

Thesis Proposal

# SPEECH DETECTION, CLASSIFICATION, AND PROCESSING FOR IMPROVED AUTOMATIC SPEECH RECOGNITION IN MULTIPARTY MEETINGS

Kofi A. Boakye

Committee: Michael Jordan, Michael Gastpar,  
Keith Johnson, and Nelson Morgan

December 18, 2006

## **Abstract**

Automatic speech recognition (ASR) in multiparty meetings presents a number of challenges owing to the complexity of the domain. The present research paradigm involves two major subtasks based on the sensors used for audio data collection: Individual microphones worn by the meeting participants and distant microphones placed in varying locations within the meeting room. In the case of the individual microphones, crosstalk speech is often the primary source of errors, making the segmentation of local speech of critical importance. I propose investigating the effectiveness of various features for an HMM based segmenter in terms of local speech detection (diarization error rate) and ASR performance (word error rate). The candidate features will mainly come from the set of cross-channel features, as they will likely introduce more robustness to crosstalk. For the distant microphones, it is overlapped speech that generates a significant number of recognition errors. Here, too, I propose as a first step using an HMM segmentation scheme for detection. The focus again will be on analyzing the effectiveness of candidate features in yielding high performance (in terms of diarization error rate) for the detection task. As a second step, I intend to investigate the ability of two speech separation techniques—harmonic enhancement and suppression and adaptive decorrelation filtering—to improve recognition word error rate.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Automatic speech recognition in multiparty meetings . . . . .	1
1.2	Crosstalk and overlapped speech . . . . .	3
1.3	Scope of Project . . . . .	4
1.3.1	Part I: Speech activity detection for nearfield microphones . . . . .	5
1.3.2	Part II: Overlap detection for farfield microphones . . . . .	5
1.3.3	Part III: Overlap speech processing for farfield microphones . . . . .	5
<b>2</b>	<b>Part I: Speech activity detection for nearfield microphones</b>	<b>6</b>
2.1	Related Work . . . . .	6
2.2	Candidate Features . . . . .	7
2.2.1	Cepstral Features . . . . .	7
2.2.2	Cross-channel Correlation . . . . .	8
2.2.3	Log-Energy Differences . . . . .	8
2.2.4	Time Delay of Arrival Estimates . . . . .	9
2.3	Feature Generation and Combination . . . . .	10
2.4	Work Plan for Part I . . . . .	10
<b>3</b>	<b>Part II: Overlap detection for farfield microphones</b>	<b>12</b>
3.1	Related Work . . . . .	12
3.2	Candidate Features . . . . .	13
3.2.1	Cepstral Features . . . . .	13
3.2.2	Cross-channel Correlation . . . . .	14
3.2.3	Pitch Estimation Features . . . . .	14
3.2.4	Spectral Autocorrelation Peak Valley Ratio . . . . .	16
3.2.5	Kurtosis . . . . .	17
3.3	Feature Generation and Combination . . . . .	18

3.4	Work Plan for Part II . . . . .	19
<b>4</b>	<b>Part III: Overlap speech processing for farfield microphones</b>	<b>20</b>
4.1	Related Work . . . . .	20
4.2	Candidate Methods . . . . .	21
4.2.1	Harmonic Enhancement and Suppression . . . . .	22
4.2.2	Adaptive Decorrelation Filtering . . . . .	24
4.3	Work Plan for Part III . . . . .	28
<b>5</b>	<b>Preliminary Experiments</b>	<b>29</b>
5.1	Experiment 1: Single feature performance . . . . .	29
5.2	Experiment 2: Initial feature combination . . . . .	31
	<b>References</b>	<b>33</b>

# 1 Introduction

In this section I provide a brief overview of the task of automatic speech recognition (ASR) in multiparty meetings, which is the area of interest for my proposed work. I give motivation through a discussion of the phenomena of crosstalk and overlapped speech which arise in the meetings domain and then describe the intended scope of the project. The scope, consisting of three main parts, outlines the subsequent sections of the proposal as well.

## 1.1 Automatic speech recognition in multiparty meetings

Perhaps more than any other domain, multiparty meetings represents a rich source of content for spoken language research and technology. From meeting data one can obtain rich transcription (transcription including punctuation, capitalization, and speaker labels), perform transcript indexing and summarization, do machine translation, or carry out high-level language and behavioral analysis with the assistance of dialog act annotation. Most of these procedures, however, rely on high quality automatic speech recognition (ASR) transcripts, and as such ASR in meetings is an important and active area of investigation.

In most typical set-ups, meeting ASR—also referred to as speech-to-text (STT) transcription—utilizes audio data obtained from various sensors located within the meeting room. The most common types are given below:

- **Individual Headset Microphone**

The individual headset microphone (IHM) is a head-mounted microphone positioned very close to the participant's mouth. The microphone is usually a cardioid or super-cardioid microphone and has the best quality signal for each speaker.

- **Lapel Microphone**

The lapel microphone (LM) is another type of individual microphone, but is placed on the participant's clothing. The microphone is generally omni-directional or cardioid and is more susceptible to interfering speech from other participants.

- **Tabletop Microphone**

The tabletop microphone is typically an omni-directional pressure-zone microphone (also called boundary microphone) and is placed between participants on a table or other flat surface. The number and

placement of such microphones varies based on table geometry and the location and number of participants.

- **Linear Microphone Array**

The linear microphone array (LMA) is a collection of omni-directional microphones with a fixed linear topology. Depending on the sophistication of the setup, the array composition can range from four to sixty-four microphones. The array is usually placed along the wall in a meeting room and enables the use of microphone beamforming techniques to obtain high signal-to-noise ratio (SNR) signals for the participants from a distance.

- **Circular Microphone Array**

The circular microphone array (CMA) combines the central location of the tabletop microphone with the fixed topology of the LMA. It consists of typically four or eight omni-directional microphones uniformly spaced around a horizontally oriented circle a few inches above table level. The array enables source localization and speaker tracking.

The first two types comprise the sensors for the *nearfield* or *close-talking microphone* condition and the last three the sensors for the *farfield* or *distant microphone* condition.

Recognition in the nearfield condition is generally performed by decoding each individual audio channel separately. The audio signal is converted into a sequence of feature vectors through a process referred to as *feature extraction*. This procedure seeks to yield a parametrization of the waveform that is robust and that captures as much of the information necessary to perform recognition while discarding the remainder, such as noise. Such sequences of feature vectors are modeled according to a hidden Markov model (HMM) with the hidden states corresponding to isolated phones or phones within a particular context. The decoding attempts to find the state (and, subsequently, word) sequence with the highest likelihood given the sequence of feature vectors. For the farfield condition, recognition is done in one of two ways. The data streams are combined either at the signal level (e.g., through some type of microphone beamforming) or at the recognition hypothesis level. The latter consists of generating hypotheses for individual channels and finding the most probable word sequence across all channels. This method tends to be much more computationally intensive and is less frequently used in practice. As is standard, the ASR performance measure for meetings is the word error rate (WER). The WER is defined as the sum

of all ASR output token errors divided by the number or scoreable tokens in a reference transcription. The errors are of three types: missed tokens (deletions), inserted tokens (insertions), and incorrectly recognized tokens (substitutions).

## 1.2 Crosstalk and overlapped speech

Automatic speech recognition in these multiparty meetings presents some specific challenges owing to the nature of the domain. The existence of multiple individuals speaking at various times leads to two phenomena in particular: crosstalk and overlapped speech. Crosstalk is a phenomenon associated only with the close-talking microphones and refers to the presence of speech on a channel that does not originate from the participant wearing the microphone. This speech is problematic because it is assumed that the speech coming from a given channel is to be attributed to the headset or lapel wearer for that channel; words generated from recognition of other participants' speech (non-local speech) are regarded as errors—in this case most likely insertion errors—for the ASR performance evaluation. In [23], for example, WER differed by 75% relative between recognition on segmented and unsegmented waveforms largely due to insertions from crosstalk. Overlapped, or co-channel, speech refers to the case when two or more participants are speaking simultaneously. Though present in both the nearfield and farfield conditions, its presence is most pronounced (and most severe) in the farfield case. Even in the nearfield case, the effects of overlapped speech can be quite significant. In a study conducted by Shriberg et al. [32], they demonstrated a 12% absolute difference between ASR performance on overlapped and non-overlapped speech segments in multiparty meetings.

The issue of crosstalk can be addressed within the framework of speech activity detection (SAD), a long-studied problem in speech processing and an important pre-processing step for ASR. The speech activity detection task consists of identifying the regions of an audio signal which contain speech from one or more speakers. This is in contrast to regions of nonspeech, which commonly includes low-level ambient noise (“silence”), laughter, breath noise, and sounds from non-human sources. For the nearfield condition, we add non-local speech (crosstalk) to the list of “nonspeech” phenomena. Though many methods exist for determining these speech activity regions, a common one—and the one of interest for this work—is to segment the audio into speech and nonspeech regions using an HMM based segmenter. The

segmenter, like an ASR system, operates using a parametrization of the audio signal into a sequence of feature vectors. The hidden states modeled here, though, correspond to speech and nonspeech. The decoding process seeks to determine the highest likelihood sequence of these states, from which a speech/nonspeech segmentation can be derived. Because of the acoustic similarity between local speech and crosstalk speech, the task of speech activity detection becomes more challenging. In particular, the features typically used in speech/nonspeech segmentation (e.g., log-energy and Mel-frequency cepstral coefficients) are insufficient in many cases to produce segmentations that yield good ASR performance.

To identify overlapped speech regions, a similar framework can be adopted. An HMM based segmenter can be used to detect not local, but co-channel speech. Again the detection task is complicated by the acoustic similarity between single-speaker and overlapped speech, so the selection of appropriate features is an area of interest. Unlike with crosstalk, excluding overlapped speech from processing by the recognizer is potentially a suboptimal approach. Ideally, these segments would be processed in a way that the speech could be more accurately recognized. Indeed, humans are able to perform such a task, a phenomenon called the *cocktail party effect*, without much difficulty. Such processing, referred to as source or speech separation, has been the subject of much study, but reported results in the literature typically come from artificial mixtures or very controlled environments (e.g., known and stationary microphone and speaker locations) and focus on speech intelligibility. It remains to be seen what performance improvements, if any, can be obtained on a state-of-the art meeting recognition system.

### 1.3 Scope of Project

This proposed thesis work divides into three parts, each in some way seeking to address the issues mentioned in the previous section. The first deals with speech activity detection for the nearfield condition and relates to the problem of crosstalk. The second discusses overlap detection for the farfield condition and the third overlap speech processing for this condition. These last two parts together make up a proposed method for addressing this co-channel speech problem.

### 1.3.1 Part I: Speech activity detection for nearfield microphones

For this part of my thesis I propose investigating the effectiveness of different features for speech activity detection using an HMM based segmenter as described in 1.2. Of particular interest is the ability of the features to aid in the exclusion of crosstalk speech from local speech detection. Performance will be measured according to two metrics: diarization error rate (DER)—a time-based metric—and word error rate—a token-based metric. The baseline features to be used for comparison are the standard cepstral features for an ASR system. Candidate features will derive mainly from cross-channel features as both intuition and evidence [38, 37] suggest that they are best suited to address the cross-channel phenomenon of crosstalk.

### 1.3.2 Part II: Overlap detection for farfield microphones

I propose using here, too, an HMM based segmenter for speech detection, though in this case it is overlapped speech from farfield microphone channels. Detection of overlapped speech is the first step in mitigating the effects of its presence on a meeting ASR system. I plan on analyzing different features for overlap detection based on performance as measured using DER. Again, the standard cepstral features will serve as baseline features. Single-channel rather than cross-channel features will dominate the pool of candidate features from which I intend to draw. Many of these, too, will be related to pitch, which will figure importantly in one of the speech separation methods I propose employing—harmonic enhancement and suppression—for speech separation in part III.

### 1.3.3 Part III: Overlap speech processing for farfield microphones

Detecting overlap, as was the focus of part II, only goes part of the way towards solving the problem posed by this phenomenon. Ideally, the signal should be processed in such a way as to enable the recognition of speech from each overlapped speaker. For this final part of my thesis I propose employing two speech separation methods—harmonic enhancement and suppression and adaptive decorrelation filtering—to see what performance improvements, if any, can be obtained on a state-of-the-art meeting recognition system. As mentioned, experimental results in the literature typically come from artificial mixtures or very controlled environments and focus on speech intelligibility, making the question of how such recognition systems should

address overlapped speech as yet not fully answered.

## 2 Part I: Speech activity detection for nearfield microphones

In this section I present the nearfield speech activity detection (SAD) component of my proposed thesis. I begin with a description of the related work in this area, and then describe the features I propose investigating. I discuss issues related to the generation and combination of these features, and then outline the work plan for this part of the thesis.

### 2.1 Related Work

Though single-channel speech activity detection has been studied in the speech processing community for some time now, the establishment of standardized corpora and evaluations for speech recognition in meetings is a somewhat recent development, and consequently the amount of work specific to multispeaker speech activity detection is rather small. The most relevant work to this thesis proposal comes from Wrigley et al. in [38] and [37]. The authors performed a systematic analysis of features for classifying multi-channel audio. Rather than look at the two classes of speech and non-speech, though, they subdivided the classes further into four: local channel speech, crosstalk speech, local channel and crosstalk speech, and no speech. They then looked at the frame-level classification accuracy (true and false positives) for each class with the various features selected for analysis. This was done for both features individually as well as combinations of features, the latter being done to find the best combination of features for a given class.

A key result from this work is that, from among the twenty features examined, the single best performing feature for each class was one derived from cross-channel correlation. This provides evidence of the importance of incorporating cross-channel information into modeling for this multi-channel detection task. This being the case, the features I intend to examine are primarily cross-channel in nature, as section 2.2 details.

Other speech activity detection work in multi-party meetings has been done by Pfau et al. in [29] and [30], along with Laskowski et al. in [14]. Both sets of results provide further evidence of the importance of using cross-channel

analysis to address the problem. Pfau et al. thresholded cross-channel correlations as a post-processing step to HMM based speech/nonspeech segmentation yielding on average a 12% relative frame error rate (FER) reduction. A different cross-channel correlation thresholding scheme by Laskowski et al. produced ASR WER performance improvements of 6% absolute over an energy-thresholding baseline. It is notable that in both cases thresholding rather than modeling was employed. Results could have potentially been further improved in a modeling based approach as was demonstrated in [4].

## 2.2 Candidate Features

The performance of various acoustically derived features for speech activity detection is the main area of investigation for this component of my thesis work and I proceed with a discussion of the features I intend to examine below.

### 2.2.1 Cepstral Features

The standard cepstral features serve as a baseline for performance of the speech activity detection system. These consist of 12th-order Mel-frequency cepstral coefficients (MFCCs), log-energy, and their first- and second-order time derivatives. The MFCCs are calculated as follows: An FFT is taken of a windowed version of the waveform. The magnitude coefficients are then binned by correlating them with each triangular filter in a Mel-scale (scaling based on human pitch-perception) filter bank. The log of these values is taken followed by a decorrelation procedure (DCT) and dimensionality reduction. These features are common to a number of speech-related fields—speech recognition, speaker recognition, speaker diarization, for instance—and so represent a natural choice for feature selection. The log-energy parameter, as well, is a fundamental component to most SAD systems and the cepstral features, being largely independent of energy, could provide information to aid in distinguishing local speech from other phenomena with similar energy levels. Breaths and coughs, for example, fall in this category and are quite prevalent on individual headset channels, especially for participants who possess poor microphone technique.

### 2.2.2 Cross-channel Correlation

For a pair of channels  $i$  and  $j$  the maximum cross-channel correlation  $C_{ij}(t)$  at time  $t$  is given by

$$C_{ij}(t) = \max_{\tau} \sum_{k=0}^{P-1} x_i(t-k)x_j(t-k-\tau)w(k) \quad (1)$$

where  $\tau$  is the correlation lag,  $x_i$  is the signal from channel  $i$ ,  $x_j$  is the signal from channel  $j$ ,  $w(k)$  represents a windowing function, and  $P$  is the window size. This is a clear first choice for a cross-channel feature to address crosstalk, and its use is documented in the literature (e.g., [20], [14], and [29]). In the examination of features for speech and crosstalk detection by Wrigley et al. [37], normalized cross-channel correlation was determined to be the most effective feature in detecting crosstalk. Normalization of this correlation value seeks to compensate for potential differences in channel gains. This tends to involve dividing the cross-channel correlation by the frame-level energy of the target channel, the non-target channel, or the square root of each. This last normalization, referred to as spherical normalization, converts  $C_{ij}(t)$  to a cosine metric that measures the angle between the vectors  $[x_i(t) \dots x_i(t-P-1)]^T$  and  $[x_j(t-\tau) \dots x_j(t-\tau-P+1)]^T$  [37]. The result is a correlation coefficient value between -1 and 1. In [37] only spherical and target channel energy normalization were examined. Curiously, preliminary results suggest that non-target channel energy normalization performs best for the data I have examined.

### 2.2.3 Log-Energy Differences

Just as energy is a good feature for detecting speech activity for a single channel, relative energy between channels should serve well to detect local speech activity using multiple channels. For example, if a single participant is speaking, his or her channel should have the highest relative energy. Furthermore, if there is crosstalk on the other channels, the energy on these channels is coupled with that of the speaker's and the relative energy over the crosstalk segment should be approximately constant. This pattern can be modeled in the HMM segmenter. Inclusion of first and second order differences as well may also help in the modeling. The log-energy difference (LED), which is proposed here, represents the log of the ratio of short-time energy between two channels. That is, for channels  $i$  and  $j$  at frame index  $t$ ,  $D_{ij}(t) = E_i(t) - E_j(t)$  where  $E$  represents log-energy. This feature

seems to be much less utilized in practice and much less discussed in the literature than cross-channel correlation, though the work I have done so far suggests it can be more robust—and consequently better performing—than correlation-based features [4]. Liu and Kubala introduce log-energy difference for segmentation in [20], but make no comparison of the feature’s performance to cross-channel correlation or any other feature.

Just as with cross-channel correlation, normalization techniques should be employed for LEDs to compensate for potential gain differences between channels. A channel-level energy normalization described by Pfau et al. in [30] is what I will adopt. This normalization occurs prior to the log-energy differencing and consists of subtracting the minimum frame log-energy of a channel from all frame log-energy values in the channel. That is, for a channel  $i$  at frame index  $t$ ,

$$E_{norm,i} = E_i(t) - E_{min,i} \quad (2)$$

where  $E$  again represents log-energy. This minimum frame log-energy serves as a noise floor estimate for the channel and has the advantage of being largely independent of the amount of speech activity in the channel.

#### 2.2.4 Time Delay of Arrival Estimates

In [7] Ellis and Liu demonstrated that speaker diarization—the task of determining who spoke when—for meetings could be performed using only time-delay of arrival (TDOA) estimates from distant microphones as features for a clustering and segmentation system. This result was improved upon by Pardo et al. [26], who then also demonstrated in [27] that the delay estimates, when used in conjunction with cepstral features, improved performance over cepstral features alone. In the case of close-talking microphones, TDOA estimates seem particularly well suited to distinguish local speech from crosstalk as crosstalk lags behind local speech. I plan to compute delay estimates using the generalized cross-correlation with phase transform (GCC-PHAT) method common to microphone array processing. For two channels  $i$  and  $j$  the GCC estimate can be expressed as

$$d_{gcc,ij}(t) = \underset{\tau}{\operatorname{argmax}} R_{ij}(\tau) \quad (3)$$

where

$$R_{ij}(\tau) = \int_{-\infty}^{\infty} W(\omega) X_i(\omega) X_j^*(\omega) e^{j\omega\tau} d\omega \quad (4)$$

$W(\omega)$  represents a weighting function and for the phase transform variation of the GCC we have

$$W_{PHAT}(\omega) = |X_i(\omega)X_j^*(\omega)|^{-1} \quad (5)$$

By dividing by the cross-spectrum magnitude,  $R_{ij}(\tau)$  depends only on the phase difference between the two signals. This weighting is suboptimal under ideal conditions, but tends to be more robust to noisy and reverberant conditions than traditional cross-correlation [5]. An issue in using TDOA estimates is the potential noisiness of the estimates. One technique to address this discussed by Anguera et al. in [2] is to compute delay estimates based not on the single largest correlation peak for the frame, but the top  $N$  and then perform Viterbi decoding to find the best sequence of delays. Other smoothing techniques from time-series analysis (e.g., median filtering) can be investigated as well.

### 2.3 Feature Generation and Combination

One key consideration in generating feature vectors for the SAD segmenter that are based on cross-channel feature values is the variable number of channels between meetings; For some standard meeting data corpora ([8], [10], [6]), the number of close-talking channels can vary between three and twelve. This variability in the number of channels must be reconciled with the need to have feature vectors of a single fixed length. I propose adopting the technique described by Wrigley et al. in [37] of using order statistics—specifically maximum and minimum—of the feature values generated on the different channels.

Along with length standardization for a given cross-channel feature, the combination of these features must also be considered. The most basic approach is to simply concatenate the features into a single vector, but this may be suboptimal in terms of system performance. I intend to analyze whether some dimensionality reduction or combination schemes may yield better results. The primary techniques that will be considered are principal component analysis (PCA), linear discriminant analysis (LDA), and multilayer perceptron (MLP) feature combination.

### 2.4 Work Plan for Part I

As mentioned, the objective of this part of my proposed thesis is to investigate the effectiveness of different features for speech activity detection using

an HMM based segmenter. To do this I intend to compare performance of HMM segmentation systems incorporating these different features. The performance will be measured in both recognition word error rate (as defined in 1.1) as well as diarization error rate (DER). Diarization error rate is the primary metric for the speaker diarization task as defined in [1] and serves as a time-based measure of the fraction of speaker time that is not attributed correctly to a speaker. In the case of SAD where only a single speaker is involved this consists of speaker false alarms and missed detections. The DER in most cases correlates closely with WER while not requiring the expensive computation of recognition hypotheses and so facilitates more exploratory experiments (e.g., those done for parameter tuning).

The data used for measuring performance will be drawn from the NIST Rich Transcription (RT) Meeting Recognition evaluation. This consists of collections of 10- to 12-minute excerpts of recordings of multiparty meetings from different sites—and, thus, with different room acoustics. As stated in section 2.2.1, the baseline performance measure for segmentation will be obtained using the standard cepstral features. The performance of each feature in isolation will be obtained as well as in conjunction with the baseline features. In terms of more extensive feature combination, I intend to determine the overall best combination of features and the best combination technique that obtains this.

A significant amount of work has already been done toward these ends, I should mention. The HMM based segmentation system has been implemented and run using baseline cepstral, cross-channel correlation, and log-energy difference features. Results for some of this work are presented in section 5. Also, some initial feature combination experiments have been performed and are presented in the section.

It should be noted, too, that though three candidate features have been given, there exist dimensions of exploration within each feature type as well. Cross-channel correlation, for example, can be computed in different ways (recall the normalizations mentioned in 2.2.2). TDOA estimates, too, can be computed and smoothed in multiple ways. Most likely here, too, the importance of having a quickly-computed alternative to the WER performance metric will emerge.

### 3 Part II: Overlap detection for farfield microphones

In this section I present the farfield overlap detection component of my proposed work. Related work is first presented followed by the candidate features I intend to investigate. Some discussion about feature generation and combination proceeds, and finally I propose a work plan for this part of the thesis.

#### 3.1 Related Work

Given that speech plus crosstalk was one of the four classes defined for the detection task by Wrigley et al., the work has relevance here as well. A significant difference between that work and my own proposed work is that Wrigley et al. performed their analysis on nearfield condition speech. The speech from the close-talking microphones has a higher SNR and suffers less from reverberant effects and so the behavior—and, consequently, performance—of the features may differ. This is certainly the case in the STT task, where word error rates for farfield consistently exceed those of nearfield (see, for example, [23] and [33]). That being said, Wrigley’s results point to energy, kurtosis, and cross-channel correlation as being the most effective features, all of which I intend to examine.

The majority of the work done on overlapped or co-channel speech detection has been done within the framework of identifying “usable” speech for speaker recognition tasks. By “usable” it is meant that the speech frame or segment can provide information to aid in determining the identity of a particular speaker. This depends largely on the ratio of target to interfering speaker energy—referred to as the TIR—of the frame or segment (see [21]). Lewis and Ramachandran in [19] compared the performance of three features—MFCCs, LPCCs, and a proposed pitch prediction feature (discussed later)—for speaker count labeling of speech frames in both a closed-set (speaker dependent) and open-set (speaker-independent) scenario. The results indicated that the proposed pitch prediction feature was superior to either the MFCCs or LPCCs. Unfortunately, no combination of features was performed to see if such an approach would yield improvements. Shao and Wang in [31] employed multi-pitch tracking for identifying usable speech for closed-set speaker recognition. Speaker count labeling was determined by the number of pitch tracks in the frame and single-speaker frames were

deemed usable. A large effort has been given by Yantorno et al. in this area as well. Spectral autocorrelation peak-valley ratio (SAPVR), adjacent pitch period comparison (APPC), and kurtosis have all been proposed and analyzed. Comparison and combination experiments have also been performed (see [39]), but this has been within the context of identifying usable overlapped speech frames and not overlapped speech frames from single-speaker speech frames.

With the exception of Wrigley et al., the work described above was all done with overlapped speech from artificial mixtures. In the case of usable speech selection this is essentially a necessity as frame-level TIRs must be known for evaluating classification accuracy. For overlap detection in multi-party meetings, however, this is not the case. Overlaps can be defined as segments based on time alignments of words. As such, the work I propose represents a significant departure from these studies.

## 3.2 Candidate Features

Much like part I, this component of the thesis focuses on an analysis of feature performance for the task of segmentation—in this case, segmentation of overlapped speech. Below is a discussion of the features I plan on investigating.

### 3.2.1 Cepstral Features

The standard cepstral features—i.e., 12 MFCCs, log-energy, and the first and second-order time derivatives—will again serve as a baseline here in the overlap detection system. The cepstral coefficients, which serve as a representation of the speech spectral envelope, should be able to provide information about whether multiple speakers are active in a time segment. Indeed, in Zissman et al. [45] a Gaussian classifier using cepstral features was able distinguish between target-only, jammer-only, and target plus jammer, speech segments with a reported accuracy of 80%. It should be noted that the detection task differed in a number of significant ways. Most importantly, the task was speaker-dependent—the training and test target and jammer speakers were the same. In addition, the overlapped speech was artificially generated by adding single-speaker speech waveforms. Thirdly, intervals of silence were removed beforehand. Because of these and other differences I anticipate significantly different performance results but a similar conclusion as to the usability of cepstral features for overlap detection.

### 3.2.2 Cross-channel Correlation

As previously mentioned, Wrigley et al. showed that, for their four-class segmentation task, the best feature for each class—one of which was speech plus crosstalk—was derived from cross-correlation. Though this is the case, it is not clear whether such features will be effective for identifying overlapped speech in the farfield condition. In the nearfield condition, because the personal microphone is significantly closer to a given participant than to any other simultaneous talker, overlapped speech segments tend to have a low cross-channel correlation. This large asymmetry in speaker-to-microphone distances is not typically present in the distant microphone case, and so the low cross-channel correlation tendency may not be present. In choosing this feature for analysis, I am looking to determine whether or not this is the case. The proposed features will be calculated according to equation (1), but between the various distant microphone signals.

### 3.2.3 Pitch Estimation Features

Features derived from various pitch estimation algorithms could be of use since the behavior of these algorithms should differ between single-speaker and multi-speaker speech segments, at least in the case of simultaneous voicing. For example, Lewis and Ramachandran in [19] proposed a pitch prediction feature (PPF) for the task of identifying temporal regions or frames as being either one-speaker or two-speaker speech. This feature was computed as the standard deviation of the distance between pitch peaks as obtained from the autocorrelation of a linear prediction (LP) residual. For a single speaker, the peaks will occur in regular sequence and correspond to glottal closures. For two speakers with different fundamental frequency, both sets of glottal closures will appear in the residual and this standard deviation of the inter-peak differences will be higher. In their work they found that the PPF outperformed both MFCC and linear prediction cepstral coefficient (LPCC) features.

The LP residual autocorrelation, too, is only one of several ways to estimate pitch. Some other major categories include:

- **Zero-Crossing Distance**

For a periodic signal the separation between zero-crossings tends toward half of the pitch period  $T$ . If one creates a histogram of these zero-crossing distances (ZCDs), there should be a peak correspond-

ing this value of  $T/2$ , from which the pitch can be estimated. This is one of the simplest methods in terms of computation, but it tends to be rather sensitive to noise and the presence of harmonics, often producing octave (pitch halving or doubling) errors.

- **Autocorrelation Function**

The autocorrelation function (ACF) for a signal  $x(n)$  is given by

$$\phi_{xx}(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n+|\tau|) \quad (6)$$

and for periodic signals produces peaks separated by the pitch period  $T$ . This method is simple in computation, but is prone to errors in the presence of noise as well.

- **Absolute Magnitude Difference Function**

The absolute magnitude difference function (AMDF) of  $x(n)$  is given by

$$\psi(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} |x(n) - x(n+\tau)| \quad (7)$$

The AMDF has the characteristic that when  $x(n)$  is similar to  $x(n+\tau)$ ,  $\psi(\tau)$  becomes small. Thus if  $x(n)$  has period  $T$ ,  $\psi(\tau)$  produces a deep notch at  $\tau = T$ . The AMDF typically exhibits better robustness to noise than the ACF and the ZCD methods.

An illustration of these pitch methods is shown in figure 1.

I intend to examine how these various pitch detectors generally behave in the presence of overlapped speech and, if that behavior differs significantly, how to encode this into one or more features. These methods can also be applied at a subband level—for example, on the output of a gammatone filterbank—which has been shown to yield improved results. This multi-band technique may be especially appropriate for the overlap detection scenario because the harmonic energy from the different speakers may be concentrated in different bands, which should aid in the detection of multiple pitches.

A possible issue in using pitch-related features for overlap detection is that they may not be well-behaved in unvoiced speech regions. One solution would be to include a feature which indicates voicing as well. Some examples include energy, zero-crossing rate, and spectral tilt. A voicing feature, however, does not address the detection of unvoiced overlap or of overlap

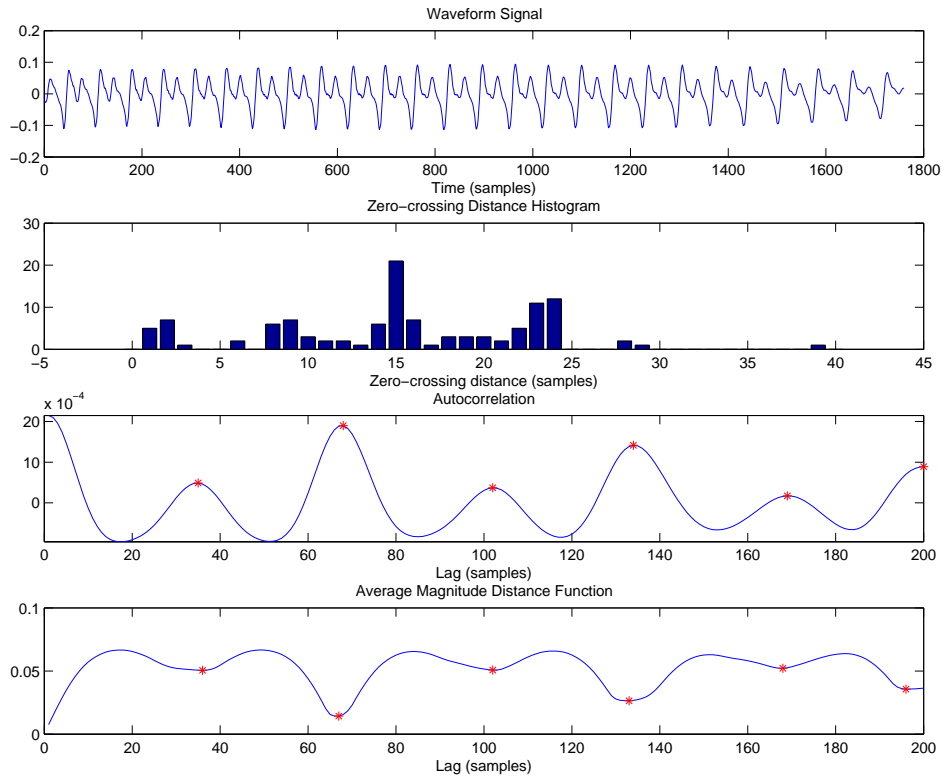


Figure 1: *Illustration of pitch detection. Top pane: Waveform of ‘o’ vowel in ‘dog’ uttered by a female speaker; Second pane: Zero-crossing distance histogram (note pitch halving); Third pane: Autocorrelation function (‘\*’ denotes automatically obtained local maximum); Bottom pane: Average magnitude distance function (‘\*’ denotes automatically obtained local minimum)*

with one-speaker voicing using pitch-related features. It may be that such features are ineffective in these cases and the results remain to be seen.

### 3.2.4 Spectral Autocorrelation Peak Valley Ratio

Related to the pitch-derived features is the spectral autocorrelation peak valley ratio (SAPVR). The spectral autocorrelation of a signal is obtained

by computing the autocorrelation of the signal’s magnitude spectrum. When computed on a frame-level basis, it can serve to reveal the existence of local harmonic structure. For a given voiced frame of speech, the spectral autocorrelation will consist of a series of peaks with progressively decreasing height and spacing equivalent to the fundamental frequency. If the speech frame is unvoiced, the spectral autocorrelation will lack any prominent peaks other than the one at zero lag. In analyzing the spectral autocorrelation of overlapped speech, Krishnamachari et al. in [11] observed that for segments of overlapped speech in which both speakers were voiced, the spectral autocorrelation contained either two distinct trains of harmonically related pulses if the speakers’ pitch separation exceeded approximately 25%, or one broad train of pulses. Based on this they proposed the SAPVR metric, which is defined as

$$\text{SAPVR} = 20 \log_{10} \frac{R(p_1)}{R(q_1)} \quad (8)$$

$R(p_1)$  is the local maximum of spectral autocorrelation other than the one for zero lag and  $R(q_1)$  is either the next local maximum that is not harmonically related to the first peak, or the local minimum between  $p_1$  and  $2p_1$ . Using this definition, the SAPVR for the two types of overlapped speech cases mentioned above should be lower than that of single speaker speech, suggesting SAPVR could be a useful feature for overlap detection.

An illustration of the spectral autocorrelation for single-speaker and overlapped speech is shown in figure 2.

### 3.2.5 Kurtosis

The kurtosis of a zero-mean random variable  $x$  is defined as

$$\kappa_x = \frac{E\{x^4\}}{\{E\{x^2\}\}^2} - 3 \quad (9)$$

where  $E\{\cdot\}$  is the expectation operator. This serves as a measure of the “Gaussianity” of a random variable, with super-Gaussian, or leptokurtic, random variables having kurtosis greater than zero and sub-Gaussian, or platykurtic, random variables having kurtosis less than zero. Speech signals, which are typically modeled as having a Laplacian or Gamma distribution, tend to be super-Gaussian. Furthermore, the sum of such distributions—in line with the central limit theorem—has lower kurtosis (i.e., is more Gaussian) than individual distributions. This was observed by LeBlanc and DeLeon in [15] and Krishnamachari et al. in [12]. As such, signal kurtosis

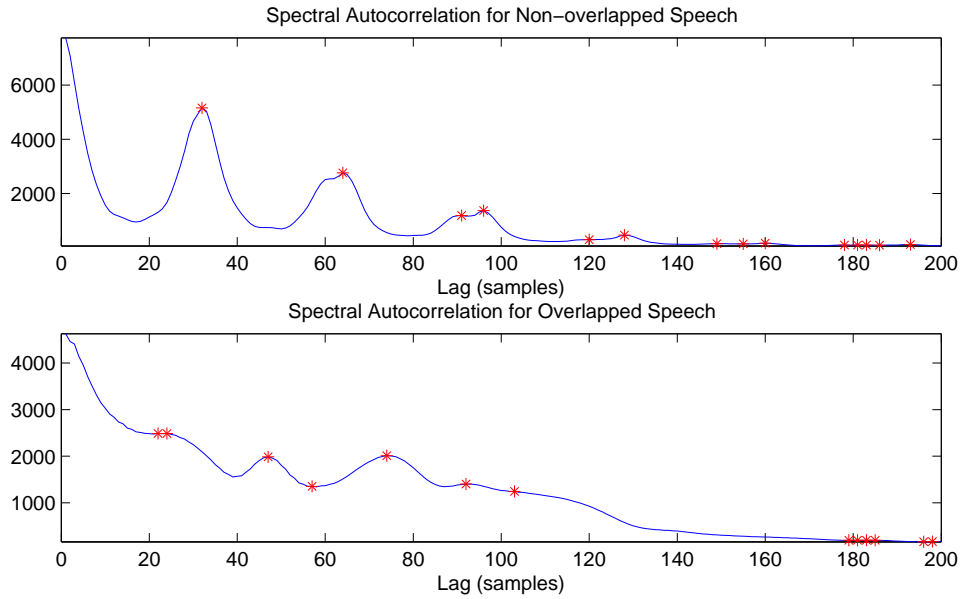


Figure 2: *Illustration of spectral autocorrelation function. Top pane: Spectral autocorrelation for non-overlapped speech (\* denotes automatically obtained local maximum); Bottom pane: spectral autocorrelation for overlapped speech (\* denotes automatically obtained local maximum)*

could serve as an effective feature for detecting overlapped speech. This was one of the features considered by Wrigley et al. and was one of the better performing features for detecting speech plus crosstalk; Indeed, it was selected by their sequential feature selection (SFS) algorithm for inclusion in the final feature ensemble for overlapped speech detection.

### 3.3 Feature Generation and Combination

In the case of nearfield condition speech activity detection we were presented with the problem of standardizing the length of the feature vector given the variable number of channels between meetings. In the farfield case, too, we have a variable number of channels between meetings, but the issues resulting from this are somewhat different. Aside from cross-channel correlation, all of the features mentioned above can be generated from the data of a single audio channel. Given that there are several channels, we now have

the challenge of using multiple (and variable numbers of) data streams to generate a single stream of features. One way to sidestep the issue is to simply select a single “best” channel, perhaps based on SNR estimates and do all processing on it; This is likely suboptimal because, for example, a delay-and-sum beamforming of the channels could significantly increase the SNR of the signal from which features would be extracted. This has been demonstrated through ASR performance in meetings and the technique is commonly used (as mentioned in section 1). An exception to the benefits of delay-and-sum beamforming might be in the case of pitch-related features. Phase differences between the signals to be summed could degrade the performance of the pitch estimation and consequently the features derived from it. Aside from using a single signal or combining signals, one could also combine feature streams generated from the individual distant microphone channels. Because of the large number of channels (potentially on the order of 60 or 70) and the related increase in computational complexity, this method is not suitable for my purposes. Selecting a single channel or delay-and-sum beamforming are more feasible for this work and I plan on employing both methods and comparing them.

Having obtained the different types of features, there is again the issue of combining them. I intend to use the same approaches as in Part I—concatenation, LDA, PCA, and MLP combination—and compare performance. In particular it will be of interest to see whether the combination results are similar between the two tasks.

### 3.4 Work Plan for Part II

As the general framework for part II closely resembles that of part I, the work plans for the two are similar. Here, too, I will compare performance of HMM based segmenters by incorporating various acoustically derived features into the systems. In this case the objective speech to detect will be overlapped farfield speech rather than local nearfield speech. In the previous case both WER and DER were selected as metrics for performance. Since the segmentation of overlapped speech for this portion of the work is only for identification purposes—no further processing of the speech will be performed—only DER serves as a meaningful performance measure. Unlike with general speech activity, human-labeled time references for overlapped speech regions are not available. As such, I will compute overlap references based on word-level forced-alignment timing of references from the nearfield microphones. Forced-alignment seeks to obtain fine time demarcations (be-

gin and end times) at the word, phone, or state level based on manually obtained reference transcriptions.

Again the data set used will come from the NIST RT meetings recognition evaluations and the baseline features will be the standard cepstral features. Feature performance comparisons as well as feature combination will proceed along the same lines as before, too.

## 4 Part III: Overlap speech processing for farfield microphones

In this section I present the overlap processing portion of my proposed work. In what follows, I describe the most relevant related work and then present the methods I plan on employing based on these studies. Within this presentation I discuss some of the key issues related to applying these techniques. Lastly, I give the work plan for this final component of my thesis.

### 4.1 Related Work

The separation of mixtures of speech signals is a much studied area of speech processing. For many cases this problem can be viewed as a subset of the general blind source separation (BSS) problem. That is, given signals  $X = [x_0 \dots x_N]^T$  produced by the mixing of signals  $S = [s_0 \dots s_M]^T$ ,  $M \leq N$  by an unknown mixing matrix  $\mathbb{A}$ , we seek to separate out the  $s_i$  components present in the  $x_i$ 's through the use of a matrix  $\mathbb{W}^T$ . If we assume the original  $s_i$ 's are independent and that at most one is Gaussian distributed, then the solving method becomes one of independent component analysis (ICA), for which a number of techniques exist and have been demonstrated [9]. For real world audio signals, however, the mixing is more accurately modeled as being convolutive rather than multiplicative. By reformulating the problem in the Z-transform domain, similar solutions are possible. Most of these are iterative approaches based on an information maximization criteria—for example, Lambert et al. [13], Lee et al. ([17, 16, 18]), and Torkkola [34]—and in the case of Lee and Bell [17] yielded improved recognition results for digit recognition in a real room environment. Another set of approaches based on minimizing cross-channel correlation—termed adaptive decorrelation filtering (ADF)—has been explored by Weinstein et al. [35, 40], and Yen et al. [41, 42, 43, 44] and has demonstrated improved recognition performance,

though in this case on a simulated mixture with coupling transfer functions estimated from a real room.

One limitation of the BSS framework for speech separation is that it is fundamentally a multi-channel approach; It cannot address the separation of speakers given only a single mixture. A set of techniques based on computational auditory scene analysis (CASA) seeks to perform this single-channel separation task by partitioning the audio spectrogram such that each partition is a speech stream for one of the overlapped speakers. The partitioning typically relies on the existence of certain types of structure in the speech signal and uses grouping cues such as pitch, continuity, and common onset and offset. Bach and Jordan in [3], for example, describe a CASA-based speech separation approach that uses spectral clustering to create speech stream partitions. The approach, though demonstrating promising results, is quite computationally intensive. A simpler though related method which exploits only the harmonic structure within speech—termed harmonic enhancement and suppression (HES)—has also been proposed for speech separation. HES utilizes pitch estimation to identify a speaker’s harmonics and enhance them, while in some cases also suppressing the harmonics of any interfering speaker. In [22] Morgan et al. provide a detailed description of the design and implementation of a HES based speaker separation system. Results on keyword spotting experiments suggest that such an approach may be useful in an ASR context as well.

## 4.2 Candidate Methods

In the following sections I describe methods I intend to explore for the separation of overlapped speech in the farfield condition. The selection of these methods was based partly on their ability to be implemented as part of a state-of-the art meeting recognition system. A key criterion for this is limited computational complexity. The harmonic enhancement and suppression (HES) method, for example, was implemented as a real-time speech separation system by Morgan et al. in [22].

It should be mentioned, too, that for this work I have chosen to focus only on the case of two-speaker speech. In their analysis of speaker overlaps in meetings, Çetin et al. in [25] reveal that about 11% of speech frames constituted two-speaker speech while only 1% constituted three-speaker speech—an order of magnitude difference. The focus is necessary because, although the methods can indeed be extended to address overlap from more speakers, this would require a significant amount of modification; this, too, would be at

the risk of hampering performance for the common case of two overlapped speakers.

#### 4.2.1 Harmonic Enhancement and Suppression

Harmonic enhancement and suppression (HES) refers to a class of single-channel speech separation methods that utilize the harmonic structure of voiced speech to separate speakers. In this approach the harmonics of a speaker are identified using pitch estimation and a signal for that speaker is generated by enhancing those harmonics. Alternatively, the time-frequency (T-F) bins of the short-time Fourier transform (STFT) that correspond to the neighborhood of the harmonics are selected, the other bins are zeroed out, and the signal is then reconstructed. This *harmonic selection* technique represents a simple form of binary masking, a procedure used for single-channel speech separation in conjunction with CASA. To obtain the signal for an additional speaker, the first speaker’s harmonics are suppressed and/or the other speaker’s harmonics are enhanced, if this speaker’s pitch can be determined. This is possible either by simultaneous multi-pitch tracking as in [28] or by pitch tracking of the signal in which the first speaker’s harmonics are suppressed as in [22]. The latter is most effective in cases where one speaker dominates the overlapped segment and consequently has the dominant pitch track for the original signal.

Below I discuss some of the key issues associated with this HES approach as it relates to the speech separation task in meetings

#### Pitch Estimation and Tracking

The effectiveness of the HES approach hinges greatly on high accuracy pitch estimation. For this reason, I plan on employing different pitch estimation techniques and comparing results. Aside from the methods mentioned in section 3.2.3, I also intend to look at maximum likelihood (ML) pitch estimation described by Wise et al. in [36]. This technique was used for HES by Morgan et al. in [22] and in [24] it was shown to outperform three other pitch estimation techniques—cepstral, harmonic matching, and the auditory synchrony model. The ML formulation for determining the optimal pitch period  $p$  over an  $N$ -point interval of speech is given by [22]:

$$E[p] = \frac{2p}{N} \sum_{l=1}^L \phi[lp] \tag{10}$$

where  $\phi[lp]$  is the autocorrelation function. The integer value  $\tilde{p}$  is chosen

that maximizes  $E[p]$  for  $L = \lfloor (N - P)/p \rfloor$  complete periods within the interval. This technique performs well in the presence of additive noise, which makes it suitable for this farfield condition. The noise robustness should also help when the interfering speech is unvoiced.

In addition to pitch estimation, good pitch tracking is also very important. This is particularly true for the co-channel speech case, as speaker dominance may alternate between the overlapped speakers from frame to frame within a segment. Morgan et al. in [22] employed a simple three-point median filter with good results. Dynamic programming techniques also exist for pitch track smoothing. This design issue will be explored further after I obtain more familiarity with the behavior of various pitch estimation algorithms in overlapped speech—for example, in working with the pitch related features of part II.

### Speech Voicing

Because HES makes use of information regarding the harmonic structure of the overlapped speech, it would seem that its intended area of application is overlap regions in which both speakers' speech is voiced. This is, however, only one of nine possible configurations for two-speaker overlap. What, then, of the other eight? Figure 3 shows a matrix for the possible combination of vocal excitation states of two-speaker overlap and the processing strategies available, as determined by Morgan et al. in [22]. HES can be applied to additional configurations, as suggested by the grid, if one performs only enhancement and suppression on a single detected pitch track. This can be achieved by conditioning the enhancement of the weaker speaker's harmonics on their detection. That is, if a second pitch track is not detected, simply suppress the first and this will be the signal for the second speaker. This demonstrates further the importance of accurate pitch detection for the algorithm.

Because unvoiced regions of speech are unable to be processed by the HES approach, the accurate detection of voicing is also important. Several features exist for voicing detection—some were mentioned in 3.2.3, for example—and their performance will need to be analyzed. Again, the results from part II may assist in determining good voicing features for this multi-channel farfield data.

### Single vs. Multiple Channels

As with overlap detection in part II, harmonic enhancement and suppression is primarily a single-channel method being applied to a multi-channel

		Interferer		
		Silence	Unvoiced	Voiced
Target	Voiced	NPN	NPN	HES NPN
	Unvoiced	NPN	X	HES
	Silence	HES NPN	HES	HES

NPN = No processing necessary  
 X = Cannot be processed  
 HES = Harmonic enhancement and suppression

Figure 3: Matrix of processing strategies available for the combinations of vocal excitation states for two simultaneous talkers.

domain. For that case I presented selecting a single best channel based on estimated SNR and using a delay-and-sum version of the signal as being the most viable approaches. The same holds for this case as well. Again, the issue with using the delay-and-sum signal arises from the estimation of pitch—that is, accurate pitch estimation may be hampered by phase differences between the summed signals. The results from Part II may steer the focus of exploration either towards or away from this delay-and-sum signal, depending on their nature.

#### 4.2.2 Adaptive Decorrelation Filtering

Adaptive decorrelation filtering (ADF) is a multi-channel signal separation technique that seeks to separate signals by adaptively determining the filters governing the coupling between two channels in a dual-channel system which models the signal mixing. Here we consider the two-source, two-channel case. Let  $s_1(t)$  and  $s_2(t)$  represent two source signals and  $y_1(t)$  and  $y_2(t)$  the signals acquired by the two microphones in the dual-channel system in figure 4(a). The signals in frequency domain are related by:

$$\begin{aligned}
 Y_1(f) &= H_{11}(f)S_1(f) + H_{12}(f)S_2(f) \\
 Y_2(f) &= H_{21}(f)S_1(f) + H_{22}(f)S_2(f)
 \end{aligned} \tag{11}$$

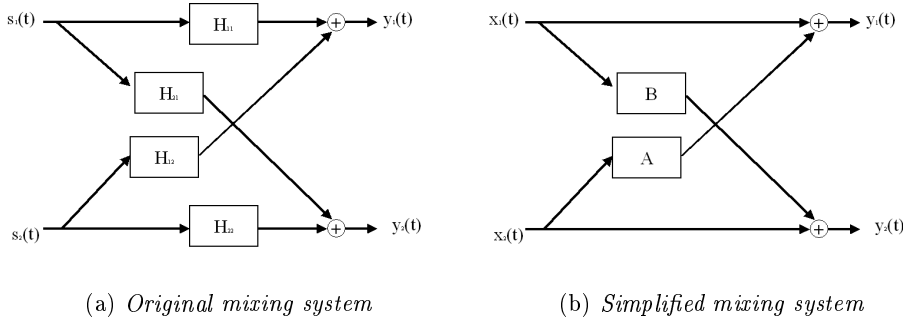


Figure 4: *Dual-channel system.*

where  $H_{ij}(f)$  is the transfer function from the source  $i$  to microphone  $j$ . This can be rewritten as:

$$\begin{aligned} Y_1(f) &= X_1(f) + A(f)X_2(f) \\ Y_2(f) &= X_2(f) + B(f)X_1(f) \end{aligned} \quad (12)$$

where

$$\begin{aligned} X_i(f) &= H_{ii}(f)S_i(f), \quad i = 1, 2 \\ A(f) &= \frac{H_{12}(f)}{H_{22}(f)} \\ B(f) &= \frac{H_{21}(f)}{H_{11}(f)} \end{aligned} \quad (13)$$

$X_i(f)$ , then, is a linearly distorted version of  $S_i(f)$ . In our case this distortion is the room response along the path from the speaker to his or her associated microphone and exists in isolated speech as well as co-channel speech; It should not, then, be detrimental to recognition. The simplified system is shown in figure 4(b). If we now process the signals  $y_1(t)$  and  $y_2(t)$  by the system in figure 5 with  $C(f) = 1 - A(f)B(f)$  and  $v_1(t)$  and  $v_2(t)$  having the relations:

$$\begin{aligned} V_1(f) &= Y_1(f) - \hat{A}(f)Y_2(f) \\ V_2(f) &= Y_2(f) - \hat{B}(f)Y_1(f) \end{aligned} \quad (14)$$

when  $\hat{A}(f) = A(f)$ ,  $\hat{B}(f) = B(f)$ , and  $C(f)$  is invertible, the signals  $x_1(t)$  and  $x_2(t)$  can be perfectly restored. Note that  $V_i(f) = C(f)X_i(f)$ , so when

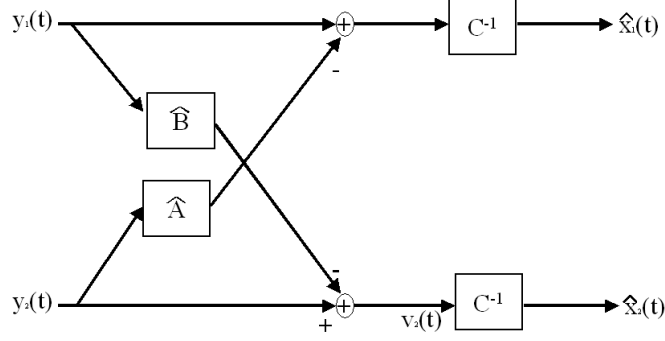


Figure 5: *Signal separation system.*

$C(f)$  is not invertible we can recover a linearly distorted versions of the  $x_1(t)$  and  $x_2(t)$ .

Since  $A(f)$  and  $B(f)$  are generally not known and are time-varying, they must be adaptively estimated. Weinstein et al. in [35] showed that if the source signals are assumed to be zero-mean and uncorrelated and if  $A$  and  $B$  are approximated by the FIR filters  $\mathbf{a} = [a_0, \dots, a_{N_a-1}]^T$  and  $\mathbf{b} = [b_0, \dots, b_{N_b-1}]$ , the filter coefficients can be iteratively estimated by the following equations:

$$\begin{aligned} \mathbf{a}^{(t)} &= \mu(t) \mathbf{v}_2(t-1) v_1(t-1)(t) \\ \mathbf{b}^{(t)} &= \mu(t) \mathbf{v}_1(t-1) v_2(t-1)(t) \end{aligned} \quad (15)$$

$$(16)$$

where

$$\begin{aligned} v_1(\tau) &= y_1(\tau) - \mathbf{y}_2(\tau)^T \mathbf{a}^{(t)} \\ v_2(\tau) &= y_2(\tau) - \mathbf{y}_1(\tau)^T \mathbf{b}^{(t)} \end{aligned} \quad (17)$$

with

$$\begin{aligned} \mathbf{y}_1(\tau) &= [y_1(\tau) \cdots y_1(\tau - N_b + 1)]^T \\ \mathbf{y}_2(\tau) &= [y_2(\tau) \cdots y_2(\tau - N_a + 1)]^T \\ \mathbf{v}_1^{(t)}(\tau) &= [v_1^{(t)} \cdots v_1^{(t)}(\tau - N_b + 1)]^T \\ \mathbf{v}_2^{(t)}(\tau) &= [v_2^{(t)} \cdots v_2^{(t)}(\tau - N_a + 1)]^T \end{aligned} \quad (18)$$

The adaptation gain  $\mu(t)$  is defined as  $\gamma/t$ .

One issue with an adaptive filtering approach such as ADF is convergence. Because overlap segments tend to be short (on the order of one or two seconds), there are a limited number of iterations through which the filter adaptation can proceed. This is especially a problem if the coupling changes rapidly within the overlap segment, but the expectation is that because of the shortness of the segments this will not be so. In either case there is a need for fast and stable convergence for good separation. Yen and Zhao in [44] derive a stability bound for the adaptation gain  $\gamma$  as being:

$$0 < \gamma < \frac{2}{N_a \text{var}\{y_2(t)\} + N_b \text{var}\{y_1(t)\}} = \Gamma \quad (19)$$

Since the variances of  $y_1(t)$  and  $y_2(t)$  can be evaluated at each frame,  $\Gamma$  can be calculated for each frame. The gain in each frame can then be chosen as

$$\mu(t) = \frac{\alpha \Gamma}{t}, \quad 0 < \alpha < 1 \quad (20)$$

to maximize convergence.

Related to convergence is the issue of initialization. The filters can always be initialized to zero, but it may be possible to speed up convergence using some other initialization scheme. Weinstein et al. propose estimating the transfer functions  $\hat{A}(f)$  and  $\hat{B}(f)$  during periods when only one of the speakers is speaking and using those estimates as initialization parameters. In the multi-speaker environment, however, though it may be possible to determine that only one speaker is speaking, without performing speaker assignment (e.g., through speaker recognition or clustering) one cannot determine if the active speaker is one of the speaker's involved in the overlap segment of interest. Depending on how rapidly the coupling between channels changes, too, it may not be beneficial to use such estimates anyway.

A final consideration for this ADF approach is microphone selection. In its formulation ADF assumes that each speaker has an associate microphone and that the nature of the overlap is akin to a "leakage" of the signal of one channel onto another. In the meeting room configurations described in 1.1 this is not really the case; farfield microphone placement is not done with the intention of associating each speaker with a microphone. It is possible to achieve this by selecting for each speaker the microphone closest to him or her. This information can be obtained via TDOA estimates from the various microphones. This procedure is more likely to be effective with centrally located microphones such as the tabletop and circular array microphones, so I intend to consider only these for this method.

### 4.3 Work Plan for Part III

This final part of my proposed thesis differs rather significantly from the previous two in that methods, rather than features, will be investigated for changes in performance measures. Using the overlap segments obtained from forced alignments as mentioned in section 3.4, the speech separation algorithms will be performed to determine if improvements can be made to the WER performance of the meeting ASR system. Here WER is the only meaningful performance measure for the purposes of the investigation. In addition, since the processing will be applied only to overlap regions only the WER for those regions will be considered for measuring performance. This being the case, it may be possible to run recognition only on waveform segments containing overlap regions. Here “segments” refers to silence separated portions of the waveform that are processed by the ASR system. One issue as yet unaddressed is how to integrate the processed speech into the waveform segment. At present I plan to extract the overlap section of the waveform and insert the processed version and smooth at the boundaries. This will be done twice—once for each of the separated signals—and so recognition will be performed twice on these segments. The data, as in the parts I and II, will be drawn from the NIST RT evaluations.

In the event that recognition performance is improved through either or both of the candidate methods, a number of analyses will be performed. One is to examine whether processing of the entire waveform segment—rather than the overlap region alone—affects results. Another is to perform processing on the overlap regions as determined by the best performing overlap detection system from part II to compare with the reference segmentation. This of course will be contingent as well on the overlap detection obtaining reasonably good results. The last is to look for patterns in performance improvement, or conversely to look for what type of errors persist. For example, in the case of HES, is the dominant speaker consistently improved more than the weaker speaker?

In the event that recognition performance is not improved, error analyses can still be performed. For example, the recognition errors made on the original and processed speech can be compared and contrasted. Also, the degree of degradation as measure by WER increase can be compared for the two.

## 5 Preliminary Experiments

This section presents results for experiments that have been performed related to this proposed work. Currently this is limited to part I— speech activity detection for the nearfield microphones—but work on part II—overlap detection for farfield microphones—is in the initial stages of experimental setup as well. The experiments were performed using the Augmented Multiparty Interaction (AMI) development set meetings for the NIST-RT05S meeting recognition evaluation. These are scenario-based meetings, elicited as described in [6], each involving four participants wearing headset or head-mounted lapel microphones. The local speech segmenter used was derived from an HMM based speech recognition system. The system was modified and simplified to consist of only two classes – “speech” and “nonspeech” – each being represented with a three-state phone model. State emission probabilities were modeled using a multivariate Gaussian Mixture Model (GMM). In the case of the baseline cepstral features the GMM consisted of 256 components while in the case of the cross-channel features only 32 components were used. This is because of the substantial difference in the number of feature vector components between the two sets: The baseline feature vectors have 39 components while the cross-channel feature vectors have only two (for maximum and minimum values across channels). For training of the segmenter, the first 10 minutes from 35 of the AMI meetings were utilized. Testing was performed on 12-minute excerpts from four additional meetings. For ASR performance results, the “fast” (two versus six decoding passes) version of the meeting recognition system field by ICSI-SRI in the NIST RT-05S evaluation was used. Details of this system can be found in [33].

### 5.1 Experiment 1: Single feature performance

In this first experiment, speech/nonspeech segmentation was performed on the test excerpts using each of the following features: baseline cepstral, normalized maximum cross-correlations (NMXCs), log-energy differences (LEDs), and normalized log-energy differences (NLEDs). The segmentation performance was then measured using both diarization and word error rate metrics. Results for diarization are given in table 1 and those for recognition in table 2. Performance in table 1 is further subdivided into missed detections (“Missed”) and false alarms (“FA”). Table 2 subdivides performance according to substitutions, insertions, and deletions.

Table 1: *DER performance comparisons for single features using AMI development data.*

System	Missed (%)	FA (%)	DER
baseline	8.94	12.15	21.09
NMXC	6.21	21.97	28.18
LEDs	15.86	3.30	19.16
NLEDs	11.29	5.46	16.74

Table 2: *ASR performance comparisons for single features using AMI development data. Results obtained using “fast” ASR system.*

System	Del	Subs	Ins	WER
baseline	17.4	13.0	7.4	37.8
NMXC	16.9	14.5	4.5	36.0
LEDs	15.7	21.1	4.0	40.8
NLEDs	16.5	18.2	4.2	38.8
reference	18.3	10.2	3.4	32.0

From table 1 we see that, for DER, LEDs and NLEDs outperform the baseline cepstral features while NMXCs do more poorly. The poor performance comes exclusively from a much higher false alarm rate, as the miss rate is lower than that of the cepstral features. The LEDs and NLEDs, though each having a miss rate exceeding that of the baseline, both have much lower false alarm rates than the baseline which accounts for their better performance. Also of note is that the NLEDs give a lower DER than the LEDs, indicating the effectiveness of the energy normalization procedure.

Regarding WER, table 2 shows that the NMXCs outperform the baseline while the LEDs and NLEDs do not. The NLEDs here, too, produce a lower error rate than the LEDs, further demonstrating the effectiveness of the normalization. All three cross-channel features notably reduce the insertion error rate by a substantial amount (between 39% and 46% relative). Given that the presence of crosstalk affects primarily insertion errors, this suggests that the cross-channel features are effective in addressing crosstalk. The use of these features, however, yields more substitution errors compared to the baseline. Further investigation is necessary to identify the cause for this as there is no direct diarization correlate for substitution errors. A final observation is that the best feature, the NMXCs, is still 4% absolute

worse than the reference segmentation given in the last row of the table. This “reference” refers to segmentation derived from the time marks in the reference for word error scoring.

## 5.2 Experiment 2: Initial feature combination

In addition to the individual performance of features for speech/nonspeech segmentation, the combination of the features is of interest. This experiment examined segmentation performance for simple concatenative combinations of the features from the previous experiment. Table 3 gives diarization results and table 4 presents recognition results.

Table 3: *DER performance comparisons for multiple features using AMI development data.*

System	Missed (%)	FA(%)	DER
baseline	8.94	12.15	21.09
base + NMXC	8.89	2.76	11.65
base + LEDs	8.65	2.62	11.28
base + NLEDs	8.63	3.16	11.79
base + NMXC + LEDs	10.37	7.13	17.50
base + NMXC + NLEDs	8.20	3.21	11.41

Table 4: *ASR performance comparisons for multiple features using AMI development data. Results obtained using “fast” ASR system.*

System	Del	Subs	Ins	WER
baseline	17.4	13.0	7.4	37.8
base + NMXC	17.3	12.1	4.6	34.0
base + LEDs	17.4	12.8	4.5	34.7
base + NLEDs	17.1	12.0	4.4	33.5
base + NMXC + LEDs	16.5	17.2	4.3	38.1
base + NMXC + NLEDs	17.5	12.5	4.6	34.6
reference	18.3	10.2	3.4	32.0

In comparing the combination of the baseline cepstral features with each of the proposed cross-channel features, we see that the performance is similar. This holds for both DER in table 3 and WER in table 4. Of particular note is that the improved performance of these features relative to the baseline

comes primarily from reductions in false alarms and, consequently, insertion errors. This provides further evidence of the effectiveness of these features in addressing crosstalk. Also notable regarding the results is the degradation in performance for the three-feature combinations. For each of these cases, the combination is worse than each two-way combination of the constituent features, and, in the case of the combination with the LEDs, performance degrades slightly beyond the baseline level. At present I am not certain as to the cause of this, but I hypothesize that the correlation between the features plays a role. Additional experiments are necessary to try and determine this and are planned. Also, as mentioned, I intend to look at other combination approaches and these may not produce the same effect.

## References

- [1] Spring 2006 (rt-06s) rich transcription meeting recognition evaluation plan, February 2006.
- [2] X. Anguera. *Robust Speaker Diarization for Meetings*. PhD thesis, Polytechnic University of Catalonia, October 2006.
- [3] F. R. Bach and M. I. Jordan. Blind one-microphone speech separation: A spectral learning approach. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 65–72. MIT Press, Cambridge, MA, 2005.
- [4] K. Boakye and A. Stolcke. Improved speech activity detection using cross-channel features for recognition of multiparty meetings. In *Proc. INTERSPEECH06-ICSLP*, 2006. Pittsburgh, PA.
- [5] M. S. Brandstein and H. F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proc. ICASSP 1997*, pages 375–378, 1997. Munich, Germany.
- [6] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meetings corpus. In *Proc. of the Measuring Behavior 2005 Symposium on “Annotating and Measuring Meeting Behavior”*, 2005. AMI-108.
- [7] D. P. Ellis and J. C. Liu. Speaker turn segmentation based on between-channel differences. In *Proc. ICASSP 2004*, 2004. Montreal, Canada.
- [8] J. Garofolo, C. Laprun, M. Michel, V. Stanford, and E. Tabassi. The NIST meeting room pilot corpus. In *Proc. LREC 2004*, 2004. Lisbon, Portugal.
- [9] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.
- [10] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Piskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In *Proc. ICASSP 2003*, 2003. Hong Kong, China.

- [11] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt. Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions. In *Proc. IEEE Int. Symp. on Intelligent Sig. Proc. and Comm. Sys.*, pages 710–713, 2000.
- [12] K. R. Krishnamachari, R. E. Yantorno, and J. M. Lovekin. Use of local kurtosis measure for spotting usable speech segments in co-channel speech. In *Proc. ICASSP 2001*, pages 649–652, 2001. Salt Lake City, UT.
- [13] R. H. Lambert and A. J. Bell. Blind separation of multiple speakers in a multipath environment. In *Proc. ICASSP 1997*, pages 423–426, 1997. Munich, Germany.
- [14] K. Laskowski, Q. Jin, and T. Schultz. Crosscorrelation-based multi-speaker speech activity detection. In *Proc. INTERSPEECH 2004 - ICSLP*, 2004. Jeju Island; Korea.
- [15] J. P. LeBlanc and P. L. D. Leon. Speech separation by kurtosis maximization. In *Proc. ICASSP 1998*, pages 1029–1032, 1998. Seattle, WA.
- [16] T.-W. Lee and A. Bell. Blind separation of delayed and convolved sources. *Advances in Neural Information Processing Systems*, pages 758–764, 1997.
- [17] T.-W. Lee and A. J. Bell. Blind source separation of real world signals. In *Proc. Int. Conf. on Neural Networks*, pages 2129–2134, 1997. Houston, TX.
- [18] T.-W. Lee and A. Ziehe. Combining time-delayed decorrelation and ica: Towards solving the cocktail party problem. In *Proc. ICASSP 1998*, pages 1249–1252, 1998. Seattle, WA.
- [19] M. A. Lewis and R. P. Ramachandran. Cochannel speaker count labelling based on the use of cepstral and pitch prediction derived features. *J. Pattern Rec. Soc.*, 34:499–507, 2001.
- [20] D. Liu and F. Kubala. A cross-channel modeling approach for automatic segmentation of conversational telephone speech. In *Proc. IEEE ASRU Workshop*, pages 333–338, 2003.

- [21] J. M. Lovekin, R. E. Yantorno, K. R. Krishnamachari, D. S. Benincasa, and S. J. Wenndt. Developing usable speech criteria for speaker identification technology. In *Proc. ICASSP 2001*, pages 421–424, 2001. Salt Lake City, UT.
- [22] D. P. Morgan, B. George, L. T. Lee, and S. M. Kay. Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Tran. on Speech and Audio Proc.*, 5(5):407–424, September 1997.
- [23] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. Meetings about meetings: Research at ICSI on speech in multiparty conversations. In *Proc. ICASSP 2003*, 2003. Hong Kong, China.
- [24] J. Naylor and S. Boll. Techniques for suppression of an interfering talker in co-channel speech. In *Proc. ICASSP 1987*, pages 205–208, 1987. Dallas, TX.
- [25] Özgür Çetin and E. Shriberg. Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap. In *Proc. ICASSP 2006*, 2006. Toulouse, France.
- [26] J. M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multi-microphone meetings using only between-channel differences. In *Proc. MLMI 2006*, 2006. to appear.
- [27] J. M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences. In *Proc. INTERSPEECH06-ICSLP*, pages 2194–2197, 2006. Pittsburgh, PA.
- [28] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Amer.*, 60(4):911–918, October 1976.
- [29] T. Pfau, D. Ellis, and A. Stolcke. Multispeaker speech activity detection for the ICSI meeting recorder. In *Proc. IEEE ASRU Workshop*, pages 107–110, 2001.
- [30] T. Pfau and D. P. Ellis. Hidden markov model based speech activity detection for the ICSI meeting project. In *Proc. Eurospeech 2001*, 2001. Aalborg, Denmark.

- [31] Y. Shao and D. Wang. Co-channel speaker identification using usable speech extraction based on multi-pitch tracking. In *Proc. ICASSP 2003*, pages 205–208, 2003. Hong Kong, China.
- [32] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversations. In *Proc. Eurospeech 2001*, 2001. Aalborg, Denmark.
- [33] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grézl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng. Further progress in meeting recognition: The ICSI-SRI Spring 2005 speech-to-text evaluation system. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005*, volume 3869 of *Lecture Notes in Computer Science*, pages 463–475. Springer, 2005.
- [34] K. Torkkola. Blind separation of convolved sources based on information maximization. In *Proc. ICASSP 1996*, pages 3509–3512, 1996. Atlanta, GA.
- [35] E. Weinstein, M. Feder, and A. V. Oppenheim. Multi-channel signal separation by decorrelation. *IEEE Trans. on Speech and Audio Proc.*, 1(4):405–413, October 1993.
- [36] J. D. Wise, J. R. Caprio, and T. W. Parks. Maximum likelihood pitch estimation. *IEEE Trans. on Acoustics, Speech, and Sig. Proc.*, ASSP-24(5):418–423, October 1976.
- [37] S. Wrigley, G. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multi-channel audio. *IEEE Trans. on Speech and Audio Processing*, 13(1):84–91, 2005.
- [38] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals. Feature selection for the classification of crosstalk in multi-channel audio. In *Proc. Eurospeech 2003*, 2003. Geneva, Switzerland.
- [39] R. E. Yantorno, B. Y. Smolenski, and N. Chandra. Usable speech measures and their fusion. In *Proc. IEEE Int. Symp. on Circ. and Sys.*, pages 734–737, 1999.
- [40] D. Yellin and E. Weinstein. Multichannel signal separation: Methods and analysis. *IEEE Trans. on Sig. Proc.*, 44(1):106–118, January 1996.

- [41] K.-C. Yen and Y. Zhao. Robust automatic speech recognition using a multi-channel signal separation front-end. In *Proc. ICSLP 1996*, pages 1337–1340, 1996. Philadelphia, PA.
- [42] K.-C. Yen and Y. Zhao. Co-channel speech separation for robust automatic speech recognition: Stability and efficiency. In *Proc. ICASSP 1997*, pages 859–862, 1997. Munich, Germany.
- [43] K.-C. Yen and Y. Zhao. Improvements on co-channel speech separation using ADF: Low complexity, fast convergence, and generalization. In *Proc. ICASSP 1998*, pages 1025–1028, 1998. Seattle, WA.
- [44] K.-C. Yen and Y. Zhao. Adaptive co-channel speech separation and recognition. *IEEE Trans. on Speech and Audio Proc.*, 7(2):138–151, March 1999.
- [45] M. Zissman, C. Weinstein, and L. Braidà. Automatic talker activity labeling for co-channel talker interference suppression. In *Proc. ICASSP 1990*, pages 813–816, 1990. Albuquerque, NM.