levels, and array size.

- **Information Kiosk:** Many museums, airports, and other locations, are interested in installing information kiosks with which users may interact through voice, touch screen and other modalities. Such kiosks are usually placed in locations where both noise and reverberation levels are both high and extremely time-variant. These kiosks could be configured with a moderate number of fixed microphones (4-16).
- **Meeting Room:** There is a lot of interest in automatic meeting transcription and summarization. This environment, typically a conference room, is usually quite reverberant. The environmental noise levels are usually fairly low, but there is often significant amounts of co-channel speech present as meeting attendees will frequently interrupt each other or speak at the same time. The number of microphones is usually high and are in fixed locations.

Several sites and organizations (*e.g.* NIST, UC Berkeley-ICSI, CMU-ISL) are currently collecting multi-channel meeting room data for the meeting transcription task. In addition, we expect to collect data from information kiosk environments. Pilot work and preliminary studies will be performed on data already available, CMU_TMS and WSJ_SIM, the two corpora used in the pilot experiments presented in Section 3.

5.2 Expected Results and Contributions of Thesis

- Array-processing algorithms to effectively compensate for noise and reverberation with little or no *a priori* knowledge of the environment.
- Evaluation of the effect of varying levels of noise and reverberation on speech recognition features and overall speech recognition performance in real environments.
- Incorporation of speech recognition confidence measures into the array-processing design paradigm.

- Development of efficient adaptation schemes to update the array processing parameters over time.
- Full development of unsupervised array-processing.
- Exploration of alternate speech recognition-based objective functions that are computationally efficient.
- Evaluation of proposed techniques in conjunction with known compensation algorithms such as CDCN, VTS, and MLLR.
- Evaluation of proposed algorithms on single-channel speech data.

5.3 Preliminary Timetable of Work

Task	Start date	End date	Duration
Refinement of Objective Function	June 2001	Aug 2001	2 months
Investigation of design strategy for reverberation compensation	Aug 2001	Nov 2001	3 months
Multi-channel data collection and baseline evaluation	Nov 2001	Jan 2002	2 months
Unsupervised array processing parameter estimation	Jan 2002	Mar 2002	2 months
Investigation/Integration of confidence measures	Mar 2002	June 2002	3 months
Formulation and evaluation of alternate objective functions	June 2002	Aug 2002	2 months
Evaluation of array-processing algorithms with other compensation methods	Aug 2002	Sept 2002	1 month
Application of algorithms to single channel data	Sept 2002	Oct 2002	1 month
Dissertation write-up	Oct 2002	Jan 2003	3 months

References

[1] Acero A., *Acoustic and Environmental Robustness in Automatic Speech Recognition*, Boston, MA: Kluwer Academic Publishers, 1993.

[2] Acero, A., Altschuler, S., and Wu, L., "Speech/noise separation using two microphones and a VQ model of speech signals," *Proc. ICSLP '00*, Beijing, China.

[3] Acero, A. and Stern, R. M., "Environmental robustness in automatic speech recognition," *Proc. ICASSP '90*, Albequerque, NM.*

[4] Allen, J. B., and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943-950, 1979.

[5] Aoshima, N., "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, vol. 69, pp. 1484-1488, May 1981.

[6] Bell, A.J. and Sejnowski, T. J., "An information-maximization approach to blind separation and blind deconvolution," Neural Computation, vol. 7, pp. 1129-1159, 1995.

[7] Brandstein, M. S., and Silverman, H. F., "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language*, vol. 11, pp. 91-126, April 1997.

[8] Chase, L., "Word and acoustic confidence annotation for large vocabulary speech recognition," in *Proc. Eurospeech '97*, Rhodes, Greece, pp. 815-818.

[9] Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G. W., "Computer-steered microphone arrays for sound transduction in large rooms," *JASA*, vol. 78, pp. 1508-1518, Nov. 1985.

[10] Frost, O. L., "An algorithm for linear constrained adaptive beamforming," *Proc. of IEEE*, vol. 60, pp. 926-935, 1972.

[11] Griffiths, L. J., and Jim, C. W., "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, AP-30(1), pp. 27-34, Jan. 1982.

[12] Hughes, T. B., Kim, H. S., DiBiase, J. H., and Silverman, H. F., "Performance of an HMM speech recognizer using a real-time tracking microphone array as input," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, pp.346-349, May 1999.

[13] Johnson, D. H., and Dudgeon, D. E., *Array Signal Processing: Concepts and Techniques*. New Jersey: Prentice Hall, 1993.

[14] Kleban, J., and Gong, Y., "HMM adaptation and microphone array processing for distant speech recognition," in *Proc. ICASSP '00*, Istanbul, Turkey, pp.1411-1414.

[15] Kurita, S., Sauwatari, H., Kajita, S., Takeda, K., and Itakura, F., "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. ICASSP '00*, Istanbul, Turkey, pp. 3140-3143.

[16] Lee, T. W., "Examples of blind source separation of recorded speech and music signals," http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html.

[17] Leggetter, C. J., Woodland, P. C. (1994), "Speaker Adaptation Of HMMs Using Linear Regression,"

Technical Report CUED/F-INFENG/ TR. 181, Cambridge University Engineering Department, Cambridge, June 1994.

[18] Liu, Q. G., Champagne, B., and Kabal, P., "A microphone array processing technique for speech enhancement in a reverberant space," *Speech Communication*, vol. 18, pp. 317-334, 1996.

[19] Miyoshi, M., and Kaneda, Y., "Inverse filtering of room acoustics," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 36, pp. 145-152, Feb. 1988.

[20] Moreno, P. J., *Speech Recognition in Noisy Environments*, Ph.D. Dissertation, Carnegie Mellon University, May 1996.*

[21] Moreno, P. J., Raj, B., and Stern, R. M., "A vector Taylor series approach for environment-independent speech recognition," *Proc. ICASSP '96*, Atlanta, GA.*

[22] Neely, S. T., and Allen, J. B., "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, pp. 165-169, July 1979.

[23] Nordholm, S., Claesson, I., and Dahl, M., "Adaptive microphone array employing calibration signals: an analytical evaluation," *IEEE Trans. on Speech and Audio Proc*., vol. 7, pp. 241-252, May 1999.

[24] Omologo, M., and Svaizer, P., "Acoustic event localization using crosspower-spectrum phase based technique," in *Proc. ICASSP '94*, pp. 273-276.

[25] Paul, D., and Baker, J., "The design of the Wall Street Journal-based CSR corpus", *Proc. DARPA Speech and Natural Language Workshop*, Harriman, New York, pp. 357-362, Feb. 1992.

[26] Polak, E., *Computational methods in Optimization*, New York: Academic Press, 1971.

[27] Putnam, W., Rocchesso, D., and Smith, J., "A numerical investigation of the invertibility of room transfer functions," *Proc. IEEE ASSP Workshop on App. of Sig. Proc. to Audio and Acoust. '95*, Mohonk, NY, pp. 249-252.

[28] Rabiner, L. R., "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp.257-286, Feb. 1989.

[29] Raghavan, P., Renomeron, R. J., Che, C., Yuk, D. S., and Flanagan, J. L., "Speech recognition in a reverberant environment using matched filter array (MFA) processing and linguistic tree maximum likelihood linear regression (LT-MLLR) adaptation," in *Proc. ICASSP '99*, Phoenix, Arizona, pp. 777-780.

[30] Silverman, H. F., Patterson, W. R., and Flanagan, J. L., "The huge microphone array (HMA)," Brown University Technical Report, May, 1996.

[31] Sullivan, T. M., *Multi-microphone correlation-based processing for robust automatic speech recognition*, Ph.D. Dissertation, Carnegie Mellon University, August, 1996.*

[32] Sullivan, T. M. and Stern, R. M., "Multi-microphone correlation-based processing for robust speech recognition," *Proc. ICASSP '93*, Minneapolis, MN.*

[33] Van Compernolle, D., "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," *Proc. ICASSP '90*, pp. 833-836.

* indicates reference is available for download from http://www.cs.cmu.edu/~robust