Audio Segmentation for Meetings Speech Processing

Kofi Agyeman Boakye

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-170
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-170.html

December 18, 2008

Audio Segmentation for Meetings Speech Processing

by

Kofi Agyeman Boakye

B.S.E. (Princeton University) 2002
M.S. (University of California, Berkeley) 2005

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Engineering—Electrical Engineering and
Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Nelson Morgan, Chair
Professor Keith Johnson
Professor Michael Jordan

Fall 2008

The dissertation of Kofi Agyeman Boakye is approved:

Professor Nelson Morgan, Chair	Date
Professor Keith Johnson	Date
Professor Michael Jordan	Date

University of California, Berkeley

Fall 2008

Audio Segmentation for Meetings Speech Processing

by

Kofi Agyeman Boakye

Abstract

Audio Segmentation for Meetings Speech Processing

by

Kofi Agyeman Boakye

Doctor of Philosophy in Engineering—Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Nelson Morgan, Chair

Perhaps more than any other domain, meetings represent a rich source of content for spoken language research and technology. Two common (and complementary) forms of meeting speech processing are automatic speech recognition (ASR)—which seeks to determine what was said—and speaker diarization—which seeks to determine who spoke when. Because of the complexity of meetings, however, such forms of processing present a number of challenges. In the case of speech recognition, crosstalk speech is often the primary source of errors for audio from the personal microphones worn by participants in the various meetings. This crosstalk typically produces insertion errors in the recognizer, which mistakenly processes this non-local speech audio. With speaker diarization, overlapped speech generates a significant number of errors for most state-of-the-art systems, which are generally unequipped to deal with this phenomenon. These errors appear in the form of missed speech, where overlap segments are not identified, and increased speaker error from speaker models negatively affected by the overlapped speech data.

This thesis sought to address these issues by appropriately employing audio segmentation as a first step to both automatic speech recognition and speaker diarization in meetings. For ASR, the segmentation of nonspeech and local speech was the objec-

tive while for speaker diarization, nonspeech, single-speaker speech, and overlapped speech were the audio classes to be segmented. A major focus was the identification of features suited to segmenting these audio classes: For crosstalk, cross-channel features were explored, while for monaural overlapped speech, energy, harmonic, and spectral features were examined. Using feature subset selection, the best combination of auxiliary features to baseline MFCCs in the former scenario consisted of normalized maximum cross-channel correlation and log-energy difference; for the latter scenario, RMS energy, harmonic energy ratio, and modulation spectrogram features were determined to be the most useful in the realistic multi-site farfield audio condition. For ASR, improvements to word error rate of 13.4% relative were made to the baseline on development data and 9.2% relative on validation data. For speaker diarization, results proved less consistent, with relative DER improvements of 23.25% on development, but no significant change on a randomly selected validation set. Closer inspection revealed performance variability on the meeting level, with some meetings improving substantially and others degrading. Further analysis over a large set of meetings confirmed this variability, but also showed many meetings benefitting significantly from the proposed technique.

Professor Nelson Morgan, Chair Date

Acknowledgments

Given the enormity of the task (or struggle?) that is the Ph.D. thesis, and the acknowledgment that none of us exists in a vacuum, many thanks are in order. First is to Nelson Morgan, my advisor and the director of the International Computer Science Institute (ICSI), that served as my research home during my graduate years. I am grateful for the guidance he provided during this project and for fostering an environment at ICSI that allowed for intellectual exploration, personal growth, and (just as important) lots of laughs. I can honestly say that this experience will be unmatched in my life.

I am also greatly indebted to my primary collaborators in this research. First is Andreas Stolcke, who provided much technical assistance with the multispeaker SAD work. By extension, I thank SRI International, for allowing me the use of many of their resources, without which this work would not have been possible. I would also like to thank Gerald Friedland, Oriol Vinyals, and Beatriz Trueba-Hornero of the ICSI Speaker Diarization group, with whom I worked closely on the overlapped speech handling component of this thesis. Their many ideas and suggestions are truly appreciated and their enthusiasm drove my efforts considerably.

A number of ICSI Speech Group visitors and staff, both past and present, were also instrumental to this thesis work. These include Xavi Anguera, Özgür Çetin, Joe Frankel, Adam Janin, Mathew Magimai-Doss, and Chuck Wooters. I would also like to thank Dan Ellis of Columbia University, who graciously hosted me in his lab in the early days of the project and has periodically advised me thereafter. I am particularly grateful for his assistance in improving my understanding of the modulation spectrogram features examined in this work.

My closest interaction by and large was with my fellow students, and to them I show much appreciation. Many thanks to Dave Gelbart, Lara Stoll, Howard Lei, Andy Hatch, Arlo Faria, and my perpetual officemate, Dan Gillick. In addition, I owe a special debt of gratitude to Mary Knox for editing this document.

Of course I would be remiss if I did not acknowledge my parents. I thank them for their patience, for always believing in me, and for setting a high standard of excellence, with word matched in deed. To them I say, "*me da mo ase ooo, me da mo ase paa!*"

Lastly, this work would not be possible without the financial support I have received during these graduate years. For this I am grateful to the AT&T Labs Fellowship program and its committee, who approved my funding year after year. I also thank the Swiss National Science Foundation, who through the research network IM2 did the same.

To my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Spoken language processing in meetings

Perhaps more than any other domain, 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 addition, because of the presence of multiple participants in these meetings, it is not only important to determine *what* was said, but *who* said it; indeed, this idea is generally part of the notion of a "transcript". Accurate speaker diarization—i.e., determining "Who spoke when?"—is therefore also of great importance to spoken language processing in meetings and has received much attention in the research community.

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 a 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. A diagram of a meeting room instrumentation with these microphones is shown in Figure 1.1.

Speaker diarization similarly uses audio from such microphones. In contrast to ASR, however, this is generally limited to the distant microphones. In theory, the speech from a nearfield microphone should be that of the wearer of the microphone, making diarization unnecessary (but not trivial because of crosstalk, as discussed below).

1.2 Crosstalk and overlapped speech

Both automatic speech recognition and speaker diarization in these meetings present 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.

1.2.1 Crosstalk

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

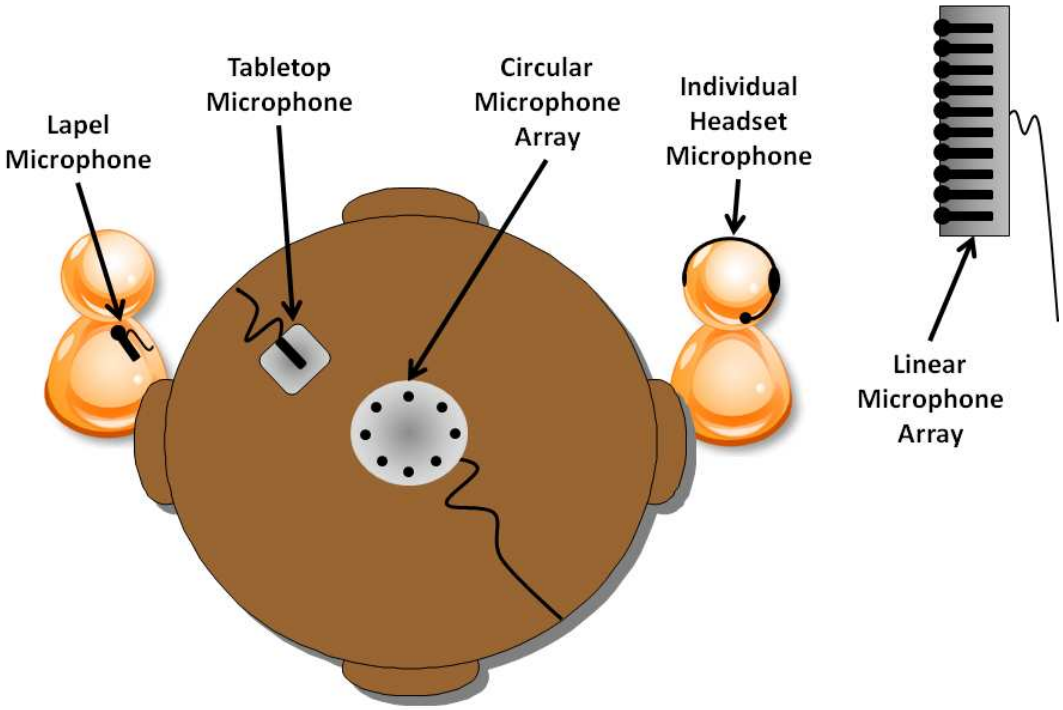


Figure 1.1: Diagram of an instrumented meeting room. The lapel and individual headset microphones correspond to the nearfield recording condition, while the tabletop microphone and the linear and circular arrays correspond to the farfield condition.

wearing the microphone. This speech is problematic because, as previously mentioned, 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 [73], for example, word error rate (WER) differed by 75% relative between recognition on segmented and unsegmented waveforms, largely due to insertions from crosstalk.

The issue of crosstalk can be addressed within the framework of speech activity detection (SAD), a long-studied problem in speech processing ([16],[93],[52],[35],[62],[95]) 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 (i.e., 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 a hidden Markov model (HMM) based segmenter. Because of the acoustic similarity between local speech and crosstalk, the task of speech activity detection in this context 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.

1.2.2 Overlapped speech

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. It is in this farfield condition, too, that overlapped speech affects the second task of interest: speaker diarization. Current state-of-the-art speaker diarization systems assign speech segments to only one speaker, thus incurring missed speech errors in regions where more than one speaker is active. For these systems, this error may represent a significant portion of the diarization error. For example, in [116] the authors reveal that 17% of the diarization error for their state-of-the-art system consisted of missed speech errors due to overlap when using a single microphone and 43% when using multiple microphones. A similar system described in [39] had 22% of its diarization error attributed to overlapped speech in the multiple microphone scenario. To be

certain, the proportions are high largely due to the low overall diarization error rate obtained by these systems. This is all the more reason, however, to address the issue of overlap, as it is now one of the most significant impediments to improved system performance. In addition, because overlap segments contain speech from multiple speakers, they should probably not be assigned to any individual speaker cluster nor included in any individual speaker model. Doing so could adversely affect the quality of the speaker models, which potentially reduces diarization performance. In [84], for example, the authors, using an oracle system which identified all overlapped speech segments, demonstrated an improvement in diarization performance by excluding these overlap regions from the input to the diarization system.

To identify overlapped speech regions, a framework similar to speech activity detection can be adopted. An audio 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.

1.3 Thesis Goals and Overview

This thesis endeavors to address the issues of crosstalk and overlapped speech described above by appropriately employing audio segmentation as a first step to both automatic speech recognition and speaker diarization in meetings. For ASR, the objective is to identify regions of local speech in the nearfield audio stream, thereby eliminating the erroneous recognition of the crosstalk speech. For diarization, we seek to identify overlapped speech for speaker segment labeling as well as for improving the speaker clustering in the system. Of primary interest, in both cases, is exploring and evaluating features suited to the detection of the audio classes of nearfield speech (and, hence, exclude crosstalk) and overlapped farfield speech to achieve high performance from

these systems.

The thesis is organized as follows. First, Chapter 2 provides background. This consists of basic information on the two target applications—automatic speech recognition and speaker diarization—as well as a review of work in the areas of multispeaker SAD and overlapped speech detection relevant to this thesis. Chapter 3 then outlines the framework for implementing and evaluating the two audio segmentation systems. Chapter 4 presents the multispeaker SAD system for improving ASR. The candidate features explored for the segmenter are discussed and the evaluation experiments performed as part of the system development are presented and analyzed. The same is done for overlapped speech handling in Chapter 5. Finally, in Chapter 6, the work is summarized and concluding remarks are given, with a focus on the contributions made by this thesis and possibilities for building upon this work.

Chapter 2

Background

As stated in Chapter 1, this thesis seeks to address issues associated with automatic speech recognition and speaker diarization in meetings. In order to do this, however, it is necessary to first have an understanding of how each of these applications works. This chapter provides an overview of the standard approaches for performing ASR and speaker diarization in the meetings setting. Another necessary step in developing the proposed systems for speech and overlap detection is a knowledge of other such attempts as found in the literature—What was done? What worked well? What did not? This chapter also gives a review of this related work with analysis from a system development perspective.

2.1 Automatic Speech Recognition in Meetings

The task of automatic speech recognition (ASR) is to determine "What was said?"—that is, to identify the sequence of words $W = w_0, w_1, \ldots, w_n$ contained within the utterances of an audio data stream. The standard practice is to adopt a probabilistic framework in which the problem becomes finding the most likely string of words $\hat{W}$

given a sequence of acoustic observations $\mathbf{Y} = \mathbf{y_0}, \mathbf{y_1}, \ldots, \mathbf{y_T}$. That is,

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|\mathbf{Y}) \tag{2.1}$$

Using Bayes' theorem, this becomes

$$\begin{aligned} \hat{W} &= \underset{W}{\operatorname{argmax}} \frac{P(\mathbf{Y}|W)P(W)}{P(\mathbf{Y})} \\ &= \underset{W}{\operatorname{argmax}} P(\mathbf{Y}|W)P(W) \end{aligned} \tag{2.2}$$

since $P(\mathbf{Y})$, the prior probability of the acoustics, is constant over all word strings. The acoustic observations $\mathbf{Y}$ are obtained from the audio signal through a process referred to as *feature extraction*. This procedure seeks to yield a parameterization 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. The first term in Equation 2.2, $P(\mathbf{Y}|W)$, represents the probability of observing the sequence $\mathbf{Y}$ given a specified word sequence W and is determined by an *acoustic model*. The second term, $P(W)$ is the a priori probability of observing W independent of the observed signal and is determined using a *language model*. These two models represent the two major components of a statistical ASR system, as shown in Figure 2.1.

Acoustic Model

Acoustic modeling seeks to provide a method of calculating the likelihood of any vector sequence $\mathbf{Y}$ given a word w. Typically, rather than explicitly model a word, however, each word is decomposed into sub-word units referred to as *phones* which are then modeled as a sequence of states $Q = q_0, q_1, \ldots, q_m$. Thus the optimization becomes

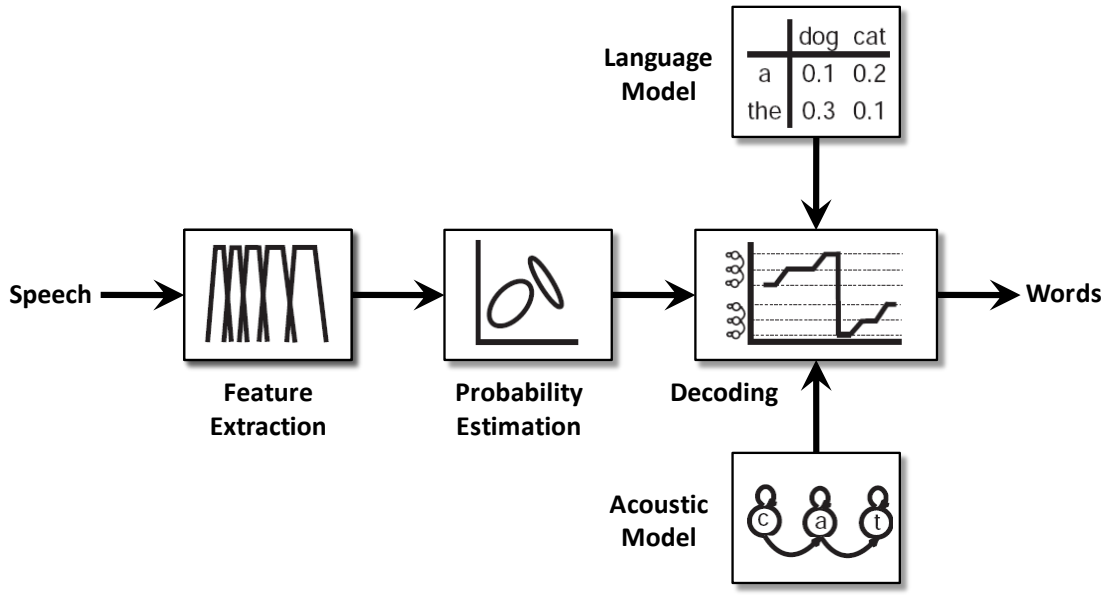


Figure 2.1: A typical ASR system. Speech audio is processed into a stream of features that, using probabilistic acoustic and language models, is decoded into a sequence of words.

$$\begin{aligned} \hat{W} &= \underset{W}{\operatorname{argmax}} \sum_{Q} P(\mathbf{Y}|Q,W)P(Q|W)P(W) \\ &\approx \underset{W}{\operatorname{argmax}} \sum_{Q} P(\mathbf{Y}|Q)P(Q|W)P(W) \end{aligned} \tag{2.3}$$

where we assume that the acoustic observations are independent of the word sequence given the state sequence. Each phone is typically represented using a hidden Markov model (HMM). An HMM is a finite state machine that models the generation of a sequence of random variables (here the vector sequence $\mathbf{Y}$) using a set of underlying hidden states (here the states of the phone). At every time step t, the finite state machine undergoes a transition to a state j and emits a speech vector $\mathbf{y_t}$ with probability $b_j(\mathbf{y_t})$. This transition from a previous state i to a state j occurs probabilistically

with discrete probability a_{ij}. In addition, a first-order Markov process is assumed, meaning the transition at every time step depends only on the state at the immediately preceding time step. Given an acoustic phone model M, the joint probability of the vector sequence $\mathbf{Y}$ and a state sequence $X = x_0, x_1, \ldots, x_T$ is calculated as the product of the transition probabilities and emission probabilities. That is,

$$P(\mathbf{Y}, X|M) \;=\; a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(\mathbf{y_t}) a_{x_t x_{t+1}} \tag{2.4}$$

where x_0 is the model entry state and x_{T+1} is the model exit state. Individual phone models are joined together by merging adjacent exit and entry states to form a composite HMM of words. These word HMMs are then joined together to model entire utterances. The acoustic model is trained by estimating the HMM parameters (the transition probabilities and the emission probability model parameters) using a procedure referred to as *Baum-Welch Re-estimation*. This procedure is discussed further in Section 3.2.1.

Language Model

The language model attempts to estimate the a priori probability of a word sequence $W = w_0, w_1, \ldots, w_n$. This joint probability can be represented as a product of conditional probabilities of the form:

$$P(w_0, w_1, \ldots, w_n) \;=\; P(w_0) \prod_{k=1}^{n} P(w_k|w_0, w_1, \ldots, w_{k-1}) \tag{2.5}$$

Thus, estimating the word sequence probability becomes a question of estimating the probability of a word w_k given the preceding words $W_0^{k-1} = w_0, w_1, \ldots, w_{k-1}$. Typically this is done by making the simplifying assumption that w_k depends only on

the preceding $n-1$ words. That is,

$$P(w_k|W_0^{k-1}) \;=\; P(w_k|W_{k-n+1}^{k-1}) \tag{2.6}$$

The sequence of N words is referred to as an *N-gram*. N-grams simultaneously encode syntax, semantics, and pragmatics without the need for explicit linguistic rules such as a formal language grammar. In addition, they concentrate on local dependencies, making them very effective for languages such as English, where word order is important and the strongest contextual effects tend to come from near-neighbors [124].

N-gram probabilities are usually estimated from frequency counts of training text data. So for the case of word bigrams ($N = 2$),

$$\hat{P}(w_k|w_{k-1}) \;=\; \frac{b(w_{k-1}, w_k)}{u(w_{k-1})} \tag{2.7}$$

where $b(w_{k-1}, w_k)$ is the bigram count for the sequence 'ab' and $u(w_{k-1})$ is the unigram count for the word 'a'. For most state-of-the-art large-vocabulary continuous speech recognition (LVCSR) systems, bigrams and trigrams are employed for language modeling.

Because some bigrams or trigrams may appear very few times or not at all in the training data, a number of smoothing techniques have been developed to improve the estimates. One involves *linear interpolation*—e.g., taking the weighted mean of unigram, bigram, and trigram probabilities. Another, referred to as *discounting*, redistributes probability mass from more frequently occurring N-grams to less frequently occurring ones. Lastly, *back-off* is a procedure in which, say, a trigram probability is replaced by a scaled bigram probability—i.e., we "back off" to a lower order N-gram model.

Decoding

Decoding attempts to find the state (and, consequently, word) sequence with the highest likelihood given the sequence of feature vectors. This search problem corresponds to replacing the summation in Equation 2.3 with a maximization, giving:

$$\hat{W} \approx \underset{W}{\operatorname{argmax}} \max_{Q} P(\mathbf{Y}|Q)P(Q|W)P(W) \tag{2.8}$$

The most common procedure for doing this is *Viterbi decoding*, which uses a dynamic programming algorithm to perform breadth-first search.

ASR in the nearfield condition is generally performed by decoding each individual audio channel separately. 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) as in [104] and [41], or at the recognition hypothesis level, as in [37]. 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 metric for meetings is the *word error rate* (WER). The WER is computed according to:

$$\text{WER} = \frac{N_{\text{deletions}} + N_{\text{insertions}} + N_{\text{substitutions}}}{N_{\text{tokens}}} \tag{2.9}$$

where the numerator is the sum of all ASR output token errors and the denominator is the number of scoreable tokens in a reference transcription. The errors are of three types: missed tokens (deletions), inserted tokens (insertions), and incorrectly recognized tokens (substitutions). The types of errors are determined based on a dynamic programming string alignment to the reference transcription that globally

minimizes a Levenshtein distance function, which can weight correct, inserted, deleted, and substituted words differently (e.g., 0, 3, 3, and 4, respectively). The algorithm is detailed in [98].

2.1.1 Related Work

Though single-channel speech activity detection has been studied in the speech processing community for some time now ([16],[93], and [52] are some older examples), 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 comes from Wrigley et al. in [119] and [118]. The authors performed a systematic analysis of features for classifying multi-channel audio. Rather than look at the two classes of speech and nonspeech, though, they subdivided the classes further into four: local channel speech, crosstalk speech, local channel and crosstalk speech (i.e., overlapped 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 for a given audio class. A key result from this work was that, from among the twenty features examined, the single best performing feature for each class was one derived from cross-channel correlation, providing evidence of the importance of incorporating cross-channel information into modeling for this multi-channel detection task.

In addition to the work by Wrigley et al., there have been a number of related efforts towards multispeaker speech activity detection. These include the work by Pfau et al., Dines et al., and Laskowski et al. and are described below.

Pfau et al in [89] proposed an ergodic HMM (eHMM) speech activity detector consisting of two states—speech and nonspeech—and a number of intermediate states to enforce time constraints on transitions. The features used for the HMM were critical band loudness values, energy, and zero-crossing rate. As a post-processing step the authors thresholded cross-channel correlations to identify and remove crosstalk speech segments, a step that yielded on average a 12% relative frame error rate (FER) reduction.

Taking cues from Wrigley et al., Dines et al. in [25] used kurtosis, mean cross-correlation, and maximum cross-correlation as auxiliary features for their nearfield speech activity detector. They also proposed a cross-meeting normalized energy feature which compared the target channel energy to the sum of the energy of all channels. Lastly, they applied a crosstalk suppression algorithm based on adaptive-LMS echo cancellation to the recordings prior to generating their baseline perceptual linear predictive (PLP) features. Using a system based on a multi-layer perceptron (MLP) classifier, the resulting segments achieved a WER within 1.3% of that obtained by manual segmentation of the audio, though with the assistance of some tuning of the speech/nonspeech class prior probabilities.

In [54], Laskowski et al., using a cross-channel correlation thresholding scheme produced ASR WER performance improvements of 6% absolute over an energy-thresholding baseline. This improved thresholding scheme later became a first-pass segmentation step in a multi-channel speech activity detection system that modeled vocal interaction between meeting participants with joint multi-participant models ([55],[53], [56], and [57]). For meeting data from the 2005 and 2006 NIST Rich Transcription evaluations, the system achieved a WER performance within 1.6% and 2.2%, respectively, of manual segmentation on a first-pass speech recognition decoding.

In all of the work described above, the use of cross-channel features played a major role in improving speech activity detection performance for the nearfield multispeaker

audio. This, of course, should not be too surprising given the nature of the phenomenon being addressed: crosstalk speech. With information about the audio on other channels, a speech activity detector should be better able to determine if speech on a target channel is local or not. Thus, the focus of this thesis regarding features for improving multispeaker SAD is exclusively features of a cross-channel nature. The question then becomes, "What specific cross-channel information should be encoded to produce these features?" The answer is given and discussed in Section 4.2.

2.2 Speaker Diarization in Meetings

Automatic speaker diarization seeks to determine, "Who spoke when?"—that is, to partition an input audio stream into segments, generally homogeneous, according to the speaker identity. These speaker identities are typically relative to a given recording (e.g., "Speaker A, Speaker B, etc.") rather than absolute, in contrast to speaker identification or speaker tracking, where a priori knowledge of the speakers' voices is provided. For most speaker diarization tasks, the number of speakers is also unknown a priori.

As shown in Figure 2.2, the standard approach for speaker diarization decomposes the task into four subtasks: speech activity detection, speaker segmentation/change-point detection, speaker clustering, and re-segmentation.

Speech Activity Detection

Speech activity detection first identifies audio regions containing speech from any of the speakers present in the recording, as these are the only relevant regions for speaker identity annotation. Depending on the domain of the data being used, the non-speech regions that are discarded may contain various acoustic phenomena, such as silence, music, room noise, or background noise. Speech activity detection is typically per-

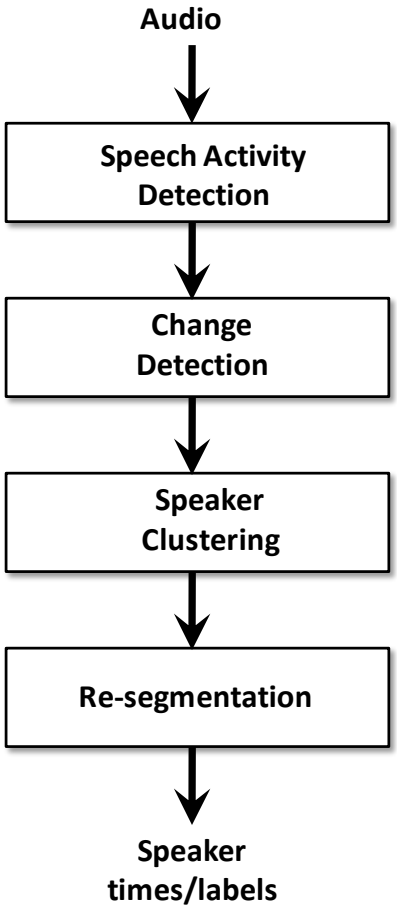


Figure 2.2: A typical speaker diarization system. Detected speech segments are analyzed for change-points and the resulting segments are clustered. The segments are refined using models from previous stages and segments with time and speaker labels are produced.

formed using maximum-likelihood (ML) classification with Gaussian mixture models (GMMs) trained on labeled training data (e.g., [117]).

Speaker Segmentation/Change-Point Detection

For segmentation/change-point detection, the standard approach is to observe adjacent windows of data and determine whether the windows originated from the same or

different speakers. This is determined by calculating a distance metric between the two windows—either using a variation on the Bayesian Information Criterion (BIC) technique introduced by Chen and Gopalakrishnam in [20] or, as pioneered by Siegler et al. in [102], representing each window as a Gaussian or mixture of Gaussians and computing the distance between the two distributions (e.g., using the symmetric KL-2 distance).

Speaker Clustering

Speaker clustering seeks to cluster segments from the same speaker together, ideally producing one cluster for each speaker present in the recording with all segments from a given speaker in a single cluster. Typically a hierarchical, agglomerative clustering procedure is used with a BIC based stopping criterion. A common distance metric for cluster merging within this BIC based scheme is the generalized likelihood ratio (GLR) [33], which compares the likelihood of a merged cluster to that of two separate clusters, each modeled by a Gaussian or mixture of Gaussians. A notable variation to this approach has been proposed by Ajmera and Wooters in [3] and consists of fixing the number of parameters between the two BIC hypotheses so as to eliminate the need for tuning the BIC penalty term. The above segmentation and clustering steps can also be performed iteratively to jointly optimize the two. This is done either using a set of GMMs as in [20], or using an ergodic HMM as in [68] and [3].

Re-segmentation

Re-segmentation attempts to refine the original segmentation boundaries and/or fill in short segments that may have been removed for more robust processing in the clustering stage. This is typically performed by Viterbi decoding (possibly in an iterative fashion) of the audio data using the final cluster and speech/nonspeech models obtained from the previous stages.

The performance of a speaker diarization system is measured using the *diarization error rate* (DER) [96]. This is defined as the sum of the per-speaker false alarm (falsely identifying speech), missed speech (failing to identify speech), and speaker error (incorrectly identifying the speaker) times, divided by the total amount of speech time in a test audio file. That is,

$$\text{DER} = \frac{T_{\text{FA}} + T_{\text{MISS}} + T_{\text{SPKR}}}{T_{\text{SPEECH}}} \tag{2.10}$$

Note that this is a time-weighted error metric and therefore intrinsically gives greater importance to more talkative speakers. The same formulation, however, can be modified to be speaker-weighted if desired, but the version in Equation 2.10 is the standard. The DER is computed by first finding an optimal one-to-one mapping of the reference, or ground-truth, speaker identities to the identities output by the diarization system and then determining the errors (false alarm, missed speech, and speaker) accordingly. Typically unscored regions determined by a "forgiveness collar" are placed at segment boundaries to address both inconsistencies in the annotation of segment times and the philosophical argument of when speech actually begins for word-initial stop consonants. The collar is generally on the order of 0.25 s (the value used in this work) and spans a region both preceding and following the segment boundary.

The standard approach for speaker diarization described above has been developed with the underlying assumption that segments are speaker-homogeneous. Change-point detection determines if two speech windows correspond to the same speaker or two distinct speakers. Speaker clustering makes a hard assignment of segments to speaker clusters, preventing a segment containing speech from multiple speakers from being assigned to multiple clusters. As a result, only one speaker label can be applied

to each speech segment. The nature of conversational speech, however, is such that invariably speakers will overlap. It is, therefore, important to explore techniques for handling overlapped speech—both its detection and its processing—and the previous work toward this effort is described below.

2.2.1 Related Work

The initial work done on overlapped, or co-channel, speech detection was 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 target-to-interferer ratio (TIR)—of the frame or segment (see [66]).

Lewis and Ramachandran in [61] compared the performance of three features—Mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs), and a proposed pitch prediction feature (PPF)—for speaker count labeling of speech frames in both a closed-set (speaker-dependent) and open-set (speaker-independent) scenario. This last feature was computed as the standard deviation of the distance between pitch peaks, as obtained from the autocorrelation of a linear prediction (LP) residual. Experiments were performed on artificial overlapped speech obtained by summing audio signals from the New England portion of the TIMIT database [31]. 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 [101] 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. This

approach was shown to consistently improve speaker recognition performance. Again, experiments were performed on artificially generated data from the New England subset of the TIMIT database.

Zissman et al. in [125] successfully distinguished target-only, jammer-only, and target-plus-jammer speech segments with a reported accuracy of 80% using a Gaussian classifier with cepstral features. A number of factors contributed to this high level of performance, however. For one, as in the previous work, artificially generated overlap segments were used for the experiments. In addition, the task was speaker-dependent—i.e., the training and testing target and jammer speakers, respectively, were the same. Lastly, intervals of silence were removed beforehand.

A large effort ([121],[123],[49],[50], [66],[65],[122],[19], and [106]) has been given by Yantorno et al. in this area as well. In [49] and [66], the authors demonstrated the effectiveness of the spectral autocorrelation ratio (SAR) as a measure of usable speech for speaker recognition. In [122], and [19] this was modified to the spectral autocorrelation peak-to-valley ratio (SAPVR) and developed as a feature for co-channel speech detection. Both the SAR and SAPVR take advantage of the structure of voiced speech in the frequency domain, namely the regularly-spaced harmonic peaks. In the time domain, a related measure—the adjacent pitch period comparison (APPC)—was explored as a usability measure in [65]. This approach was useful in spotting approximately 75% of usable segments for speaker identification as determined by TIR. Combination experiments were performed on these two features too in [123], increasing the correctly identified segments to 84%. In [50], between 83% and 92% of usable speech segments were found to be bracketed by spikes in speech amplitude kurtosis, indicating a method of identifying such segments. The authors make a point to note that the kurtosis "by itself does not point to usable clusters, rather a coarse location where usable clusters may be searched in a co-channel utterance". Wrigley et al., however, obtained good performance (67.5% true positives at the

equal error rate point) for nearfield overlapped speech detection using kurtosis as a feature. Lastly, the work in [106] demonstrated that linear predictive coding (LPC) analysis—specifically, identifying the number of peaks within the typical range for the first formant (0-1kHz)—was also a viable method of determining usable speech.

In all of the work described above, the overlapped speech data was artificially generated by mixing single-speech audio data. This was necessary since the TIR of the speech segments was of interest and this information can only reliably be obtained if the speech energy of the individual speakers is available. In addition, the common data set, TIMIT, being clean, nearfield recordings, differs significantly from the reverberant, farfield signals obtained in meetings. There are, nevertheless, a number of things to take away from this work. The primary one is that, in seeking to identify overlapped speech, the features/measures exploited the structure of single-speaker voiced speech, both in the time domain (e.g., APPC, and PPF) and the frequency domain (e.g., SAPVR and formant peak count). This structure relates to both the harmonics and the spectral envelope of the speech signal. In the presence of an overlapping speaker, this structure is typically altered, and so by encoding this information in the form of one or more features, the two forms of speech can be distinguished.

Even more so than multispeaker SAD, the work on overlapped speech detection specific to meetings is quite recent. Yamamoto et al. in [120], using microphone array processing techniques, detected overlapped speech by applying support vector regression on the eigenvalues of the spatial correlation matrix. These values give the relative power of the sources estimated, from which a decision about whether there is a single or multiple sources can be made. Applying the technique to a single meeting in a room with a reverberation time of 0.5 s and recorded using a circular microphone array, the approach detected around 50% of the overlapping segments. It should be noted, however, that the results were obtained using an optimal threshold. Asano and Ogata in [7] detected multiple speech events—and, consequently, overlapped speech segments—

using an adaptive beamforming framework. The approach first uses an adaptation of the MUSIC method of source localization [100] to identify peaks in the spatial spectrum based on half-second audio segments. The segments are then clustered using K-means (the number of speakers must be known a priori) to determine the spatial range of each speaker. Speech events for each speaker are detected and overlapped speech is identified by finding overlapping speech events. Though overlapped speech detection performance was not given, the authors demonstrated an 8% improvement of phoneme accuracy for ASR using the combined speech event detection and separation procedure.

The nearfield multi-channel audio work of Wrigley et al. described in section 2.1.1 has some relevance to overlapped speech detection as well. This is because one of the four classes defined for the detection task was speech plus crosstalk—i.e., overlapped speech. One potentially significant difference, though, is the TIR of overlapped segments in the nearfield case. Since the individual headset microphone location is much closer to the target speaker than the interfering speaker, it is likely that the TIR will be high. It is also likely, however, that features which work well detecting overlapped speech in high TIR conditions should work well in low TIR conditions. The other significant difference is the nearfield versus farfield condition. 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 for nearfield (see, for example, [73] and [104]). That being said, Wrigley's results point to energy, kurtosis, and cross-channel correlation as being the most effective features for overlap detection.

The first mention of overlap detection specifically for speaker diarization comes from van Leeuwen and Huijbregts in [111]. The authors attempted to directly integrate overlap detection into a diarization system by building two-speaker HMM state models

for each possible pair of overlapping speakers using speech from pairs of clusters from the diarization engine. The segmentation HMM was then modified to allow transitions between single-speaker and two-speaker states and a final re-segmentation of the audio data was performed. The features used in the system consisted of 12 perceptual linear prediction (PLP) coefficients plus log-energy, along with first differences. The authors state that the approach, though producing overlapped speech segments, did not improve diarization performance and only correctly identified the overlapping speakers in about one third of the overlap segments.

Otterson and Ostendorf, however, demonstrated the potential benefit of utilizing overlap information in speaker diarization in [84]. In oracle experiments with perfect overlap detection, the authors demonstrated a 17% relative improvement in diarization error rate on data from the NIST Rich Transcription meeting evaluations by both excluding overlapped speech from speaker clustering and assigning additional speakers in overlapped speech regions of the diarization system output. In his Ph.D. thesis [83], Otterson also investigated methods of automatic overlap detection for speaker diarization using both monaural features and multi-microphone location ones.

In the monaural case, he examined a number of features based on Hough transform pitch detection. These include Hough image entropy, synthesized spectral match, harmonic screen, normalized square synthesis error, spectral flatness, peak-to-valley-ratio spectral flatness, and envelope correlation. On synthesized overlapped speech generated using a subset of the TIMIT corpus, a generalized linear model (GLM) classifier with these features in combination with MFCCs was able to detect 74.7% of overlapped speech frames with a false detect rate of 12.3%. He also observed that the pitch-based overlap detection features performed roughly as well as MFCCs and the combination of the features improved performance. Similar experiments on real meeting data, unfortunately, yielded poor results.

In the multi-microphone case, Otterson examined features derived from microphone delays. Using a GMM classifier with these location features and MFCCs, he obtained a precision of 0.24, a recall of 0.58 and an F-score of 0.34 for data from the NIST meeting evaluations. By feeding these GMM posteriors into a multi-layer perceptron (MLP), the performance was altered to a precision of 0.67, a recall of 0.19, and an F-score of 0.30—an improvement since high precision is preferred for this task (this is discussed later in Section 3.3).

The limited work on overlapped speech detection in meetings seems to point down two general paths. On one hand, there are the multi-microphone approaches, which utilize array processing techniques and have demonstrated moderate success. On the other, there are the monaural, or single-microphone, approaches, which bear similarity to the usable speech detection methods, but have thus far been less successful under realistic conditions. In this thesis, the latter is pursued over the former, with the intention of making significant inroads to performing this challenging task. The methods and measures devised to do so are discussed in the chapter to come.

Chapter 3

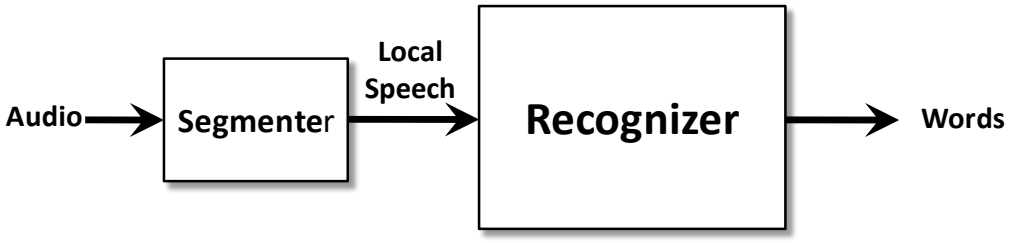
Experimental Framework

In this chapter, the groundwork is lain for the related tasks of handling crosstalk and overlapped speech in meetings. The chapter begins with an explanation of the audio segmentation procedure to be employed, followed by a discussion of the system development process. Lastly, the evaluation paradigm is presented and the evaluation procedure for each task detailed.

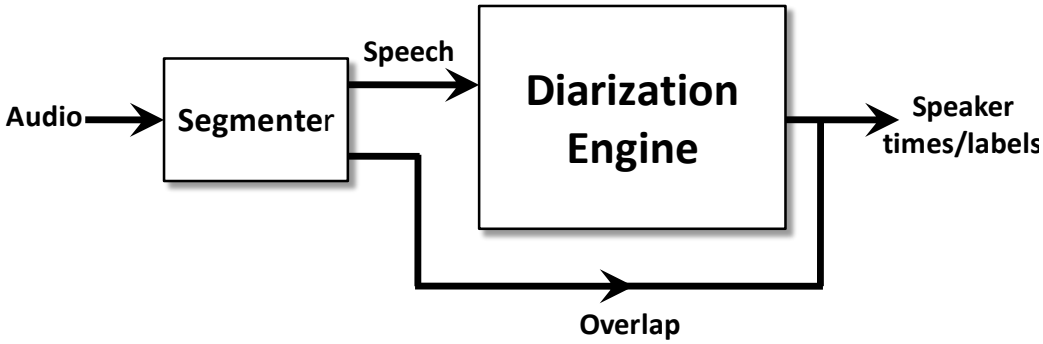
3.1 Audio Segmentation: The Heart of the Matter

Crosstalk and overlapped speech, as discussed in Section 1.2, both represent undesirable acoustic events from the perspective of speech processing systems. The presence of crosstalk speech on nearfield audio channels leads to insertion errors for ASR systems, which mistakenly attempt to recognize speech in these audio regions. Overlapped speech, on the other hand, produces missed speech errors, as diarization systems hypothesize only one speaker per speech region. In addition, these overlap segments, containing speech from multiple speakers, potentially have a negative effect on speaker modeling. A reasonable approach to address both of these problems is to employ an audio segmentation scheme.

Audio segmentation, in general, consists of segmenting a continuous audio stream into acoustically homogeneous regions, where the notion of homogeneity is task-dependent. In the case of crosstalk, the issue is sufficiently addressed by segmenting the local speech regions that are to be processed by the recognizer. Thus, a system which focuses on local speech by utilizing features which distinguish it from its counterpart, crosstalk, is what is desired. In reality, this involves improving an already existing component of the ASR system: the speech activity detector. For overlapped speech in diarization, the objective is somewhat different. In this case we seek to explicitly identify these overlap regions. By doing so, additional speaker labels can be applied and the overlap regions can be further processed to improve speaker clustering. One possibility would be to process the audio in the overlap segments to separate the speakers, and a number of source separation techniques exist such as those based on independent component analysis (ICA) [59], adaptive decorrelation filtering [114], and harmonic enhancement and suppression [72]. A simple first step, however, is to exclude these overlap regions from the speaker clustering phase of the diarization procedure. Since overlapped speech is simply a subset of speech in general, it again makes sense to view this as a modification to the already existing speech activity detection component of the diarization engine. Rather than segment speech and nonspeech, an alternative "overlap-aware" audio segmenter would segment (single-speaker) speech, nonspeech, and overlapped speech. To do this, of course, would once again require the use of additional features, ones that are suited to distinguishing overlap from single-speaker speech and nonspeech. Figure 3.1 diagrams the use of audio segmentation as described above for the applications of ASR and speaker diarization.



(a) Automatic speech recognition. Only local speech is processed by the recognizer.



(b) Speaker diarization. Single-speaker speech is processed by the diarization engine while overlap segments are assigned additional speaker labels as a post-processing step.

Figure 3.1: Diagram of the interface of the proposed audio segmenters for (a) automatic speech recognition and (b) speaker diarization.

3.2 Segmentation System Development

To develop a segmenter as described above requires a number of key design decisions. First and foremost, of course, is the basic implementation of the segmenter. For this work, an HMM approach was adopted, as it represented a natural choice for segmentation; the HMM, for instance, combines both classification and sequence modeling, unlike, say, a support vector machine, which is limited to frame-level classification. Even more important to this work, however, is the selection of features

for identifying the audio classes of interest. The following sections discuss these two major components of the segmentation system development, highlighting some of the associated design issues and noting how they were addressed here.

3.2.1 The HMM Segmenter: An Overview

In Section 2.1, the hidden Markov model was identified as one of the central components to statistical speech recognition as it is generally framed. The modeling of a sequence of words is quite similar to the modeling of a sequence of audio classes, though, and for many cases the HMM approach is used in audio segmentation as well.

We begin with an analogous optimization problem: finding the most likely sequence of audio classes $C = c_0, c_1, \ldots, c_n$ (which can be as abstract as desired) given a sequence of observation vectors $\mathbf{Y}$:

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, P(\mathbf{Y}|C)P(C) \tag{3.1}$$

The first term corresponds to the acoustic model for the audio classes and the second the "language" model, which in this case can be a simple unigram or bigram model.

As before, for the acoustic model, it is useful to divide the classes into (even more abstract) sub-units and model these using an HMM. That is, we assume the sequence of observation vectors $\mathbf{Y}$ is the result of an underlying process that consists of state transition and vector emission where the states $Q = q_0, q_1, \ldots, q_m$ are hidden. The subdivision of the audio classes potentially improves the modeling of quasi-stationary acoustic phenomena. Using three states for each class as done here, for example, permits one state to model the onset of the phenomenon, one to model the main stationary component, and one to model the end. Sub-units are also useful in enforcing minimum duration constraints for audio classes, as in the segmentation component of the diarization system described in Section 3.3.4.

Since the formulation is the same as with ASR, the acoustic model here is also of the parametric form given in Equation 2.4. Previously, little mention was made of the output probability $b_j(\mathbf{y_t})$. In reality, the choice of this probability function is crucial since it needs to model all of the relevant variability in the feature space, in this case acoustically-derived features such as MFCCs. Though alternatives exist (e.g., vector quantization), the most common approach for speech and audio is to model the features using a continuous density Gaussian mixture model (GMM). That is,

$$b_j(\mathbf{y_t}) = \sum_{m=1}^{M} c_{jm} N(\mathbf{y_t}; \mu_{jm}, \Sigma_{jm}) \tag{3.2}$$

where c_{jm} is the weight of the mixture component m in state j and $N(\mathbf{y_t}; \mu_{\mathbf{jm}}, \Sigma_{jm})$ denotes a multivariate Gaussian of mean μ and covariance Σ. Furthermore, in many cases this covariance matrix is constrained to be diagonal, greatly reducing the number of model parameters and thus the amount of training data needed to estimate these parameters. The GMM allows for the modeling of complex distributions with multiple modes, the complexity being determined by the number of mixtures used. As with all parametric modeling, one seeks a number of mixtures (and, thus, parameters) that balances the trade-off between model accuracy and generalization on unseen data.

As mentioned in Section 2.1, the model parameters of the HMM are trained using Baum-Welch re-estimation, an instance of the well-known expectation maximization (EM) algorithm. The algorithm consists of the following steps:

1. Initialize the parameter values.

2. Calculate the forward and backward probability of each state.

3. Update the HMM parameters using the newly obtained probabilities from 2.

4. Repeat steps 2 and 3 until no significant increase in model likelihood is obtained.

In step 1, the HMM transition probabilities are typically initialized using alignment information from the training data, while the GMM parameters are initialized using a clustering procedure to partition the feature space into K regions and compute means for the K mixtures. In step 2, the forward probability is defined as

$$\alpha_j(t) = P(\mathbf{y_1}, \ldots, \mathbf{y_t}, x(t) = j|M) \tag{3.3}$$

and represents the joint probability of observing the first t speech vectors and being in state j at time t. The backward probability $\beta_j(t)$, is

$$\beta_j(t) = P(\mathbf{y_{t+1}}, \ldots, \mathbf{y_T}|x(t) = j, M) \tag{3.4}$$

and represents the probability of observing the vectors from $t+1$ to the end of the sequence, given we are in state j at time t. The two quantities are efficiently computed using a recursion, the *forward-backward algorithm*. The forward and backward probabilities allow for the computation of the state occupation likelihoods $\gamma_j(t)$, which are used to update the HMM parameters. The EM algorithm is locally optimal and is guaranteed to not decrease the model likelihood.

Viterbi Decoding

For decoding, the maximum likelihood state sequence—i.e., the sequence which maximizes $P(\mathbf{Y}, X|M)$ in Equation 2.4—is used as a computationally efficient approximation to the complete likelihood, given by

$$P(\mathbf{Y}, X|M) = \sum_X a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(\mathbf{y_t}) a_{x_t x_{t+1}} \tag{3.5}$$

$$\approx \max_X a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(\mathbf{y_t}) a_{x_t x_{t+1}} \tag{3.6}$$

This latter likelihood, in turn, can be computed iteratively in what is known as the *Viterbi algorithm*. For a given model M, let $\psi_j(t)$ represent the maximum likelihood of observing vector $\mathbf{y_0}$ to $\mathbf{y_t}$ and being in state j at time t. This partial likelihood can be computed using the recursion

$$\psi_j(t) = \max_i\{\psi_i(t-1)a_{ij}\}b_j(\mathbf{y_t}) \tag{3.7}$$

where $\psi_1(1) = 1$ and $\psi_j(1) = a_{1j}b_j(\mathbf{y_1})$ for $1 < j < N$. The maximum likelihood is then given by

$$\psi_N(T) = \max_i\{\psi_i(T)a_{iN}\} \tag{3.8}$$

To prevent underflow, log-probabilities are generally used and the products in the equations above become summations.

The Viterbi algorithm is visualized as finding a path through a trellis where the vertical dimension corresponds to the possible states and the horizontal to frames. The log-probability for any path is computed by summing the log transition probabilities and the log output probabilities along that path. The paths are grown from left-to-right column-by-column. At time t, each partial path is known for all states and the log form of Equation 3.7 is used to extend the partial paths by one time step. Since the state at each time is known using this procedure, segment start and end times can be obtained by identifying frames where state transitions take place.

3.2.2 Feature Fusion

A major focus of this thesis is investigating which features work well for the audio segmentation tasks of local and overlapped speech detection when used in the HMM based segmenter as described above. For most classification tasks, features are typically used in combination rather than in isolation, so the combination of the candidate

features must also be considered. Several schemes exist for combining features in a statistical classifier, each with its benefits and drawbacks.

Concatenation and Transformation

The simplest and most straightforward method, *serial fusion*, consists of concatenating the individual features or feature vectors into a supervector. One issue with this method is that it can result in feature vectors of very large size, leading one to fall prey to Bellman's *curse of dimensionality* [9]—that is, a sparsity of data and an increase in noise with the addition of dimensions. It may also be the case that the features are highly correlated, which violates the independence assumption of classifiers such as diagonal-covariance multivariate GMMs (discussed in Section 3.2.1). These issues can be addressed, however, using dimensionality reduction techniques, which represent the next level of complexity in feature combination methods.

The simplest of these techniques is principal component analysis (PCA) [29], in which the supervector data is decorrelated and projected into a lower-dimensional space based on the variance of each feature vector dimension. This approach has a number of potential issues, however. First, PCA assumes the features obey a (uni- or multivariate) Gaussian distribution. This can be addressed by Gaussianizing the data, a technique discussed in Section 3.2.4. Secondly, this approach is based on the assumption that higher-variance dimensions provide more information for classification. This assumption may be violated if the features have different dynamic ranges. This, too, can be addressed; one can apply either Gaussianization or simple variance normalization. Thirdly, the PCA projection only ensures the global decorrelation of features. For classification, though, it is usually most important for the features representing a given class (e.g., an HMM state) to be decorrelated. This last issue is addressed by the alternate technique of linear discriminant analysis (LDA) [26].

LDA is a feature transformation technique that, like PCA, can be used to obtain a linear projection of decorrelated feature supervectors with reduced dimension. LDA, however, uses class label information to find a projection that maximizes the ratio of between-class to within-class covariance and so preserves class-separating dimensions rather than simply high-variance ones. This makes feature scaling such as variance normalization unnecessary. LDA is based on the assumption that the covariances for all classes is the same, though, which is often violated in speech-related classification problems. To address this, a generalization of LDA, the *heteroscedastic* linear discriminant analysis (HLDA) [51] is often employed. HLDA utilizes class-conditional covariance matrices to produce a transformation matrix for supervector projection. This, of course, requires that there be sufficient data for each class to reliably estimate these parameters, a potential issue for minority classes and small data sets. It is possible, though, to interpolate between the global and class-conditional parameters in this case, as in [17] to yield smoothed estimates and improve performance.

Despite the improvements over concatenation that may be obtained using any of the above transformations, the dimensionality reduction techniques all have the problem of dimensionality selection; that is, how does one determine the dimensionality of the projection space? Though several automatic methods exist for selection (e.g., [67],[115],[110], and [92]), most yield results inferior to manual selection and are not widely used. Manual selection may prove time and computationally intensive, however, in particular for complex classification systems with large amounts of data such as in speech-related areas. All of this suggests the benefits of feature transformation are far from guaranteed.

Multi-stream Likelihood/Posterior Combination

An alternate class of combination techniques, termed *parallel fusion*, uses parallel streams of features in synchrony and merges information from the various streams by

way of frame-level likelihoods or posteriors from independently trained probabilistic models such as a GMM or MLP. In doing so, the techniques sidestep many of the issues faced by the concatenative class of combination methods. The multi-stream approach is particularly well-suited for speech processing, as there exists an understanding of band-level processing within the human auditory system that provides a basis for many of the processing techniques and tools in use today (MFCCs, PLPs, Gammatone filterbanks, among others). This gives rise to multi-band systems such as [13], [107], [69], and [81]. To merge stream information, the simplest and most widely used methods are the sum and the product rule—the arithmetic and geometric mean of the posteriors, respectively. Generalizations of these two approaches exist in which streams are not equally weighted, but determining appropriate weights can prove challenging, especially for large numbers of streams. Though more principled approaches exist, a common technique is to manually tune the weights using held-out data. As with the manual dimensionality selection associated with feature transformation methods, this procedure may take a lot of time and computational resources. In addition to the weights of the streams, the appropriate number of streams and the features which are assigned to each stream may not be obvious either.

Given that parallel fusion, too, has its own set of issues, it is not surprising that performance of this approach relative to simple concatenation varies. In [27], for example, using separate streams for PLP and MSG features yielded a significant improvement in WER (76.5% vs. 63%) for the Aurora task while in [21], performance using concatenation and multiple streams was similar (5.9% vs. 5.8%) on the Numbers95 task.

Having analyzed the various methods of feature fusion, it was decided that simple concatenation was sufficient for the purposes of this thesis. As feature selection also figures heavily in this work, it was anticipated that some of the potential pitfalls of concatenation (such as redundant correlated features) would be avoided.

3.2.3 Feature Selection

Having determined a method of combining features, it is important to realize that using all features in combination may be suboptimal for the desired segmentation systems. Some features may simply be unnecessary, neither improving nor worsening performance given the feature ensemble. This may be the case even if the feature works well in isolation or in a smaller ensemble. Other features may actually worsen performance. Again, this may occur even if the feature displays discriminative properties in isolation or in a smaller collection of features. It becomes necessary, then, to search the space of feature combinations to find the "best" combination according to some criterion. For N features, an exhaustive search yields 2^N possible combinations (if the set of no features is included), which quickly becomes intractable as N increases, especially if the evaluation of the feature combination is computationally intensive, as in the case of the target applications for this work. As is often the case, optimality may need to be sacrificed for efficiency and a "good" feature combination may have to suffice. The methods by which this combination is determined fall under the category of feature selection, a topic which has received much attention in the machine learning, pattern recognition, statistics, and several other communities.

Though defined by many authors, the most relevant definition of feature selection to this work comes from Koller and Sahami in [48] as follows: a process which aims to choose a subset of features for improving prediction accuracy or decreasing the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features. Dash and Liu in [23] outline the four major components of a feature selection method:

1. A *search/generation procedure* to generate the next candidate subset;

2. An *evaluation function* to evaluate the subset under examination;

3. A *stopping criterion* to decide when to stop; and

4. A *validation procedure* to check whether the subset is valid.

Broadly speaking, feature selection methods fall within either the *filter model*—in which selection is performed as a filtering preprocessing step to the induction algorithm (e.g., machine learning classifier)—or the *wrapper model*—in which the subset selection algorithm exists as a wrapper around the induction algorithm. Though the filter method is generally less computationally intensive, John et al. in [42] argue for the use of wrapper approaches, with the reasoning that the induction method that ultimately uses the feature subset should better estimate the accuracy than a separate measure that may have a different inductive bias. Perhaps consequently, these approaches have received more attention and tend to dominate in machine learning and related fields. The wrapper model was adopted here, too, for this work.

Figure 3.2 gives a diagram of the wrapper approach to feature subset selection. In this approach, the induction algorithm is considered a "black box". The algorithm is run on the dataset, usually partitioned into training and test/cross-validation sets, with different sets of feature removed from the data (*search/generation*). Subsets are iteratively evaluated by the induction algorithm (*evaluation function*) until all subsets have been evaluated (*stopping criterion*). The subset of features with the highest evaluation is chosen as the final set on which to run the induction algorithm. The classifier is then evaluated on an independent test set that was not used during the search (*validation procedure*).

In terms of the search/generation procedure for feature selection, many common AI algorithms have been employed. The most commonly used, however, fall under one of two types of greedy algorithms: forward selection and backward elimination.

In forward selection, features are successively added starting from an empty set. Each of the N features is evaluated and the feature yielding the best performance is

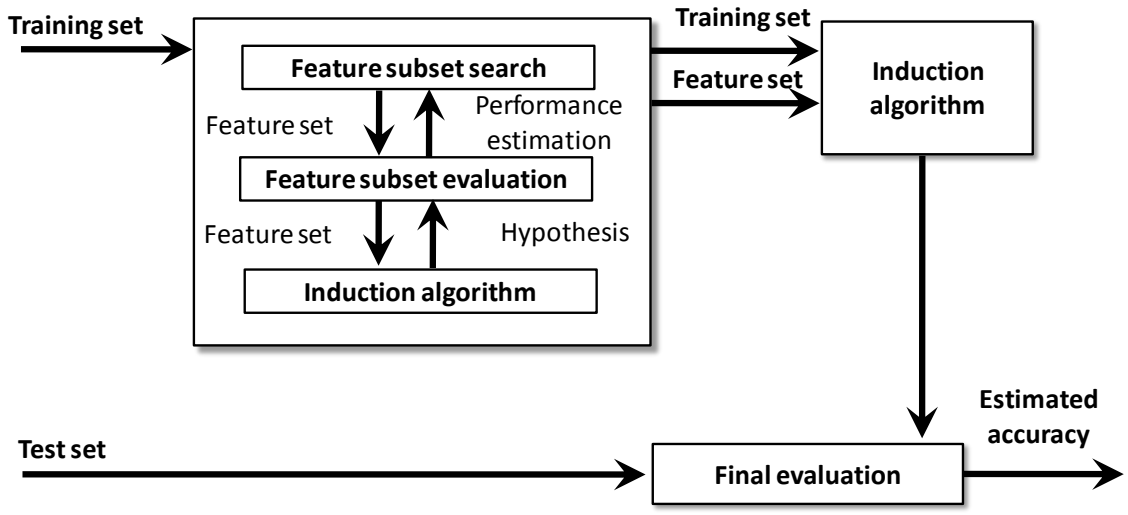


Figure 3.2: The wrapper approach to feature selection. The induction algorithm, considered a "black box" by the subset selection algorithm, is repeatedly applied and the results evaluated as part of the selection process.

first added. This feature is then combined with each of the remaining $N-1$ features and the best of these two-feature combinations is selected. Note that this is only the best two-feature combination involving the single best feature; all other two-feature combinations are disregarded. The process iterates until the addition of a feature causes a performance degradation; until no improvement in performance is obtained; or until all features have been added, in which case the best performing set in the series is chosen.

One potential issue with forward selection is the fact that each addition of a new feature may render one or more of the already included features irrelevant. This may arise, for example, in the case of correlated features. Backward elimination avoids this problem by starting with the full ensemble of features and successively eliminating the worst performing features, one at a time. In the first stage, each of the N features is individually removed to create N sets of $N-1$ features. Each set is evaluated and the best performing set is selected. Each of the remaining $N-1$ features in this set is then

removed and the best $N-2$-feature subset of the $N-1$ features is determined. The process continues until performance degrades; until performance ceases to improve; or until no features remain, in which case the best performing set in the series is once again chosen.

Backward elimination has drawbacks as well. Sometimes features are dropped that would be significant when added to the final reduced feature set. In addition, if the total number of features N is very large, evaluation of the feature sets may be prove too computationally expensive for some classifiers (e.g., support vector machines or neural networks).

For this thesis, two different feature selection approaches were adopted for the two different tasks. In the case of the multispeaker SAD system, the number of features explored was small enough that exhaustive search of all 2^N combinations could be performed to find the globally optimal set of features. This was not so for overlapped speech detection, for which a backward elimination method was adopted. Though the number of features in this case is large enough to merit the use of a suboptimal greedy algorithm, it was not large enough to be prohibitive for the backward elimination procedure.

3.2.4 Feature Transformation

A common issue for classification systems is robustness to data mismatches. For speech and audio data, the mismatch typically relates to the channel (e.g., headset vs. distant microphone) or the recording environment (e.g., different rooms or microphone locations). The most severe form of this mismatch occurs between training and test sets, but another form can arise within either (or both) of these sets and also results in reduced performance. In the case of meeting speech processing, for example, having training or test data from a variety of sites—and, hence, different recording

environments—creates an internal mismatch that can affect the performance of the trained models. This is particularly the case with farfield audio data, which is much more susceptible to the reverberant and convolutive effects of the room in which it was recorded.

One possible technique to address this issue is to perform meeting-level transformations so that the distributions of features in the various meetings is similar. The simplest method of doing this is mean-variance normalization, in which feature values are first shifted by the estimated sample mean and then scaled by the inverse of the estimated sample standard deviation to produce a distribution with zero mean and unity standard deviation for each feature in each meeting. This approach, however, can only address linear differences in data, which may be insufficient in real-world environments with complex acoustic and channel effects.

Consequently, nonlinear normalization techniques have also been proposed and have demonstrated good results ([71],[24], and [99]). One such procedure was adopted for this work. For each meeting, feature space Gaussianization is performed independently on each feature vector component by using a nonlinear warping constructed from histograms and an inverse Gaussian function. The result is a feature component normalized to a zero-mean unity-variance Gaussian distribution. The technique was motivated by Saon et al. in [99], in which the authors applied feature space Gaussianization to obtain an 11% relative WER improvement on an in-car database over standard linear feature space transformations. In addition, they motivated the choice of normal target distribution by suggesting it facilitates the use of diagonal covariance Gaussians in the acoustic models. To further facilitate this, a decorrelation procedure via the Karhunen-Loeve Transform (KLT) was applied after the Gaussianization. As previously mentioned, the distortion effects are most significant in the farfield condition, and so the transformation was only applied in the overlapped speech detection system.

3.3 System Evaluation: Are We Making a Difference?

A critical component to the development of any system is its evaluation. Evaluation provides a clear way of testing experimental hypotheses and determining if objectives have been met. Furthermore, having a well-defined evaluation plan can simplify the research and development process by providing better direction for, and consistency between, efforts. The history of speech technologies provides a good example of this. In the early days of the field, research sites were effectively islands to themselves; though working on the same problems, each site used its own data sources and metrics for evaluation. The outcome of this was reported results that were neither reproducible nor fully comparable. With the advent of standardized evaluation plans—and with them, standardized data corpora—efforts became more directed and collective, and the pace of advancement increased.

Since 2002, the National Institute of Standards in Technology (NIST) has run a rich transcription (RT) evaluation for meetings speech processing. The evaluation measures the progress towards developing systems that "provide the basis for the generation of more usable transcriptions of human-human speech in meetings for both humans and machines" [76][77][78][79]. The tasks involved have included speaker localization, speech activity detection, speech-to-text (commonly referred to as ASR), speaker diarization, and speaker attributed speech-to-text (essentially a combination of ASR and speaker diarization). The evaluation plan prescribes data for system development and testing and defines performance metrics for the various tasks. The NIST RT evaluation, then, was a natural choice for an evaluation framework and was consequently adopted for this thesis.

Ultimately this work seeks to improve meeting ASR and speaker diarization. However, the proposed method of doing so involves audio segmentation preprocessing for each application. It is important, then, to make the distinction between the

performance—and, thus, evaluation—of the intermediate application and the target application in each case.

As mentioned in Section 2.1, the primary performance metric for ASR is word error rate, which can be decomposed into the constituent metrics of substitution, deletion, and insertion error rates. Speech/nonspeech segmentation performance, in contrast, is often measured using speech diarization error rate (SDER). This metric is computed similarly to the speaker diarization error rate in Equation 2.10, but with the speaker error term removed, as no speaker labels are applied. Because ASR output is greatly affected by the input speech segmentation (as is evident, for example, from the insertion errors caused by crosstalk), SDER and WER tend to be highly correlated. For this work, WER improvement was the evaluation metric selected, but SDER performance was measured as well to better analyze the connection between segmentation and recognition; knowing the false alarm rate and its relation to the insertion rate, for example, makes it easier to identify a system improvement as being due to reduced crosstalk.

For speaker diarization, the primary metric is, of course, diarization error rate. As a rather uncommon task, the detection of overlapped speech itself lacks a standard metric, but common segmentation metrics are applicable. Here, as well, DER could be modified to an "overlap diarization error rate" (ODER), analogous to SDER and computed in the same fashion. A serious issue with this is the relatively small (though, as has been argued, certainly significant) amount of overlapped speech involved. For a meeting such as those in the evaluation test sets, the total amount of overlapped speech can be on the order of a few minutes. Related to this, the overlap segments are generally very short in duration. Figure 3.3 shows a histogram of the overlap segment durations for the AMI corpus. The median overlap duration is 0.46 s. The result of this is very low false alarms, but very high miss rates and, with them, high overlap diarization error rates. If we consider the target application, however, we

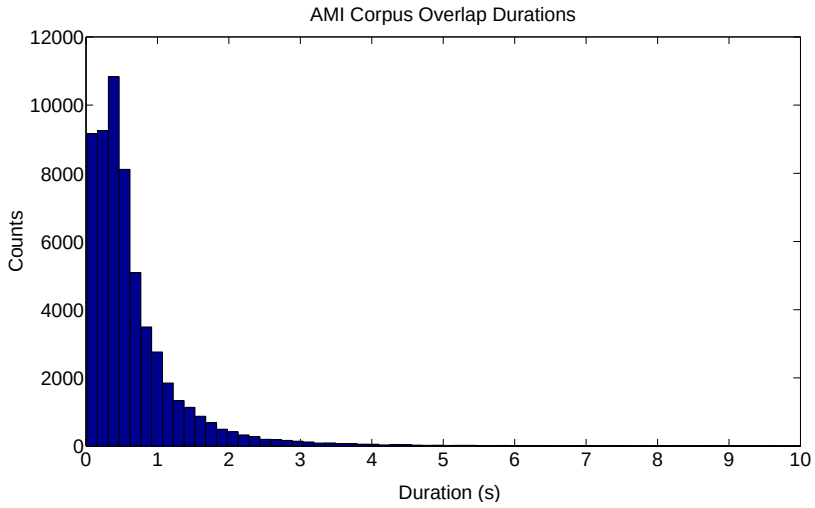


Figure 3.3: Histogram of overlapped speech segment durations in the AMI meeting corpus. The median value is 0.46 s

realize that the two constituent errors of the ODER impact the diarization system quite differently. False alarms generated by the overlapped speech segmenter carry through to increase the diarization false alarm error, while misses have no effect on the baseline DER. The importance of a low false alarm operating point for the segmenter is more clearly conveyed through the metrics of precision, recall, and F-score. For segmentation, precision can be computed as

$$\text{Precision} \;=\; \frac{T_{\text{OVERLAP}} - T_{\text{MISS}}}{T_{\text{OVERLAP}} - T_{\text{MISS}} + T_{\text{FA}}} \tag{3.9}$$

and measures in frames the number of true positives divided by the sum of true and false positives. A low false alarm rate, as desired for this task, translates to a high precision. Recall measures the number of true positives over the sum of true positives and false negatives and is calculated as

$$\text{Recall} \;=\; \frac{T_{\text{OVERLAP}} - T_{\text{MISS}}}{T_{\text{OVERLAP}}} \tag{3.10}$$

The F-score is a summary metric obtained by dividing the geometric mean of the precision and recall by the arithmetic mean. These three metrics were utilized in addition to relative DER improvement, the primary figure of merit.

The sections to follow further describe the evaluation of the audio segmentation systems, detailing the evaluation data as well as the recognition and diarization systems used for measuring target application performance.

3.3.1 Multispeaker Speech Activity Detection

The performance of the nearfield SAD system was evaluated using data from the 2004 and 2005 NIST RT meeting evaluations. This consists of collections of 11- to 12-minute excerpts of recordings of multiparty meetings from different sites—and, thus, with different room acoustics. The audio was obtained from individual headset microphones with a 16 kHz sampling rate. Table 3.1 lists the meetings from the 2004 and 2005 evaluations, along with relevant statistics such as number of speakers, excerpt length, and amount of speech. The 2004 evaluation served as a development test set, where system modifications—primarily the selection of candidate features—could be performed prior to a final validation on the 2005 data. This is in line with the wrapper model approach outlined in Figure 3.2.

To measure the significance of the different word error rate results produced by the recognizer for the segmentations generated by the SAD system, the matched pairs sentence-segment word error (MAPSSWE) test was used. This test, suggested for ASR evaluations by Gillick in [32] and implemented by NIST [85], looks at the number of errors occurring in sentence-segments specific to the output of the two systems being compared. For MAPSSWE the segments are sequences of words that include recognition errors in at least one of the two systems being compared, bounded on both sides by two or more words correctly identified by both systems. This contrasts

the standard matched pairs sign test, which looks at entire sentences rather than these shorter sentence-segments. The rationale for the MAPSSWE approach is to increase the sample size and thus obtain better statistics. This is justified by asserting that the errors in two adjacent segments are independent because they are separated by at least two correctly recognized words, and most ASR systems do not exploit more than a trigram context. Given the large number of segments, the central limit theorem states that the distribution of the number of errors is normally distributed. The MAPSSWE test, then, is a t-test for estimating the mean difference of normal distributions with unknown variances. To perform the test, the mean and variance of the segment-level error difference, $\hat{\mu}_z$ and $\hat{\sigma}_z^2$, respectively, are computed. The null hypothesis asserts that $\mu_z = 0$ and the probability of this hypothesis is calculated according to $P = 2Pr(Z \geq |w|)$, where $Z \sim N(0,1)$ and w is a realization of W, given by

$$W = \frac{\hat{\mu}_z}{\hat{\sigma}_z/\sqrt{n}} \tag{3.11}$$

Systems were deemed significantly different if $P < 0.05$.

3.3.2 The ICSI-SRI RT-05S Meeting ASR System

To evaluate the performance of ASR using segments obtained from the multispeaker SAD system, the ICSI-SRI system [104] from the NIST Spring 2005 Rich Transcription meeting recognition evaluation (RT-05S) [77], was used. The system is based on the SRI-ICSI-UW conversational telephone speech (CTS) recognizer [105] for the NIST Fall 2004 Rich Transcription evaluation (RT-04F) [75] with various adaptations for the meeting domain.

	Meeting	Number of Speakers	Excerpt Length (mins)	Amount of Speech (mins)
RT-04S	CMU_20030109-1530	4	11.02	10.62
	CMU_20030109-1600	4	11.10	10.99
	ICSI_20000807-1000	6	11.37	10.60
	ICSI_20011030-1030	10	11.50	11.01
	LDC_20011121-1700	3	11.03	10.41
	LDC_20011207-1800	3	11.62	10.09
	NIST_20030623-1409	6	11.23	10.66
	NIST_20030925-1517	4	11.03	9.81
RT-05S	AMI_20041210-1052	4	12.18	10.76
	AMI_20050204-1206	4	11.91	10.59
	CMU_20050228-1615	4	12.03	11.39
	CMU_20050301-1415	4	11.97	10.86
	ICSI_20010531-1030	7	12.18	11.02
	ICSI_20011113-1100	9	11.99	11.03
	NIST_20050412-1303	9	12.12	9.99
	NIST_20050427-0939	4	11.93	10.75
	VT_20050304-1300	5	11.98	10.62
	VT_20050318-1430	5	12.08	9.05

Table 3.1: Test meetings for the RT-04S and RT-05S evaluations.

Basic System

The system diagram is shown in Figure 3.4. The "upper" tier of decoding steps is based on MFCC features. The "lower" tier of decoding steps uses PLP features. The outputs from the two tiers are combined twice using word confusion networks (the crossed ovals in the diagram). With the exception of the first stage of each tier, the acoustic models are adapted to the output of the previous step from the other tier (i.e., cross-adapted) using maximum likelihood linear regression (MLLR) adaptation [60]. The final output is the result of a three-way system combination of non cross-word MFCC (MFCC-nonCW), cross-word MFCC (MFCC-CW), and crossword PLP (PLP-CW) decoding branches and runs in under 20 times real time.

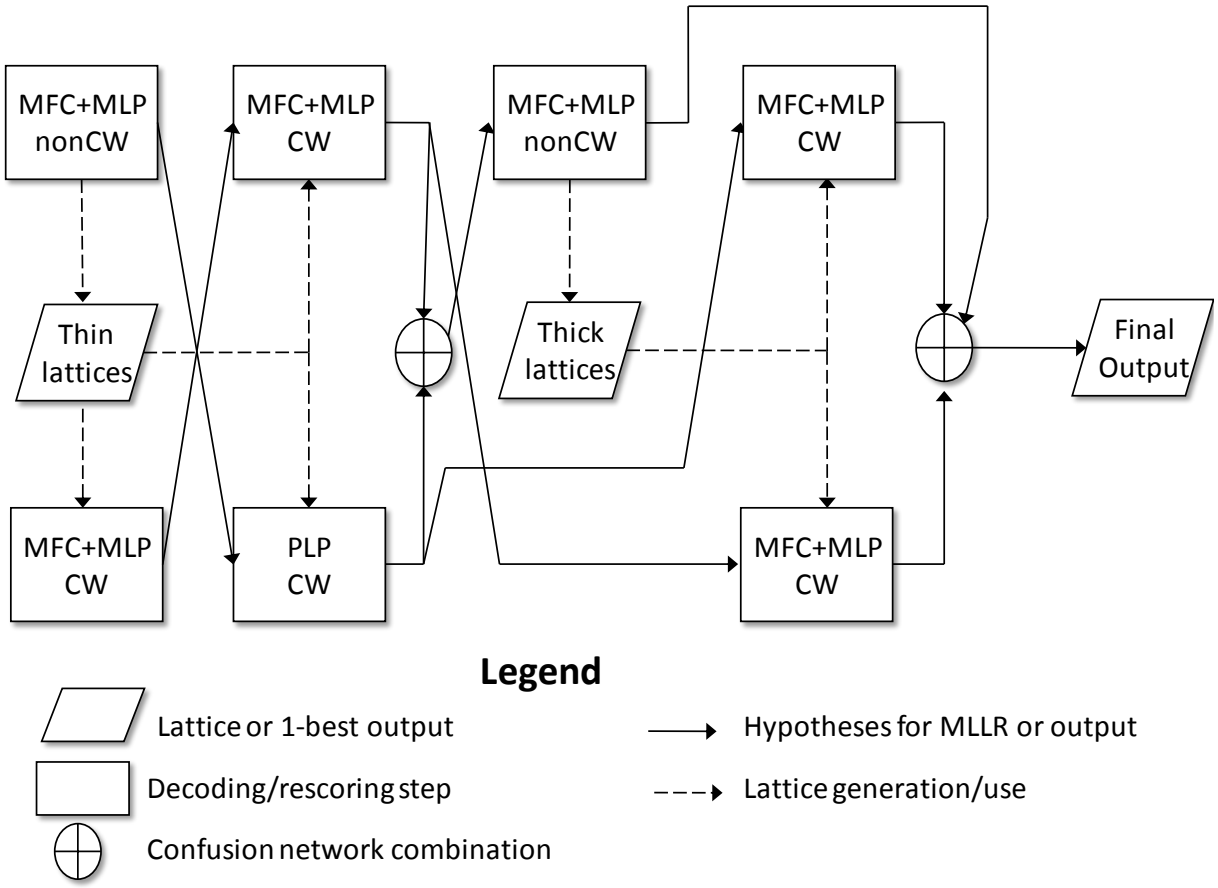


Figure 3.4: Diagram of the ICSI-SRI RT-05S meeting recognition system. The "upper" tier of decoding steps is based on MFCC features, while the "lower" tier uses PLP features.

Acoustic Features

The MFCC features consist of 12 cepstral coefficients, energy, first-, second-, and third-order differences, and 2 x 5 voicing features developed by Graciarena et al. and described in [34]. The cepstral features were normalized using vocal tract length normalization (VTLN) [113] as well as zero-mean and unit-variance normalization per speaker/cluster. The features were then transformed using heteroscedastic linear discriminant analysis (HLDA). Finally, a 25-component Tandem/HATs [74] feature vector estimated by multilayer perceptrons was appended. The same sequence of

processing steps occurred with the PLP coefficients, with the exception of appending voicing and MLP features.

Acoustic Models

The acoustic models were trained using the minimum phone error (MPE) criterion [91] on about 1400 hours of conversational telephone speech data. These models were then adapted for the meeting domain using maximum a posteriori (MAP) adaptation of about 100 hours of meeting data.

Language Models

Decoding involved the use of three language models: a multiword bigram for generating lattices, a multiword trigram for decoding from lattices, and a word 4-gram for lattice and N-best rescoring. The language models were estimated from a mix of telephone conversations, meeting transcripts, broadcast, and Web data. The vocabulary consisted of about 55,000 words.

3.3.3 Overlapped Speech Detection

The farfield overlapped speech detection system was evaluated using test sets obtained from the AMI meeting corpus [18]. The corpus is a multi-modal data set consisting of 100 hours of meeting recordings. The meetings are primarily scenario-based (about two-thirds are of this form), ranging in duration from 11 to 90 minutes, and each having four participants. As with the RT data, recordings were produced at multiple sites with different room acoustics. These recordings contain on average 15% overlapped speech, making them highly useful for this task.

A rather different approach was taken for selecting the evaluation test sets from this data. Two of the main sources of acoustic variability which affect the performance

of audio segmentation are noise and differences in room acoustics. To analyze the influence of each of these sources, test sets were constructed to control for them. The first test condition consisted of 12 meetings recorded at a single site, IDIAP, using audio data obtained by summing the nearfield headset channels of the participants, referred to as "mixed-headset" audio. With high SNR and minimal environment variability, the results for this set were intended to represent an upper bound of system performance. This is analogous to the nearfield microphone condition for meeting ASR, which serves as an upper bound for farfield recognition performance. This subset of meetings contains approximately 17% overlapped speech. The second test condition consisted of the same 12 meetings, but using farfield audio. The third (and final development) condition consisted of 10 meetings randomly selected from the corpus and using farfield audio. This multi-site test set represented the most realistic—and, consequently, most challenging—of the development set test conditions. This meeting subset also contains around 17% overlapped speech on average. The final validation test set consisted of the farfield audio of 10 meetings that were also randomly selected from the corpus, but done so as not to include any meetings from the development data. For this data, overlapped speech comprises about 17% of the total speech. Table 3.2 lists the meetings for the various test sets along with relevant statistics such as the amount of speech and percentage of overlapped speech.

Statistical significance of DER results was determined using a modification to the MAPSSWE test for ASR described in Section 3.3.1. By sampling the output segmentation of the diarization system at a regular interval (here, one second), "words" corresponding to speaker labels were produced of the form:

```
Reference: <sil> Dave Dave Dave Joe  Joe  Joe  Mark Mark <sil> Dave
System 1:  <sil> spk1 spk1 spk1 spk2 spk2 spk2 spk3 spk3 <sil> spk1
System 2:  <sil> spkA spkA spkC spkC spkB spkB spkC spkC <sil> spkB
```

Afterwards, an optimal mapping was made between system speaker names and reference names (as in standard DER scoring) and the MAPSSWE test was performed. Here, too, the threshold for significance was set at $P < 0.05$.

Another part of the overlap segmenter experimental setup which merits explanation is the use of the single distant microphone (SDM) audio condition. As listed in Section 1.1, there are various sensors used to obtain audio data in meetings. This is particularly true for the distant microphone condition, where multiple microphones can be employed for signal enhancement, source localization, or source separation. The availability of multiple microphones allows for processing techniques which improve the performance of spoken language technology systems, such as those for automatic speech recognition and speaker diarization. Microphone beamforming, for example, can increase the signal-to-noise ratio of the speech signal, leading to lower WERs (on the order of a 15% relative improvement in the case of [41], for example). The use of delay features computed from multiple distant microphones, too, has yielded significant gains (also about 15% relative as in [86]) for speaker diarization.

The benefits of multiple microphones is certainly clear. Indeed, the MDM condition is the primary condition for the NIST RT evaluations. There are some cases, however, when using multiple microphones is problematic, and overlapped speech detection within the framework presented here is among them. The segmentation of overlapped speech regions requires the selection of the appropriate acoustically-derived features for the task, a primary focus of this work. The delay-and-sum beamforming signal enhancement commonly performed in MDM processing, by computing delays relative to the dominant speaker, however, suppresses speaker overlap and in doing so negatively affects any features derived from the resulting signal. Further still, the delay features obtained as a by-product of this procedure produce speaker clustering improvements that effectively obviate the overlap exclusion method. Otterson and Ostendorf in [84], for example, showed that overlap exclusion yielded no improvements on top of those

	Meeting	Amount of Speech (mins)	Percent Overlap(%)
AMI Single-site	IS1000a	14.47	13.60
	IS1001a	9.84	15.79
	IS1001b	25.13	9.43
	IS1001c	16.84	9.69
	IS1003b	16.90	12.44
	IS1003d	26.71	34.22
	IS1006b	26.67	16.11
	IS1006d	24.24	39.03
	IS1008a	11.68	4.84
	IS1008b	21.26	6.67
	IS1008c	19.74	13.61
	IS1008d	18.84	12.93
AMI Multi-site	EN2003a	26.39	9.00
	EN2009b	31.57	19.21
	ES2008a	11.50	5.47
	ES2015d	24.76	29.04
	IN1008	44.71	9.02
	IN1012	44.17	27.67
	IS1002c	26.91	11.38
	IS1003b	16.90	12.44
	IS1008b	21.26	6.67
	TS3009c	31.41	26.53
AMI Validation	EN2006a	35.29	18.81
	EN2006b	29.64	22.24
	ES2002a	12.16	10.83
	ES2003c	28.08	5.04
	IN1014	49.31	13.85
	IS1002d	15.19	17.14
	IS1004d	22.17	19.57
	TS3006c	33.65	21.76
	TS3007a	17.01	13.25
	TS3010c	20.22	12.11

Table 3.2: Test meetings for the AMI single-site, multi-site, and validation evaluation test sets.

obtained by the use of delay features. A single distant microphone does not have these issues, however, and so this alternate condition was used for this work. Here the SDM channel was obtained by randomly selecting from one of the several microphone array (both circular and linear) channels available. As with the nearfield audio, the data was sampled at 16 kHz and noise-reduced via Wiener filtering.

This is not to say, of course, that MDM overlapped speech detection is not feasible. Recall that Asano and Ogata in the work described in Section 2.2.1, for example, detected overlapped speech events using adaptive beamforming. The SDM condition, however, benefits much more from the HMM based segmentation framework. In addition, with diarization error rates some 60% higher than with multiple distant microphones ([116]), the SDM diarization task may also benefit more from the focus of attention.

3.3.4 The ICSI RT-07S Speaker Diarization System

Similar to the ASR task, a state-of-the-art speaker diarization system was used to evaluate the performance of the overlapped speech detection system. The ICSI system [116] fielded in the speaker diarization component of the NIST Spring 2007 Rich Transcription meeting recognition evaluation (RT-07S) [79] served this purpose. The system is based on agglomerative clustering of segments with merging using a modification of the Bayesian Information Criterion in which the number of parameters between the two BIC hypotheses is constant. This modification, mentioned in Section 2.2 and proposed in [3], eliminates the need for a BIC penalty term and thus one of the parameters that must be tuned.

Basic System

A diagram of the system is shown in Figure 3.5. Front-end acoustic processing is first performed followed by speech/nonspeech detection using energy information. After starting with a uniform segmentation of speech segments that corresponds to a large number of clusters, the system then proceeds with several iterations of cluster merging along with model re-training and re-alignment. The final output consists of speech segments with speaker labels corresponding to the final N speaker models obtained through clustering.

Front-end Processing

The acoustic processing consists of Wiener filtering of the audio data followed by feature extraction. The Wiener filtering seeks to reduce corrupting noise—assumed to be additive and stochastic—based on noise spectral estimates in nonspeech regions. The nonspeech regions were determined by a voice activity detection (VAD) component in the Aurora 2 front-end proposed by ICSI, OGI, and Qualcomm in [1]. The feature extraction generates 19 MFCCs every 10 ms with a 30 ms analysis window. In speech processing the use of higher order cepstra (i.e., beyond 12) has been shown to improve performance on speaker-specific tasks (e.g., speaker recognition [97] [22]) and suggests that these components capture more speaker-specific information.

Speech/Nonspeech Detection

The speech/nonspeech detector performs iterative training and re-segmentation of the audio into three classes: speech, silence, and audible nonspeech. To bootstrap the process, an initial segmentation is created with an HMM that contains a speech and silence GMM trained on broadcast news data. The silence region is then subdivided into two classes—regions with low energy and regions with high energy and high zero-crossing rates—and new GMMs are trained. A GMM is then also trained for

the speech class. Since it is possible that a recording may not have a significant amount of audible nonspeech, the system checks to see if the speech and audible nonspeech models are similar by comparing the BIC score of a single model to that of two separate models. In the event that they are similar, the audible nonspeech model is discarded and a speech model is trained on the pooled data of the two classes. The features used for the process consist of 12 MFCCs, zero-crossing rate, and first- and second-order differences. For this thesis, reference segmentation was utilized in lieu of the speech/nonspeech detector. This was done so as not to confound the false alarm error contributions of the overlap handling and diarization systems.

Diarization Algorithm

The agglomerative clustering algorithm initially splits the data into K clusters (where K is chosen to be much greater than the number of true speakers), and then iteratively merges the clusters until a stopping criterion is met. The acoustic data is modeled using an ergodic HMM, where the initial number of states is equal to the initial number of clusters (K). The final number of states corresponds to the hypothesized number of speakers, with each state modeling a distinct speaker. In addition, each state has a set of substates (all of which share the same GMM) that serve to impose a minimum duration on the model. Here, the minimum duration was chosen to be 2.5 s. The overall diarization procedure is as follows:

1. Run front-end acoustic processing.
2. Run speech/nonspeech detection.
3. Extract acoustic features from the data and remove nonspeech frames.
4. Create models for the K initial clusters via linear initialization.

5. Perform several iterations of segmentation and training to refine the initial models.

6. Perform iterative merging and retraining as follows:

   (a) Run a Viterbi decoding to re-segment the data.

   (b) Retrain the models using the Expectation-Maximization (EM) algorithm and the segmentation from step (a).

   (c) Select the cluster pair with the largest merge score (based on ΔBIC) that is > 0.

   (d) If no such pair of clusters is found, stop and output the current clustering.

   (e) Merge the pair of clusters found in step (c). The models for the individual clusters in the pair are replaced by a single, combined model.

   (f) Go to step (a).

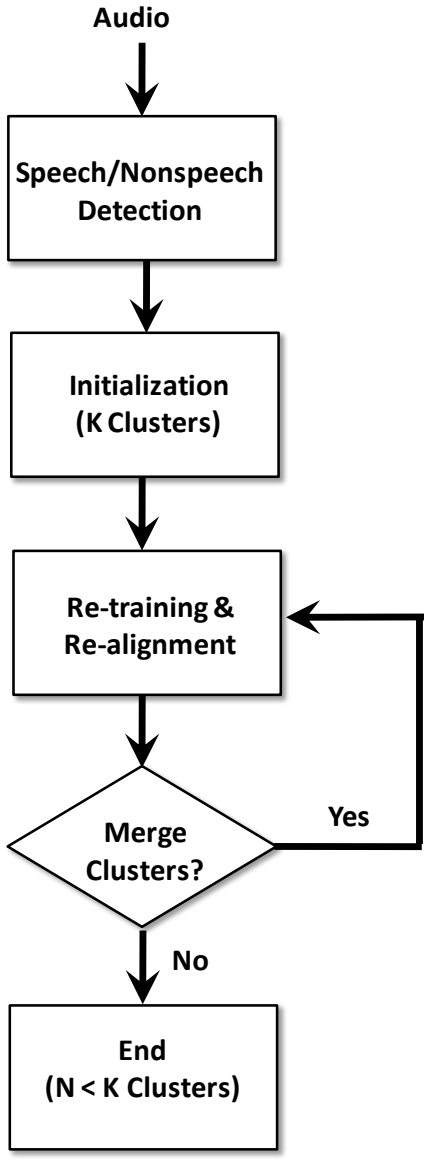


Figure 3.5: Diagram of the ICSI RT-07S meeting diarization system. The system performs iterative clustering and segmentation on detected speech regions starting with a uniform segmentation corresponding to K clusters. The merging decision is based on a modified version of the Bayesian Information Criterion (BIC).

Chapter 4

Multispeaker SAD for Improved ASR

The segmentation of an audio signal into regions of speech and nonspeech is a critical first step in the task of automatic speech recognition. This is especially the case within the context of multispeaker meetings with nearfield recordings obtained using lapel or individual headset microphones. For these meetings, a significant amount of crosstalk (considered nonspeech for the desired recognition task) may be present on the channel, leading to insertion errors produced by the ASR system. This chapter describes the development of a multispeaker SAD system utilizing cross-channel features to address the phenomenon of crosstalk and consequently improve the accuracy of speech recognition.

4.1 System Overview

As mentioned in chapter 3, the key component to performing the audio segmentation is the HMM based segmenter. In this section, an overview of the speech/nonspeech segmenter used for speech activity detection is given.

4.1.1 HMM Architecture

The speech/nonspeech segmenter was derived from an HMM based speech recognition system, namely the SRI DECIPHER recognizer. The system was modified and simplified to consist of only two classes: 1)"local speech" (sp); and 2) "nonspeech" (nsp), here a broad class which includes (but is not limited to) low-level ambient noise (i.e., silence), laughter, breath noise, and, most importantly, crosstalk. Each class is represented with a three-state phone model. State emission probabilities are modeled using a multivariate Gaussian Mixture Model with 256 components and diagonal covariance matrices. The Gaussians for the states in the same class are shared, with separate mixture weights for each state. This mixture tying procedure is quite common and is done for statistical efficiency; i.e., to provide robustness to small amounts of training data [8]. For training, the GMMs are initialized using iterative Gaussian splitting and the Linde-Buzo-Gray (LBG) vector quantization algorithm [63]. This algorithm provides a method of partitioning the feature space into K regions in each of which a Gaussian mean is to be estimated. After these statistics are collected, the Baum-Welch algorithm is performed to re-estimate the GMM parameters and estimate the HMM transition probabilities. For testing, segmentation is carried out by decoding the full IHM channel waveform. The decoding is potentially performed multiple times, with decreasing transition penalty between the speech and nonspeech classes, so as to generate segments that do not exceed 60 seconds in length.

4.1.2 Segmenter Post-processing

In designing a multispeaker SAD system, it is necessary to be mindful of the final application of the system, in this case automatic speech recognition. Specifically, it is important to be aware of the behavior of the ASR system to different types of SAD system errors. If the SAD system has a false alarm error, depending on the

nature of the audio in the false alarm region, it may be possible for the ASR system to recover from this error. Indeed, the ICSI-SRI meeting recognition system includes a "`REJECT`" model to deal with audio regions where no plausible word hypothesis can be generated. This is useful for addressing some forms of nonspeech audio such as coughs, paper shuffling, etc.; but not others, such as crosstalk. If the SAD system, however, has a missed detection error, it is not possible for the ASR system to recover; the system cannot recognize what is not there. Even in the case where an utterance is "clipped" (i.e., has the beginning or end speech cut off), the recognition could be severely impacted since the decoding of one word influences the decoding of other words in its vicinity through language modeling. In light of this, the speech segments output by the SAD system were post-processed to make them more suitable for the ASR system. To mitigate the effect of "clipped" segments as described above, the segments were padded on both ends by a fixed amount. Similarly, adjacent speech segments having a separation by nonspeech smaller than a given threshold were merged to "smooth" the segmentation. The merged segments were limited to a maximum duration of 60s.

4.1.3 Parameter Tuning

As with most systems, the speech/nonspeech segmenter has a number of parameters which need to be tuned to optimize performance. The segment padding amount and merging threshold are two such parameters. Another is the language model weight, which scales the language model log-probabilities. These parameters were optimized using a grid search technique on held-out data. Because of the high time and computational costs associated with performing recognition, coupled with the number of evaluations necessary to estimate optimal parameters, minimum speech diarization error rate (SDER) was used as the optimization criterion. As mentioned

in Section 3.3, this metric tends to correlate highly with WER, and thus serves as a reasonable substitute metric.

4.2 Candidate Features

As with any classifier, the selection of the appropriate features for this HMM based audio segmenter is of critical importance to its intended task. In this section, the candidate features for inclusion in the system are presented and analyzed.

4.2.1 Fixing Component Length

In this thesis, a supervised approach to SAD has been adopted which involves training on several meetings and testing on others. As should be expected, these meetings vary in number of participants and, consequently, the number of channels of audio data. This presents a potential issue given the desire to use cross-channel features to improve performance under crosstalk conditions. How do we produce a parameterization of this cross-channel data with fixed component length—a necessity for the Gaussian mixture modeling—given a variable number of channels per meeting?

One possible approach is to determine the feature vector length based on the meeting in the training and test data with the fewest channels. If, for example, the fewest channels in a collection of meetings is four, then three cross-channel computations can be performed for each reference channel in the collection. This approach raises a number of issues, however. First is the ordering of the features; it is not clear which channel (in addition to the reference) should be chosen for computing the first component, the second component, and so on. One choice would be a random or arbitrary ordering per meeting. Another possibility would be to do a sorted ordering (ascending or descending) by feature value per frame. A second issue which arises

is the selection of channels in meetings with greater than this minimum number of channels. Again, simple random selection could be applied. Perhaps a better method would be to select channels based on an estimated signal-to-noise ratio. This approach was used by Anguera et al. in [5] to select the reference channel for TDOA feature calculation in their farfield microphone beamforming algorithm. Better still would be to compute features for all channels and select the channels based on some form of mutual information. This is, however, significantly more computationally intensive than the other approaches. In each case, though, it seems as if information is lost as channels are completely discarded. A third issue is that the approach is potentially limited in dealing with new or unseen data. It is reasonable to expect that in some cases the number of channels of the test meetings is unavailable prior to testing. If a test meeting had fewer channels than any training meeting, proper feature generation and, consequently, speech segmentation could not be performed. In a similar case, if new training data became available which again had fewer channels than any previous training meeting, features would potentially have to be recomputed for *all* of the training data—clearly an undesirable undertaking. The only way to prevent either situation would be to use the absolute minimum number of channels—two channels and, hence, one feature vector component—for all meetings. This is clearly suboptimal, however, in that as many as 10 channels of audio data would simply be discarded.

Another approach would be to use per frame statistics—order statistics such as minimum and maximum, as well as mean, variance, etc.—of the features to fix the vector length, as proposed by Wrigley et al. in [119]. This approach addresses the ordering and channel selection issues raised by the previous one. In addition, since these statistics (most notably, the mean) are a function of all channels, more information is preserved. There is one issue, with this approach, however, and it relates to the small number of values over which these statistics are calculated. Some meetings have as few as three channels and thus two values from which frame-level statistics are

computed. This limits us to the simplest of statistics—minimum, maximum, mean, range, and variance—all of which are highly sensitive to outliers, and some of which are quite unreliable given such few samples. Nevertheless, using order statistics presents significantly fewer complications than the previously described approach and so was adopted for this work. Specifically, the statistics stated above (with the exception of variance) are examined here.

4.2.2 Cepstral Features (MFCCs)

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 differences. 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, and 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. These features were computed over a window of 25 ms advanced by 20 ms and cepstral mean subtraction (CMS) was performed as a waveform-level normalization.

4.2.3 Normalized Maximum Cross-Correlation (NMXC)

Cross-channel correlation is a clear first choice for a feature to address crosstalk and this is evidenced by its prevalence in the literature (e.g., [118],[119],[90],[89], and [54]). A crosstalk speech signal on the channel of a speaker who is not talking is likely to have a high correlation with the local speech signal on the channel of the active speaker. Thus, it is possible to identify crosstalk regions by observing the maximum cross-correlation of two nearfield microphone channels within a small window. Maximum cross-correlation alone, however, is insufficient for two reasons. First, the maximum cross-correlation value between signals on two channels i and j is the same regardless of which is the reference channel. As a result, one cannot distinguish between the local speech channel and the crosstalk channel during instances of crosstalk speech. Indeed, in [118], cross-correlation was effective in identifying speech plus crosstalk (i.e., overlapped speech), but much less so in identifying either speech or crosstalk individually. Second, differences in microphone gains may cause maximum cross-correlation values between different microphone pairs to be less comparable to one another. To address these issues, the cross-correlation is typically normalized by dividing by the frame-level energy of the target channel [119], the non-target channel [54], or the square root of each [119]. For this thesis, non-target normalization was used, as initial experiments indicated superior performance to the other two. Thus, we define the normalized maximum cross-correlation between a target (i.e., reference) channel i and non-target channel j to be

$$\Gamma_{ij} = \frac{\max_\tau \phi_{ij}(\tau)}{\phi_{jj}(0)} \tag{4.1}$$

where $\phi_{ij}(\tau)$ represents the cross-correlation at lag τ and $\phi_{jj}(0)$ is the non-target channel autocorrelation for lag 0 (i.e., its short-time energy). Cross-correlation and autocorrelation values were computed over a context window of 50 ms using a Hamming

window function with an advance of 20 ms.

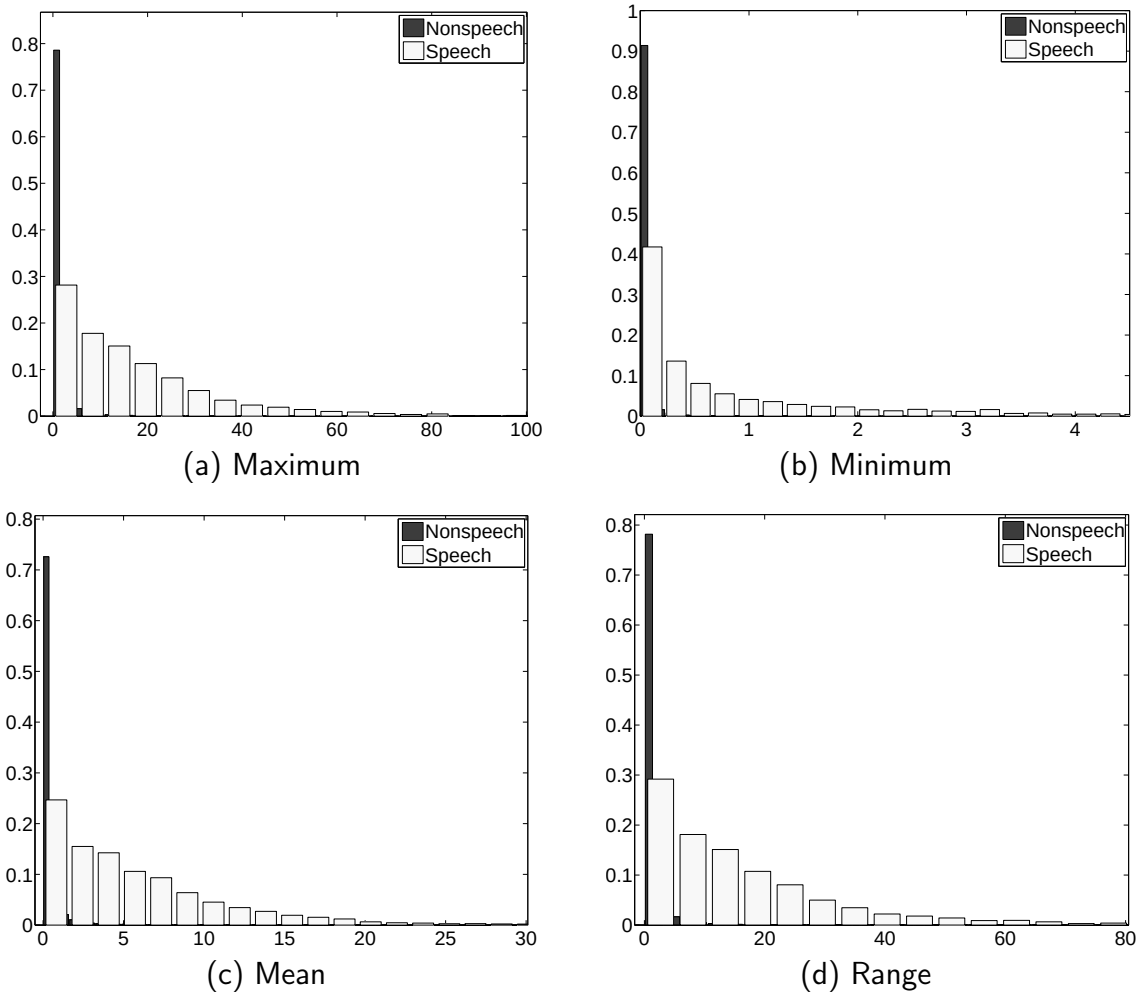


Figure 4.1: Normalized histograms of the (a) maximum, (b) minimum, (c) mean, and (d) range of the normalized maximum cross-correlation (NMXC) in meeting Bdb001.

Figure 4.1 shows the normalized histograms of the maximum, minimum, mean, and range of the NMXC feature for the two classes of (local) speech and nonspeech in a selected meeting, `Bdb001`, from the ICSI meeting corpus [40]. The first thing to note is that the speech class histogram appears to have an exponential distribution for each of the four statistics, with the maximum number of values near zero and

an exponential drop-off. The values for the nonspeech class, in contrast, are *very* concentrated near zero with many fewer beyond this region. To illustrate, the median nonspeech class value for each of the statistics is 0.57, 0.01, 0.21, and 0.55, respectively; compare this with 12.65, 0.34, 4.37, and 11.63, respectively, for speech. The two class distributions, at first blush, do not seem to coincide with initial expectations; crosstalk, considered here to be nonspeech, was hypothesized to have high correlation. In fact, the reason for the significant difference in the two class distributions comes from the energy normalization. In the case where the target channel has crosstalk and the non-target channel has local speech, the feature value becomes small because of the significantly larger energy content on the non-target channel. When the situation is reversed, the lower energy content of the non-target channel produces a large feature value. The result is two class distributions which are quite distinguishable; applying a simple threshold, for example, would produce high classification accuracy on this data.

4.2.4 Log-Energy Difference (LED)

In the previous section, we saw that the utility of the NMXC feature came from the channel normalization. The difference in relative channel energies produced features of very different values for the two classes. It seems reasonable, then, to directly use this relative channel energy as a feature, which is what is done with the next candidate feature—the log-energy difference (LED).

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. The

log-energy difference, examined 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,

$$D_{ij}(t) \;=\; E_i(t) - E_j(t) \tag{4.2}$$

where E represents short-time log-energy. As with the baseline features, the short-time energy is computed over a window of 25 ms with an advance of 20 ms.

Liu and Kubala introduced log-energy difference for segmentation in [64], but made no comparison of the feature's performance to cross-channel correlation or any other feature. Similarly, inspired by the idea of using relative channel energy, Dines et al. [25] proposed a cross-meeting normalized energy feature which compares the target channel energy to the sum of the energy of all channels as mentioned in 2.1.1. Energy ratios have also been shown to complement TDOA location features for improved speaker diarization [82].

Figure 4.2 shows the normalized histograms of the maximum, minimum, mean, and range of the LED feature for the two classes of (local) speech and nonspeech in the same meeting as Section 4.2.3. First of note is that the speech class for each statistic is unimodal. This would also be the case for the nonspeech class, but a second peak is visible for each of the maximum, minimum (only very slightly), and mean statistics, creating a bimodal distribution. The multiple peaks are due to differences in gain from the various personal microphones; if one microphone has a smaller gain than another, the difference in log-energy will be greater for the first than the second. It is possible to address this by subtracting the minimum frame log-energy of the channel from all values for that channel as in [10]. This minimum 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. The approach was not adopted here, however, as initial experiments showed no signs of significant improvement. A

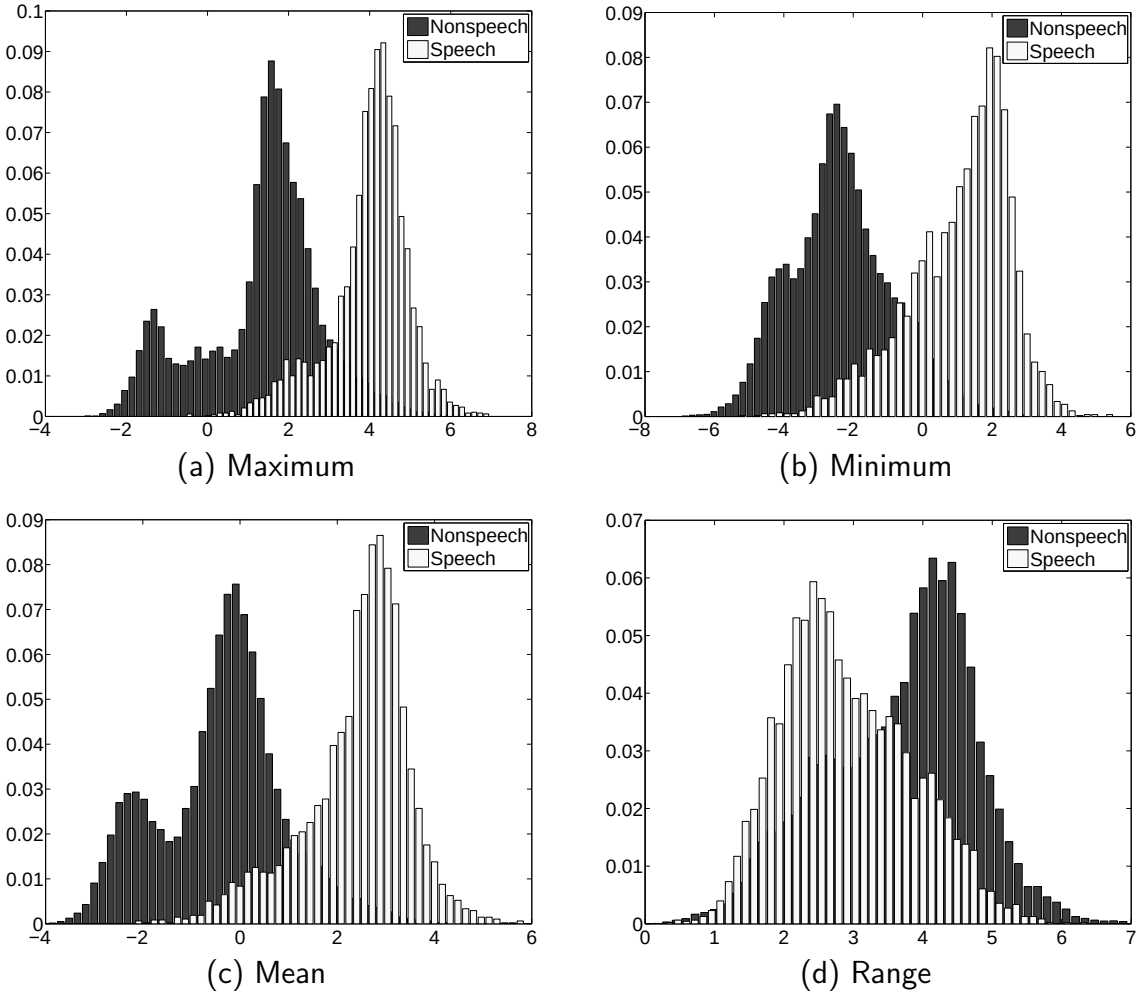


Figure 4.2: Normalized histograms of the (a) maximum, (b) minimum, (c) mean, and (d) range of the log-energy difference (LED) in meeting Bdb001.

possible reason for this is the use of additional statistics ([10] used only maximum and minimum). Another is better optimization of tuning parameters (the parameters were not re-optimized for the LED feature in the other work). At any rate, even though some of the distributions are bimodal, the classes should still be modeled without difficulty since a GMM is utilized.

Also of note is the separation between the two classes. The maximum log-energy difference shows the smallest overlap in distributions, followed by the mean, the minimum, and the range. The distance between the two class means, however, gives a different ranking. The minimum statistic gives the largest distance (3.4), followed by the mean (2.8), the maximum (2.6), and lastly the range (0.8). In general, the separation between the classes for maximum, minimum, and mean are similar while for the range it is much smaller. Nevertheless, each statistic appears capable of yielding good classification accuracy. The theoretical threshold of zero, in addition, would only be a reasonable choice for the minimum log-energy difference.

4.2.5 Time Difference of Arrival Values (TDOA)

For a collection of personal microphones, let $d_{ij}(t)$ correspond to the time difference of arrival (TDOA) of a speech signal $s(t)$ between microphones i and j computed at time t. If $s(t)$ corresponds to the speech of the individual associated with microphone i, then it is clear that

$$d_{ij}(t) > 0 \qquad \forall\ j \neq i \tag{4.3}$$

That is, the speech of the individual should arrive at his or her personal microphone before any other microphone. This relationship suggests that it may be possible to distinguish crosstalk speech from local speech using such TDOA values.

TDOA values are the fundamental feature of most speaker localization algorithms (e.g., [15], [103], and [38]). The TDOA values are also a key component to delay-and-sum beamforming, a signal enhancement technique and pre-processing step for both automatic speech recognition and speaker diarization in meetings. In the multiple distant microphone (MDM) condition, the various signals are combined using delay-and-sum beamforming to produce an enhanced audio signal with a higher signal-to-noise ratio than any single channel. In addition, the work of Ellis and Liu [28] first

demonstrated that TDOA values can serve as useful features for speaker turn detection, a component of speaker diarization. Later work by Pardo et al. improved upon these results by using the features in a full MDM diarization system (segmentation and clustering) in isolation ([87]) and in combination with cepstral features ([88] and [86]). Otterson in [84] later revealed that these TDOA features also improve speaker clustering for diarization in the presence of overlapped speech.

For this work, the TDOA values were computed using the generalized cross-correlation phase transform method (GCC-PHAT). For two channels i, and j, the generalized cross-correlation is given by

$$\hat{G}(f) \;=\; X_i(f)X_j^*(f)\Psi_{ij}(f) \tag{4.4}$$

Where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of the two signals and $\Psi_{ij}(f)$ is a weighting function. The estimated delay is then computed as

$$\hat{d}(i,j) \;=\; \underset{\tau}{\operatorname{argmax}}(\hat{R}(\tau)) \tag{4.5}$$

Where $\hat{R}(\tau)$ is the inverse Fourier transform of Equation 4.4. The GCC-PHAT, as presented by Knapp and Carter in [46], represents a version of the generalized cross-correlation in which the weighting function is given as

$$\Psi_{ij}^{\mathrm{PHAT}}(f) \;=\; \frac{1}{|X_i(f)X_j^*(f)|} \tag{4.6}$$

That is, the cross-spectrum is whitened prior to computing the cross-correlation. With this particular weighting, the cross-correlation is computed using only phase information, hence the name *phase transform*. The GCC-PHAT method is a popular technique for source localization in reverberant environments, as motivated by Brandstein and Silverman in [14].

For a given meeting, one nearfield microphone channel is arbitrarily selected as a reference and TDOA values are computed between this channel and all other nearfield channels. The TDOA values for a given non-reference channel are then derived by identifying the estimated delay between that channel and the reference, and then computing the delay between the other channels based on the known delays with the reference. This drastically reduces computational complexity, as only $N-1$ delays need to be computed rather than $\binom{N}{2}$. In addition, to improve the quality of the estimates, a procedure based on the work in [6] and [4] was adopted. For a given channel, the N-best TDOA candidates (for this work $N = 4$) between the target and reference channels are selected for every frame and a 1-best sequence of delays is obtained using Viterbi decoding. For the Viterbi algorithm, the states are taken as the N possible delay values at each time step and the emission probabilities as the GCC-PHAT value for each delay. The transition probability between two states i and j is defined as

$$a_{ij}(t) \;=\; \frac{\text{max\_diff}(i,j) - |\text{TDOA}_i(t) - \text{TDOA}_j(t-1)|}{\text{max\_diff}(i,j)} \tag{4.7}$$

where $\text{max\_diff}(i,j) = \max_{i,j}(|TDOA_i(t) - TDOA_j(t-1)|)$. This formulation produces values only between 0 and 1, assigning a probability of 0 to the most distant delay pair. GCC-PHAT values are computed over a window of 500 ms (similar to [88]) with an advance of 20 ms.

Figure 4.3 shows the normalized histograms of the maximum, minimum, mean, and range of the TDOA feature for the two classes of (local) speech and nonspeech for the same meeting, `Bdb001`. The class distributions here are quite different from those of the NMXC and LED feature statistics. The speech class distributions are characterized by two distinct and narrow peaks with very low values elsewhere. These peaks correspond to speech activity by the dominant speakers of the meeting. For

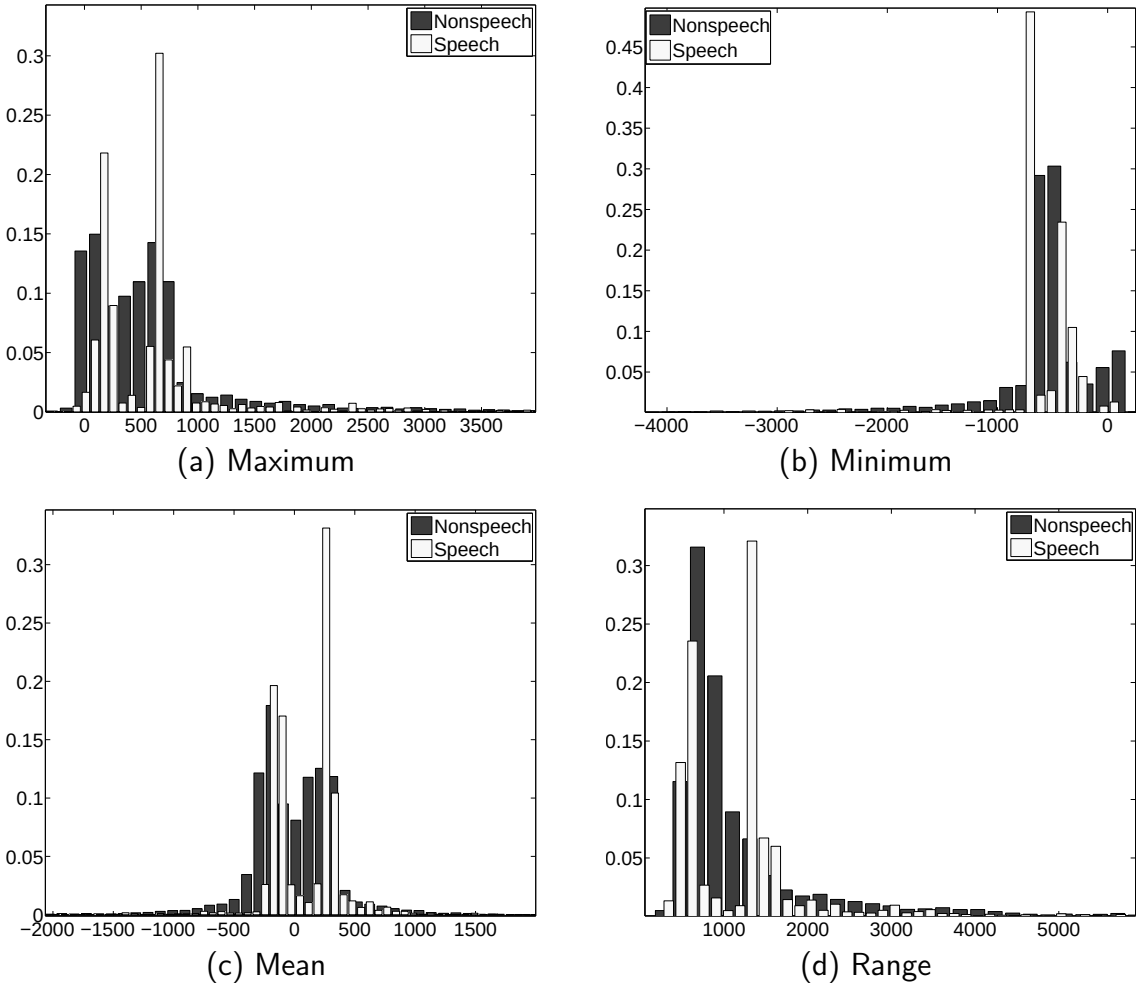


Figure 4.3: Normalized histograms of the (a) maximum, (b) minimum, (c) mean, and (d) range of the time-difference-of-arrival (TDOA) in meeting Bdb001.

nonspeech, there is one main peak for the minimum and range statistics and two small and much less distinct peaks for the maximum and mean statistics. A reason for this spread may be that TDOA estimates are noisier and, consequently, less reliable for nonspeech. Though the distributions in this case defy general categorization, a significant difference between class distributions exists for the statistics, suggesting usefulness for classification.

4.3 Experiments

This section details the experiments associated with evaluating the candidate features mentioned above for use in the multispeaker SAD system. This consisted of

1. Single-feature combination to observe the performance of each feature when combined with the baseline features;

2. Feature selection, in which all feature subset combinations were evaluated and an optimal one selected for the final system; and

3. Evaluation of the final system on unseen test data.

4.3.1 Single-Feature Combination

In this first set of experiments, each cross-channel feature was combined with the baseline cepstral features to determine its ability to improve both SAD and ASR performance as well as to compare the performance of the different features with one another. As a part of this, the combination of the various statistics for the cross-channel features was analyzed, with the assumption that certain statistics may prove more useful for some features than others. The systems in these experiments were evaluated using the 2004 NIST RT evaluation test data ("Eval04" for short) listed in Table 3.1. The speech activity detector was trained on the nearfield audio data consisting of the first ten minutes of 73 meetings from the ICSI meeting corpus and 15 meetings from the NIST Meeting Room pilot corpus [30] for a total of 88 hours of data. The tuning parameters were optimized using held-out data from the AMI meeting corpus. A list of the training and tuning meetings is found in Appendix A.

Results

The results for the various combinations are presented in Table 4.1. The first column gives the feature combination, with "MFCC" being the baseline cepstral features, "NMXC" the normalized maximum cross-correlation, "LED" the log-energy difference, and "TDOA" the time-delay-of-arrival values. The "Reference segmentation" row refers to using manually-obtained time marks obtained for word error scoring. This provides an upper bound for the ASR word error rate performance obtained from the different automatic segmentations. The other two major column divisions contain the speech diarization error rate and ASR word error rate performance results, with a breakdown into the constituent metrics of each. The final column, "R.I.", gives the relative improvement of the multispeaker SAD system in terms of word error rate compared to using the baseline features.

System	SDER			WER				R.I.
	FA	Miss	Total	Subs.	Del.	Ins.	Total	(%)
MFCC	9.51	33.07	43.54	12.8	13.2	3.9	29.9	-
+ NMXC (max,min)	0.66	36.94	37.73	12.3	13.2	1.3	26.8	10.4
+ NMXC (max,min,μ)	0.65	37.08	37.81	12.4	13.1	1.3	26.9	10.0
+ NMXC (max,min,range)	0.85	37.00	37.97	12.3	13.4	1.4	27.1	9.4
+ NMXC (max,min,μ,range)	0.92	37.08	38.14	12.3	13.3	1.4	27.0	9.7
+ LED (max,min)	1.27	35.06	36.71	12.6	12.2	1.8	26.6	11.0
+ LED (max,min,μ)	1.01	35.07	36.44	13.0	12.2	1.6	26.8	10.4
+ LED (max,min,range)	0.96	35.24	36.46	12.4	12.2	1.6	26.2	12.4
+ LED (max,min,μ,range)	0.87	35.09	36.15	12.6	12.0	1.5	26.1	12.7
+ TDOA (max,min)	7.73	34.22	42.69	12.7	12.6	2.3	27.6	7.7
+ TDOA (max,min,μ)	7.48	34.33	42.59	13.0	13.1	2.9	28.9	3.3
+ TDOA (max,min,range)	8.33	33.95	43.03	13.1	12.7	3.2	29.0	3.0
+ TDOA (max,min,μ,rnge)	8.57	33.69	43.12	13.1	12.8	3.0	28.9	3.3
Reference segmentation	-	-	-	12.6	10.5	2.1	25.1	16.1

Table 4.1: Performance comparisons on Eval04 data for single-feature combination systems. The "+" indicates the following feature is concatenated with the baseline MFCCs and the "(...)" indicates which statistics are included in the feature.

First and foremost of note is that each feature combination yields an improvement over the baseline, indicating the benefit of using these cross-channel features. It is also true (but not shown here) that the combination performance exceeds that of each cross-channel feature in isolation; preliminary experiments on similar data, for example, showed a 15% relative improvement for the LED feature when combined with MFCCs. The baseline and cross-channel features, then, are both necessary for good segmentation performance. For the data shown here, the smallest improvement in terms of WER comes from the "(max,min,range)" combination of TDOA values. The WER reduction is from 29.9 to 29.0. Using the MAPSSWE test shows a statistically significant difference for even this case.

So how exactly are the features improving the system? For SDER, miss errors increase in every case, but false alarm errors are significantly reduced. This is most evident with the NMXC feature, for which false alarms are reduced from 9.51% to below 1% for every combination of statistics. The smallest reduction comes from a combination of TDOA value statistics, "(max,min,mean, and range)", with a false alarm rate of 8.57%. Generally speaking, for false alarms, the NMXC feature performs best with the LED feature performing slightly less well and the TDOA features much more poorly than the other two. Looking at WER, the picture is slightly less clear. For substitutions, the NMXC feature yields reductions in all cases and the LED feature in three of the four cases. The TDOA values, however, produce no significant reductions in substitutions. Deletions are reduced significantly for the LED and three of the four TDOA combinations, but not in any of the NMXC combinations. Clear improvements are obtained for insertion errors with all feature combinations, however. Indeed, it was hypothesized in Section 2.1.1 that the cross-channel features would address the crosstalk phenomenon and, consequently, the insertion errors that came about as a result of this crosstalk. The results seem to support this. Generally speaking, for insertions, again the NMXC feature performs best with the LED feature closely behind.

The TDOA feature again performs significantly more poorly than the other two. The trends seen for the false alarms and insertions, however, do not carry over to either DER or WER. The TDOA feature performs most poorly for both DER and WER, and significantly so. The LED feature, though, performs best overall with the NMXC feature performing in between the LED and TDOA features. The reason for this is that the LED feature yields a significant reduction (about 1% absolute) in deletion errors while the NMXC feature does not. This is somewhat unexpected, as the SDER miss errors are not reduced using this feature.

To select the best combination of statistics for each feature, the reduction in WER was used. By this criterion, maximum and minimum provide the best combination for the NMXC feature; maximum, minimum, mean, and range work best for the LED feature; and maximum and minimum again are the best for the TDOA feature. The relative WER improvements are 10.4%, 12.7%, and 7.7%, respectively. Note that the best system goes much of the way towards eliminating the performance gap between the baseline and reference systems.

Metric Correlations

In Section 3.3 it was stated that the SDER of a SAD segmentation has a strong positive correlation with the ASR WER obtained from a recognition run using this segmentation. In this way SDER serves as a reasonable predictor of ASR performance in cases where a full recognition run is not feasible, such as the optimization for parameter tuning mentioned in Section 4.1.3. Figure 4.4 shows a plot of DER-WER pairs for the various feature combinations listed in Table 4.1. The dashed line represents the best-fit line for the data obtained via regression. The plot suggests that there is, indeed, a strong positive correlation between these two metrics, helping to verify the claim. The correlation is not perfect, however, as evidenced by the outlier to the far right. This point, incidentally, represents the best combination for the TDOA feature;

the WER is much lower than should be expected for the given DER. Computing Pearson's linear correlation coefficient on the data gives a value of 0.94 which confirms the strong correlation.

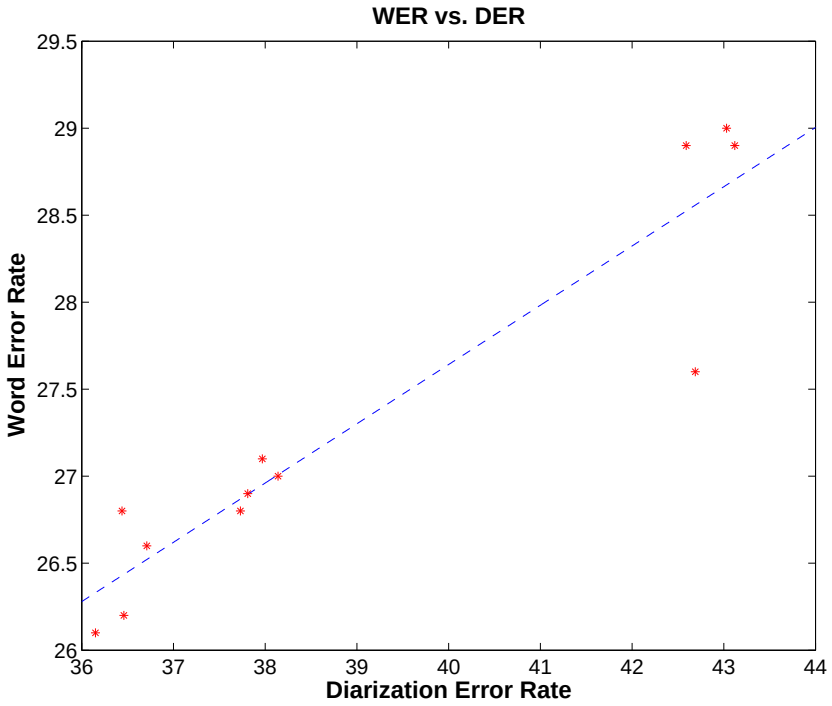


Figure 4.4: Plot of WER versus DER for the systems in Table 4.1. The dashed line represents the best linear fit to the data in the least-squared sense. A strong linear relationship is apparent and is confirmed by the linear correlation coefficient of 0.94.

Given the strong connection between SDER and ASR WER, there likely exists correlations between the constituent metrics of each—namely, false alarms, misses, deletions, insertions, and substitutions. It seems reasonable to presume that false alarm errors in speech activity detection lead to insertion errors for speech recognition, particularly in the case of crosstalk, as has been asserted. Nonspeech regions which are falsely detected as speech—possibly due to phenomena containing significant energy content such as breathing, coughing, laughter, or crosstalk—provide the recognizer with opportunities to hypothesize words that are either nonexistent or not uttered by

the wearer of the personal microphone. This does not always occur, however, since, as previously mentioned, some such phenomena can be addressed by a "REJECT" model, as employed in the ICSI-SRI recognizer. It also seems plausible to presume that missed detections result in deletion errors by the recognizer. Without the opportunity to process the audio within a missed speech region, the recognizer will not hypothesize words within that region and should produce deletion errors. A priori, it is not clear how strongly either of the two speech diarization metrics should correlate with substitutions. With the addition or removal of speech audio regions, whole sequences of hypothesized words may change, resulting in insertions, substitutions, and deletions. As with DER and WER, the correlations of false alarms and misses to insertions,

Diarization Metric	**ASR Metric**	**Correlation Coefficient**
False Alarms	Insertions	0.96
False Alarms	Substitutions	0.76
Misses	Substitutions	-0.87
Misses	Deletions	0.53
DER	WER	0.94

Table 4.2: Correlation between diarization metrics and related ASR metrics.

deletions, and substitutions was computed for the data in Table 4.1. The results are quite interesting. As anticipated, there exists a high positive correlation of 0.96 between false alarm errors in diarization and insertion errors in ASR. This supports the claim that the reduction in insertion errors obtained using the cross-channel features is evidence of crosstalk being addressed by the approach.

There also appears to be a rather strong positive correlation of 0.76 between false alarms and substitutions. This arises because of a particular error pattern, illustrated in the scoring example for meeting CMU_20030109-1600 in Figure 4.5. The figure shows three text columns, corresponding to the error decision, the reference token, and

the ASR hypothesis token, respectively. Along with each tokens is the forced-alignment time for its beginning and end. For this example, the recognizer correctly hypothesizes

```
COR         that's(530.945-531.115)    that's(530.940-531.120)
COR         good(531.115-531.365)      good(531.120-531.370)
INS                                    @reject@(531.450-533.270)
INS                                    yeah(534.780-535.100)
INS                                    @reject@(538.150-539.550)
INS                                    @reject@(539.980-541.450)
SUB         wow(549.610-549.820)       @reject@(551.350-552.310)
SUB         it's(549.820-549.950)      @reject@(552.580-553.010)
SUB         crazy(549.950-550.300)     @reject@(553.650-555.150)
SUB         or(555.250-555.460)        wear(555.250-555.470)
COR         safety(555.460-555.900)    safety(555.470-555.910)
COR         for(555.900-556.030)       for(555.910-556.030)
COR         that(556.030-556.170)      that(556.030-556.170)
COR         matter(556.170-556.560)    matter(556.170-556.510)
```

Figure 4.5: Scoring example from meeting CMU_20030109-1600. The error pattern seen here contributes to the positive correlation between false alarms and substitutions.

the first two tokens, but then proceeds to insert tokens in a nonspeech (i.e., false alarm) region from 531.365 s-549.610 s. Note that three of the four hypothesized tokens in this region correspond to the `REJECT` model; this is common in crosstalk false alarm regions where the crosstalk speech is of lower volume and thus the recognizer cannot properly identify the words. The error generation propagates into speech regions as the `@reject` token continues to be hypothesized. The result is a series of substitution errors following the insertion errors. Eventually, the errors cease and the recognizer resumes correctly identifying the words. The existence of this pattern is further supported by the high correlation between insertions and substitutions (0.85) for this data as well.

As for misses and substitutions, we see that a strong negative correlation of -0.87 exists between the two. This is at least in part explained by the error pattern revealed

in Figure 4.5 combined with the fact that misses generally trade off with false alarms (their correlation is -0.79). As misses increase, false alarms—which play a role in both insertion *and*, it seems, substitutions—are reduced.

Lastly, unlike the one between false alarms and insertions, the correlation between misses and deletions of 0.53 seems surprisingly low. Recall, however, that a reduction in deletion errors was observed for the LED feature relative to the baseline, even though misses actually increased. Indeed, if the LED feature data is removed, the correlation jumps up to a high value of 0.84. Though the low correlation can be explained from a data perspective, the results leave something to be desired as far as intuition is concerned. How can increased misses yield fewer deletions? Unfortunately, error scoring analysis in this case revealed no clear pattern. Comparing the output of systems with lower miss and higher deletion rates to those with higher miss and lower deletion rates showed that deletions for one can become correct tokens for the other, but the mechanism by which this occurs was not evident. An example is shown in Figure 4.6 for meeting `ICSI_20011030-1030`. The first output corresponds to segmentation produced using the baseline cepstral features and the second to segmentation from the maximum, minimum, mean, and range statistics of the LED feature. In the first case, one of the "there's" tokens is absent in the hypothesis, leading to a deletion. Despite no obvious changes in speech activity region, the token is properly recognized in the second case and, furthermore, a substitution in the first case is corrected.

Another interesting area of analysis is the variance relationship between the scores of correlated metrics, presented in Table 4.3. The variance for false alarms is several times larger than for any other metric. The correlated ASR metrics of insertions and deletions, however, are rather small. This implies a relative lack of sensitivity by either to changes in false alarms and, perhaps, a robustness of the ASR system to this type of segmentation error. The use of a "`REJECT`" model, is likely a contributing factor

```
COR         okay(391.450-391.740)        okay(391.300-391.740)
COR         so(392.426-392.686)          so(392.420-392.690)
DEL         there's(392.686-392.936)
DEL         a(392.936-392.976)
COR         there's(393.106-393.296)     there's(392.690-392.940)
DEL         a(392.936-392.976)
COR         there's(393.106-393.296)     there's(392.690-392.940)
SUB         a(393.296-393.326)           this(393.110-393.310)
SUB         little(393.326-393.636)      whole(393.310-393.610)
COR         thing(393.756-393.986)       thing(393.750-393.990)
-------------------------------------------------------------
COR         okay(391.450-391.740)        okay(391.440-391.740)
COR         so(392.426-392.686)          so(392.420-392.690)
COR         there's(392.686-392.936)     there's(392.690-392.940)
DEL         a(392.936-392.976)
COR         there's(393.106-393.296)     there's(393.110-393.310)
DEL         a(392.936-392.976)
COR         there's(393.106-393.296)     there's(393.110-393.310)
COR         a(393.296-393.326)           a(393.310-393.340)
SUB         little(393.326-393.636)      whole(393.340-393.620)
COR         thing(393.756-393.986)       thing(393.750-393.980)
```

Figure 4.6: Scoring example from meeting ICSI_20011030-1030. The deleted token, "there's" in the first case is properly recognized in the second and the substituted "this" is recognized correctly as "a".

here. Relatively speaking, deletions and substitutions appear to be more sensitive to missed speech errors, but for the case of deletions, the weaker correlation with misses obscures the relationship somewhat.

	FA	Ins.	Subs.	Miss	Del.	Subs.
Variance	12.41	0.51	0.10	1.68	0.24	0.10

Table 4.3: Variances of correlated SDER and WER metrics.

4.3.2 Feature Selection

Having compared and contrasted the various statistics combinations for the three cross-channel features, the next step was to observe the three features in combination. As discussed in Section 3.2.3, several methods exist for selecting a good combination of features. Though it was stated that exhaustive search can become intractable in many cases, the small number of features used here made this approach feasible and attractive, since it is guaranteed to find the globally optimal subset of features. For this experiment, the same training, tuning, and evaluation data was used as in Section 4.3.1. Each of the $2^3 = 8$ feature combination subsets was evaluated (including the empty set, which corresponds to the cepstral baseline).

Results

The results for the feature selection experiments are given in Table 4.4 and are presented in the same format as the previous experiments.

System	SDER			WER				R.I.
	FA	Miss	Total	Subs.	Del.	Ins.	Total	(%)
MFCC	9.51	33.07	43.54	12.8	13.2	3.9	29.9	-
+ NMXC+LED+TDOA	2.86	34.06	37.40	12.9	11.7	1.4	25.9	13.4
+ LED+TDOA	0.84	35.02	36.00	12.6	12.1	1.4	26.1	12.7
+ NMXC+LED	3.15	33.47	37.14	12.9	11.5	1.5	25.9	13.4
+ NMXC+TDOA	5.80	34.42	41.02	12.8	12.3	1.8	27.0	9.7
+ NMXC	0.66	36.94	37.73	12.3	13.2	1.3	26.8	10.4
+ LED	0.87	35.09	36.15	12.6	12.0	1.5	26.1	12.7
+ TDOA	7.73	34.22	42.69	12.7	12.6	2.3	27.6	7.7
Reference segmentation	-	-	-	12.6	10.5	2.1	25.1	16.1

Table 4.4: Performance comparisons on Eval04 data for systems representing all possible feature combinations. The "+" indicates the following feature is concatenated with the baseline MFCCs .

Here, too, we first note that all feature combinations improve over the baseline performance. The smallest relative WER improvement is 7.7% and is obtained using only the TDOA feature with the MFCCs. The MAPSSWE test once again determines this difference to be significant. Generally speaking, the same improvement trends emerge for the multiple feature combination systems as they did for single-feature combination. In all cases, false alarm errors are significantly reduced with the additional features while missed speech errors remain at the same level or increase slightly. In terms of WER, the most significant improvements appear in the insertion errors followed by deletions and lastly substitutions, which change very little. The trend of improved deletions was less consistent in the single-feature combinations, where the various NMXC statistics combinations produced no significant improvement, the TDOA value statistics a small improvement, and the LED feature the largest one. The seemingly curious phenomenon of reductions in deletions accompanied by increases in misses, however, does emerge here, unfortunately with no additional clues to the cause.

It is noteworthy that here, too, the largest reductions in deletion errors involve feature combinations including the LED feature. Following this logic, one might anticipate the best improvements in insertions to come from combinations involving the NMXC feature, as it consistently yielded the largest reductions in the previous experiments. This, however, is not categorically true, as the combination of NMXC and TDOA features produces the second-lowest reduction in insertion errors. It is the TDOA feature in single combination with MFCCs, though, that yields the lowest reduction, so it seems plausible that the combination of best and worst feature might produce this poor performance. Looking more closely at the two-feature combinations, we see that the combination of the first- and second-best individual features in terms of WER—the LED and NMXC features—produces the best combination. The combination of the second- and third-best features, similarly, leads to the poorest

performance. In addition, though the WER difference between the first- and second- and second- and third-ranked individual features is about the same, the improvement from combining numbers one and two is much smaller than the degradation from combining numbers two and three. Interestingly, the full three-feature set produces no gains over any subset. The performance essentially matches that of the subset in which the TDOA feature is removed, suggesting that the feature provides no information in this combination, though in single-feature combination can do so. Again using WER improvement as the optimization criterion, the subset of NMXC and LED features was chosen as the feature combination to be used in the final SAD system. Note that this system narrows the performance gap with the reference segmentation to a low 0.8 for this data and obtains a substantial relative WER improvement of 13.4%.

4.3.3 Final System

For the final phase of the adopted experimental paradigm, it was necessary to validate the multispeaker SAD system by evaluating on an independent test set. A modified version of the data from the 2005 NIST RT evaluations, shown in Table 3.1 of Section 3.3.1, was used for this task. The modification consisted of removing one meeting, `NIST_20050412-1303`, which contained a participant on speakerphone who, consequently, had no associated personal microphone channel. This situation would result in a large number of insertion errors triggered by crosstalk. A method to address this using a single distant microphone as a stand-in for the speakerphone participant's personal microphone channel was presented in [10] and was shown to be effective. However, to simplify comparisons, the meeting was excluded for this work. This was also done by Laskowski et al. in [53], who referred to the set as `rt05s_eval*`, for similar purposes.

For this test, the speech activity detector was trained using all of the training data from the previous experiments, plus the first ten minutes of 35 meetings from the AMI meeting corpus (also listed in Appendix A). For optimizing the various tuning parameters, the Eval04 test data was used, with the expectation that the diversity of the meetings in the tuning set would improve generalization to this unseen test data.

Results

The results for the final system are presented in Table 4.5 in the same format as the previous experiments.

System	SDER			WER				R.I.
	FA	Miss	Total	Subs.	Del.	Ins.	Total	(%)
MFCC	8.66	6.71	15.37	10.7	9.7	3.6	24.0	-
+ NMXC+LED	1.44	6.65	8.09	10.9	9.2	1.7	21.8	9.2
Reference segmentation	-	-	-	11.1	6.0	1.8	18.8	21.7

Table 4.5: Performance comparisons on Eval05* data for systems representing the baseline and best feature combinations.

The first thing to note is that the baseline and reference error rates are much lower—24.0 and 18.8, respectively as compared to 29.9 and 25.1, respectively. The performance gap between the two, however, is slightly wider. In terms of SDER, we see that the final system, which includes the NMXC and LED features, produces significant reductions in false alarms, leaving misses essentially unchanged but nearly halving the overall diarization error rate. On the ASR side, this translates into substitutions being relatively unchanged, deletions decreasing slightly (by 0.2% absolute), and insertions dropping significantly (by almost 2% absolute). The final system, then, appears to be highly tuned to reducing insertion errors and thus addressing crosstalk. Both the deletion and insertion rate reductions on the development data were greater, though, and the system achieves a relative WER improvement of 9.2% on this data

as compared to 13.4% on the development data. In addition, the performance gap in this case is 3% as compared to 0.8% previously. This could partly be explained by the reference segmentation being better for this test set than the other, which is possible since the gap between baseline and reference is higher here. It could also partly be explained by the limits to which such systems can generalize; it is rare that performance of a system will not degrade when tested on unseen data.

4.4 Discussion

This chapter presented the work towards addressing the issue of crosstalk for nearfield microphone meeting recognition. Using the framework outlined in Chapter 3, a multispeaker speech activity detection system was developed that incorporated cross-channel features to improve the segmentation of local speech and nonspeech, and subsequently improve recognition performance. The results here showed that all three proposed features–normalized maximum cross-correlation, log-energy difference, and time-difference-of-arrival values improved performance, in particular by reducing the insertion errors typically associated with crosstalk. The best combination of features proved to be the NMXC and LED features, reducing WER by 13.4% relative on the development data and 9.2% relative on the validation set. This compares to a possible relative improvement of 16.1% and 21.7%, respectively, as determined by reference segmentation.

These numbers suggest that there are still possible gains to be made by the segmentation system. Where, exactly, do they lie? One clue can be found by doing a site-level performance analysis, as shown in Table 4.6. Here the baseline, final, and reference WER results are presented by site for the Eval05* test set.

A number of interesting things are revealed by these numbers. First, the AMI and ICSI data give similar performance results for both the baseline and proposed systems

System	WER					
	ALL	AMI	CMU	ICSI	NIST	VT
MFCC	24.0	21.6	22.8	20.7	22.9	32.5
+ NMXC+LED	21.8	21.8	21.7	20.8	20.9	23.5
Reference segmentation	18.8	18.8	19.8	15.7	19.6	20.8

Table 4.6: Site-level performance comparisons on Eval05* data for baseline and best-combination systems.

while a significant gap exists between these and the reference segmentation. For the AMI meetings, the difference is about 3% absolute, while for the ICSI data it is as high as 5%. The reason for this is revealed by looking at the WER breakdown, shown in Figure 4.7. Both systems have similar performance to the reference for insertions and substitutions, but do much more poorly for deletion errors. Indeed the differences in deletions are about the same as for overall WER, indicating that deletions almost completely explain the discrepancy. For CMU, NIST, and VT the proposed system narrows the performance gap by an increasing amount. For CMU and NIST this comes from reduced deletions, as the insertion performance here, too, is similar for the baseline, proposed, and reference. A *huge* reduction in insertion errors, however, is obtained on the VT meetings (from 10.6 to 1.4), vividly demonstrating the utility of the cross-channel features and the proposed system. As in the other cases, the remaining error consists primarily of deletions. This trend is also apparent over all sites, as displayed in Table 4.5.

Given that deletions appear to be the majority of the remaining errors, it would be useful to see if any pattern exists in the occurrence of these errors. To do this, the recognition output obtained using the proposed system segmentation was scored against the output using the reference segmentation and the deleted tokens were identified. Figure 4.8 presents the counts of the most common deletions occurring in the data.

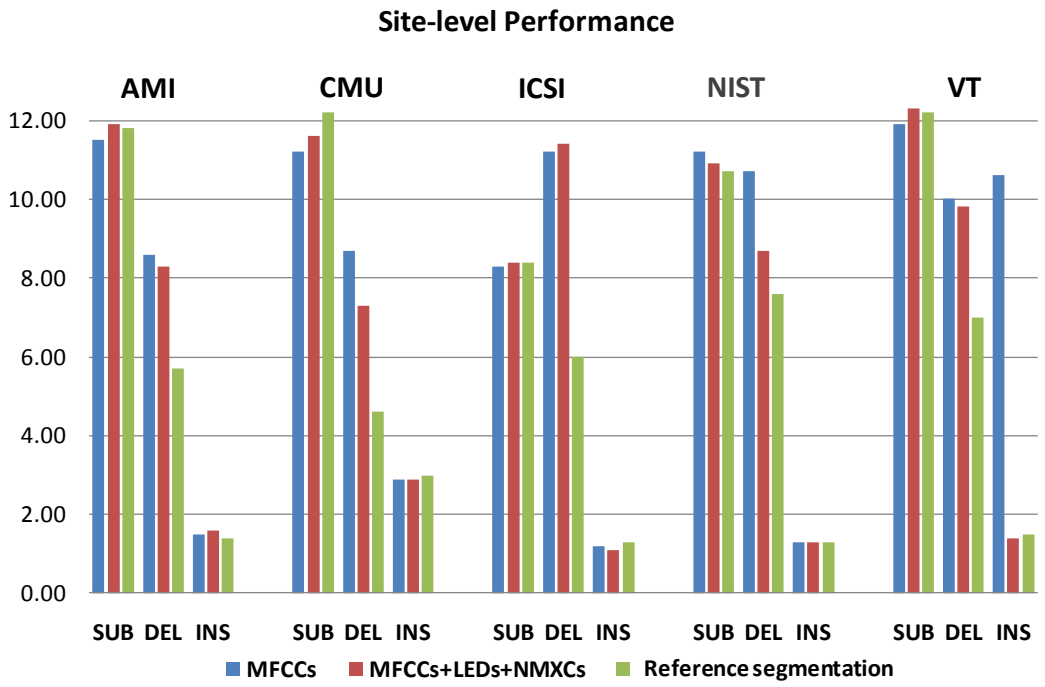


Figure 4.7: Bar graph of site-level performances on Eval05* data for the systems presented in Table 4.6

The graph shows that the top four deleted tokens all fall within the category of backchannels—words that provide feedback to the dominant speaker, indicating that the non-dominant speaker is still engaged in conversation. The majority of the backchannels, in addition, consist of the top two tokens "uhhuh" and "yeah", with a significant drop-off thereafter. Because of their function in conversation (i.e., encouraging another participant to continue speaking), the backchannels are typically uttered at a lower volume than words uttered when the speaker has the floor. Consequently, the audio associated with these tokens has lower energy content. Since the speaker does not have the floor, these words also often appear in isolation and sometimes overlap the speech of the dominant speaker. It is easy to see how such speech regions, then, may go undetected by automatic systems still largely based on energy, such as those described in this chapter. One possible change that may help

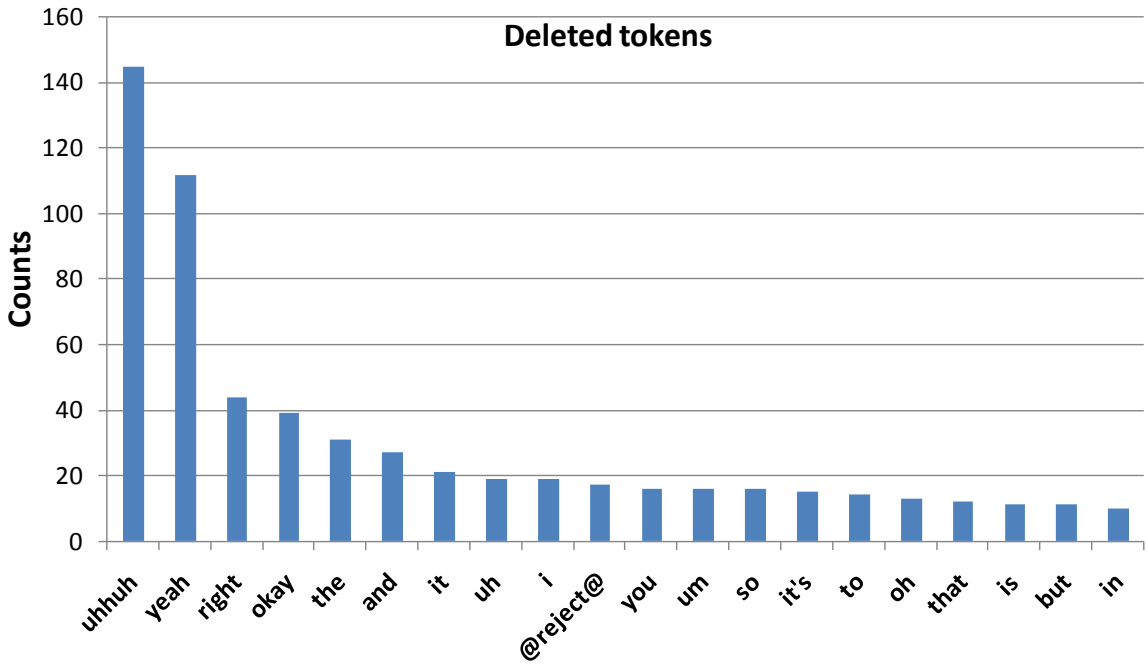


Figure 4.8: Histogram of deleted tokens when scoring ASR output using reference segmentation against output using automatic segmentation.

address the problem is to adjust the segmenter's operating point via the language model weighting to increase the amount of speech detected. This may allow for more short segments to be detected, some of which may be backchannels. The false alarm error will likely increase, but the recognizer, with its "`REJECT`" model, could be robust to this. Another possibility is to improve the modeling of these backchannel speech regions by increasing their number in the training data, either in absolute or relative terms.

Of course, all of this sidesteps the question of the importance of these deleted words. Generally occurring in isolation, these deletions should not affect other word hypotheses greatly. Somewhat related to this, the semantic and syntactic importance of these words is rather low. They also provide no information about meeting content, making them of little use for downstream processing such as summarization or topic

identification. One exception is the word "yeah", not as a backchannel, but as an affirmation. One could imagine a meeting—e.g., for example a budget proposal—where this word could be of great significance—e.g., indicating the budget was, in fact, approved. It is not clear how often this context is encountered, however. Thus, though the deletion errors remaining present a challenge for further system development, these deletions can be considered the "least harmful" for many target applications.

Chapter 5

Overlapped Speech Handling for Improved Speaker Diarization

The presence of overlapped, or co-channel, speech in meetings is a common occurrence and a natural consequence of the spontaneous multiparty conversations which arise within these meetings. This speech, in addition, presents a significant challenge to automatic systems that process audio data from meetings, such as speaker diarization systems. Specifically, in regions where more than one speaker is active, missed speech errors will be incurred and, given the high performance of some state-of-the-art systems, this can be a substantial portion of the overall diarization error. Furthermore, because overlap segments contain speech from multiple speakers, including them in any one speaker model may adversely affect the quality of the models, which potentially reduces diarization performance. This chapter describes the development of a monaural overlapped speech detection, labeling, and exclusion system which seeks to improve diarization performance by addressing these issues.

5.1 System Overview

As with multispeaker SAD, the HMM based segmenter is a major component to the overlapped speech handling system. This section gives an overview of this overlapped speech segmenter, highlighting key differences with the speech/nonspeech segmenter of Chapter 4. The pre-processing overlap exclusion for speaker clustering and post-processing segment labeling—the other major components of this system—are also described.

5.1.1 HMM Architecture

The overlap detector is an HMM based segmenter consisting of three classes—"speech" (sp), "nonspeech" (nsp), and "overlapped speech" (olap). As with the multispeaker SAD system, each class is represented with a three-state model and emission probabilities are modeled using a 256-component multivariate GMM with diagonal covariances. For each class HMM, mixtures are shared between the three states, with separate mixture weights for each state. The models are trained using an iterative Gaussian splitting technique with successive re-estimation. In this case, the mixture splitting proceeds by copying the mixture component with the largest weight and perturbing the two mixtures by plus or minus 0.2 standard deviations. Once the final number of mixtures is reached, a single Baum-Welch iteration is performed for GMM parameter re-estimation and transition probability estimation. In contrast to the multispeaker SAD system, decoding consists of a single Viterbi pass of the SDM channel waveform.

Rather than use a language model for class transitions, a word network was used. This word network is a higher-level finite state machine that models class transitions via a network of arcs (the set of possible transitions) and associated probabilities (the transition probabilities). This method is useful for representing simple grammars such as the three-class "grammar" of the overlap detector. In particular, since the network of

arcs is explicitly specified, it is possible to prohibit certain transitions. For overlapped speech detection, it is almost never the case that two speakers spontaneously and simultaneously overlap completely. Rather, one speaker will begin to speak and the other will overlap briefly to interrupt (as with a backchannel or floor-grabber) or the two speakers will overlap as the floor passes from one speaker to another. In either case, overlapped speech is preceded and followed by speech, not nonspeech. This constraint is imposed by the word network used, shown in Figure 5.1.

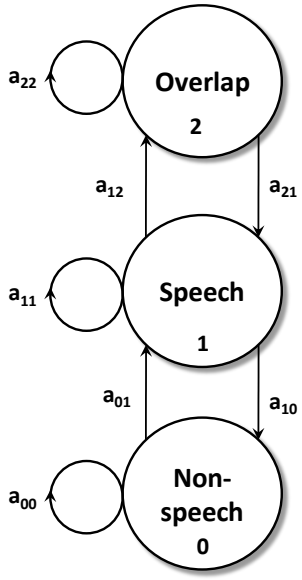


Figure 5.1: Finite state machine representing the HMM word network. The transition between speech and overlap, a_{12}, is the sole tuning parameter of the system.

5.1.2 Parameter Tuning

Whereas the nearfield speech/nonspeech segmenter had three tunable parameters, the overlapped speech segmenter has only one: the transition penalty between the speech and overlapped speech classes. This single parameter allows for tuning (though somewhat coarsely) the trade-off of misses versus false alarms—or recall versus precision—for

overlapped speech detection. As mentioned in Section 3.3, a high precision operating point is what is desired, since false alarms from the detector increase the baseline diarization error rate, whereas misses have zero effect. This is the reason, too, why the segment padding and merging post-processing performed for the speech/nonspeech segmenter are not performed (and thus not tuned) here. Though high precision is desirable, ultimately improved DER performance of the diarization engine is the goal, and this was the optimization criterion for tuning, with held-out data being used as in Chapter 4.

5.1.3 Overlap Speaker Labeling

To apply the segment information obtained from the overlap detector to the diarization system, the following procedure is employed. For each frame y, the diarization system produces speaker likelihoods, $p(y|C_k)$, based on each cluster model C_k. Speaker cluster posteriors are then computed according to

$$\begin{aligned} p(C_k|y) &= \frac{p(y|C_k)p(C_k)}{p(y)} \\ &= \frac{p(y|C_k)p(C_k)}{\sum_k p(y|C_k)p(C_k)} \end{aligned} \tag{5.1}$$

where $p(C_k)$ is calculated as the proportion of data frames assigned to cluster C_k. By summing the speaker posteriors over the frames of the identified overlap segment, a single "score" for each speaker is obtained. Typically, the diarization system will have assigned the segment to the speaker with the highest score, in which case the speaker with the second highest score is chosen as the other speaker. In the event that the system has chosen another speaker, then this highest-scoring speaker is selected as the additional speaker. Note that this procedure limits the number of possible overlapping speakers to two, but for the corpora of interest two-speaker overlap typically comprises

80% or more of the instances of overlapped speech [109].

5.1.4 Overlap Exclusion

In addition to using overlap segment information in a post-processing procedure as described above, this information was utilized in a pre-processing step in which these segments were excluded from the speaker clustering process of the diarization system. The expectation was that this would result in purer speaker clusters and thus improve the diarization system performance by reducing speaker error. Because the speaker label assignment in the post-processing step utilizes speaker posteriors—which may improve as a result of the purer clusters—it was hypothesized that this procedure would benefit from the pre-processing as well.

5.2 Candidate Features

Based on related work (detailed in Section 2.2.1) as well as previous work (see [12] and [11]), several features were identified as candidates for use in the segmenter. These include cepstral features, RMS energy, zero-crossing rate, kurtosis, LPC residual energy, spectral flatness, and modulation spectrogram features. Each feature is presented and discussed in this section, with a focus on providing motivation for the feature as well as analyzing potential performance.

5.2.1 Cepstral Features (MFCCs)

As with the speech/nonspeech segmenter, the baseline features for the overlapped speech segmenter were derived from Mel-frequency cepstra, specifically 12th-order MFCCs along with first differences. 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. Recall, for example, from Section 2.2.1 that Zissman et al. [125] distinguished between single-speaker and overlapped speech using such features. Many significant differences exist between the experimental setups, however, and so different performance results were anticipated. The procedure for the computing the MFCCs was the same as in Section 4.2.2, though here they were computed using a Hamming window of 60 ms with an advance of 10 ms.

5.2.2 RMS Energy

The energy content of a speech segment will likely be affected by the presence of additional speakers. At the most basic level, the superposition of speech from two speakers will produce a signal with higher energy than that of either individual speaker. This effect is enhanced, however, by the nature of conversational dynamics: in many cases one or more of the overlapping speakers in a meeting will speak more loudly to be heard and understood by the other participants. This is most true for the floor-grabbing scenario, where the interjecting speaker tries to establish dominance; in the case of backchannels, this is less likely to occur. For this work, the short-time root-mean-squared energy was used. This is computed according to

$$E_{RMS} = \sqrt{\frac{\sum_{i=0}^{N-1} x[i]^2}{N}} \tag{5.2}$$

Energy was computed using a Hamming window of 60 ms with an advance of 10 ms. Also, to compensate for potential channel gain differences, signal waveforms were normalized based on overall RMS channel energy estimates prior to energy computation.

Figure 5.2 shows a plot of the normalized histogram of the RMS energy feature for the three classes of nonspeech, speech, and overlapped speech using the mixed-headset audio in a selected meeting, `IS1004c`, from the AMI meeting corpus [18].

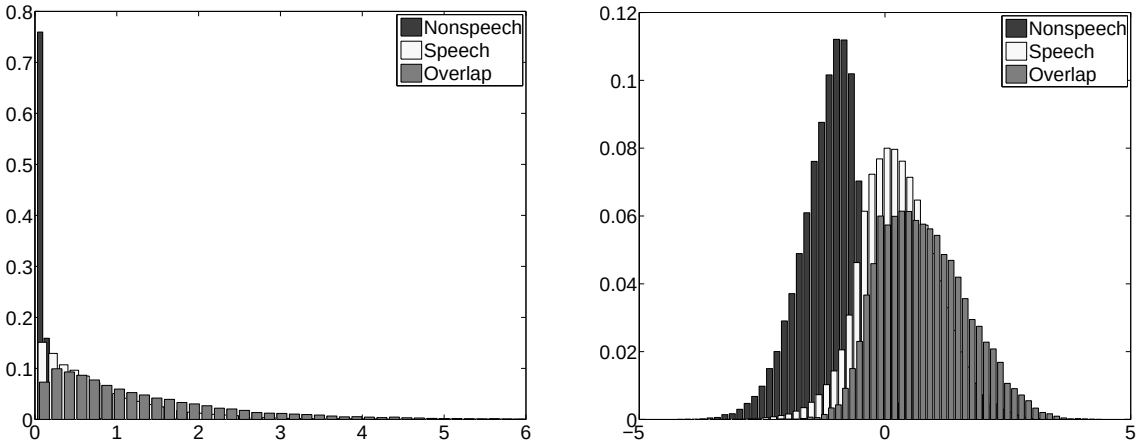


Figure 5.2: Normalized histograms of the (a) raw and (b) Gaussianized RMS energy for meeting IS1004c.

Figure 5.2 (a) shows the raw feature distribution while 5.2 (b) shows the distribution after the Gaussianizing feature transformation. The raw features appear to have an exponential distribution for each of the three classes, with the nonspeech class being most concentrated toward zero, followed by the speech and then the overlap class; this seems to confirm the hypothesis about overlapped speech energy content above. After Gaussianization, each of these features resembles a normal distribution, demonstrating the effectiveness of the technique. In addition, the positional relationship of the three classes is much more clear after the transformation. For example, we see that the separation between the speech and overlapped classes is much smaller than either with the nonspeech class.

Given that the transformation effectively produces Gaussian distributions, a useful measure of separation between the classes (in particular for feature comparison) is the symmetric Kullback-Leibler distance (KL-2). The KL-2 distance is a symmetrized version of the Kullback-Leibler (KL) divergence [43], which measures the difference between two probability distributions P and Q. For a continuous random variable,

the KL divergence is given as

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{5.3}$$

The KL-2 distance Δ_{KL} is obtained by summing the divergence between P and Q with the divergence between Q and P; that is, $\Delta_{KL} = D_{KL}(P||Q) + D_{KL}(Q||P)$. Note that, though symmetric, the KL-2 distance does not obey the triangle inequality and as such is not a true distance. For the data shown here, the minimum KL-2 distance is between the speech and overlap classes and has value 0.386. Compare this to a KL-2 distance of 4.376 between speech and nonspeech and 7.44 between overlap and nonspeech.

5.2.3 Zero-Crossing Rate

The zero-crossing rate (ZCR)—the rate of sign changes along a signal—is commonly used in speech processing for audio classification. Some common classification tasks include voiced/unvoiced classification [2], endpoint detection [94], and speech/music discrimination [108]. The periodicity of voiced speech and music tends to produce a lower zero-crossing rate than unvoiced speech (e.g., fricatives) and background noise, making the value useful for such tasks. In the case of simultaneously voiced speech from overlapping speakers, it is plausible that the superposition of the speech signals, in addition to increasing energy, will increase the zero-crossing rate. The value may then be of use for the overlapped speech detection task. The feature was computed according to

$$ZCR = \frac{1}{2N} \sum_{i=1}^{N-1} |\operatorname{sgn}(x[i]) - \operatorname{sgn}(x[i-1])| \tag{5.4}$$

where N is the number of samples in the applied Hamming window, here set to 50 ms.

Figure 5.3 shows the normalized histograms of the zero-crossing rate for the same

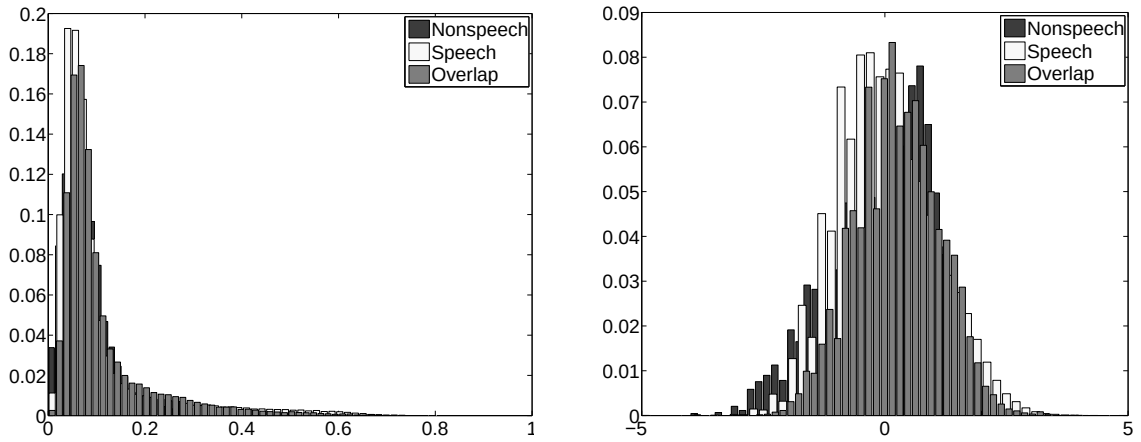


Figure 5.3: Normalized histograms of the (a) raw and (b) Gaussianized zero-crossing rate for meeting IS1004c.

meeting as before. The distributions of the raw features exhibit a strong positive skew, with the majority of values between 0 and 0.2 and the tail extending to about 0.8. In addition, there appears to be very little difference in the positions of the three class distributions. On the right, the Gaussianization succeeds in making the distributions more symmetric and centered about zero. This plot also reveals a slight difference in distribution positions with nonspeech the farthest to the left followed closely by speech which is in turn followed closely by overlap. The difference, however, does not appear significant enough to confirm the hypothesis. In terms of KL-2, the smallest distance here is between speech and nonspeech (0.014), followed by speech and overlap (0.126) and lastly nonspeech and overlap (0.216). Note that these results differ by an order of magnitude with those of the RMS energy feature.

5.2.4 Kurtosis

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

$$\kappa_x = \frac{E[x^4]}{[E[x^2]]^2} - 3 \tag{5.5}$$

where $E[\cdot]$ is the expectation operator. This serves as a measure of the "Gaussianity" of a random variable, with super-Gaussian, or leptokurotic, random variables having kurtosis greater than zero and sub-Gaussian, or platykurtotic, 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 [58] and Krishnamachari et al. in [50]. As such, signal kurtosis 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. For this work, signal kurtosis was computed using a Hamming window of 50 ms with an advance of 10 ms. The normalized histograms of the kurtosis feature for the three audio classes are presented in Figure 5.4. As with the zero-crossing rate, the raw kurtosis distributions exhibit a strong positive skew. Here, too, the distributions are almost completely overlapping. A great distinction between the classes, however, is apparent upon Gaussianizing the data. The nonspeech class separates from the speech and overlap classes with KL-2 distances of 0.764 and 1.06, respectively. The speech and overlap classes, however, exhibit little separation, as is confirmed by their KL-2 distance of 0.05. Generally speaking, these distances are larger than those of the zero-crossing rate, but much smaller than those of the RMS energy.

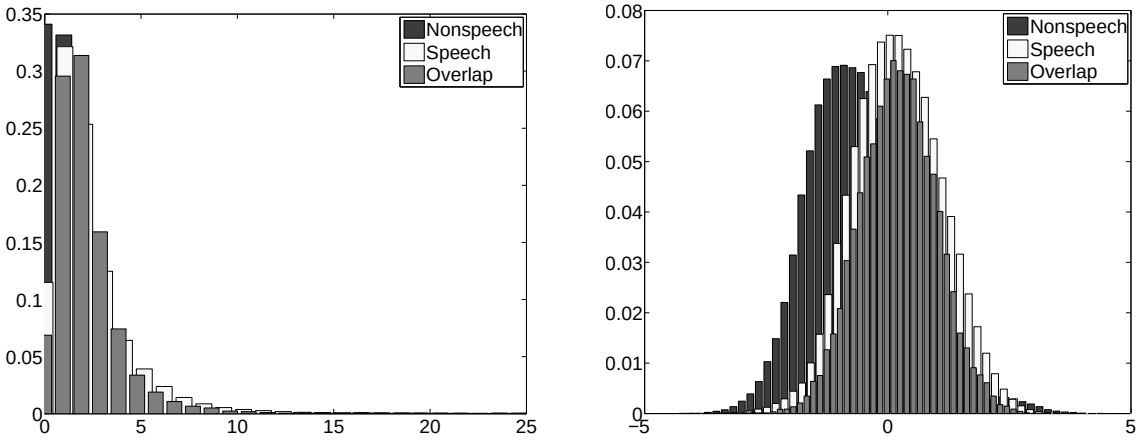


Figure 5.4: Normalized histograms of the (a) raw and (b) Gaussianized kurtosis for meeting IS1004c.

5.2.5 LPC Residual Energy

Linear predictive coding (LPC) analysis is an important speech processing technique used in many applications such as speech coding [36], speech activity detection [80], as well as speech and speaker recognition. LPC is based on the *source-filter* model of human speech production, in which glottal excitations act as the source signal that is filtered by the vocal tract (mouth, tongue, and lips) to produce the output speech signal. Mathematically, this is represented as

$$Y(z) = X(z)H(z) \tag{5.6}$$

where $Y(z)$ represents the output signal, $X(z)$ the excitation source signal, and $H(z)$ the vocal tract filter. The vocal tract is modeled as an all-pole digital filter of the form

$$\hat{H}(z) = \frac{G}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}} = \frac{S(z)}{E(z)} \tag{5.7}$$

where p is the model order, G is the gain, and $E(z)$ is the excitation input. The prediction coefficients $a_1, a_2, \ldots, a_p$ encode information about the formants—the resonances of the vocal tract. By inverse filtering $Y(z)$ with $\hat{H}^{-1}(z)$, the excitation signal can be obtained. Since $\hat{H}(z)$ does not perfectly model the vocal tract, the resulting signal, termed the *residual* will also contain a component representing the error associated with the modeling.

The source-filter model, and consequently the LPC model, is based on single-speaker speech production. In the case of, say, two overlapped speakers, the output speech signal will contain formants for each of the speakers, potentially doubling the number of peaks. Depending on the model order, the LPC coefficients may not suitably represent this formant structure. In [106], for example, the authors demonstrated a case where an 8th-order LPC model was unable to model the two peaks between 0 kHz and 800 kHz of overlapped speakers whereas a 16th-order model properly did so. The result of poor modeling is greater prediction error and, consequently, more energy in the residual signal. It was hypothesized, then, that LPC residual energy could serve as a useful feature for this overlapped speech detection task. For this work, a 12th-order LPC model was used, the coefficients being computed every 25 ms with an advance of 10 ms.

Figure 5.5 shows the normalized histograms for this residual energy feature. Similar to the RMS energy of Section 5.2.2, the raw residual energy class distributions appear exponential in nature. Any separation between the classes, however, is not visible for this plot. The feature Gaussianization, in addition, was less effective in this case. The overlap class is noticeably more symmetric, but the other two, especially nonspeech, have strong positive skew. In addition, the class means are rather distant from the target value of zero. The reason for both is the large number of raw feature values very close to zero. The transformation, however, does reveal a noticeable separation between the classes, with nonspeech farthest to the left, followed by speech, and

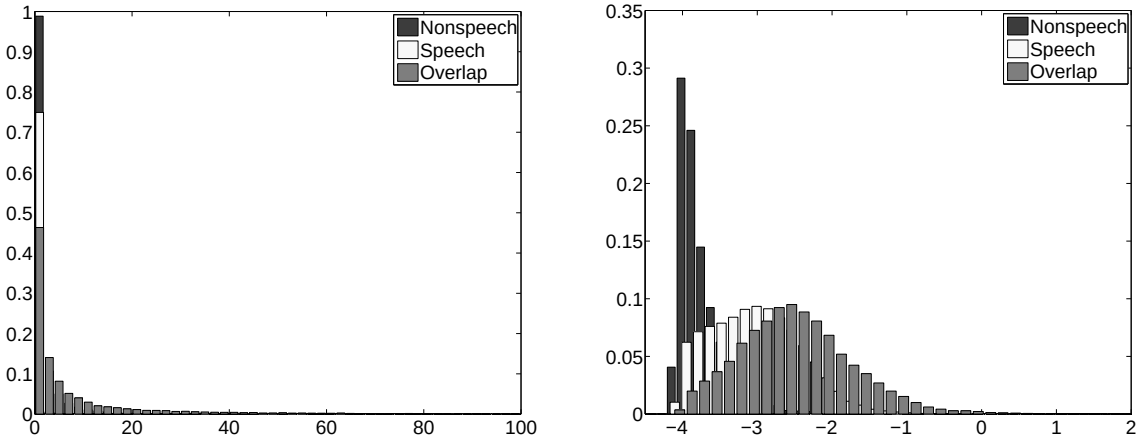


Figure 5.5: Normalized histograms of the (a) raw and (b) Gaussianized LPC residual energy for meeting IS1004c.

then overlap. The separation, too, supports the hypothesized trend of higher residual energy for overlapped speech than single-speaker speech. This separation measures a KL-2 distance of 0.82, larger than any of the above listed distances for these two classes. The distance between speech and nonspeech is 4.34 and between nonspeech and overlap is 11.54, the latter also the largest distance for the two classes.

5.2.6 Spectral Flatness

The spectral flatness measure, SFM_{dB}, in dB is given as:

$$SFM_{dB} = 10 \log_{10} \frac{Gm}{Am} \tag{5.8}$$

$$Gm = \sqrt[N]{\prod_{i=0}^{N-1} \mathrm{X}(i)} \text{ and } Am = \frac{1}{N} \sum_{i=0}^{N-1} \mathrm{X}(i) \tag{5.9}$$

where Gm is the geometric mean, Am is the arithmetic mean, $X(i)$ is the magnitude of the spectral line i, and N is the number of FFT points or spectral lines. Signals

that have power evenly distributed across all spectral bands—the most extreme example being white noise—tend to have high spectral flatness; those that have power concentrated in a small number of bands—the most extreme example being a simple sinusoid—tend to have low spectral flatness. As such, spectral flatness, like zero-crossing rate, is often used as a measure of voicing (e.g., [122]) in speech signals. Because the measure is related to the shape of the spectrum, spectral flatness may also be of use in distinguishing single-speaker speech from overlapped speech. In the case of simultaneous voiced speech, the harmonics of the overlapping speakers will produce many more concentrated energy bands than for a single speaker. The extent to which this effect is observed depends on the difference in fundamental frequencies of the speakers as well as the relative energies, but for many cases the result should be reduced spectral flatness. The spectral flatness measure was computed over a Hamming window of 50 ms and using the first 100 bins of a 1024-point FFT.

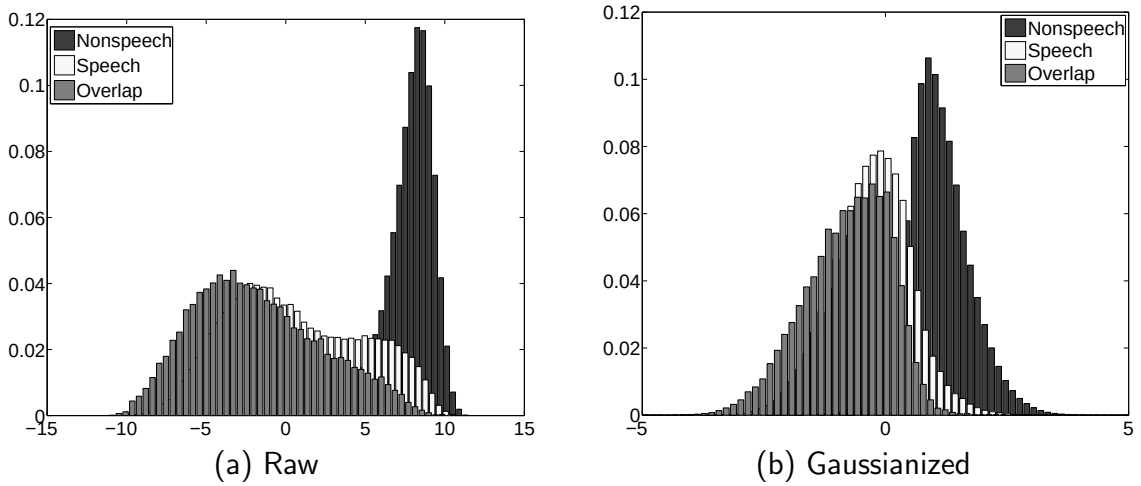


Figure 5.6: Normalized histograms of the (a) raw and (b) Gaussianized spectral flatness for meeting IS1004c.

In contrast to the preceding features, the normalized histograms of the raw spectral flatness feature (shown in Figure 5.6(a)) have very different distributions for the three

classes. The nonspeech class—positioned farthest to the right—resembles a normal distribution with small variance, while the overlap class—positioned farthest to the left—resembles a large-variance Gaussian with some positive skew. The speech class is bimodal—the sum of Gaussians representing voiced and unvoiced speech. In addition, the separation between nonspeech and overlap is quite large: a KL-2 distance of 15.56. The separation between speech and overlap, however, is small—in particular for voiced speech (the leftmost of the two Gaussian "humps"). For this case, feature transformation has the seemingly undesirable effect of bringing the class distributions closer together: the KL-2 distance between nonspeech and overlap is reduced to 7.29. The previously bimodal speech data, however, is successfully transformed into a normal distribution and the lower spectral flatness of overlap compared to speech is evident. The KL-2 distance between these two classes is 0.379 and between speech and nonspeech is 4.184. These distances are quite similar to those for the RMS energy feature, making for an interesting comparison in terms of feature performance.

5.2.7 Harmonic Energy Ratio

As described in Section 5.2.6, the frequency-domain structure of simultaneous voiced speech in many cases differs from that of single-speaker speech. Specifically, the harmonic structure of the speech from the various speakers produces concentrations of energy in frequency bins associated with the integer multiples of each speaker's fundamental frequency. By explicitly analyzing the energy distribution between harmonic and non-harmonic frequency bands, one should be able to distinguish single-speaker and overlapped speech. The harmonic energy ratio (HER), which encodes this information, represents the ratio of harmonic to non-harmonic energy for a frame as determined by a pitch detector. In the case of overlapped speech, the pitch detector estimates the fundamental frequency of the dominant speaker. Since the harmonic

energy of any other speakers will likely lie outside of the regions of harmonic energy for this speaker, the ratio will be lower than for the single-speaker scenario. The ratio was computed by selecting each harmonic FFT band (as determined by pitch detection) plus two adjacent bands, computing the energy and dividing by the energy of the remaining bands. In the event no pitch was detected, the average pitch value was used and the ratio computed accordingly. The HER was computed over a window of 50 ms.

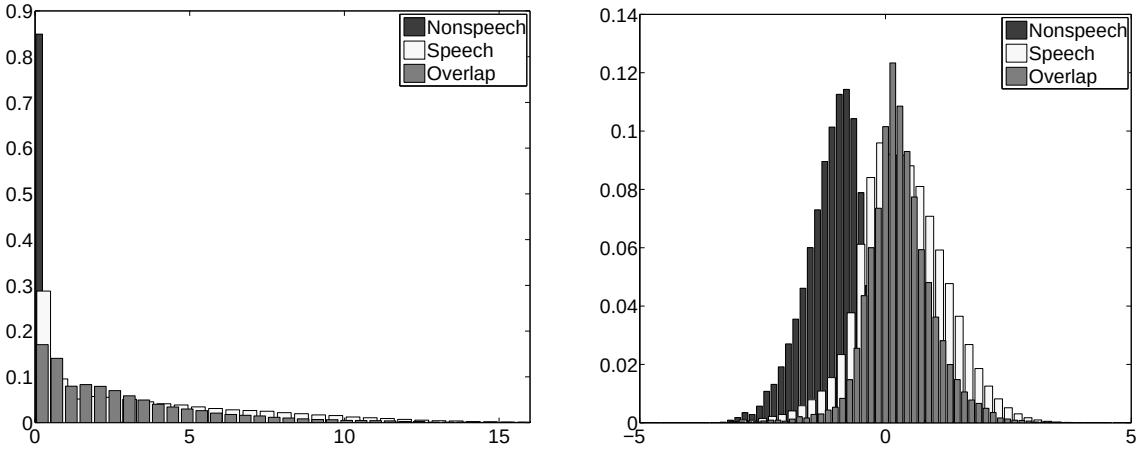


Figure 5.7: Normalized histograms of the (a) raw and (b) Gaussianized harmonic energy ratio for meeting IS1004c.

Figure 5.7 shows the normalized histograms of the HER for the example meeting. As with many other features, the raw harmonic energy ratio exhibits an exponential distribution for all three classes, with a large number of values concentrated near zero. This concentration is especially true for the nonspeech class and is explained by the fact that the "harmonic" bands selected when no pitch is detected contain little energy since no real harmonics exist. For this plot it also appears that the overlap probability mass lies slightly to the left of the speech mass, matching the hypothesis about HER values for these classes. This is slightly more evident when observing the

Gaussianized feature distributions in Figure 5.7 (b). The nonspeech class is clearly to the left of speech and overlap, while the overlap class appears slightly to the left of speech ($\mu_{\text{olap}} = 0.30$, $\mu_{\text{sp}} = 0.35$). This small separation between speech and overlap is also revealed by the KL-2 distance of 0.176, one of the smaller values for the features examined.

5.2.8 Diarization Posterior Entropy

The diarization posterior entropy (DPE) [109][11], represents the entropy of frame-level speaker posteriors as determined by speaker likelihoods output by the diarization system. The posteriors, computed according to Equation 5.1, indicate the confidence of the system in identifying the speaker. For single-speaker speech, confidence should be high and a single speaker model should give high probability while the remainder give significantly lower values and low entropy. In overlapped speech, by comparison, there should be lower, more evenly distributed probabilities among the overlapping speakers and, as a result, the entropy should be higher. To compute the feature, the frame-level posteriors were first filtered with a Hamming window of 500 ms. At each frame y, the resulting values were used to compute the entropy according to

$$H(p(C_k|y)) = \sum_k p(C_k|y) \log \frac{1}{p(C_k|y))} \tag{5.10}$$

Lastly, the entropy was normalized by the maximum possible entropy given the M speaker classes, $\log(M)$. This was done to make values comparable between meetings, which varied in number of hypothesized speakers.

As with the previous features, normalized histograms for the diarization posterior entropy are shown for meeting `IS1004c`. Unlike the other features, the raw DPE class distributions possess a large negative skew, particularly in the case of speech and overlap. In terms of positioning, the nonspeech class is farthest to the right,

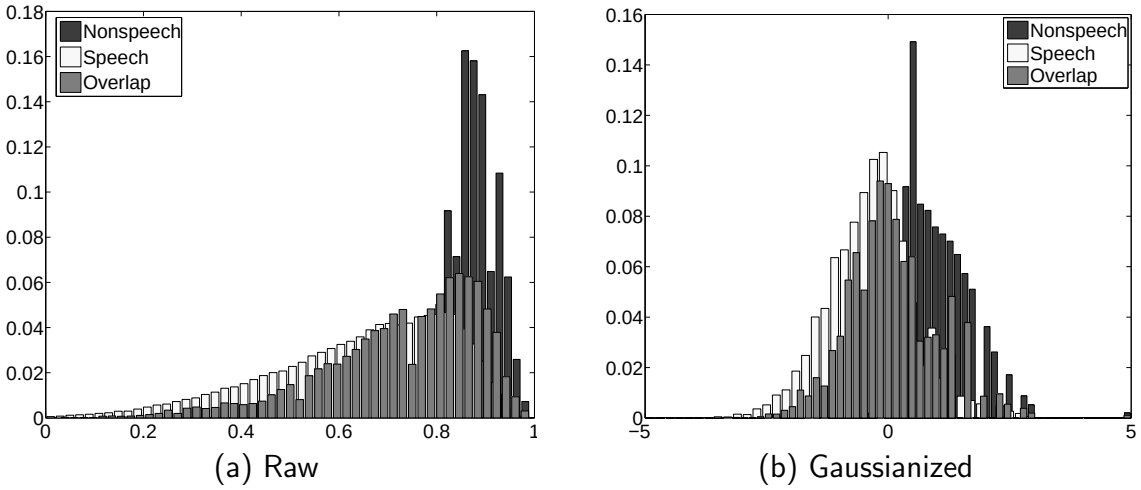


Figure 5.8: Normalized histograms of the (a) raw and (b) Gaussianized diarization posterior entropy for meeting IS1004c.

followed by overlap and, lastly, speech. The ordering conforms to the initial hypothesis: overlapped speech exhibits a slightly higher entropy than single-speaker speech. The nonspeech class has a much higher, though lower-variance, entropy than both speech and overlap since such data is modeled poorly by all speaker models. As with the spectral flatness measure, the Gaussianization procedure, though making the data more normally distributed, appears to bring the class distributions closer to one another. The KL-2 distance between speech and overlap for this transformed feature is 0.188, comparable to that of the harmonic energy ratio.

5.2.9 Modulation Spectrogram Features

The modulation spectrogram (MSG) provides an alternative and complementary representation of the speech signal with a focus on temporal structure. Developed by Kingsbury et al. and detailed in [44], the spectrogram was originally designed to capture key aspects of the auditory cortical representation of speech. These include critical-band frequency resolution, adaptation in the form of automatic gain control,

sensitivity to low-frequency amplitude modulations, and the enhancement of spectro-temporal peaks. Using features derived from this spectrogram, the authors showed the stability of the representation to noise and reverberation for automatic speech recognition. The features were later adopted for use in other speech processing tasks such as speaker verification [45], speaker diarization [112], and laughter detection [47]. Given the relation these tasks have to overlapped speech detection, and given that noise and reverberation are serious issues for this detection task, it seems reasonable that the MSG features may be of use here as well.

The features are computed as follows. The spectrogram of the signal is generated using an FFT with step size of 10 ms and an analysis window of 25 ms. The spectrogram is then divided and integrated into 18 subbands according to the Bark scale. A square root is applied to the sequence of framewise subband energies, which are then processed by two different filters—a 0-8 Hz filter and an 8-16 Hz filter—each of analysis length equal to 210 ms. To complete the process, feedback automatic gain control is applied to individual subbands and the features are locally normalized to be zero-mean and unit-variance. For each frame, the MSG features capture the low-pass and band-pass behavior of the spectrogram of the signal within each of the 18 subbands, resulting in a total of 36 features per frame.

5.3 Experiments

A series of experiments were conducted to evaluate the performance of the various candidate features mentioned above. As with multispeaker SAD, the experiments are organized into three groups, as follows:

1. Single-feature combination to observe the performance of each feature when combined with the baseline features;

2. Feature selection, in which feature combinations were determined using the backward elimination algorithm described in Section 3.2.3; and

3. Evaluation of the final system on unseen validation test data.

5.3.1 Single-Feature Combination

For this initial set of experiments, each candidate feature was combined with the baseline MFCCs to observe and analyze the improvement obtained by each with regard to both detection of overlapped speech and speaker diarization. The various systems were evaluated using the summed nearfield and SDM conditions of the single-site test set ("AMI Single-site") defined in Table 3.2 of Section 3.3.3 and the SDM condition of the multi-site test set ("AMI Multi-site") defined in the same table. In the single-site case, training of the segmenter was performed using 22 meetings, ranging in duration from 13 to 40 minutes for a total of 10 hours of audio. The tuning parameter was optimized using held-out data consisting of 4 additional meetings from this site. For the multi-site evaluation, the training data consisted of 40 meetings ranging in duration from 13 to 58 minutes and totaling 20 hours. The tuning data in this case consisted of 10 meetings. The list of meetings for each set can be found in Appendix A. The overlapped speech segments identified for training, tuning, and testing were obtained using forced alignments of nearfield speech to reference transcriptions, a procedure performed by the SRI DECIPHER recognizer.

Results

The results for the three experimental conditions are found below and presented separately.

Single-site nearfield

Table 5.1 shows the results for the single-site nearfield condition. The first column gives the system, as identified by the features in use. The "Δ" refers to the inclusion of the feature first differences, which was done for all combinations. The "Reference segmentation" presented in the final row corresponds to using the reference overlap segmentation information obtained as described above. The next major column division contains the detection performance metrics of precision, recall, and F-score. The last column division contains various DER improvement metrics. "Labeling" refers to the improvement obtained solely by adding the additional overlap speaker labels to the detected segments (segment post-processing). "Exclusion" refers to the improvement obtained solely by excluding detected overlapped speech from the speaker clustering portion of the diarization algorithm (segment pre-processing). "Both" refers to the gain made by doing both procedures. This last set of metrics is also presented in the bar graph of Figure 5.9. The first thing to note is the high precision performance of

System	Prec.	Recall	F-Score	Rel. DER Imp. (%)		
				Labeling	Exclusion	Both
MFCC + Δ	0.76	0.15	0.26	5.14	-5.32	-0.66
MFCC + RMS_Eg + Δ	0.78	0.15	0.25	5.58	3.34	8.48
MFCC + Spec_Flatness + Δ	0.80	0.17	0.29	6.95	1.76	8.53
MFCC + HER + Δ	**0.83**	0.15	0.26	6.64	4.00	10.29
MFCC + LPC_Eg + Δ	0.77	0.15	0.25	5.10	3.60	8.44
MFCC + MSG + Δ	0.82	**0.21**	**0.33**	**8.09**	**4.26**	**11.47**
MFCC + DPE + Δ	0.69	0.19	0.30	4.66	-5.27	-0.97
MFCC + Kurtosis + Δ	0.82	0.04	0.07	1.36	1.93	3.16
MFCC + ZCR + Δ	0.78	0.12	0.21	4.09	-2.02	1.67
Reference segmentation	-	-	-	26.59	10.99	37.10

Table 5.1: Performance comparisons on single-site nearfield data for single-feature combination systems. The "Δ" indicates first differences were included in the feature set.

all systems. Recall from Section 3.3 that high precision was important for systems

performing this particular segmentation task. These results show that, in seeking to maximize DER improvement (the tuning optimization criterion), a high precision operating point is achieved, highlighting the connection between the two. The highest precision of 0.83 is obtained by the HER feature while the lowest, 0.69, is obtained by the DPE feature. In contrast to precision, recall for the systems is very low, a consequence of the need to trade off between the two. The recall values are generally very similar, with the exception of the maximum of 0.21 from the MSG features and the minimum of 0.04 from kurtosis, the two major outliers. These MSG features also seem to strike the best balance between precision and recall as indicated by F-score, since they achieve the maximum of 0.33 for this evaluation condition. Though F-scores

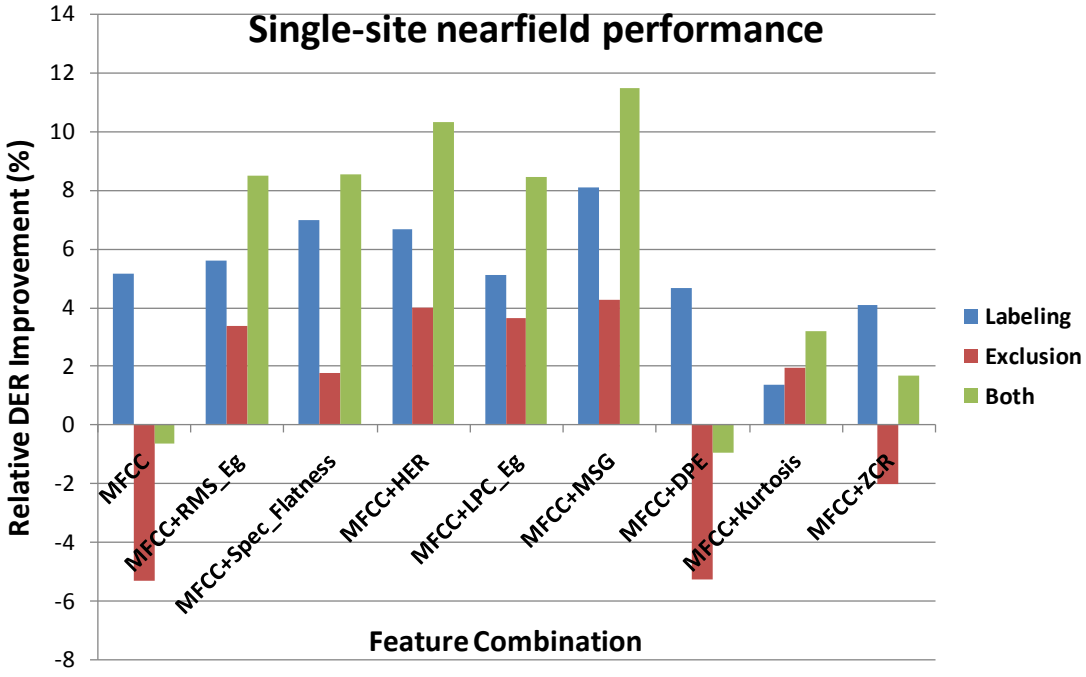


Figure 5.9: Bar graph of performance results from Table 5.1 (reference segmentation results omitted).

for the different features are generally similar, the results in terms of relative DER improvement vary significantly. For segment labeling, the best performing system

is the one with MSG features, yielding a relative improvement of 8.09% while the worst performing feature, kurtosis, yields an improvement of only 1.36%, a difference not deemed statistically significant. The performance differences are even greater where overlap exclusion is concerned, with some features (MFCC, DPE, and ZCR) actually worsening the diarization. This is somewhat curious as all three features produce significant improvements in the labeling case. For the DPE feature this could be explained by the relatively low precision, but the hypothesis clearly does not hold for the other two. Further investigation reveals that in each case a single meeting—`IS1008d` for MFCC and `IS1000a` for DPE and ZCR—degrades significantly (more than doubles) in terms of speaker error when overlap exclusion is performed. This indicates a certain sensitivity of the diarization algorithm to data selection for these meetings. For cases in which both labeling and exclusion produce positive improvements, even larger improvements are gained when both are employed. The MSG features system remains the best under this scenario with an improvement of 11.47% relative while the worst is the DPE feature system with -0.97%, though this is entirely due to the poor overlap exclusion results. Comparing the three scenarios in both the table and in Figure 5.9 reveals that more of the overall improvement comes from the labeling than the exclusion. Last of note here is that the MSG system, which was best for all three cases, lags far behind the performance for ideal segmentation with the references.

Single-site farfield

Table 5.2 and Figure 5.10 present the results for the single-site farfield condition in the same format as the previous experiments. For this condition, the ZCR feature obtains the highest precision of 0.85 while the DPE feature once again has the lowest precision, in this case 0.54. This range in precision is wider than the previous one and the average value across the features is lower. This trend is to be expected, though,

since the farfield condition introduces greater variability in the form of noise and room reverberation. The recall values split into three groups, with low performers around 0.08, the majority around 0.15, and high performers around 0.19. The highest recall is obtained by the RMS energy feature while the lowest is by the HER and DPE features. As with recall, the F-scores divide into three groups around 0.16, 0.24, and 0.3. One of the high performers is the set of MSG features, which once again appears to strike a good balance between precision and recall. With regard to relative

System	Prec.	Recall	F-Score	Rel. DER Imp. (%)		
				Labeling	Exclusion	Both
MFCC + Δ	0.76	0.13	0.23	3.19	4.81	7.94
MFCC + RMS_Eg + Δ	0.71	**0.20**	**0.31**	3.58	3.32	6.93
MFCC + Spec_Flatness + Δ	0.78	0.15	0.25	3.61	4.46	8.04
MFCC + HER + Δ	0.70	0.08	0.15	1.07	3.64	4.98
MFCC + LPC_Eg + Δ	0.57	0.15	0.24	-0.36	1.72	1.27
MFCC + MSG + Δ	0.82	0.19	**0.31**	**5.11**	**5.63**	**10.54**
MFCC + DPE + Δ	0.54	0.08	0.14	-0.26	2.08	1.89
MFCC + Kurtosis + Δ	0.81	0.09	0.16	2.34	1.89	4.36
MFCC + ZCR + Δ	**0.85**	0.14	0.23	4.26	1.85	5.76
Reference segmentation	-	-	-	18.71	7.42	27.07

Table 5.2: Performance comparisons on single-site farfield data for single-feature combination systems. The "Δ" indicates first differences were included in the feature set.

DER improvement, the MSG features, as before, yield the greatest improvement in all three categories. The least improvement in all three categories is also obtained by one feature, the LPC residual energy. Previously this feature had average performance relative to the others. This significant degradation suggests a relatively higher sensitivity to reverberation and noise than other features. The improvements made by overlap labeling in this case are lower than with the nearfield audio. Indeed, two features—LPC residual energy and diarization posterior entropy—now increase the DER slightly (though not to a significant level). For exclusion, we no longer

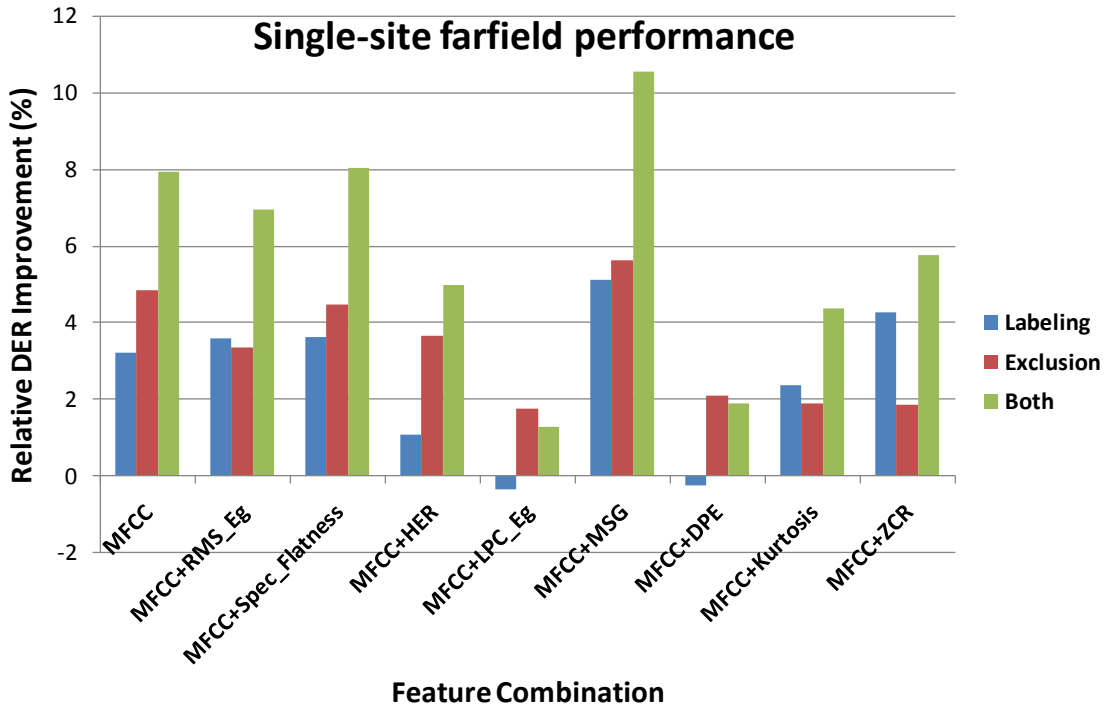


Figure 5.10: Bar graph of performance results from Table 5.2 (reference segmentation results omitted).

see negative improvements and, related to this, the average rate of improvement has increased. Also, as is most evident in Figure 5.10, the contribution to the overall improvement by labeling and exclusion is now roughly the same. This is not true, however, when using the reference segmentation, where the ratio is similar to that of the nearfield condition. The reference segmentation here has a smaller performance gap with the best system, but the ideal gains to be made have also been reduced.

Multi-site farfield

The final single-feature combination results are shown in Table 5.2 and Figure 5.11 and pertain to experiments under the multi-site farfield evaluation condition defined in Section 3.3.3. In terms of precision, the highest performance is obtained by the

modulation spectrogram features and the lowest, as in the other two cases, by the DPE feature. There is also a noticeable reduction in the general precision performance

System	Prec.	Recall	F-Score	Rel. DER Imp. (%)		
				Labeling	Exclusion	Both
MFCC + Δ	0.53	0.07	0.13	-0.43	5.74	5.19
MFCC + RMS_Eg + Δ	0.61	**0.28**	**0.38**	-0.18	5.92	5.83
MFCC + Spec_Flatness	0.67	0.19	0.30	1.28	4.55	6.59
MFCC + HER + Δ	0.57	0.07	0.12	0.03	7.75	7.51
MFCC + LPC_Eg + Δ	0.65	0.21	0.31	0.95	6.19	7.42
MFCC + MSG + Δ	**0.71**	0.15	0.25	**1.31**	0.85	1.89
MFCC + DPE + Δ	0.51	0.18	0.26	-1.46	6.38	5.13
MFCC + Kurtosis + Δ	0.64	0.07	0.13	0.43	**8.42**	**8.79**
MFCC + ZCR + Δ	0.67	0.08	0.15	0.79	5.19	6.04
Reference segmentation	-	-	-	15.75	12.48	27.62

Table 5.3: Performance comparisons on multi-site farfield data for single-feature combination systems. The "Δ" indicates first differences were included in the feature set.

as compared to the other two conditions, with the highest precision being 0.71 and the lowest 0.51. This is again to be expected with the additional increase in variability due to the different recording environments of the meetings. For recall, the best feature is once again RMS energy, with a performance of 0.28, while MFCCs, HER, and kurtosis all have the lowest performance of 0.07. Furthermore, this high recall for the RMS energy is better than that of any of the previous systems. This can also be said of the F-score of this same feature, the result of such a high recall. Similarly, the low recall of the HER feature causes it to yet again produce the lowest F-score.

The results for relative DER improvement are rather different for this condition than the previous two. In the other two cases, the MSG features yielded the largest improvement across all three categories. Here, the MSG features give the best results for overlap segment labeling (though marginally), but the features actually perform quite poorly for overlap exclusion and for the two together. In addition, it is the

kurtosis feature—one of the lowest performers previously—that yields the greatest improvement for these categories.

Further investigation reveals that the problem comes from the tuning of the speech-to-overlap transition penalty. Running an oracle experiment using the MSG features in which the optimal penalty was selected yielded relative DER improvements of 1.77%, 9.06%, and 10.22%, corresponding to overlap labeling, overlap exclusion, and the two together. In addition, the selected tuning parameter value differed greatly from the optimal in this multi-site case whereas for the single-site case the selected and optimal were the same. Of course, this is not surprising, given the better match of the tuning data to test data in the latter scenario, but the sensitivity of the features to tuning should be of concern. The feature transformations performed were in part intended to address this issue, but to some extent it still persists. The problem of generalizing to unseen data, as mentioned in Section 4.3.3, is always present, however, so the challenge becomes employing techniques to mitigate its effects.

That being said, the majority of the candidate features actually perform fairly well in this multi-site condition. With the exception of the MSG features, the lowest relative DER improvement with overlap exclusion is 4.55% and overall is 5.13%, which are higher than in the single-site condition. The improvements made by overlap labeling have decreased further still compared to the other two conditions, but the improvements from overlap exclusion have increased, resulting in only a small decrease in overall improvement. The overall potential gains, as indicated by the reference segmentation results, remain about the same as in the farfield single-site condition, but the contribution by labeling has decreased while the contribution by exclusion has increased, just as with the automatic systems. Lastly, the performance gap between reference and the best system has increased dramatically for the overlap segment labeling, but only slightly for exclusion and overall.

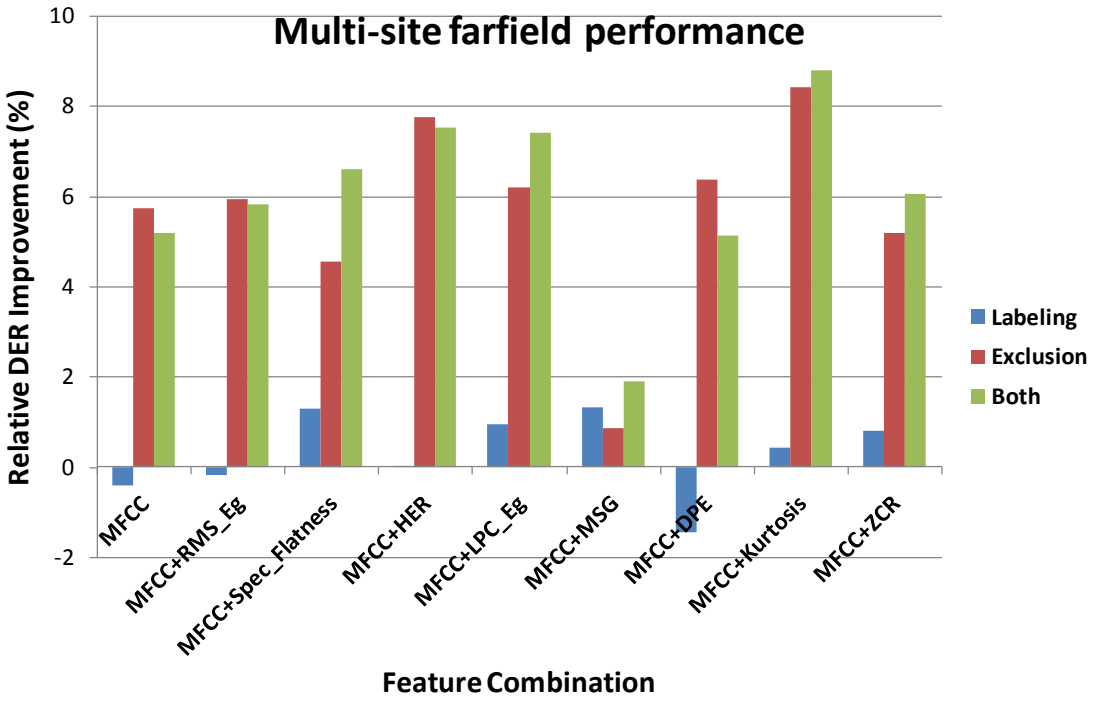


Figure 5.11: Bar graph of performance results from Table 5.3 (reference segmentation results omitted).

5.3.2 Feature Selection

The results of the single-feature combination experiments generally confirmed the utility of the candidate features for overlapped speech detection. Additional experiments were then performed to observe the features in combination and determine a good subset of features for the overlap detection system. As explained in Section 3.2.3, the backward elimination algorithm was chosen to determine this subset. Beginning with the full set of features, the worst-performing feature was removed from the collection at each step, one at a time. The evaluation criterion was again relative DER improvement, in this case using both overlap segment labeling and overlap exclusion. Also note that, since the cepstral features were regarded as the baseline, they were not considered for removal. Given the relatively small number of features, the process was

run until only these baseline features remained and the best performing combination was then identified. The same training, tuning, and evaluation data was used for these experiments as for the ones in Section 5.3.1.

Results

As with single-feature combination, experiments were performed under three experimental conditions and are presented separately.

Single-site nearfield

Table 5.4 shows the results for the single-site nearfield evaluation condition. The table structure is similar to the one for Section 5.3.1, the one difference being the first column here identifies the removed feature for a step in the elimination process. The steps proceed sequentially down the rows, starting with the full set of features (represented as "None" being removed), until all but the baseline MFCCs have been removed. As with the previous experiments, first differences are included as features for all systems.

As can be seen in the table, the algorithm begins with all of the features and successively removes spectral flatness, zero-crossing rate, kurtosis, RMS energy, the modulation spectrogram features, diarization posterior entropy, LPC residual, and finally the harmonic energy ratio. This order of removal, it should be noted, only roughly follows the the reverse order of performance for the features in single-feature combination. DPE (the worst performing feature), for example, is removed quite late in the process while spectral flatness (one of the best features) is removed first. A reason for this could be that the posteriors, being not directly generated from the acoustic data, are less correlated to other features and thus possess utility; other features, such as RMS and LPC residual energy, are highly correlated and thus may not both be necessary. Looking at overlap detection metrics, an interesting trend is

Removed Feature	Prec.	Recall	F-Score	Rel. DER Imp. (%)		
				Labeling	Exclusion	Both
None	0.78	**0.31**	**0.45**	**10.02**	-12.70	-3.16
Spec_Flatness	0.79	0.25	0.38	7.91	**8.09**	15.03
ZCR	0.78	0.29	0.42	8.84	7.38	14.51
Kurtosis	0.77	0.27	0.40	7.87	3.78	11.52
RMS_Eg	0.80	0.25	0.38	9.14	7.87	**15.78**
MSG	0.79	0.20	0.31	7.25	3.82	9.89
DPE	0.78	0.19	0.30	7.16	-3.25	3.21
LPC_Eg	**0.83**	0.15	0.26	6.64	4.00	10.29
HER	0.76	0.15	0.26	5.14	-5.32	-0.66
Reference segmentation	-	-	-	26.59	10.99	37.10

Table 5.4: Performance comparisons on single-site nearfield data for feature combinations determined using backward elimination. Each row represents a subset created by the removal of the one feature—listed in the first column—that produces the greatest gain or least degradation.

observable. The precision performance stays relatively stable as features are removed, while recall and, consequently, F-score performance decrease with the removal of features. The one notable exception is the high precision of 0.83 achieved when the LPC residual energy feature is removed.

As with recall and F-score, the relative DER improvement from labeling generally decreases as features are removed, with the highest performance obtained by using all features. The trend is much less consistent with exclusion, however; this is most notable in the case of the negative improvement of -12.70% when the same set of all features is used. The removal of the spectral flatness feature, a modest performer in the corresponding single-feature combination condition, restores the trend and gives the best performing combination of 8.09%. The combination giving the best overall improvement is achieved after removing the RMS energy and consists of MFCCs, MSG, DPE, LPC residual energy, and HER. The combination yields a relative DER improvement of 15.78%, higher than the best single-feature combination system—including

the MSG features—which yielded an improvement of 11.47%. Indeed, for each of the DER improvement categories, the best performing system exceeded the performance of the best single-feature combination system for that category and, more generally, the multiple-feature combinations outperformed the single-feature ones. This served to narrow the performance gap between the automatic and reference results, in particular for overlap exclusion. Lastly, it is interesting to note that the best overall relative DER improvement follows a decrease in improvement from a previous step. This shows that stopping when no improvement or negative improvement is obtained, as is an alternative to running the removal process to completion, can sometimes produce poorer results.

Single-site farfield

The results using the farfield single-site data are presented in Table 5.5 in the same format as above. Again we see that the single-feature combination performance does not reliably predict order of removal, as evidenced once again by the DPE feature and, rather surprisingly, the MSG features. The removal of the MSG features in the previous experiment nearly halved the overall DER improvement, but here its removal leads to the best performing combination in terms of exclusion and overall improvement: MFCCS, LPC residual energy, kurtosis, DPE, ZCR, HER, and spectral flatness. As with the previous case, a noticeable trend of decreasing recall and F-scores with the removal of features exists. The pattern for precision consists of a steady level until the maximum value of 0.84 is obtained when the DPE feature is removed, followed by a small decrease thereafter. The maximum recall and F-score of 0.25 and 0.37, respectively, are both achieved with the removal of the RMS energy feature. Generally speaking, the values for overlap detection metrics of this farfield data are lower than those of the corresponding nearfield set. To some extent this is to be expected, as the farfield data presents a greater challenge in terms of acoustic variability from noise

Removed Feature	Prec.	Recall	F-Score	Rel. DER Imp. (%)		
				Labeling	Exclusion	Both
None	0.74	0.23	0.35	4.20	-5.40	-0.55
RMS_Eg	0.73	**0.25**	**0.37**	4.55	8.07	12.82
MSG	0.71	0.23	0.34	3.48	**12.85**	**16.33**
LPC_Eg	0.72	0.21	0.33	3.29	7.84	10.64
Kurtosis	0.70	0.20	0.31	3.35	7.16	10.61
DPE	**0.84**	0.18	0.30	**5.07**	7.68	12.43
ZCR	0.80	0.17	0.28	4.20	4.29	8.49
HER	0.78	0.15	0.25	3.61	4.46	8.04
Spec_Flatness	0.76	0.13	0.23	3.19	4.81	7.94
Reference segmentation	-	-	-	18.71	7.42	27.07

Table 5.5: Performance comparisons on single-site farfield data for feature combinations determined using backward elimination. Each row represents a subset created by the removal of the one feature—listed in the first column—that produces the greatest gain or least degradation.

and reverberation. As with the detection metrics, the relative DER improvement metric for overlap labeling exhibits reduced performance for this data. Recall, though, that in the single-feature combination experiments, labeling performance also declined between nearfield and farfield conditions. This was accompanied by an increase in overlap exclusion performance, which also occurs here to some extent. There is also the trend of decreased relative DER improvement as features are removed, with the notable exceptions of the maximum for each of the improvement categories. The maximum exclusion DER improvement is most notable, as it appears that the automatic system outperformed the reference segmentation. Indeed, a number of exclusion performance values exceed the presumed "best" performance obtained using this reference segmentation.

How can this be? The answer, after detailed analysis, it seems, is that not all overlap segments are created equal. Though adulterated by one or more additional speakers, some segments of overlapped speech provide useful information for speaker

modeling and, as a result, when these segments are excluded, the speaker error is higher. Indeed, this phenomenon, was the basic principle of the "usable" speech co-channel speech detection work described in Section 2.2.1 and appears to hold true in the context of diarization as well. This is particularly plausible here given the method with which overlapped speech was identified: using nearfield forced alignment. The relative energy content of each overlapping speaker is not known, so the extent to which the dominant speaker is corrupted could be quite small. The automatic system, detecting many fewer overlap segments in general, removes such "usable" segments less frequently and the DER improvement is thus higher. In some cases, then, the reference segmentation is less than "ideal".

Multi-site farfield

Table 5.6 shows the results for the multi-site farfield evaluation condition. It is interesting to see that the best performance for precision, recall, and F-score is produced in each case with a very small number of the candidate features (two for precision, one for recall, and one for one of the two best F-scores). The combination giving the best recall and F-score performance, in addition, represents an exception to the repeated trend of reduced performance in these categories as features are removed. This combination occurs after the removal of the MSG features, which, in contrast to the other evaluation conditions, produced very low DER improvements for this data. These features do appear to be useful for DER improvement, however, as they remain in the feature combinations giving the best labeling, exclusion, and overall relative DER improvement. The combinations giving the best exclusion and overall DER improvement also utilize a small number of the features for this condition.

As with the single-site farfield condition, a number of the combinations give performance on overlap exclusion which exceeds that of the reference segmentation. Further investigation reveals that the same phenomenon is at work here, namely

Removed Feature	Prec.	Recall	F-Score	Rel. DER Imp. (%)		
				Labeling	Exclusion	Both
None	0.70	0.24	0.36	2.38	13.27	16.20
ZCR	0.72	0.25	0.37	2.50	12.45	15.04
LPC_Eg	0.70	0.26	**0.38**	**2.56**	9.25	11.93
Kurtosis	0.70	0.21	0.33	2.35	11.38	14.59
Spec_Flatness	0.66	0.24	0.35	1.89	16.39	18.19
DPE	0.70	0.19	0.30	1.83	**21.21**	**23.25**
HER	**0.77**	0.15	0.25	2.50	5.25	8.33
MSG	0.61	**0.28**	**0.38**	-0.18	5.92	5.83
RMS_Eg	0.53	0.07	0.13	-0.43	5.74	5.19
Reference segmentation	-	-	-	15.75	12.48	27.62

Table 5.6: Performance comparisons on multi-site farfield data for feature combinations determined using backward elimination. Each row represents a subset created by the removal of the one feature—listed in the first column—that produces the greatest gain or least degradation.

that usable—and, more importantly, useful—overlapped speech is being excluded. The result is that the overall performance of the best system has a relatively small performance gap with the "ideal" system results. The feature combination for the final overlapped speech handling system was chosen using the best combination for this condition, namely MFCCs, RMS energy, MSG features, and the harmonic energy ratio.

5.3.3 Final System

As with the multispeaker SAD system, the overlapped speech handling system was evaluated on an independent test set for validation. This set consisted of 10 randomly selected meetings from the AMI corpus (the "AMI Validation" set listed in Table 3.2) using single distant microphone recordings. For this test the overlapped speech detector was trained using the best feature combination from the multi-site farfield condition—namely, MFCCs, RMS energy, harmonic energy ratio, and modulation

spectrogram features. In addition, the same 40 training and 10 tuning meetings for the condition were used here.

Results

The results for the final system are presented in Table 5.7. In addition, results for the baseline MFCCs and the reference segmentation are included. As with the development data, the baseline MFCCs produce low performance in terms of precision, recall, and F-score. Previously, a small gain was made from performing overlap exclusion with these features, but in this case little difference results.

System	Prec.	Recall	F-Score	Rel. DER Imp. (%)		
				Labeling	Exclusion	Both
Baseline MFCCs	0.56	0.04	0.07	0.23	-0.06	-0.09
Combination	0.58	0.19	0.28	0.23	-0.09	-0.66
Reference segmentation	-	-	-	13.09	9.91	23.46

Table 5.7: Performance results on validation data for the baseline MFCC features and the best feature combination in Table 5.6.

With regard to the combined-feature system, a major difference is evident. Precision, recall, and F-score are all greater for this feature set than the baseline features, but the precision performance for the system here is considerably lower than for any feature combination with the development data. The result is that no significant gains are made by the labeling or exclusion procedures. This stands in contrast to the reference segmentation, where gains comparable to the ones made on the previous multi-site test set are achieved.

How do we account for this difference? Investigation of the tuning parameter with oracle experiments revealed no significant difference between the automatically obtained and the optimal parameter values. Evaluation of other feature combinations in the selection process similarly yielded no improvements. An analysis of performance

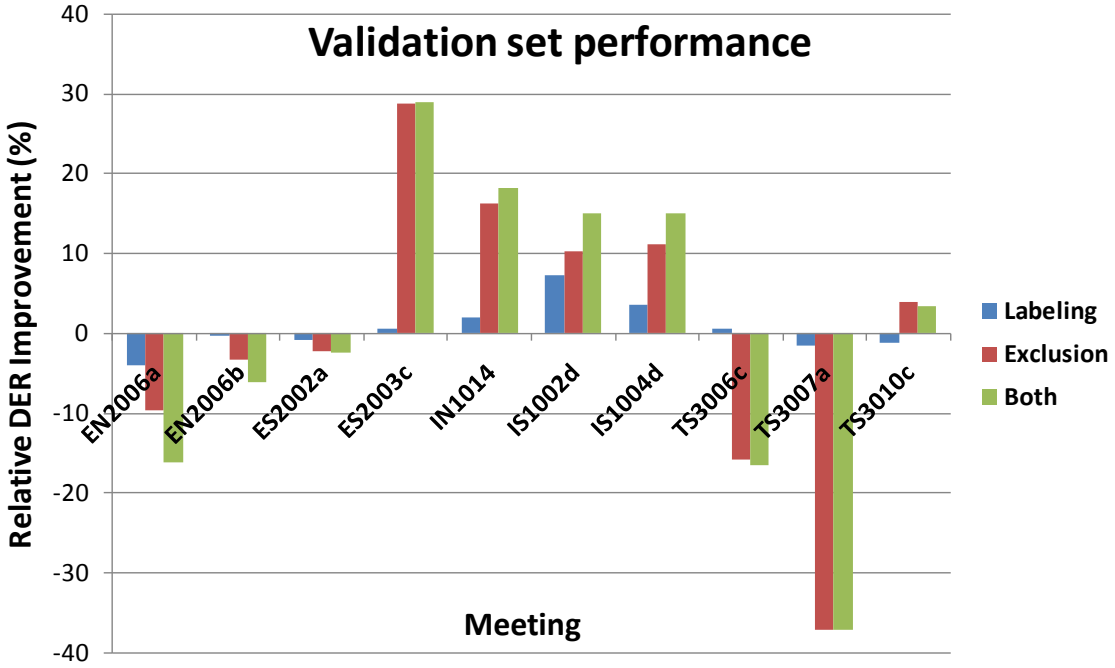


Figure 5.12: Bar graph of meeting-level performance on validation data for the "Combination" system of Table 5.7.

on the meeting level, as shown in Figure 5.12, however, did help to explain the results. The bar graph shows that for four meetings—ES2003c, IN1014, IS1002d, and IS1004d—substantial performance improvements are obtained with the overlap handling procedure, the major contribution once again being from overlap exclusion. This first meeting, in particular, improves by nearly 30% relative. For three other meetings—EN2006b, ES2002a, and TS3010c—no significant change is observed. The remaining three meetings degrade in performance, most notably TS3007a, which worsens by nearly 40% and thus strongly influences the average performance shown in Table 5.7.

For comparison, Figure 5.13 shows a similar bar graph for the multi-site development meetings. Here, too, there are select meetings—namely, EN2009b, IN1008, IN1012, and TS3009c—which benefit substantially (indeed, more than for the valida-

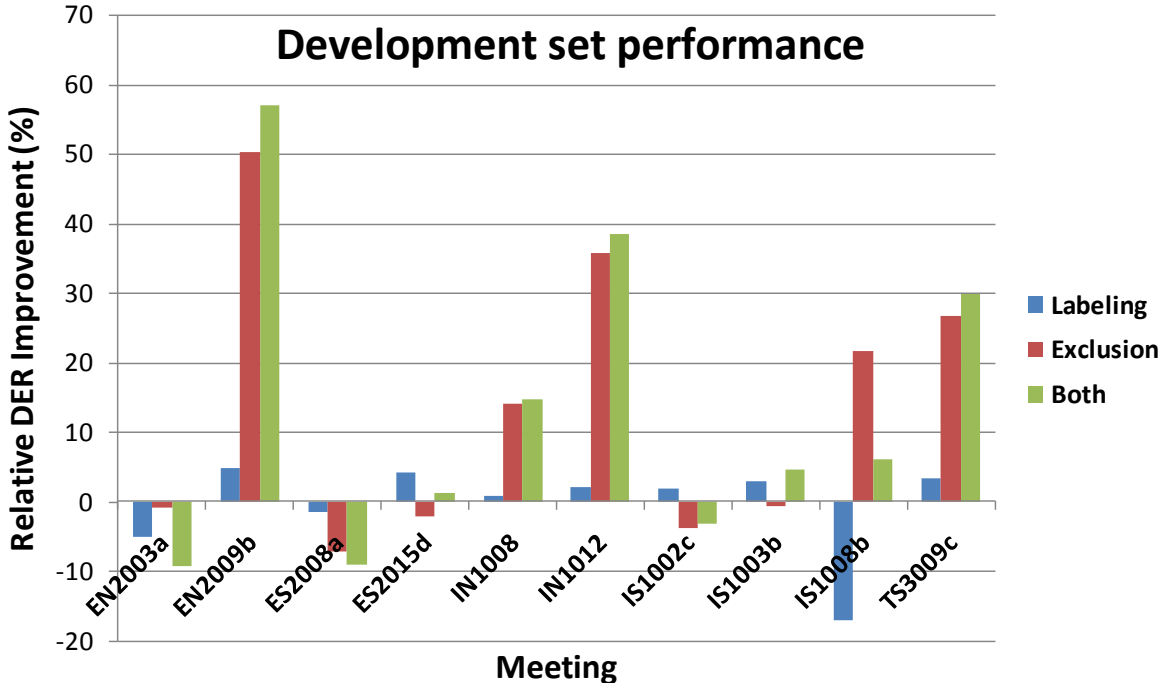


Figure 5.13: Bar graph of meeting-level performance on development data for the best feature combination system of Table 5.6.

tion set) from this overlap handling procedure. Interestingly one meeting, `IS1008b`, improved from overlap exclusion while significantly worsening from overlap labeling. As with the validation set, some meetings degrade in performance—here, `EN2003a` and `ES2008a`—but in this case only moderately. Lastly, there exist meetings here as well which are not significantly affected by the procedure.

This trend is further illustrated by testing on still more meetings. Figure 5.14 shows scatter plots of relative DER improvement due to labeling (Figure 5.14 (a)) and exclusion (Figure 5.14 (b)) versus percent overlapped speech for several meetings including those from the multi-site development and validation test sets. For labeling, a strong positive correlation exists between DER improvement and the overlapped speech percentage in the meeting, as should be expected for a reasonably performing system. In addition, though the percentage changes achieved by labeling alone are

generally too small to be significant, the changes that are significant tend be positive and, hence, improvements. Regarding exclusion, no such correlation is evident. The development data meetings appear concentrated in the positive portion of the relative DER improvement axis while the validation meetings are less so; this trend is also seen for overlap labeling. For the data as a whole, though, the majority of changes in DER are the result of improvements. The distribution of these improvements is bimodal, with a peak in the 2% area (not statistically significant) and one around 13%.

5.4 Discussion

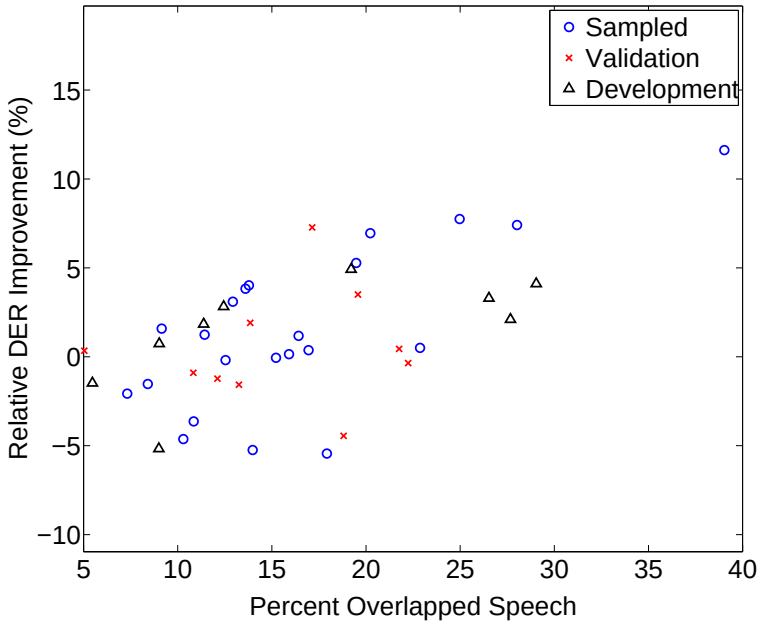
The work presented in this chapter sought to address the issue of overlapped speech in meeting audio processed using speaker diarization. As in Chapter 4, an HMM based audio segmenter was utilized as an enhanced pre-processing component to the target application. In addition to MFCCs, nine features were investigated for the segmentation task across three conditions: single-site nearfield, single-site farfield, and multi-site farfield. The rankings for the single-feature combinations for the three conditions are shown in Table 5.8. The results reveal that for many features it is difficult to predict relative performance when moving from one condition to the other, evidence of the significance of the variability differences between the conditions. Variability also played a role in the final validation results, as it was shown that significant variation existed between relative DER improvements for a large number of test meetings, in particular for overlap exclusion.

This large variation is likely evidence of the sensitivity of the diarization algorithm to meeting variability, a phenomenon documented in the literature. Otterson in [83], for example, when analyzing the effect of various location features for improving speaker diarization noted that the range between the best and worst scores using a

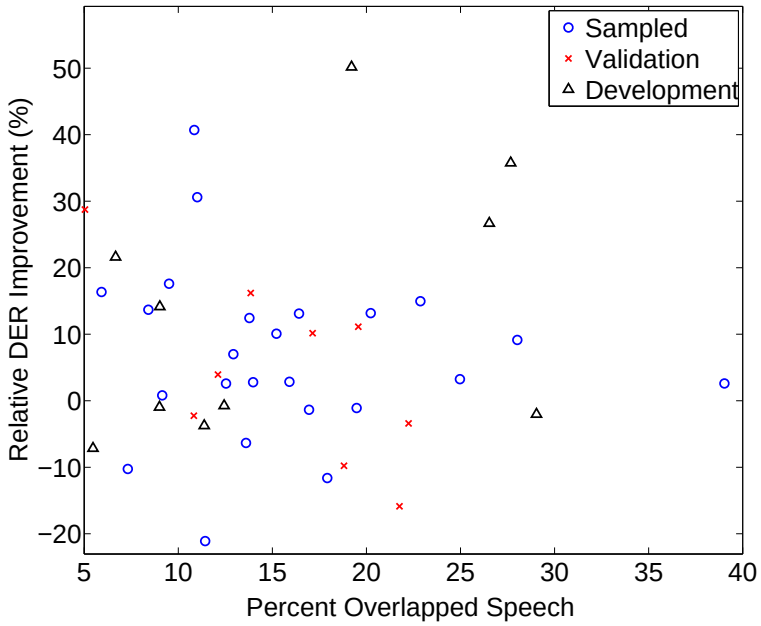
Single/Near	Single/Far	Multi/Far
MSG	MSG	Kurtosis
HER	Spec_Flatness	HER
Spec_Flatness	MFCC	LPC_Eg
RMS_Eg	RMS_Eg	Spec_Flatness
LPC_Eg	ZCR	ZCR
Kurtosis	HER	RMS_Eg
ZCR	Kurtosis	MFCC
MFCC	DPE	DPE
DPE	LPC_Eg	MSG

Table 5.8: Rankings of features in the single-feature combination scenario for the three development testing conditions.

feature was up to 50% and no lower than 20.7%. Mirghafori and Wooters in [70], identified two types of audio files which bring about this sensitivity: "nuts", which exhibit unusually high DER (and, thus, are hard to "crack"), and "flakes", which are very sensitive to tuning parameters. Nuts were characterized by many speakers and a large number of speaker turn changes, while flakes were not so easily identified. In addition to tuning parameters, meetings may also be sensitive to data selection, as in the case of overlap exclusion. This is partly explained by the removal of usable overlapped speech as discussed in Section 5.3.2, but this accounts for degradations in performance and not improvements as witnessed here. Regardless, the results in Figure 5.14 indicate that significant improvements can be obtained, at least for overlap exclusion, on a number of meetings. The challenge becomes identifying the factors which limit the range of meetings for which this occurs.



(a) Labeling



(b) Exclusion

Figure 5.14: Scatter plots of relative DER improvement for (a) Segment labeling and (b) Overlap segment exclusion versus percent overlapped speech for several meetings from the AMI corpus. "Sampled" refers to a sampling of meetings across the percent overlapped speech spectrum; "Validation" denotes the validation meetings; and "Development" refers to meetings from the multi-site development test set of Section 5.3.2.

Chapter 6

Conclusion

The work in this thesis emerged from the identification of two issues, distinct but related, that affect the processing of speech in the multiparty meeting domain:

- Crosstalk—the presence of speech on a personal microphone channel not originating from the microphone wearer—can erroneously be processed by an automatic speech recognition system, often producing word hypotheses that represent insertion errors from the local speech perspective. These errors can propagate beyond the crosstalk region as well due to the context dependencies of N-gram language models.

- Overlapped speech, in failing to be identified by state-of-the-art diarization systems, produces missed speech errors which can constitute a significant portion of the error of these well-performing systems. In addition, this speech can negatively affect speaker modeling in the clustering process, increasing speaker error as a result.

To address these issues, the common approach of audio segmentation was adopted. In the case of crosstalk, local speech on nearfield audio was to be segmented for

multispeaker SAD. In the case of overlapped speech, this speech was to be segmented separately from single-speaker speech to (on the pre-processing end) be excluded from speaker clustering and (on the post-processing end) be identified for overlapped speaker labeling. The HMM framework formed the basis of the two audio segmenters and a primary focus was to identify appropriate features for segmenting the relevant audio classes for improving ASR and speaker diarization.

6.1 Multispeaker SAD

For ASR (Chapter 4), an HMM based segmenter was designed and implemented that incorporated both standard cepstral features as well as cross-channel features. The latter were believed to be important to distinguish local speech from crosstalk due to the cross-channel nature of the phenomenon. The features explored consisted of the normalized maximum cross-channel correlation, the log-energy difference, and time-delay-of-arrival values. To address the issue of a variable number of channels in the meetings, simple statistics (maximum, minimum, mean, and range) of the various feature values were used rather than the values themselves.

In a series of development experiments, each individual feature in combination with the baseline MFCCs achieved significant improvements over this baseline, and the optimal combination of MFCCs, NMXCs, and LEDs produced still more gains. For final validation, the improvements were reduced, but still significant at 9.2% relative WER improvement. Furthermore, analysis of errors and the correlation between error metrics indicated that the source of improvement came from reduced insertion errors and thus reduced crosstalk false alarms by the segmenter. Lastly, additional analysis revealed that a significant portion of the remaining errors are due to deletions of backchannels, which may be difficult to detect and lower in importance than other tokens.

6.2 Overlapped Speech Handling

For speaker diarization, a similar segmenter was designed and implemented. In this case, however, features related to energy, harmonics, and the spectral envelope of the speech signal were explored in addition to the baseline MFCCs. Such features may be used to exploit the difference in structure of single-speaker and overlapped speech. The specific features investigated were RMS energy, zero-crossing rate, kurtosis, LPC residual energy, spectral flatness, harmonic energy ratio, diarization posterior entropy, and modulation spectrogram features, and evaluation occurred across three conditions—single-site nearfield, single-site farfield, and multi-site farfield—with the intention of controlling for variability as occurs across sites and across channels.

Again, a series of system development experiments were performed to identify effective features for the task. Feature performance varied across condition, however, to the extent that the best feature in one case—the modulation spectrogram features—became the worst in another. Nevertheless, significant improvements of around 10% could be made when combining individual features with the baseline MFCCs. When combining multiple features, this number rose as high as 23%, but testing on validation data initially indicated no such gains. A closer analysis of the data revealed that a high variability exists for performance across meetings—in particular for the overlap exclusion pre-processing technique—but a number of meetings tend to benefit significantly from the procedure.

6.3 Contributions and Future Work

This thesis furthered the work on audio segmentation in meetings, providing a systematic analysis of features for the segmentation of both local speech for nearfield audio and overlapped speech for farfield audio. The use of TDOA values, though

common for multi-microphone farfield speech processing, appears to not have been successfully employed in nearfield segmentation previously. The use of the harmonic energy ratio and modulation spectrogram features for overlapped speech detection, too, seems to appear first here. This overlap work, in addition, stands out, as little success has been made thus far in monaural overlapped speech detection for farfield audio, especially in the meeting domain. As discussed in Section 2.2.1, most efforts toward overlapped speech detection have involved artificial mixtures of clean speech (such as from the TIMIT database). Though the amount varied significantly, improvements to speaker diarization from overlap handling were demonstrated using single distant microphone audio from real recordings. The most comparable work to this is [83]. But, as previously mentioned, results were poor when the algorithm was applied to real recordings.

As the primary interest lay in the investigation of appropriate features, the audio segmentation systems presented in this thesis stand to gain from many refinements. As discussed in Section 3.2.2, for example, feature fusion is a large area of study that, for this thesis, was not fully explored. In addition, modifications to the basic HMM architecture—for instance, employing a hybrid HMM/ANN or TANDEM approach—have not been treated here.

The results on speaker diarization point in a number of directions. First, we saw that, for the most realistic condition, the majority of the performance gain was due to overlapped speech exclusion. The high precision criterion of the system, however, was largely an optimization for overlap speaker labeling, since it was based on the assumption of solely false alarms being detrimental. The trade-off between false alarms and misses for speaker clustering after overlap exclusion has yet to be examined. It is conceivable that a diarization system could be robust to false alarms and benefit from reduced misses from this perspective. Second, the issue of "nuts" and "flakes", in particular regarding overlap exclusion, needs to further be explored. Increasing

the consistency of performance improvement is important to having this proposed technique be adopted. Finally, since the primary source of improvement is reduced speaker error, it may be possible to achieve similar improvements by incorporating the features directly into the diarization system. This is akin to Otterson and Ostendorf's use of location features to improve speaker clustering in the MDM condition in [84]. Vinyals and Friedland in [112], for example, have already demonstrated improved diarization in the SDM condition with the modulation spectrogram features examined here. Other features from the candidate list may prove useful as well.

Appendix A

Training and Tuning Meetings

Eval04

	Meetings				
Train	20011115	20011211	20020111	20020213	20020304
	20020627	20020731	20020815	20020904	20020911
	20021003	20030702	20030729	20031204	20031215
	Bdb001	Bed002	Bed003	Bed004	Bed005
	Bed006	Bed008	Bed009	Bed010	Bed011
	Bed012	Bed013	Bed014	Bed015	Bed016
	Bed017	Bmr001	Bmr002	Bmr003	Bmr005
	Bmr006	Bmr007	Bmr008	Bmr009	Bmr010
	Bmr011	Bmr012	Bmr014	Bmr015	Bmr016
	Bmr019	Bmr020	Bmr021	Bmr022	Bmr023
	Bmr024	Bmr025	Bmr026	Bmr027	Bmr028
	Bmr029	Bmr030	Bmr031	Bns001	Bns002
	Bns003	Bro003	Bro004	Bro005	Bro007
	Bro008	Bro010	Bro011	Bro012	Bro013
	Bro014	Bro015	Bro016	Bro017	Bro018
	Bro019	Bro021	Bro022	Bro023	Bro024
	Bro025	Bro026	Bro027	Bro028	Bsr001
	Btr001	Btr002	Buw001		
Tune	ES2009b	ES2009d	IS1009a	IS1009c	

Eval05*

The Eval05* training set consists of the same meetings as the Eval04 training set with the addition of the 35 meetings presented in the table below. Tuning for Eval05* was performed using the Eval04 test meetings, listed in Table 3.1.

	Meetings				
Train	ES2006a	ES2006b	ES2006c	ES2007a	ES2007b
	ES2007c	ES2007d	IS1000a	IS1000b	IS1000d
	IS1001a	IS1001b	IS1001c	IS1001d	IS1002b
	IS1002c	IS1002d	IS1003a	IS1003b	IS1003c
	IS1003d	IS1004a	IS1004b	IS1004c	IS1004d
	IS1005a	IS1005b	IS1005c	IS1006a	IS1006b
	IS1006c	IS1006d	IS1007a	IS1007b	IS1007c

AMI Single-site

	Meetings				
Train	IS1000b	IS1000c	IS1001d	IS1002b	IS1002c
	IS1003a	IS1003c	IS1004a	IS1004b	IS1004c
	IS1005a	IS1005b	IS1005c	IS1006a	IS1006c
	IS1007a	IS1007b	IS1007c	IS1009a	IS1009b
	IS1009c	IS1009d			
Tune	IS1000d	IS1002d	IS1004d	IS1007d	

AMI Multi-site

	Meetings				
Train	EN2002d	ES2002d	ES2003b	ES2004b	ES2005b
	ES2005d	ES2006a	ES2006b	ES2007a	ES2007d
	ES2008d	ES2009b	ES2011a	ES2012a	ES2012b
	ES2014a	ES2014b	ES2016a	ES2016c	IB4002
	IB4005	IN1001	IN1009	IS1000c	IS1001a
	IS1001b	IS1004a	IS1005b	IS1006a	IS1006b
	IS1007a	IS1009a	TS3003c	TS3004d	TS3006b
	TS3008a	TS3008b	TS3009b	TS3010a	TS3010b
Tune	EN2004a	ES2013c	IS1001c	IS1001d	IS1005a
	IS1007b	IS1007c	TS3006a	TS3007c	TS3012b

AMI Validation

The training and tuning sets for the validation test set were the same as those for the AMI Multi-site test set.

Bibliography

[1] Andre Adami, Lukáš Burget, Stephane Dupont, Hari Garudadri, Frantisek Grezl, Hynek Hermansky, Pratibha Jain, Sachin Kajarekar, Nelson Morgan, and Sunil Sivadas. Qualcomm-ICSI-OGI features for ASR. In *Proceedings of ICSLP*, pages 21–24, 2002. Denver, CO.

[2] Sassan Ahmadi and Andreas S. Spanias. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Transactions on Speech and Audio Processing*, 7(3):333–338, May 1999.

[3] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Proceedings of the IEEE ASRU Workshop*, pages 411–416, 2003. St. Thomas, US Virgin Islands.

[4] Xavier Anguera. *Robust Speaker Diarization for Meetings*. PhD thesis, Polytechnic University of Catalonia, October 2006.

[5] Xavier Anguera, Chuck Wooters, and Javier Hernando. Speaker diarization for multi-party meetings using acoustic fusion. In *Proceedings of the IEEE ASRU Workshop*, pages 426–431, 2005. San Juan, Puerto Rico.

[6] Xavier Anguera, Chuck Wooters, and Jose M. Pardo. Robust speaker diarzation for meetings: ICSI rt06s meetings evaluation system. In S. Renals, S. Bengio, and J. Fiscus, editors, *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 346–358. Springer Berlin/Heidelberg, 2006.

[7] Futoshi Asano and Jun Ogata. Detection and separation of speech events in meeting recordings. In *Proceedings of INTERSPEECH-ICSLP*, pages 2586–2589, 2006. Pittsburgh, PA.

[8] Jerome R. Bellegarda and David Nahamoo. Tied continuous parameter modeling for speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(12):2033–2045, December 1990.

[9] Richard E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton,NJ, 1961.

[10] Kofi Boakye and Andreas Stolcke. Improved speech activity detection using cross-channel features for recognition of multiparty meetings. In *Proceedings of INTERSPEECH-ICSLP*, pages 1962–1965, 2006. Pittsburgh, PA.

[11] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *Proceedings of ICASSP*, pages 4353–4356, 2008. Las Vegas, NV.

[12] Kofi Boakye, Oriol Vinyals, and Gerald Friedland. Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech. In *Proceedings of INTERSPEECH*, pages 32–35, 2008. Brisbane, Australia.

[13] H. Bourlard and S. Dupont. A new asr approach based in independent processing and recombination of partial frequency bands. In *Proceedings of ICSLP*, pages 426–429, 1996. Philadelphia, PA.

[14] Michael S. Brandstein and Harvey F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proceedings of ICASSP*, pages 375–378, 1997. Munich, Germany.

[15] M.S. Brandstein, J.E. Adcock, and H.F. Silverman. A closed-form location estimator for use with room environment microphonearrays. *IEEE Transactions on Speech and Audio Processing*, 5(1):45–50, January 1997.

[16] K. Bullington and J.M. Fraser. Engineering aspects of TASI. Technical report, Bell System Technical Journal, March 1959.

[17] Lukáš Burget. Combination of speech features using smoothed heteroscedastic linear discriminant analysis. In *Proceedings of INTERSPEECH - ICSLP*, pages 2549–2552.

[18] 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 *Proceedings of of the Measuring Behavior 2005 Symposium on "Annotating and Measuring Meeting Behavior"*, 2005. AMI-108.

[19] Nishant Chandra and Robert E. Yantorno. Usable speech detection using the modified autocorrelation peak to valley ratio using the LPC residual. In *Proceedings of IASTED 2002*, pages 146–149, 2002.

[20] S.S. Chen and P.S. Gopalakrishnam. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, 1998. Lansdowne, VA.

[21] Heidi Christensen, Borge Lindberg, and Ove Andersen. Employing heterogeneous information in a multi-stream framework. In *Proceedings of ICASSP*, pages 1571–1574, 2000. Istanbul, Turkey.

[22] Andres Corrada-Emmanuel, Michael Newman, Barbara Peskin, Lawrence Gillick, and Robert Roth. Progress in speaker recognition at dragon systems. In *Proceedings of ICSLP*, pages 1355–1358, 1998. Sydney, Australia.

[23] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997.

[24] S. Dharanipragada and M. Padmanabhan. A nonlinear unsupervised adaptation technique for speech recognition. In *Proceedings of ICSLP*, volume 4, pages 556–559, 2000. Beijing, China.

[25] John Dines, Jithendra Vepa, and Thomas Hain. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proceedings of INTERSPEECH-ICSLP*, pages 1213–1216, 2006. Pittsburgh, PA.

[26] R.O. Duda and P.B. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[27] Daniel P.w. Ellis. Stream combination before and/or after the acoustic model. Technical Report TR-00-007, International Computer Science Institute, April 2000.

[28] Daniel P.W. Ellis and Jerry C. Liu. Speaker turn segmentation based on between-channel differences. In *Proceedings of NIST ICASSP 2004 Meeting Recognition Workshop*, 2004. Montreal, Canada.

[29] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Inc., New York, 1990.

[30] John Garofolo, Christophe Laprun, Martial Michel, Vincent Stanford, and Elham Tabassi. The NIST meeting room pilot corpus. In *Proceedings of LREC*, pages 1411–1414, 2004. Lisbon, Portugal.

[31] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren. The DARPA TIMIT acoustic-phonetic continuous speech corpus. CDROM.

[32] L. Gillick and S.J. Cox. Some statistical issus in the comparison of speech recognition algorithms. In *Proceedings of ICASSP*, pages 532–535, 1989. Glasgow, UK.

[33] Herbert Gish, Man-Hung Siu, and Robin Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *Proceedings of ICASSP*, pages 873–876, 1991. Toronto, Canada.

[34] M. Graciarena, H. Franco, J. Zheng, D. Vergyri, and A. Stolcke. Voicing feature integration in SRI's decipher LVCSR system. In *Proceedings of ICASSP*, pages 921–924, 2004. Montreal, Canada.

[35] J.A. Haigh and J.S. Mason. Robust voice activity detection using cepstral features. In *Proceedings of TENCON*, 1993. Beijing, China.

[36] Simon Haykin. *Adaptive Filter Theory*. Prentice Hall, 2002.

[37] Jing Huang, Etienne Marcheret, Karthik Visweswariah, Vit Libal, and Gerasimos Potamianos. The IBM rich transcription 2007 speech-to-text systems for lecture meetings. In *Multimodal Technologies for Perception of Humans*, volume 4625 of *Lecture Notes in Computer Science*, pages 429–441. Springer Berlin/Heidelberg, 2008.

[38] Y. Huang, J. Benesty, G.W. Elko, and R.M. Mersereau. Real-time passive source localization: A practical linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing*, 9(8):943–956, January 2001.

[39] Marijn Huijbregts and Chuck Wooters. The blame game: Performance analysis of speaker diarization system components. In *Proceedings of INTERSPEECH*, pages 1857–1860, 2007. Antwerp, Belgium.

[40] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The ICSI meeting corpus. In *Proceedings of ICASSP*, pages 364–367.

[41] Adam Janin, Andreas Stolcke, Xavier Anguera, Kofi Boakye, Özgür Çetin, Joe Frankel, and Jing Zheng. The ICSI-SRI spring 2006 meeting recognition system. In S. Renals, S. Bengio, and J. Fiscus, editors, *MLMI 2006, Lecture Notes in*

Computer Science, volume 4299, pages 444–456. Springer Berlin/Heidelberg, 2006.

[42] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of ML-94*, pages 121–129. Morgan Kaufmann, 1994.

[43] Don H. Johnson and Sinan Sinanovi'c. *Symmetrizing the Kullback-Leibler Distance*. Rice University, 2001.

[44] B.E.D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25:117–132, August 1998.

[45] Tomi Kinnunen, Kong-Aik Lee, and Haizhou Li. Dimension reduction of the modulation spectrogram for speaker diarization. In *Proceedings of Odyssey*, 2008. Stellenbosch, South Africa.

[46] Charles H. Knapp and G. Clifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(4):320–327, August 1976.

[47] Mary Tai Knox, Nelson Morgan, and Nikki Mirghafori. Getting the last laugh: Automatic laughter segmentation in meetings. In *Proceedings of INTERSPEECH*, pages 797–800, 2008. Brisbane, Australia.

[48] Daphne Koller and Mehran Sahami. Toward optimal feature selection. pages 284–292. Morgan Kaufmann, 1996.

[49] Kasturi R. Krishnamachari, Robert E. Yantorno, Daniel S. Benincasa, and Stanley J. Wenndt. Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions. In *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pages 710–713, 2000. Honolulu, HI.

[50] Kasturi R. Krishnamachari, Robert E. Yantorno, and Jereme M. Lovekin. Use of local kurtosis measure for spotting usable speech segments in co-channel speech. In *Proceedings of ICASSP*, pages 649–652, 2001. Salt Lake City, UT.

[51] Nagendra Kumar and Adreas G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 26:283–297, December 1998.

[52] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon. An improved endpoint detector for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(4):777–785, August 1981.

[53] Kornel Laskowski, Christian Fügen, and Tanja Schultz. Simultaneous multispeaker segmentation fro automatic meeting recognition. In *Proceedings of EUSIPCO*, pages 1294–1298, 2007. Poznan, Poland.

[54] Kornel Laskowski, Qin Jin, and Tanja Schultz. Crosscorrelation-based multispeaker speech activity detection. In *Proceedings of INTERSPEECH - ICSLP*, pages 973–976.

[55] Kornel Laskowski and Tanja Schultz. Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. In *Proceedings of ICASSP*, pages 993–996, 2006. Toulouse, France.

[56] Kornel Laskowski and Tanja Schultz. A geometric interpretation of non-target-normalized maximum cross-channel correlation for vocal activity detection in meetings. In *Proceedings of NAACL-HLT2007*, pages 89–92, 2007. Rochester, NY.

[57] Kornel Laskowski and Tanja Schultz. Modeling vocal interaction. In *Proceedings of ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, pages 148–155, 2008. Columbus, OH.

[58] James P. LeBlanc and Philip L. De Leon. Speech separation by kurtosis maximization. In *Proceedings of ICASSP*, pages 1029–1032, 1998. Seattle, WA.

[59] Te-Won Lee and Anthony J. Bell. Blind source separation of real world signals. In *Proceedings of Int. Conf. on Neural Networks*, pages 2129–2134, 1997. Houston, TX.

[60] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer, Speech, & Language*, 9(2):171–185, April 1995.

[61] Michael A. Lewis and Ravi 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.

[62] Qi Li and Augustine Tsai. A matched filter approach to endpoint detection for robust speaker verification. In *Proceedings of the IEEE Workshop on Automatic Identification*, October 1999. Summit, NJ.

[63] Yoseph Linde, Andrés Buzo, and Robert M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[64] Daben Liu and Francis Kubala. A cross-channel modeling approach for automatic segmentation of conversational telephone speech. In *Proceedings of the IEEE ASRU Workshop*, pages 333–338, 2003. St. Thomas, US Virgin Islands.

[65] Jereme Lovekin, Kasturi R. Krishnamachari, and Robert E. Yantorno. Adjacent pitch period comparison (appc) as a usability measure of speech segments under co-channel conditions. In *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pages 139–142, 2001.

[66] Jereme M. Lovekin, Robert E. Yantorno, Kasturi R. Krishnamachari, Daniel S. Benincasa, and Stanley J. Wenndt. Developing usable speech criteria for speaker identification technology. In *Proceedings of ICASSP*, pages 421–424, 2001. Salt Lake City, UT.

[67] Edmund R. Malinowski. Determination of the number of factors and the experimental error in a data matrix. *Analytical Chemistry*, 49(4):612–617, April 1977.

[68] Sylvain Meignier, Jen François Bonastre, and Stéphane Igounet. E-hmm approach for learning and adapting sound models for speaker indexing. In *Proceedings of Odyssey*, pages 175–180, 2001. Crete, Greece.

[69] Nikki Mirghafori. *A Multi-Band Approach to Automatic Speech Recognition*. PhD thesis, University of California, Berkeley, 1998.

[70] Nikki Mirghafori and Chuck Wooters. Nuts and flakes: A study of data characteristics in speaker diarization. In *Proceedings of ICASSP*, pages 1017–1020, 2006. Toulouse, France.

[71] S. Molau, M. Pitz, and H. Ney. Histogram based normalization in the acoustic feature space. In *Proceedings of ASRU*, pages 21–24, 2001. Madonna di Campiglio, Italy.

[72] David P. Morgan, Bryan George, Leonard T. Lee, and Steven M. Kay. Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Transactions on Speech and Audio Processing*, 5(5):407–424, September 1997.

[73] Nelson Morgan, Don Baron, Sonali Bhagat, Hannah Carvey, Rajdip Dhillon, Jane Edwards, David Gelbart, Adam Janin, Ashley Krupski, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. Meetings

about meetings: Research at ICSI on speech in multiparty conversations. In *Proceedings of ICASSP*, pages 740–743, 2003. Hong Kong, China.

[74] Nelson Morgan, Barry Chen, Qifeng Zhu, and Andreas Stolcke. TRAPping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition. In *Proceedings of ICASSP*, pages 536–539, 2004. Montreal, Canada.

[75] National Institute of Standards and Technology. *Fall 2004 Rich Transcription (RT-04F) Evaluation Plan*, 2004. http://www.nist.gov/speech/tests/rt/2004-fall/docs/rt04f-eval-plan-v14.pdf.

[76] National Institute of Standards and Technology. *Spring 2004 (RT-04S) Rich Transcription Meeting Recognition Evaluation Plan*, 2004. http://www.nist.gov/speech/tests/rt/2004-spring/documents/rt04s-meeting-eval-plan-v1.pdf.

[77] National Institute of Standards and Technology. *Spring 2005 (RT-05S) Rich Transcription Meeting Recognition Evaluation Plan*, 2005. http://www.nist.gov/speech/tests/rt/2005-spring/rt05s-meeting-eval-plan-V1.pdf.

[78] National Institute of Standards and Technology. *Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation Plan*, 2006. http://www.nist.gov/speech/tests/rt/2006-spring/docs/rt06s-meeting-eval-plan-V2.pdf.

[79] National Institute of Standards and Technology. *Spring 2007 (RT-07S) Rich Transcription Meeting Recognition Evaluation Plan*, 2007. http://www.nist.gov/speech/tests/rt/2007/docs/rt07s-meeting-eval-plan-v2.pdf.

[80] Elias Nemer, Rafik Goubran, and Samy Mahmoud. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*, 9(3):217–231, March 2001.

[81] Shigeki Okawa, Enrico Bocchieri, and Alexandros Potamianos. Multi-band speech recognition in noisy environments. In *Proceedings of ICASSP*, pages 641–644, 1998. Seattle, WA.

[82] Scott Otterson. Improved location features for meeting speaker diarization. In *Proceedings of INTERSPEECH*, pages 1849–1852, 2007. Antwerp, Belgium.

[83] Scott Otterson. *Use of Speaker Location Features in Meeting Diarization*. PhD thesis, University of Washington, 2008.

[84] Scott Otterson and Mari Ostendorf. Efficient use of overlap information in speaker diarization. In *Proceedings of the IEEE ASRU Workshop*, pages 683–686, 2007. Kyoto, Japan.

[85] D.S. Pallet, W.M. Fisher, and J.G. Fiscus. Tools for the analysis of benchmark speech recognition tests. In *Proceedings of ICASSP*, pages 97–100, 1990. Albuquerque, NM.

[86] Jose M. Pardo, Xavier Anguera, and Charles Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transactions on Computers*, 56(9):1212–1224, September 2007.

[87] Jose M. Pardo, Xavier Anguera, and Chuck Wooters. Speaker diarization for multi-microphone meetings using only between-channel differences. In *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 257–264. Springer Berlin/Heidelberg, 2006.

[88] Jose M. Pardo, Xavier Anguera, and Chuck Wooters. Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences. In *Proceedings of INTERSPEECH-ICSLP*, pages 2194–2197, 2006. Pittsburgh, PA.

[89] T. Pfau, D.P.W. Ellis, and A. Stolcke. Multispeaker speech activity detection for the ICSI meeting recorder. In *Proceedings of the IEEE ASRU Workshop*, pages 107–110, 2001. Trento, Italy.

[90] Thilo Pfau and Daniel P.W. Ellis. Hidden markov model based speech activity detection for the ICSI meeting project. In *Proceedings of EUROSPEECH*, pages ??–??, 2001. Aalborg, Denmark.

[91] D. Povey and P.C. Woodland. Minimum phone error and I-Smoothing for improved discriminative training. In *Proceedings of ICASSP*, pages 105–108, 2002. Orlando, FL.

[92] S. Joe Qin and Ricardo Dunia. Determining the number of principal components for best reconstruction. *Journal of Process Control*, 10(2-3):245–250, April 2000.

[93] Lawrence R. Rabiner and Marvin R. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(4):338–343, 1977.

[94] L.R. Rabiner and M.R. Sambur. An algorithm for determining the endpoints of isolated utterances. Technical report, The Bell System Technical Journal, February 1974.

[95] Philippe Renevey and Andrzej Drygajlo. Entropy based voice activity detection in very noisy conditions. In *Proceedings of EUROSPEECH*, pages 1887–1890, 2001. Albie's, Denmark.

[96] D.A. Reynolds and P. Torres-Carrasquilo. Approaches and application of audio diarization. In *Proceedings of ICASSP*, pages 953–956, 2005. Philadelphia, PA.

[97] Doug Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643, October 1994.

[98] David Sankoff and Joseph B. Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, Reading,MA, 1983.

[99] G. Saon, S. Dharanipragada, and D. Povey. Feature space gaussianization. In *Proceedings of ICASSP*, volume 1, pages 329–332, 2004. Montreal, Canada.

[100] R.O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions Antennas Propagation*, AP-34(3):276–280, March 1986.

[101] Yang Shao and DeLiang Wang. Co-channel speaker identification using usable speech extraction based on multi-pitch tracking. In *Proceedings of ICASSP*, pages 205–208, 2003. Hong Kong, China.

[102] Matthew A. Siegler, Uday Jain, Bhiksha Raj, and Richard M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of DARPA Speech Recognition Workshop*, pages 97–99, 1997. Chantilly, VA.

[103] J.O. Smith and J.S. Abel. Closed-form leaset-squares source location estimation from range-difference measurements. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(12):1661–1669, December 1987.

[104] 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 Steve Renals and Samy 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.

[105] Andreas Stolcke, Chuck Wooters, Nikki Mirghafori, Tuomo Pirinen, Ivan Bulyko, Dave Gelbart, Martin Graciarena, Scott Otterson, Barbara Peskin, and Mari Ostendorf. Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system. In *Proceedings of NIST ICASSP 2004 Meeting Recognition Workshop*, 2004. Montreal, Canada.

[106] Nithya Sundaram, Robert E. Yantorno, Brett Y. Smolenski, and Ananth N. Iyer. Usable speech detection using linear predictive analysis - a model based approach. In *Proceedings of ISPACS*, pages 231–235, 2003. Awaji Island, Japan.

[107] Sangita Tibrewala and Hynek Hermansky. Multi-band and adaptation approached to robust speech recognition. In *Proceedings of EUROSPEECH*, pages 2619–2622, 1997. Rhodes, Greece.

[108] Kari Torkkola. Real-time discrimination of broadcast speech/music. In *Proceedings of ICASSP*, pages 993–996, 1996. Atlanta, GA.

[109] B. Trueba-Hornero. Handling overlapped speech in speaker diarization. Master's thesis, Universitat Politecnica de Catalunya, May 2008.

[110] Sergio Valle, Weihua Li, and S. Joe Qin. Selection of the number of principal components: The variance reconstruction error criterion with a comparison to other methods. *Industria Engineering Chemical Research*, 38(11):4389–4401, September 1999.

[111] David van Leeuwen and Marijn Huijbregts. The AMI speaker diarization system for NIST rt06s meeting data. In *Proceedings of of the Rich Transcription 2006 Meeting Recognition Evaluation Workshop*, pages 371–384, 2006. Washington, D.C.

[112] Oriol Vinyals and Gerald Friedland. Modulation spectrogram features for speaker diarization. In *Proceedings of INTERSPEECH*, pages 630–633, 2008. Brisbane, Australia.

[113] Hisashi Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(2):183–192, 1977.

[114] Ehud Weinstein, Meir Feder, and Alan V. Oppenheim. Multi-channel signal separation by decorrelation. *IEEE Transactions on Speech and Audio Processing*, 1(4):405–413, October 1993.

[115] S. Wold. Cross validatory estimation of the number of components in factor and principal component analysis. *Technometrics*, 20:397–406, 1978.

[116] C. Wooters and M. Huijberts. The ICSI RT07s speaker diarization system. In *Proceedings of of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop*, 2007. Baltimore, MD.

[117] Chuck Wooters, James Fung, Barbara Peskin, and Xavier Anguera. Toward robust speaker segmentation: The ICSI-SRI fall 2004 diarization system. In *Proceedings of NIST ICASSP 2004 Rich Transcription Workshop (RT-04)*, 2004. Palisades, NY.

[118] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multi-channel audio. *IEEE Transactions on Speech and Audio Processing*, 13(1):84–91, 2005.

[119] Stuart N. Wrigley, Guy J. Brown, Vincent Wan, and Steve Renals. Feature selection for the classification of crosstalk in multi-channel audio. In *Proceedings of EUROSPEECH*, pages 469–472, 2003. Geneva, Switzerland.

[120] Kiyoshi Yamamoto, Futoshi Asano, Takeshi Yamada, and Nobuhiko Kitawaka. Detection of overlapping speech in meetings using support vector regression. In *Proceedings of IWAENC 2005*, pages 37–40, 2005. Eindhoven, The Netherlands.

[121] Robert E. Yantorno. Co-channel speech and speaker identification study. Final Report for Summer Research Faculty Program, September 1998.

[122] Robert E. Yantorno, Kasturi R. Krishnamachari, and Jereme Lovekin. The spectral autocorrelation peak valley ratio (SAPVR) - a usable speech measure employed as a co-channel detection system. In *Proceedings of the IEEE Workshop on Intelligent Signal Processing*, pages 193–197, 2001. Hungary.

[123] Robert E. Yantorno, Brett Y. Smolenski, and Nishant Chandra. Usable speech measures and their fusion. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 734–737, 1999.

[124] Steve Young. Large vocabulary continuous speech recognition: a review. Technical report, Cambridge University Engineering Department, April 1996.

[125] M.A. Zissman, C.J. Weinstein, and L.D. Braida. Automatic talker activity labeling for co-channel talker interference suppression. In *Proceedings of ICASSP*, pages 813–816, 1990. Albuquerque, NM.