Analysis and Transcription of General Audio Data

by

Michelle S. Spina

B.S., Rochester Institute of Technology (1991)
S.M., Massachusetts Institute of Technology (1994)

Submitted to the Department of Electrical Engineering
and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2000

Author .....
Department of Electrical Engineering and Computer Science
May 22, 2000

Certified by .....
Victor W. Zue
Senior Research Scientist
Thesis Supervisor

Accepted by .....
Arthur Smith
Chairman, Departmental Committee on Graduate Students

Analysis and Transcription of General Audio Data

by

Michelle S. Spina

Submitted to the Department of Electrical Engineering
and Computer Science
on May 22, 2000, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In addition to the vast amount of text-based information available on the World Wide Web, an increasing amount of video and audio based information is becoming available to users as a result of emerging multimedia computing technologies. The addition of these multimedia sources of information have presented us with new research challenges. Mature information retrieval (IR) methods have been developed for the problem of finding relevant items from a large collection of text-based materials given a query from a user. Only recently has there been any work on similarly indexing the content of multimedia sources of information.

In this work, we focus on general audio data (GAD) as a new source of data for information retrieval systems. The main goal of this research is to understand the issues posed in describing the content of GAD. We are interested in understanding the general nature of GAD, both lexically and acoustically, and in discovering how our findings may impact an automatic indexing system. Specifically, three research issues are addressed. First, what are the lexical characteristics of GAD, and how do they impact an automatic recognition system? Second, what general sound classes exist in GAD, and how well can they be distinguished automatically? And third, how can we best utilize the training data to develop a GAD transcription system?

In our attempt to answer these questions, we first developed an extensive GAD corpus for study in this work. We collected and transcribed over 100 hours of data for lexical analysis. Ten hours were additionally transcribed for acoustic analysis and recognition experiments. Next, we studied the properties of the GAD vocabulary. This analysis discovered some potential problems for a general large vocabulary continuous speech recognition approach to the transcription of GAD. We found that even for large training set sizes and vocabularies, new words were still regularly encountered. With a training set of nearly one million words (resulting in over 30,000 unique vocabulary words), the out of vocabulary rate was just over 2%. A part-of-speech analysis suggested that the new words were predominately proper nouns and nouns, which would be very important to recognize if we were describing the content of this data. We found that this problem was magnified when we investigated the more realistic scenario of constructing a training set from an out-of-domain source. We then examined the acoustic characteristics of GAD and developed a sound recognition system to segment the audio into its salient sound classes. We subjectively

identified seven acoustically distinct classes based on visual and aural examination of the data. We achieved a 79.4% recognition accuracy for these seven classes on unseen data, using relatively straightforward acoustic measurements and pattern recognition and smoothing techniques. A speech/non-speech recognizer achieved an accuracy of over 92.4%. Next, based on the results of our lexical analysis, we proposed a subword approach to the lexical transcription of GAD. Specifically, we developed a phonetic recognizer for GAD and investigated the use of environment-specific models. Overall, we found that a multiple recognizer system achieved performance similar to a single recognizer, trained in a multi-style fashion. Upon closer inspection of the results, we found that the multi-style system primarily benefitted from the increased amount of data available for training.

Thesis Supervisor: Victor W. Zue
Title: Senior Research Scientist

Acknowledgments

First, I would like to sincerely thank my thesis advisor, Victor Zue, for his guidance, encouragement, support, and patience throughout my years in the SLS group. In addition to being a very helpful advisor, he has put together a phenomenal group of people and cultivated a wonderful atmosphere in which to work. I would also like to thank my thesis committee, consisting of Victor Zue, Jim Glass and Ken Stevens. They provided valuable advice at my committee meetings, and on drafts of the written thesis. I thank them for their time and understanding of my circumstances.

I would also like to mention a few professors at RIT who encouraged me to go to graduate school in the first place. Profs. Unnikrishnan, Salem, Walker and Madhu, thank you for your encouragement those many years ago.

Many people in the SLS group deserve sincere thanks. I have had a number of officemates who have provided invaluable advice and hours of fun over the years. I would especially like to mention Alex Manos for being a wonderful officemate and friend, and Mike McCandless for his friendship and never ending help with SAPPHIRE. Thanks also to Jane Chang, Giovanni Flammia, Chao Wang, Jon Yi, Kenney Ng, Grace Chung, Lee Hetherington, TJ Hazen, Stephanie Seneff, Ray Lau, Joe Polifroni, Jim Glass, Sally Lee and Vicky Palay for all of your help, and for making SLS an incredibly fun place to work.

Thanks also to my family, who have always been very supportive of my graduate career, and for never (or at least rarely) asking, "Aren't you done yet?". Their love and support has meant a lot to me through the years. Thanks especially to Mom and Carl for always helping above and beyond the call of duty.

Finally, I have to somehow try to thank my husband Robert. He has been a pillar of support over the past 9 years. He always knew that I could do this, even when I did not. Thank you my love, for everything. Finally, I'd like to thank our new son James, for giving me the kick in the pants that I needed to finish this thesis.

This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center, by a research contract from BellSouth Intelliventures, and by a fellowship from Intel Corporation.

To Robert and James

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The last few years have been an exciting time in the "information age." We have seen an enormous growth in the amount of information available electronically to users, and as the popularity of the World Wide Web continues to grow, we will continue to see further increases. Until recently, the vast majority of this information has been text-based, from sources such as quarterly reports, text-based web pages, catalogs, theses, conference proceedings, weather reports, etc. Recently, in addition to the increase in the *amount* of information available to users, we have also seen an increase in the *type* of information available. In addition to text-based data, images, video and audio data from sources such as television, movies, radio and meeting recordings are fast becoming available. Access to these multimedia sources of information would allow us to fulfill such requests as "Play me the speech in which President Kennedy said 'Ich bin ein Berliner'," "Show me the footage from the last presidential debate," or "Excerpt Victor's conclusions from the last staff meeting."

These multimedia sources of information have presented us with new research challenges. Much research has been done on the problem of selecting relevant documents from a large collection of text-based materials [38, 66, 68]. Traditionally, key words present in the text documents are used to index and describe the content of the documents, and information retrieval techniques have been developed to efficiently search through large collections of such data. Only recently has there been work addressing the retrieval of information from other media such as images, audio, video or

speech [17, 21, 39, 47, 62]. Unlike text-based data, however, multimedia data sources do not have such a direct way to index or describe their content. In particular, audio materials from sources such as recorded speech messages, radio broadcasts and television broadcasts have traditionally been very difficult to index and browse [71]. However, these audio materials are becoming a large portion of the available data to users. Therefore, it is critical to the success of future information systems to have an ability to automatically index and describe the content of this information.

In this chapter, we first provide background into the area of information retrieval. Next, we describe some of the challenges presented by the inclusion of audio as a data source. We then review areas of related research, and describe the goals and contributions of this thesis. Finally, we give a chapter by chapter overview of the thesis.

1.1 Introduction to Information Retrieval

The goal of an information retrieval (IR) system is to retrieve relevant documents from a stored database in response to a user's request. The user is not looking for a specific fact but is interested in a general topic or subject area and wants to find out more about it. For example, a user may be interested in receiving articles about "the 1992 presidential debates" from a collection of newspaper articles. The goal of the IR system is to inform the user of the existence of documents in the stored data collection that are relevant to his or her request.

Figure 1-1 illustrates the three major components of a typical IR system, enclosed in the dashed box. First, all of the documents in the database must be compactly represented. This representation must capture all of the important information contained in the document, in a form that is compatible with the retrieval process. This is the *indexing* process. Similarly, the user's request must undergo a transformation that extracts the important information present in the request and converts it to a form that is compatible with the retrieval system. This is the *query formation* process. Finally, the *retrieval* system compares the query with the indexed information

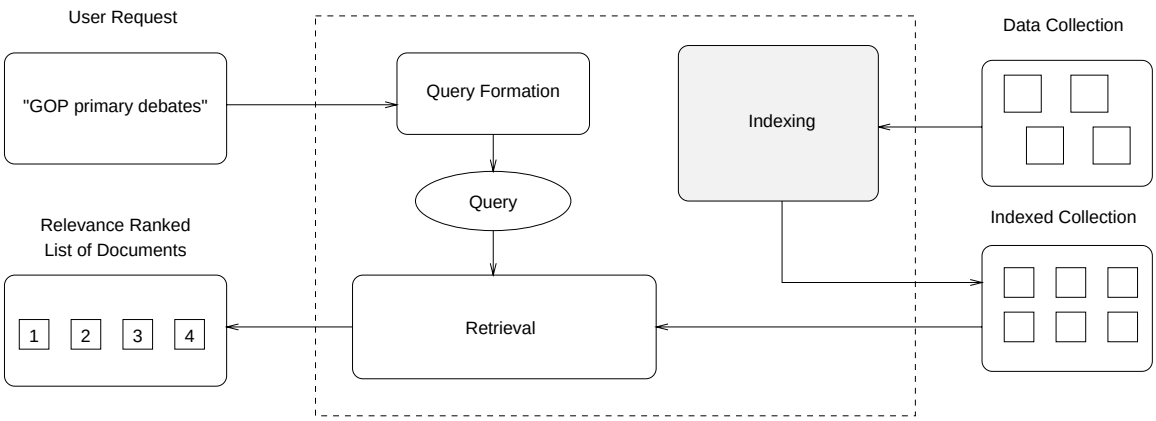

Figure 1-1: Illustration of the major components of an information retrieval system.

documents and retrieves those that are found to be relevant.

While the vast majority of documents used by current information retrieval systems are text in nature, in theory, there is no restriction on the type of documents that can be handled. However, traditional IR models are not well suited to other media sources. Traditional IR models [68] represent the documents and queries as vectors, where each vector component is an indexing term. For text-based documents, the terms are typically the words present in the document. Each term has an associated weight based on the term's occurrence statistics both within and across documents. The weight reflects the relative discrimination capability of that term. A similarity measure between document and query vectors is then computed. Using this similarity measure, documents can then be ranked by relevance and returned to the user. Powerful IR techniques have been developed to accomplish this task [68]. Unlike text-based data, however, other types of media such as images or audio do not have such a direct way to describe their content. Methods to automatically derive indexing terms from non-text documents are required to facilitate their inclusion into IR systems.

1.2 Describing the Content of General Audio Data

In this work, we focus on general audio data (GAD) as a new source of data for information retrieval systems. Given that the amount of GAD as an information

source continues to grow, the development of methods to index the content of this data will become more important. A manual approach to this problem seems intractable, due to its tedious and time-consuming nature. Therefore, automatic methods of producing indices are desirable. In addition to providing a mechanism to include GAD documents into an IR system, the generation of indices for GAD will facilitate access to this data in rich new ways. By nature, audio is difficult to browse and search. Traditionally, to browse audio, one is restricted to real time, sequential listening. Indices to important acoustic cues in the audio would allow users to listen to just those portions of a long discussion which involve a given subset of speakers, or to instantly skip ahead to the next speaker.

General audio data from sources such as radio, television, movies, meeting recordings, etc., can be characterized by a variety of acoustic and linguistic conditions. The data may be of high-quality, full-bandwidth, or it may have been transmitted over a telephone channel. The speech material may be interspersed with music or other interfering sounds. The speech is produced by a wide variety of speakers, such as news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, or non-native speakers. The linguistic style ranges from prepared, read speech to spontaneous speech that may contain incomplete or mispronounced words. GAD may also contain non-speech segments, such as music. To fully describe the content of GAD, a complete audio description must be created in addition to transcribing the speech material. This description should indicate regions of speech and non-speech, identify segments spoken by particular speakers, indicate the speaking environment, etc.

Figure 1-2 illustrates this multi-level description of GAD. In this example, two speakers (spkr1 and spkr2 in the figure) present the local news after an introductory musical segment. The second speaker is speaking over background music. A third speaker (spkr3 in the figure) then presents the weather report. In addition to providing a transcription of the words spoken (e.g., "Good morning, today in Boston..."), a complete description of this simple example should indicate major acoustic changes with appropriate labels (e.g., music segments, specific speaker segments, speaking

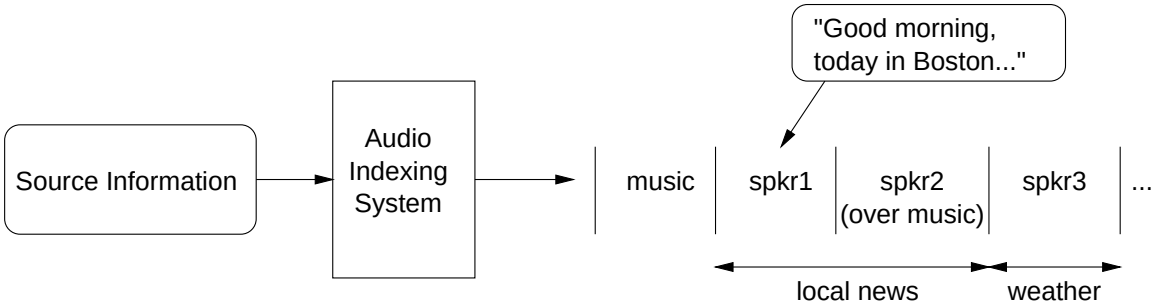

Figure 1-2: Describing the content of GAD.

environment, etc.) and the story topics and boundaries (e.g., local news, weather, etc.).

Therefore, to *fully* describe the content of GAD, an audio indexing system requires two major components. A lexical transcription component is required to generate the linguistic description of the speech data present in the audio. A second transcription component is required to generate the complete acoustic description of the audio data. These components and their subsystems are shown in Figure 1-3. Sections 1.2.1 and 1.2.2 describe these components in more detail.

1.2.1 Transcription of Linguistic Content

The prevailing approach that has been taken in the task of spoken document retrieval is to transform the audio data into text using a large vocabulary speech recognizer and then use a conventional full-text retrieval system [39, 48, 83]. In this approach, the main research challenge is on improving the speech recognition system so it can operate accurately under very diverse acoustic conditions. Although this has been the dominant approach in recent National Institute of Standards and Technology (NIST) sponsored Text REtrieval Conferences (TREC) [26, 27], it has several drawbacks. First, as we illustrate in Chapter 3, the vocabulary generated from GAD is large, diverse, and continually growing [75], while current recognizer technology has practical limits on the size of the recognizer vocabulary. Second, the recognition system must be able to gracefully handle the introduction of new words. As we will show, new vocabulary words are commonly proper nouns, which are important for content

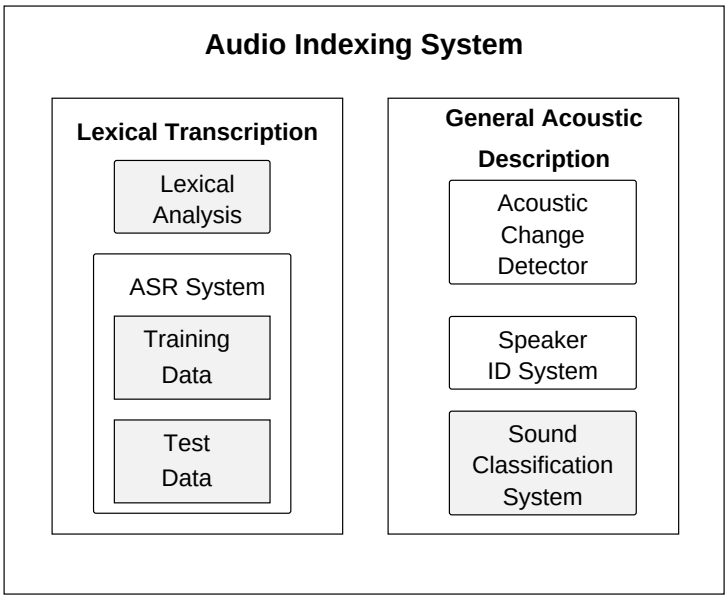

Figure 1-3: Illustration of the major components of an audio indexing system.

description and information retrieval purposes. Finally, recognition systems typically rely on domain-specific data for training the large vocabulary models. Data may not be readily available for this purpose. As illustrated as a subsystem of the lexical transcription component in Figure 1-3, a thorough lexical analysis must be completed to determine what recognition units should be used in a GAD transcription task. In Chapter 3, we perform this analysis to determine whether or not a large vocabulary speech recognition approach is the most appropriate for this task.

In addition to determining *what* recognition units should be used in a GAD transcription system, we must also explore *how* the training and testing data can best be utilized in the automatic speech recognition (ASR) subsystem of the lexical transcription component. Preliminary analysis of GAD has shown that it contains a number of diverse speaking conditions, and that recognizers do not perform equally well in each of these conditions [63, 76]. Segmenting GAD into acoustically homogeneous blocks and using appropriate models for each segment has been shown to improve overall recognition accuracies [76]. Improved recognition results have also been obtained in a word-based approach by clustering the segments and applying adaptation techniques on the resulting homogeneous clusters [10, 51, 73, 81, 85]. In Chapter 5, we explore

different training and testing methods to determine how we can best utilize the GAD data.

1.2.2 Description of General Acoustic Content

While an automatic speech recognition system can provide us with the linguistic content of GAD, the collection of possible audio signals is much wider than speech alone. Considering the range of sounds that people might want access to (e.g., different musical genres, sound effects, animal cries, synthesizer samples), it is clear that purely speech-based methods alone are inadequate to fully describe the content of GAD [21]. An indexing system for GAD requires an additional component that is capable of indexing the variety of acoustic events that occur in the data. When examining GAD we quickly see that there are many different levels of segmentation that can be constructed. We can visualize a very coarse segmentation that indicates boundaries between speech and non-speech sounds, another that indicates boundaries between different background acoustic environments, another that indicates boundaries between different speakers, etc. Each of these possible segmentations is useful for different reasons and should be included in any representation of GAD. As shown in Figure 1-3, a number of subsystems would be required to construct this full acoustic description. For example, an acoustic change detector is required to mark instances of significant acoustic differences between neighboring segments. A speaker identification system is required to label segments spoken by particular speakers. A sound classification system is required to label segments with specific sound tags (e.g., music, speech over the telephone, silence, etc.). In addition to providing valuable information to the acoustic description system, this sound classification system could also provide useful information to the speech transcription system. As we mentioned in the previous section, ASR systems can benefit from knowledge of the speaking environment.

Providing a complete multi-level representation of GAD is beyond the scope of this work. Here, we concentrate on segmenting the data into a single-level representation, based on general acoustic classes. This is explored in Chapter 4. In addition to

contributing to the general acoustic description of GAD, this segmentation may also be useful for the automatic speech recognition system. This topic is explored in Chapter 5.

1.3 Related Research

Although the topic of describing the content of GAD is relatively new, there has been some work in this area recently. In this section, we review approaches that have been taken in the automatic speech recognition of GAD, speaker segmentation and identification, music and audio analysis, and audio interfaces.

1.3.1 Automatic Speech Recognition of GAD

A number of research sites have begun to address the problem of transcribing GAD. The Defense Advanced Research Projects Agency (DARPA) sponsored Hub-4 [61] task was developed in 1995 with the purpose of encouraging research in the ability of ASR systems to adapt to varying conditions of input, whether in acoustic characteristics or content. Hub-4 was also designed to focus on the problems of processing speech materials which have not been created specifically for the purpose of speech system development or evaluation, such as television or radio. The corpus developed for this transcription task consists of a collection of audio files from a number of television and radio news shows such as ABC *Nightline News*, CNN *Headline News*, and National Public Radio *Marketplace*. Each of these shows contain a wide variety of speaking styles (e.g., news texts read by anchors or other studio announcers, more casual speech from correspondents and interviewees), and dialects (both regional and foreign-accented English). In addition, the speech material sometimes contains background noise and other channel effects (such as reduced bandwidth speech). A number of research sites have been involved with the Hub-4 transcription task since its inception in 1995, with varying results. All sites have found that word recognition error rates vary widely across different speaking conditions, from 9.9% for clean, wideband read speech to 29.9% for spontaneous, accented speech in the presence of

background noise [61]. These systems typically use a very large vocabulary (on the order of 64,000 words) speech recognition system, and multiple passes with successively more powerful language models.

We have argued that the transcription of GAD would benefit from a preprocessing step that first segmented the signal into acoustically homogeneous chunks [75, 76], because such preprocessing would enable the transcription system to utilize the appropriate acoustic models and perhaps even to limit its active vocabulary. Some participating Hub-4 sites have used some type of segmentation and clustering of the test utterance to improve their recognition results. Following is a brief description of a few of the data segmenting approaches taken for the Hub-4 task.

The BBN Byblos system [51] first segments the monolithic broadcast news input using a context-independent 2-gender phone decoder. The chopped segments are then clustered automatically to pool data from each speaker for a following adaptation step. Gender-dependent, speaker-independent (SI) models are then refined by Speaker Adapted Training (SAT) [2, 59]. The goal of SAT is to model the speaker-specific variation separately from the other phonetically relevant variation of the speech signal. This produces acoustic models with reduced variance relative to the original general models. They found that the SAT-adapted system achieved an overall relative recognition accuracy gain of 10% over their previous SI system.

Woodland et al., using HTK, takes a more supervised approach to their segmentation step [85]. First, the audio data is classified into three broad categories: wide-band speech, narrow-band speech and music. After rejecting the music, a gender-dependent phone recognizer is used to locate silence portions and gender change points, and after applying a number of smoothing rules, the final segment boundaries are determined. After the initial classification of the data, maximum likelihood linear regression (MLLR) [56] adaptation transforms are computed for each class. The MLLR approach translates, rotates and scales the mean vectors of the density functions used by the general acoustic models, so that there is a better match between the models and the class data. After computing the MLLR transforms, the decoding is repeated using the adapted models. This approach is able to discard 70% of the non-speech

material, while only erroneously discarding 0.2% of the speech. Speech segments are then clustered separately for each gender and bandwidth combination for use with MLLR adaptation of SI acoustic models. They found that this segmentation and adaptation approach achieved recognition results at least as good as data-type specific models.

The CMU Sphinx-3 system [73] takes an approach similar to that taken by HTK. Before recognition, the unannotated broadcast news audio is automatically segmented at acoustic boundaries. Each segment is classified as either full-bandwidth or narrow-bandwidth in order that the correct acoustic models may be applied. Segments are then clustered together into acoustically-similar groups, for use in a following MLLR adaptation step.

While many of the Hub-4 sites found that some sort of segmentation was useful as a preprocessing step for a subsequent speech recognition system, they weren't necessarily concerned with the absolute performance of the segmentation system. For example, while the HTK system [85] cited improved recognition results with the use of their segmentation step, it has a limitation in that it only proposes speaker change boundaries when a change in gender is encountered. This would obviously be a problem if an accurate acoustic description of the data was desired in addition to the lexical transcription.

1.3.2 Speaker Segmentation and Identification

Much research has been done in the area of speaker segmentation and speaker identification (speaker ID). This research is applicable in the acoustic description component of a GAD indexing system. Following is a brief description of a few approaches to this task.

Xerox Palo Alto Research Center (PARC) has developed a system to segment and analyze recorded meetings by speaker [49, 82]. A time-line display was developed to show when particular speakers were talking during the meeting, as well as random access to play back a desired portion of the recording. In addition, non-speech segments such as silence and applause were located, providing important cues for a

more general segmentation system. To accomplish this, they created a hidden Markov model (HMM) for each speaker or acoustic class (e.g., music, applause, silence). The HMM for each class was trained using a maximum likelihood estimation procedure. Each class HMM was then combined into a larger network, and the Viterbi algorithm [79] was used to determine the maximum likelihood state sequence through the network to determine the final segmentation of the audio data. They found this system worked very well (error rate of under 1%) when it was used on a formal recorded panel discussion. Results were degraded to a segmentation error of 14%, however, when the system was used on a more informal recorded meeting. The characteristics of the informal meeting were quite different from those of the panel discussion. The panel discussion had speakers speaking in turn (i.e., no interruptions) with average utterance lengths of 20 seconds. In contrast, the recorded meeting had one third of its utterances interrupted, resulting in average utterance lengths of only 3 seconds. In addition, the speakers in the panel discussion were individually miked, while the recorded meeting simply had two microphones placed on the meeting table. This is a potential limitation to their approach, because many GAD sources are likely to be of this more informal nature.

Segmenting audio data based on speaker change can also help multimedia segmentation applications. The detection of speaker changes in the soundtrack of a video or multimedia source may indicate a scene or camera change. Wyse and Smoliar investigate the use of a "novelty measure" based on the cepstral difference between a short and long analysis window [86]. Differences are only computed in similar regions of the feature space to prevent intra-speaker variation. When this difference exceeds a threshold, a speaker change is detected. Chen et al. [11] also used a completely data-driven approach for the segmentation problem. They modeled the input audio stream as a Gaussian process in the cepstral space. A maximum likelihood approach was used to detect turns in this Gaussian process based on the Bayesian information criterion [72]. They analyzed their data in terms of insertions and deletions of boundaries. They achieved a very low insertion rate (4.1%), and a 33.4% deletion rate. The majority of their deletions occurred when a segment was less than 2 sec-

onds in length. These most likely occurred because there wasn't sufficient data to adequately develop the Gaussian model for these segments. However, we have found this to be a potentially serious limitation of this approach. In our analysis of GAD, we found that while the average segment length is over 4.8 seconds, nearly 20% of the segments are less than 2 seconds in length.

1.3.3 Music and Audio Analysis

A very general problem in audio analysis is to detect regions of non-speech. This has a direct implication for speech recognition systems, as non-speech data can be eliminated from computation by the ASR system. Saunders [69] uses a straightforward approach to the discrimination of speech and music. A simple multivariate Gaussian system is trained using features that were determined to discriminate between music and speech. He found that using statistics computed from the zero crossing rate, he could achieve a classification accuracy averaging 90%. By including additional information about the energy contour, he improved his accuracy to 98%. Scheirer and Slaney [70] report similar results on their speech/music discriminator. The discriminator was based on various combinations of 13 features such as 4-Hz modulation energy, zero crossing rate, and spectral centroid. They investigated a number of classification strategies, such as Gaussian mixture models and K-nearest-neighbor classifiers. When looking at long-term windows (2.4 seconds), they achieved an error rate of 1.4%. In Chapter 4, we develop a more extensive sound classification system which classifies audio into one of seven basic sound classes. Using a maximum *a posteriori* approach on mel-frequency cepstral coefficients, we were able to achieve a classification accuracy of over 85%. In addition, we develop a speech/non-speech classifier that achieves an accuracy of over 95%.

1.3.4 Audio Interfaces

A number of research sites have developed complete audio information retrieval systems. The Informedia Digital Library Project [40] at Carnegie Mellon University is

creating a large digital library of text, images, videos and audio data, and are attempting to integrate technologies from the fields of natural language understanding, image processing, speech recognition and video compression to allow a user to perform full content retrieval on multimedia data. An automatic speech recognition system is used on audio data to create time-aligned transcripts of the data. Traditional text-based information retrieval techniques are then used to locate segments of interest. They have found that automatic speech recognition even at high error rates is useful. However, they have also found that a segmentation step that first identifies regions of speech and non-speech would be helpful.

Muscle Fish [60] has developed a retrieval by similarity system for audio files. Their approach analyzed sound files for a set of psychoacoustic features. Attributes computed from the sound files included pitch, loudness, bandwidth, and harmonicity [84]. A covariance-weighted Euclidean distance is then used to measure similarity between audio files. For retrieval, the distance is computed between the given sound sample and all other sound samples. Sounds are then ranked by distance, with the closer ones being most similar.

Foote [22] has developed a similar retrieval system, using a different approach. He computes distance measures between histograms derived from a discriminatively trained vector quantizer. Mel-frequency cepstral coefficients are first computed from the audio data. A learning algorithm then constructs a quantization tree that attempts to put samples from different training classes into different bins. A histogram of an audio file is made by looking at the relative frequencies of samples in each quantization bin. Histograms from different audio samples are then used as the feature vectors for a simple Euclidean measure computation to determine the similarity between them. This approach has also been used by Foote for speaker identification [23, 24], and music and audio retrieval [22].

1.4 Goals and Contributions

As reviewed in the previous section, strides have been made in a number of the components that are required to describe the content of GAD. The Hub-4 task has encouraged speech recognition researchers to tackle the difficult problem of general audio, rather than just clean speech, as has been the case with the majority of past speech recognition research. Audio information retrieval is also now a sub-task of TREC, encouraging those in the text-retrieval community to consider audio as well. Speaker identification and segmentation systems have been successfully used to provide indices into meeting recordings, allowing for easier browsing of the audio, and more general sound segmentation systems have been developed to allow for even more detailed indexing. While the success in each of these subsystems is encouraging, the analysis and transcription of general audio data is still a very new topic. The main goal of this research is to understand the issues posed in describing the content of GAD. We are interested in understanding the general nature of GAD, both lexically and acoustically, and in discovering how our findings may impact an automatic indexing system. Specifically, three research issues are addressed:

1. What are the lexical characteristics of GAD, and how do they impact an automatic speech recognition system?

2. What general sound classes exist in GAD, and how well can they be distinguished automatically?

3. How can we best utilize the training data to develop a GAD transcription system?

In our attempt to answer these questions, we first develop an extensive GAD corpus for study in this work. We collect and transcribe over 100 hours of data for lexical analysis. Ten hours are additionally transcribed for acoustic analysis and recognition experiments. Next, we study the properties of the GAD vocabulary. We are interested in determining the size of the vocabulary and in observing how the vocabulary grows as additional data is encountered. We then look more closely at the

out-of-vocabulary occurrence rate under two conditions. First, we study the best-case scenario, that is, building a vocabulary using task-specific training data. Second, we determine how the out-of-vocabulary rate is affected when we use training data from a similar corpus. A part of speech analysis is then performed to further understand the lexical properties of this data. Shifting to the acoustic description of GAD, we investigate its general acoustic properties to determine what salient sound classes exist in the data. We then study their general characteristics and distributions. Next, we determine how well we can automatically segment the sound stream into these acoustic classes. After we develop a recognition system to accomplish this task, we evaluate the results to determine if our subjectively defined acoustic classes need further refinement. Next, based on the results of our lexical analysis, we propose a subword approach to the lexical transcription of GAD. Specifically, we develop a phonetic recognizer for GAD. Our acoustic analysis revealed that GAD contains a number of different acoustic speaking environments. Since the performance of ASR systems can vary a great deal depending on speaker, microphone, recording conditions and transmission channel, we investigate the use of environment-specific models for the phonetic recognition of GAD.

In this thesis, we make the following contributions to research in the area of GAD analysis and transcription:

- **Development of GAD Corpus:** To complete this work, over 100 hours of data were collected and orthographically transcribed for lexical analysis. Ten hours were additionally transcribed for acoustic analysis and recognition experiments. This corpus will be valuable to others working on research issues in GAD.

- **Lexical Analysis of GAD:** We performed a lexical analysis to understand the general lexical characteristics of GAD. This analysis discovered some potential problems for a general LVCSR approach to the transcription of GAD. We found that even for large training set sizes and vocabularies, new words are still regularly encountered, and that these words are primarily high content words (i.e.,

proper nouns) and therefore would need to be correctly recognized to describe the linguistic content of GAD. We proposed a subword based approach to the recognition of GAD.

- **Acoustic Analysis and Development of Sound Recognition System:** We performed an acoustic analysis to determine what sound classes exist in GAD, and discovered the characteristics of the classes. A sound recognition system was developed which would benefit both an acoustic description system, and a recognition system.

- **Discovery of Optimal Recognition Strategies for GAD:** We investigated a number of different training and testing strategies for the phonetic recognition of GAD. We found that knowledge of the speaking environment is useful for phonetic recognition.

1.5 Overview

The remainder of this thesis is organized in five chapters. Chapter 2 describes the background for the experimental work presented in this thesis. This includes information about the NPR-ME corpus that was developed for this work, information about the TIMIT corpus, and a description of the SUMMIT speech recognition system used in Chapter 5. Chapter 3 describes the lexical analysis completed in our study of GAD. We begin with an analysis of the general orthographic properties of the corpus. Next, we thoroughly examine the behavior of the vocabulary of our GAD corpus, and determine how it compares with similar corpora. Finally, we determine how this behavior may affect the development of an ASR system. Chapter 4 describes the development of our sound recognition system. We begin the chapter with an acoustic analysis of GAD which subjectively determines what sound classes are present in the data. Next, we present the development and results of the sound recognition system. Chapter 5 describes the experiments we conducted in the phonetic recognition of GAD. Specifically, we explore a variety of training and testing methods to determine

how to best utilize our training data, and how to best process test data in a phonetic recognition task. Finally, Chapter 6 summarizes the work, draws some conclusions, and mentions directions for future work.

Chapter 2

Experimental Background

This chapter contains background information for the work presented in the remainder of this thesis. This includes information about the corpora used in the experiments and a description of the SUMMIT speech recognition system.

2.1 NPR-ME Corpus

We have chosen to investigate the nature of GAD by focusing on the National Public Radio (NPR) broadcast of the *Morning Edition* (ME) news program. NPR-ME is broadcast on weekdays from 6 to 9 a.m. in the US, and it consists of news reports from national and local studio anchors as well as reporters from the field, special interest editorials and musical segments. We chose NPR-ME after listening to a selection of radio shows, noting that NPR-ME had the most diverse collection of speakers and acoustic conditions and would therefore be the most interesting for study.

The following sections describe the data collection and processing procedures that were followed to develop the NPR-ME corpus, and a description of the data sets that were created for use in the analysis and experiments presented in this thesis.

2.1.1 NPR-ME Data Collection and Processing

The NPR-ME data was recorded from an FM tuner onto digital audio tape at 48 kHz. Since some of the news segments are repeated hourly, we chose to record approximately 60 minutes of the program on a given day. A copy of the original recordings was given to a transcription agency in Cambridge, Massachusetts, who produced orthographic transcriptions of the broadcasts in electronic form. In addition to the words spoken, the transcripts also include side information about speaker identity, story boundaries, and acoustic environment. The convention for the transcription follows those established by NIST for the spoken language research community. The details of the transcription convention can be found in Appendix A.

Table 2.1 shows an example segment of a transcribed show. This example starts with an introduction by speaker "A", speaking over background music, which is indicated with the [music/] tag. This introduction is followed by a segment of music (e.g., [musical_interlude]), then speaker "A" continues with the introduction, again speaking over background music. The full introduction is marked with start and end story tags. Following the introduction, a second speaker begins another story. Words that are unclear in the broadcast are indicated in the transcription with surrounding double parentheses (e.g., ((Hausman))). Words whose proper spelling is unknown are indicated with a preceding "@" symbol (e.g., @Cassell). These words are later manually reviewed and a common spelling is adopted for all of the broadcasts.

An additional file containing specific information about the speakers found in the show was also generated by the transcription agency. For each speaker, their name, role (e.g., NPR-ME anchor, BBC reporter, traffic reporter, etc.), gender and age (adult or child) was specified. If the speaker's name was not provided in the broadcast, it was indicated as "unknown".

Shows were further processed for use in our acoustic analysis and sound recognition system development presented in Chapter 4, and speech recognition system development presented in Chapter 5. First, the hour long shows were downsampled to 16 kHz and transferred to computer disk. A DAT-Link+ [77] digital audio interface

<broadcast id="morning_edition.121396">

<story id=1 topic="Upcoming Headlines on Morning Edition">
A: [music/] President Clinton is expected to fill more cabinet positions at a news conference later today. Attorney General Janet Reno is said to be staying on board. It's Friday, the Thirteenth of December. This is Morning Edition. [/music]

[musical_interlude]

A: [music/] This hour on Morning Edition, world trade talks between one hundred twenty eight nations underway in Singapore produce lower prices for computers. The European Union bickers over what its money should look like. And new evidence suggests babies learn language much earlier than believed. Cloudy today. Some drizzle possible this morning. Rain this afternoon. In the forties today. At ninety point nine, this is W_B_U_R. [/music]
</story>

<story id=2 topic="Hostage Stand-Off in Paris">
B: From National Public Radio News in Washington, I'm Carl @Cassell. A hostage stand-off is underway in Paris. French police say a gunman has taken about thirty-five people hostage at an office building in the French capital. He reportedly has shot and wounded two people. A police spokesman says the man took the group hostage on the third floor of a building on Boulevard ((Hausman)). Police say the man is fifty-five years old and had a grievance against his former employer, a security delivery firm.
</story>

Table 2.1: Example segment of a transcribed NPR-ME program.

was used to downsample and transfer the data. Since our analysis and recognition tools were unable to accommodate files of such length, each of the hour-long data files were automatically segmented into manageable sized waveform files at silence breaks. Boundaries between non-silence and silence were proposed at locations where the value of the average energy, computed every 5 ms using a 5 ms window, exceeded a threshold. The threshold was determined manually by examining the resulting waveform files of a sample show.

Set	# Unique Speakers	# Hours
Train	286	8
Test	83	2

Table 2.2: Number of speakers and hours of data in the NPR-ME training and test sets used for analysis and system training and testing.

2.1.2 NPR-ME Data Sets

A total of 102 hours of NPR-ME was recorded and transcribed between July, 1995 and July, 1998. All 102 recorded and transcribed shows were used for the lexical analysis presented in Chapter 3. Ten of the shows were transferred to computer disk and segmented into waveform files as described in Section 2.1.1. This data was divided into two sets: one for training and tuning the sound classification and speech recognition systems, and another for use as test data. The test set was comprised of two shows chosen at random from the collection of ten shows. The training set was comprised of the remaining eight shows. Table 2.2 indicates the number of unique speakers and the number of hours of data in the NPR-ME training and test sets.

Because the NPR-ME data consists of different broadcasts of the same radio show, there will be some recurring speakers, such as the studio announcers, from one show to another. As a result, there will be a number of speakers that appear in both the training and test set. This may give rise to improved speech recognition results for those utterances in the test set spoken by speakers found in the training set. Therefore, for speech recognition purposes, we can not consider the NPR-ME test data to be speaker independent. Rather, the data should be considered as multi-speaker.

2.2 TIMIT Corpus

To facilitate comparison of the NPR-ME recognition results with other phonetic recognition results, experiments were performed on the commonly used TIMIT corpus [20, 25, 54]. The following sections describe the sets used in training and testing

Set	# Speakers	# Utterances	# Hours
Train	462	3,696	3.14
Core Test	24	192	0.16

Table 2.3: Number of speakers, utterances, and hours of speech in the TIMIT training and core test sets.

and the phones used in the TIMIT transcriptions.

2.2.1 TIMIT Data Sets

The TIMIT acoustic-phonetic continuous speech corpus contains speech from 630 speakers representing eight major dialects of American English, each speaking ten phonetically-rich sentences. There are 438 male speakers and 192 female speakers. Each speaker read ten utterances, which included two dialect sentences (SA) designed to reveal the dialectical differences among the speakers, five phonetically compact sentences (SX) designed to cover all phoneme pairs, and three phonetically diverse sentences (SI) selected from existing text sources.

NIST has divided the SX and SI data into independent training and test sets that do not overlap either by speaker or by sentence [20, 25, 54]. The core test set contains 192 SX and SI utterances read by 24 speakers (two male speakers and one female speaker from each of the eight dialects). All TIMIT experiments in this thesis report error rate on this set.

The NIST "complete" test set contains 1344 SX and SI utterances read by the 168 speakers who read any of the core test sentences. The training set consists of the 462 speakers which are not included in either the core or the complete test set. There is no overlap between the utterances read by the training and testing speakers.

Table 2.3 summarizes the number of speakers, the number of utterances, and the number of hours of speech in the TIMIT data sets used in this thesis.

2.2.2 TIMIT Phones

TIMIT was phonetically transcribed using a set of 61 phones [54]. Table 2.4 shows these phones with their corresponding IPA and ARPAbet symbols, followed by an example sound. This phone set was also used to phonetically transcribe the NPR-ME data.

2.3 Speech Recognition System

The SUMMIT speech recognition system, developed by the MIT Laboratory for Computer Science's Spoken Language Systems Group [32], was used to complete the recognition experiments presented in Chapter 5. The system uses a probabilistic segment-based approach that differs from conventional frame-based hidden Markov model approaches [65]. In frame-based approaches, speech is represented as a temporal *sequence* of feature vectors. The feature vectors are typically computed at a fixed rate, such as every 10 ms. In segment-based approaches, speech is represented as a temporal *graph* of variable-length segments. Acoustic features extracted from these segmental units have the potential to capture more of the acoustic-phonetic information encoded in the speech signal, especially those that are correlated across time. To extract these acoustic measurements, explicit segmental start and end times are needed. The SUMMIT system uses a segmentation algorithm [31] to produce the segmentation hypotheses. First, a spectral representation of the signal is computed every 5 ms using a 21 ms analysis window. Major segment boundaries are hypothesized at locations where the spectral change between neighboring measurements exceeds a pre-defined global threshold. Then, minor boundaries are hypothesized between the major boundaries based again on spectral change, but this time using a local threshold that is computed from the signal between the major boundaries. Finally, all boundaries between the major boundaries are fully interconnected to form a network of possible segmentations on which the recognition search is performed. The size of this network is determined by the pre-defined global threshold.

Traditional frame-based approaches compute measurements every frame from the

IPA	ARPAbet	Example	IPA	ARPAbet	Example
[ɑ]	aa	b*o*b	[ɨ]	ix	deb*i*t
[æ]	ae	b*a*t	[iʸ]	iy	b*ee*t
[ʌ]	ah	b*u*t	[j]	jh	*j*oke
[ɔ]	ao	b*ough*t	[k]	k	*k*ey
[ɑʷ]	aw	b*ou*t	[k□]	kcl	k closure
[ə]	ax	*a*bout	[l]	l	*l*ay
[əʰ]	ax-h	p*o*tato	[m]	m	*mom*
[ɚ]	axr	butt*er*	[n]	n	*noon*
[ɑʸ]	ay	b*i*te	[ŋ]	ng	si*ng*
[b]	b	*b*ee	[ɾ̃]	nx	wi*nn*er
[b□]	bcl	b closure	[oʷ]	ow	b*oa*t
[č]	ch	*ch*oke	[oʸ]	oy	b*oy*
[d]	d	*d*ay	[p]	p	*p*ea
[d□]	dcl	d closure	[□]	pau	pause
[ð]	dh	*th*en	[p□]	pcl	p closure
[ɾ]	dx	mu*dd*y	[ʔ]	q	glottal stop
[ɛ]	eh	b*e*t	[r]	r	*r*ay
[l̩]	el	bott*le*	[s]	s	*s*ea
[m̩]	em	bott*om*	[š]	sh	*sh*e
[n̩]	en	butt*on*	[t]	t	*t*ea
[ŋ̍]	eng	Wash*ing*ton	[t□]	tcl	t closure
[▯]	epi	epenthetic silence	[θ]	th	*th*in
[ɝ]	er	b*ir*d	[ʊ]	uh	b*oo*k
[eʸ]	ey	b*ai*t	[uʷ]	uw	b*oo*t
[f]	f	*f*in	[ü]	ux	t*oo*t
[g]	g	*g*ay	[v]	v	*v*an
[g□]	gcl	g closure	[w]	w	*w*ay
[h]	hh	*h*ay	[y]	y	*y*acht
[ɦ]	hv	a*h*ead	[z]	z	*z*one
[ɪ]	ih	b*i*t	[ž]	zh	a*z*ure
-	h#	utterance initial and final silence			

Table 2.4: IPA and ARPAbet symbols for phones in the TIMIT corpus with example occurrences

speech signal, which results in a sequence of observations. Since there is no overlap in the observations, every path through the network accounts for all observations. However, segment-based measurements computed from a segment network lead to a network of observations. For every path through the network, some segments are on the path, and some are off the path. To maintain probabilistic integrity when comparing different paths it is necessary for the scoring computation to account for all observations by including both on-path and off-path segments in the calculation. This is accomplished using a single non-lexical acoustic model, referred to as the "not" model, or the "antiphone," to account for all off-path segments [32].

The recognizer uses context-independent segment and context-dependent boundary (segment transition) acoustic models. The feature vector used in the segment models has 77 measurements consisting of three sets of 14 Mel-frequency cepstral coefficient (MFCC) and energy averages computed over segment thirds, two sets of MFCC and energy derivatives computed over a time window of 40 ms centered at the segment beginning and end, log duration, and a count of the number of internal boundaries proposed in the segment. The boundary model feature vector has 112 dimensions and is made up of eight sets of MFCC averages computed over time windows of 10, 20, and 40 ms at various offsets ($\pm 5, \pm 15$, and ± 35 ms) around the segment boundary. Cepstral mean normalization [1] and principle components analysis [78] are performed on the acoustic feature vectors.

The distribution of the feature vectors is modeled using mixture distributions composed of multivariate Gaussian probability density functions (pdf) [64]. In the experiments presented in this work, the covariance matrix of the Gaussian pdf was restricted to be diagonal. In comparison with full covariance models, the use of diagonal models allows the use of more mixture components because there are many fewer parameters to train per component. The number of mixture components is determined automatically based on the number of training tokens available.

The Gaussian mixtures were trained by a two-step process. In the first step, the K-means algorithm [19] was used to produce an initial clustering of the model feature vectors. In the second step, the results of the K-means algorithm were used

to initialize the Expectation-Maximization (EM) algorithm [16, 19] which iteratively maximizes the likelihood of the training data and estimates the parameters of the mixture distribution. The EM algorithm converges to a local maximum, with no guarantee of achieving the global optimum. Therefore, the EM algorithm is highly dependent on the initial conditions obtained from the K-means algorithm. In order to improve the robustness of the mixture models, a technique called aggregation [41] was used. Specifically, five separate acoustic models were trained using different initializations, and these models were then combined into a single, larger model using a simple linear combination with equal weights for each model.

To determine the final hypothesis string, a forward Viterbi search [79] with a statistical bigram language model was used to determine the final recognition hypothesis.

2.3.1 Performance Evaluation

Speech recognition performance is typically measured in terms of the error rate (in percent) resulting from the comparison of the recognition hypotheses with the reference transcriptions. All phonetic error rates are computed using the NIST alignment program [20]. This program finds the minimum cost alignment, where the cost of a substitution is 1.0, and the cost of a deletion or an insertion is 0.75. The total recognition error rate is computed from the sum of the substitution, deletion and insertion errors that occur. Following convention, recognition results are reported in terms of error rate in this thesis. In accordance with common practice [55], we collapsed the 61 TIMIT labels into 39 labels before computing error rates. This mapping is shown in Table 2.5.

2.4 Summary

In this chapter we have presented background information for the work presented in the remainder of this thesis. To investigate the nature of GAD, we developed a corpus from the National Public Radio broadcast of the *Morning Edition* news program. Over 100 hours of data was recorded and orthographically transcribed. Ten of the

1	iy	20	n en nx
2	ih ix	21	ng eng
3	eh	22	v
4	ae	23	f
5	ax ah ax-h	24	dh
6	uw ux	25	th
7	uh	26	z
8	ao aa	27	s
9	ey	28	zh sh
10	ay	29	jh
11	oy	30	ch
12	aw	31	b
13	ow	32	p
14	er axr	33	d
15	l el	34	dx
16	r	35	t
17	w	36	g
18	y	37	k
19	m em	38	hh hv
39	bcl pcl dcl tcl gcl kcl q epi pau h# not		

Table 2.5: Mapping from 61 classes to 39 classes for scoring of results, after Lee[55].

shows were further processed for use in the acoustic study and sound recognition experiments presented in Chapter 4 and the phonetic recognition experiments presented in Chapter 5.

In addition to describing the details of the NPR-ME corpus, we also described the TIMIT corpus, which will be used for comparison in our phonetic recognition work. Finally, the details of the SUMMIT speech recognition system were presented.

Chapter 3

Lexical Analysis

One objective of this thesis is to understand the nature of general audio data. Two aspects of GAD that we are particularly interested in are its lexical and acoustic properties. In this chapter, we will present a study of the nature of the lexical properties of GAD, while the acoustic properties will be studied in Chapter 4.

First, we study the general characteristics of GAD to gain a better understanding of the data, and to see how this data compares with that typically used in the ASR community. Next, we study the properties of the GAD vocabulary. We are interested in determining the size of the NPR-ME vocabulary and in observing how the vocabulary grows with time. The vocabulary growth characteristics will indicate if it is likely that new words will be encountered as more data is accumulated. We then look more closely at the out of vocabulary occurrence rate under two conditions. First, we study the best-case scenario, that is, building a vocabulary using task-specific training data (i.e., training data collected from NPR-ME). Second, we determine how the out of vocabulary rate is affected when we use training data from a similar corpus (i.e., training data collected from news broadcasts other than NPR-ME). Finally, we look more carefully at the cumulative, common and out of vocabulary vocabularies to determine their part of speech characteristics. We are interested in determining if these vocabularies contain common words that could be obtained from a standard lexicon, or if they contain very task-specific words that would require a vast amount of task-specific data to construct a vocabulary for this corpus.

In addition to providing us with a greater understanding of the lexical properties of GAD, the results of this analysis are important for the development of a large vocabulary speech recognition system (LVCSR). As we outlined in Chapter 1, a transcription component is required to generate a linguistic description of the speech data present in GAD. The prevailing approach for this task is the use of a LVCSR system to convert the speech data to text. In the development of a recognition system one important consideration is the system's vocabulary. Specifically, it is important to understand the lexical properties of the data to be transcribed. We must determine the size of the vocabulary to verify that it is within the capabilities of current speech recognition technology. In addition, we must determine how the recognizer vocabulary should be constructed. Can we use a standard lexicon to develop the vocabulary, or is a large amount of task-specific data required to fully cover the range of words encountered in GAD? Are new words likely? Would they be difficult to obtain from standard sources? The results of the analysis presented in this chapter will attempt to answer these questions.

3.1 Other Corpora

In addition to the NPR-ME corpus described in Chapter 2, we examined the orthographic transcriptions of three other corpora in the experiments presented in this chapter. The additional corpora, which are similar in nature to the NPR-ME corpus, were included to facilitate comparisons with the NPR-ME data, and to study cross-corpus effects. Specifically, one additional speech-based corpus (Hub4) and two text-based corpora (WSJ and LA-Times) were used. The Hub4 corpus is similar to NPR-ME, consisting of utterances from 97 hours of recordings from radio and television news broadcasts, gathered between June 1997 and February 1998, from sources such as ABC *World News Tonight*, CNN *Headline News*, PRI *The World*, etc. This data contains both prepared and spontaneous speech, as does the NPR-ME data. This data was provided by the Linguistic Data Consortium (LDC) [57] to supplement the 1996 Broadcast News Speech collection. While both the NPR-ME

and Hub4 corpora were collected as speech, only the orthographic transcriptions were used in the experiments in this chapter.

The WSJ [18] and LA-Times corpora consist of text from the *Wall Street Journal* and *Los Angeles Times* newspapers, respectively. The text for WSJ was made available by the ACL Data Collection Initiative [4] and represents three years (1987-1989) of newspaper text. The LA-Times data was used in the text retrieval task in TREC-6 [38], and represents two years (1989-1990) of newspaper text.

3.2 Data Preparation and Vocabulary Creation

The orthographic transcriptions from the NPR-ME and Hub4 corpora, and the text from the WSJ and LA-Times corpora required processing with regard to capitalization, punctuation, numbers and compound words. Since case specification in speech recognition is meaningless (i.e., Bill and bill are acoustically indistinguishable to an ASR system), all words were converted to lower case. All punctuation was removed, except for the apostrophe, which was left alone. Therefore, words like "couldn't" and "Robert's" remained as they were transcribed. Spoken numbers, such as "fifty-three" were broken into their component words (e.g., "fifty" and "three"), to prevent the artificial creation of a large number of words. However, compound words created from a string of letters (e.g., U.S.A., F.B.I., etc.) were left as is since they are commonly used words.

With such large corpora, there are bound to be spelling errors, and we generally did not attempt to correct them. The exception to this was the tagged words in the NPR-ME corpus. Recall that the transcribers tagged a word with an unknown spelling with a preceding "@" symbol. All of the NPR-ME transcriptions were modified to generate a common spelling for each of the "@" words. For example, the words "@Karl" and "@Carl" were changed to the common spelling "Carl" to prevent the artificial creation of additional words due to spelling differences.

Since two of the topics under investigation are the behavior of vocabularies and out of vocabulary occurrences, it is important to understand how words are defined,

and how vocabularies are created. After the previously mentioned processing, a word is defined to be a string of characters deliminated by spaces. Vocabularies were determined automatically by processing a collection of text (the training set), and placing all words that occur at least n times in the vocabulary list. For all of our experiments, $n = 1$, meaning that our vocabulary consisted of all unique words in the training set.

3.3 General Analysis

In this section we are interested in discovering the general characteristics of GAD and comparing them with data typically used in the ASR community. To accomplish this, we studied the transcriptions for the 102 NPR-ME shows. As we described in Chapter 2, in addition to the words spoken, the transcribers also noted musical segments, speakers and story boundaries. To compute the timing information presented here, we also utilized results from the acoustic analysis presented in Chapter 4.

Table 3.1 shows the general characteristics of an NPR-ME show. The number of music segments, speakers, stories and turns were computed from the transcriptions of each show. There are an average of 5.5 musical segments, which we found usually occur at story boundaries. The number of speakers for an hour-long show ranges from 21 to 65, with an average of more than 43 per show. Since there are about 28 stories in a show, each story typically involves 2-3 speakers. In this analysis, we define a turn to be a continuous segment of speech spoken by a given speaker. We found that there are over 150 such turns in an average NPR-ME hour-long show.

Figure 3-1 shows the distribution of speech and non-speech (music, silence, etc.) in the NPR-ME data. We found that the fraction of a typical show containing speech was approximately 89%, or just over 53 minutes. This suggests that each turn (i.e., a contiguous segment of speech spoken by a given speaker) is just over 21 seconds. The speaking rate, inferred from the number of word tokens (nearly 10,000) and the fraction of the show containing speech, is about 180 words per minute.

The characteristics of the NPR-ME corpus (and the similar Hub4 corpus described

	Average (ave ± std)
# music segs	5.5 ± 3.5
# speakers	43.4 ± 6.1
# stories	27.5 ± 5.3
# turns	151.0 ± 19.3
# words spoken	9706.9 ± 590.7
# vocab words	2605.1 ± 227.9

Table 3.1: Summary of general characteristics of the NPR-ME corpus, averaged over 102 shows. The table indicates the average value and standard deviation for each entry.

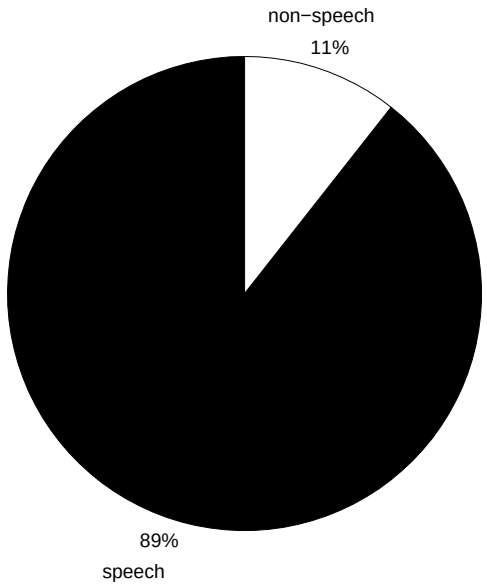

Figure 3-1: Distribution of speech and non-speech in the NPR-ME data.

in Section 3.1) are very different than those of other corpora collected by the speech recognition community. Three general types of data have typically been used in speech recognition research. First, data has been collected for core recognition development. The TIMIT corpus, described in Chapter 2, is an example of this type of data. This data is recorded over controlled acoustic conditions, using a high-quality, noise canceling microphone. Single utterances are read by each speaker, and there are no spontaneous speech effects or non-speech events. The average utterance duration is 3.0 seconds. Second, data has been collected for human/computer interactive problem solving development. The JUPITER corpus, developed at the MIT Labo-

Corpus	Recording Method	Channel Conditions	Non Speech	Type of Speech	Presentation Style
TIMIT	known	high quality	no	read	one-sided
JUPITER	unknown	telephone	yes	spont	one-sided
SWITCHBOARD	known	telephone	no	spont	dialog (2 speakers)
Hub4	unknown	mix	yes	mix	dialog (many speakers)
NPR-ME	unknown	mix	yes	mix	dialog (many speakers)

Table 3.2: Summary of the characteristics of corpora collected for use in speech recognition system development. Characteristics considered are: recording method (i.e., do we know *a priori* how the data was recorded - known or unknown), channel conditions (high quality, telephone, background noise, or mix of conditions), presence of non-speech events (yes or no), type of speech (read, spontaneous, or mix), and presentation style (one-sided or multiple speakers).

ratory of Computer Science Spoken Language Systems Group, is an example of this type of data [33, 87]. JUPITER is a telephone-based weather information system, which allows a user to access and receive on-line weather information over the phone. The data collected for the development of this system contains spontaneously spoken utterances from both novice and expert users. Non-speech events are occasionally encountered in this data, and the average utterance length is 3.3 seconds. Third, data has been collected to study human/human dialogs. The SWITCHBOARD corpus is an example of this type of data. It consists of spontaneous human/human dialogs collected by Texas Instruments [34]. The data was collected under known acoustic conditions (telephone), and contains spontaneous speech between two speakers engaged in dialog. Table 3.2 summarizes the characteristics of the corpora described here.

Unlike the TIMIT, JUPITER, and SWITCHBOARD corpora described above, our analysis of GAD suggests that it contains a broad collection of speaking conditions, both speech and non-speech segments, and multiple speakers speaking in turn. Like the SWITCHBOARD corpus, the speakers are not directly interacting with a speech recognition system. These characteristics make GAD a more challenging data set from a speech recognition stand-point.

There are a few speakers that appear regularly in the NPR-ME corpus, such as the national and local hosts. While no single speaker dominates the broadcast, the

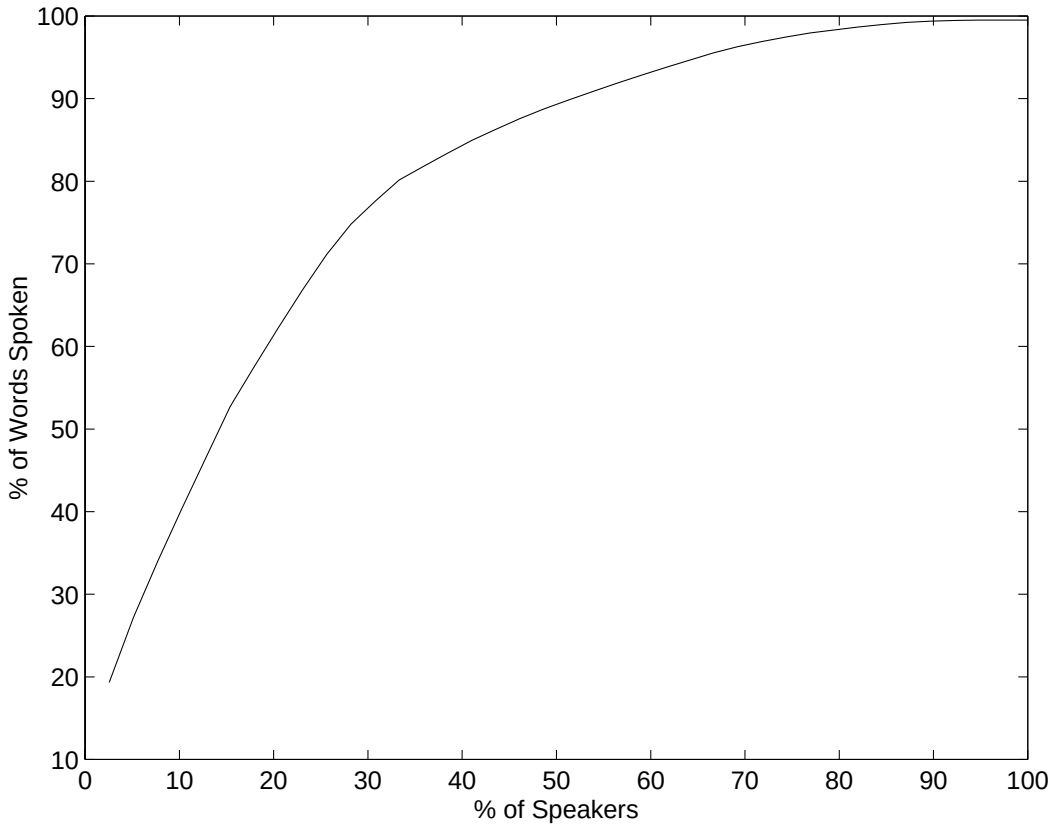

Figure 3-2: The percentage of words spoken in the NPR-ME corpus as a function of the percentage of speakers encountered. Speakers were added in order of amount of speech material (based on word count), starting with the most prolific speakers.

national and local hosts typically comprise 20% of the speech data. The local host's speech typically comprises 15% of the speech data in an entire show, while the national host typically comprises 5% of the speech data. Regularly appearing reporters typically provide anywhere from 2-5% of the total speech in a show. Figure 3-2 illustrates the behavior of the percentage of words spoken as speakers are encountered in a typical NPR-ME broadcast. Speakers were added in order of amount of speech material (based on word count), starting with the most prolific speakers. We see that nearly 90% of all the words spoken are provided by 50% of the speakers.

On average, over 76% of the speakers in a given show have not appeared in previous shows. Even if we restrict our analysis to the most prominent speakers in a given show (those whose speech comprises over 5% of the speech data in the entire show), only 56% of these speakers are found in other shows. This result suggests that we

could not take a completely speaker-dependent (SD) approach when transcribing this data using an ASR system. Speaker-dependent approaches generally yield increased performance over speaker-independent approaches. However, a collection of training data is required to develop the SD acoustic models [64]. Our analysis shows that we could only generate training data for 24% of the speakers in a given show. Even among the most prominent speakers in a show, training data could only be collected for just over half of them.

3.4 Vocabulary Analysis

In this section, we investigate the characteristics of the GAD vocabulary. We specifically are interested in the behavior of the vocabulary over time, the out of vocabulary rate as a function of training set and vocabulary size, the part-of-speech characteristics of the vocabulary, and the effects on the out of vocabulary rate if task-specific training data is not available. In addition to providing us with an understanding of the lexical characteristics of GAD, each of these dimensions are important when developing a vocabulary for an ASR system. The vocabulary growth analysis will determine the potential size of an ASR system's vocabulary and will indicate if out of vocabulary words are likely. If the vocabulary size tends to level off after enough training data has been processed, we can assume that out of vocabulary words will not occur very frequently. If the size does not level off in time, then we are likely to encounter out of vocabulary words. The out of vocabulary rate analysis will specifically indicate how often new words are encountered, as a function of training set and vocabulary size.

3.4.1 Vocabulary Growth

In Section 3.3, we found the working vocabulary of a typical NPR-ME hour-long show was just over 2600 words. The frequency of usage of these words is highly skewed. As illustrated in Figure 3-3, the most frequently occurring 20% of the vocabulary words account for over 90% of the words spoken. However, as we will show in Section 3.5, the

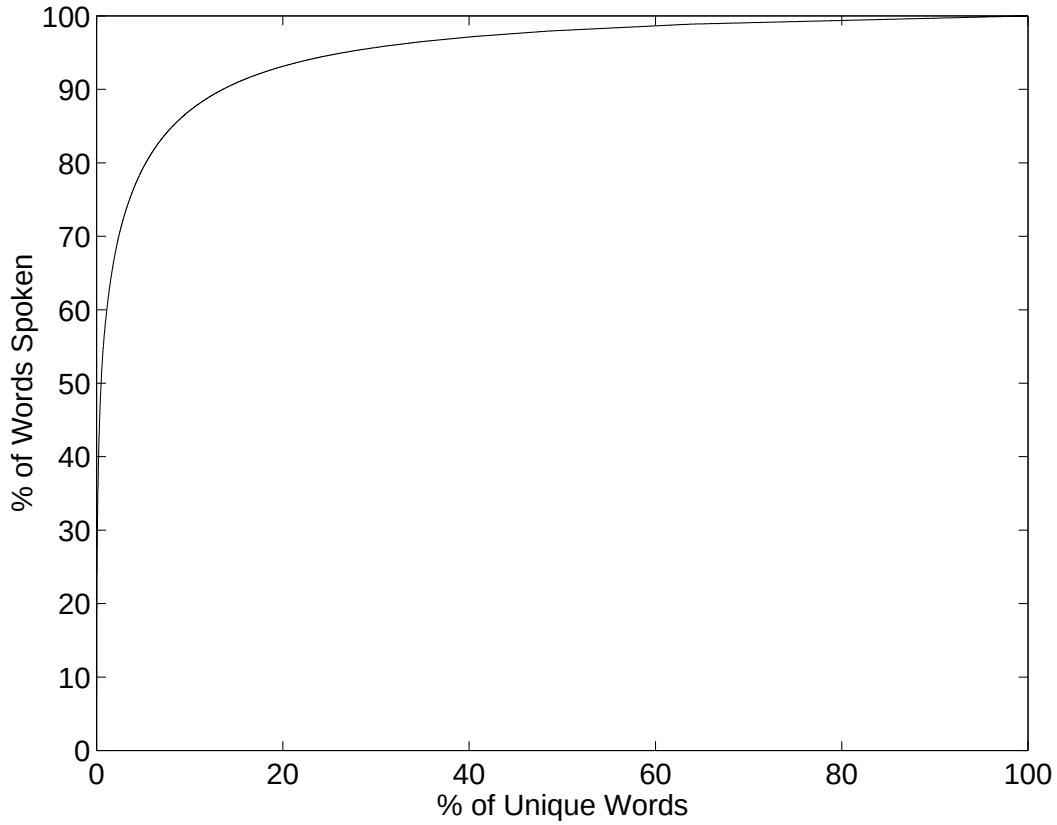

Figure 3-3: The percentage of words spoken in the NPR-ME corpus as a function of the percentage of vocabulary words considered.

least frequently occurring vocabulary words are potentially the most important for understanding the content of the utterances (e.g., names, cities, etc.), and therefore would be most important to recognize in an automatic transcription system.

A vocabulary size of 2600 words is quite manageable for current speech recognition systems. However, closer examination of the data reveals otherwise. Figure 3-4 plots the relationship between the number of distinct words encountered (i.e., the recognizer's vocabulary) versus the size of the training set as the training set size is increased. The training set is increased by adding in NPR-ME shows incrementally, in chronological order, until all 102 hours of data has been added.[1] We are interested in observing the behavior of the vocabulary as a function of training set size (rather than as a function of the number of shows included) for two reasons. First, this allows us

[1]We have found that the trends revealed in this figure are independent of the order in which the shows are added.

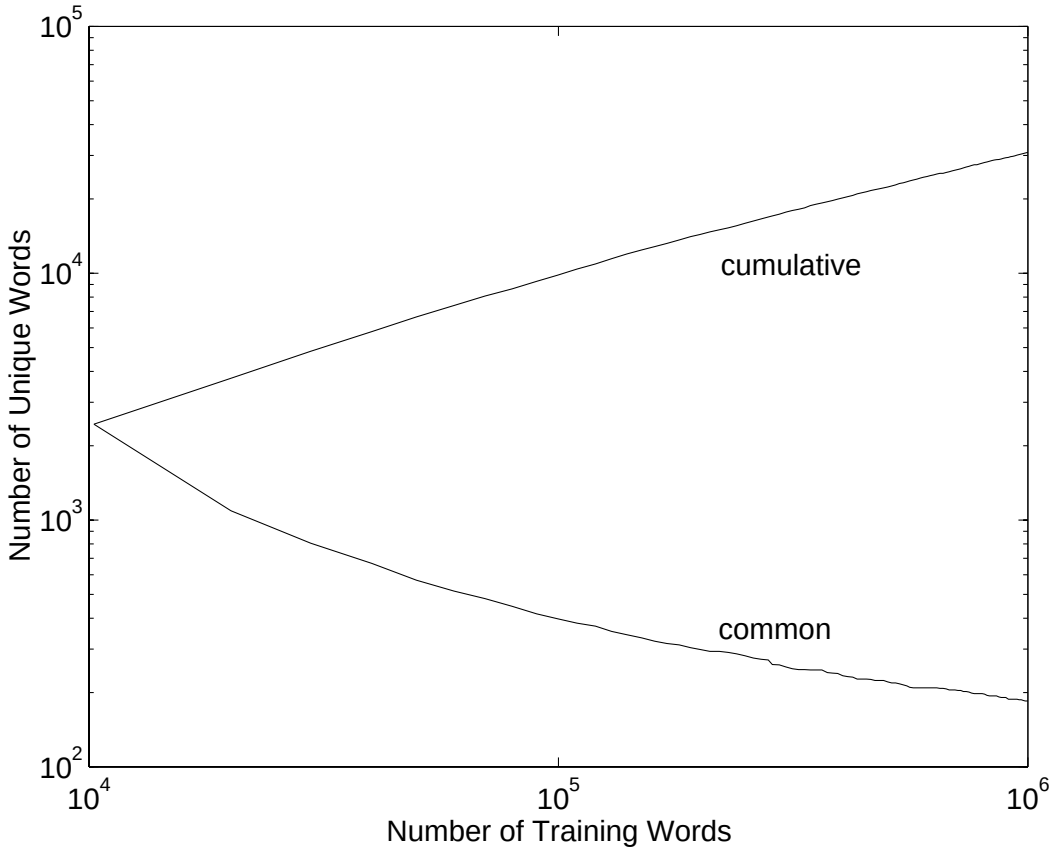

Figure 3-4: The number of distinct words as a function of the number of words encountered in the NPR-ME corpus.

to directly compare these results with the behavior of other corpora. Second, it allows us to understand the quantity of training data that would be required to construct vocabularies of given sizes.

The upper curve of Figure 3-4 shows the cumulative sum of all the distinct words, and therefore represents the potential vocabulary of the recognizer. While the actual size of the vocabulary after 102 shows (over 30,000 words) is within the capabilities of current-day ASR systems, it is quite alarming that the *growth* of the vocabulary shows no sign of abating. If this trend were to continue, then the vocabulary that an ASR system must contend with will reach 100,000 words if a whole year's worth of just this one show is to be transcribed and indexed.

We found that this trend is similar to those of the other large vocabulary corpora. Figure 3-5 shows the vocabulary size as a function of training words encountered for

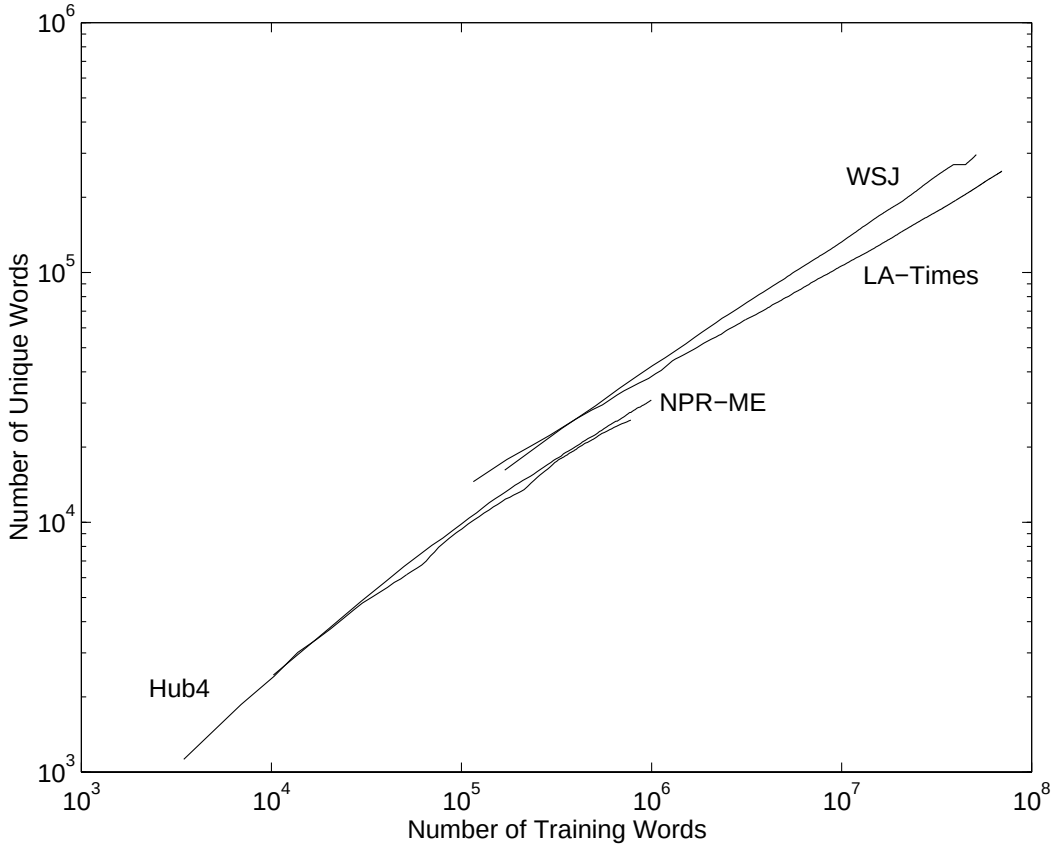

Figure 3-5: The number of distinct words as a function of the number of words encountered in the NPR-ME, Hub4, LA-Times, and WSJ data sets.

the NPR-ME, Hub4, LA-Times, and WSJ data sets. This figure illustrates that the two speech-based broadcast news corpora, NPR-ME and Hub4, have very similar vocabulary growth characteristics, while the text-based corpora, LA-Times and WSJ, have similar characteristics. The speech-based corpora have a slightly smaller vocabulary for a given training set size, but otherwise, the behavior of the two types of data are quite similar. Even after a substantial amount of data has been collected (over 33 million words after one year of LA-Times data), new vocabulary words continue to be encountered, and the vocabulary size grows to over 250,000 words after two years of LA-Times data is collected. Since the NPR-ME and Hub4 curves display a similar growth trend, we can assume that their vocabularies will grow to a similar size. This poses two problems for speech recognition systems. First, a vocabulary size of 250,000 words is well beyond the capabilities of current state-of-the-art ASR systems.

Second, it is alarming that new words are *still* being encountered (as indicated by the continued growth of the vocabulary), despite the use of massive amounts of training data and vocabulary size.

As more shows are included, the size of the *common* vocabulary across the shows will presumably decrease, as news topics change as time passes. This is illustrated by the lower curve in Figure 3-4, which indicates that less than 200 words occur in all of the 102 shows, most of them being function words and generic corpus-dependent words such as "news," "traffic," and "forecast." Although there are relatively few of these common words (they comprise less than 0.5% of the complete vocabulary), they account for over 47% of the total words spoken in the corpus.

3.4.2 Out of Vocabulary Rate

While Figure 3-4 indicates how fast the NPR-ME vocabulary grows as more training data is added, it doesn't directly reveal how well its vocabulary covers unseen data. In other words, it doesn't reveal the likelihood of encountering out of vocabulary words, or the out of vocabulary rate. In another experiment, we measured the coverage of cumulatively constructed vocabularies on a set of unseen data (one held out NPR-ME show). To generate Figure 3-6, we measured the vocabulary coverage of the NPR-ME test show as we built up a vocabulary incrementally by adding in NPR-ME training shows.[2] Figure 3-6 shows the probability of encountering an out of vocabulary word in the NPR-ME corpus for a given training set size. We estimated this probability by measuring the fraction of words in the test set that were not covered by the constructed vocabularies. Figure 3-6 illustrates that as more training data is accumulated, the out of vocabulary rate falls from a maximum value of over 22% (with a training set size of just over 10,000 words), to a minimum value of 1.9% (with a training set size of nearly 1,000,000 words).

We are often interested in understanding the relationship between the out of vocabulary rate and vocabulary size rather than training set size, since the size of

[2]We have found that the trends revealed in this figure are independent of the order in which the shows are added.

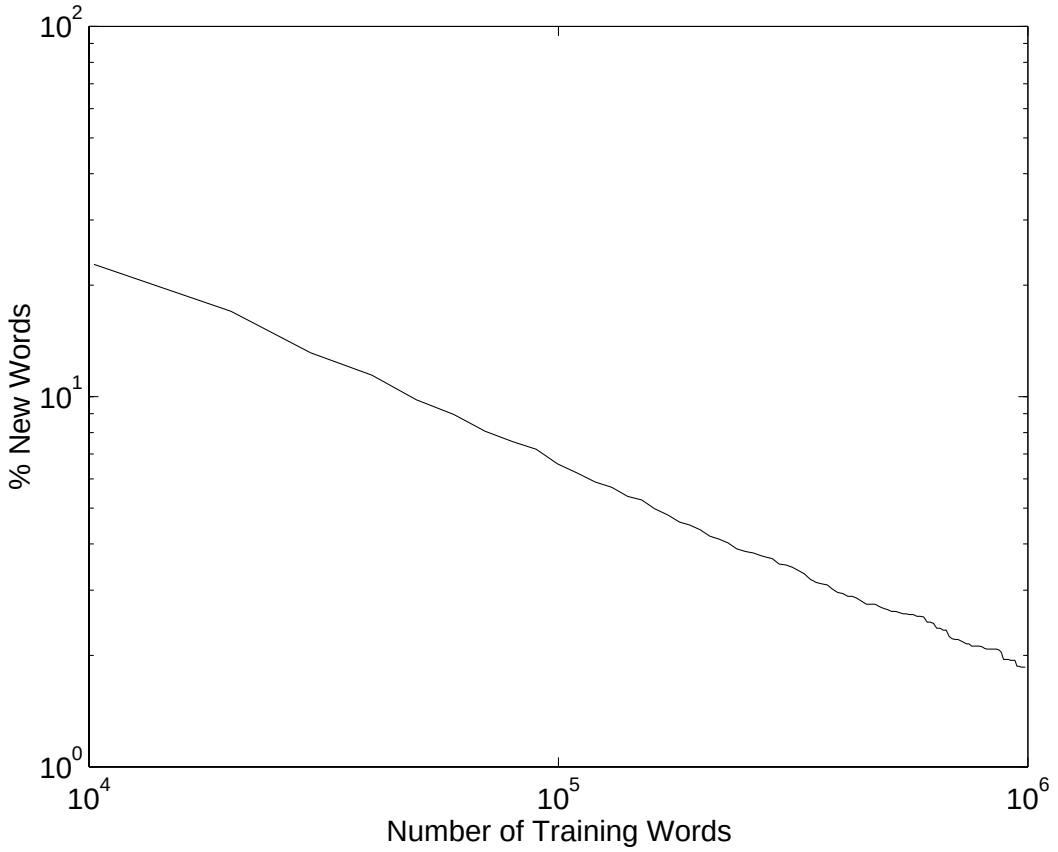

Figure 3-6: NPR-ME out of vocabulary rate as a function of the quantity of NPR-ME training data.

the vocabulary is an important parameter in an ASR system. Figure 3-7 shows the out of vocabulary rate versus vocabulary size instead of amount of training data, as in Figure 3-6. In this figure, two methods of determining vocabulary size were used. For the top curve, labeled *by adding shows*, the vocabulary size was determined by incrementally adding in new training shows. For the bottom curve, labeled *by frequency count*, we assume that we have the entire training set available to us at the outset, and we build a vocabulary of a given size (v) by computing word frequencies over all of the training set and adding the v most frequent words to the vocabulary. Since we are using word-frequency information to determine the vocabulary for the latter case, we would expect this curve to yield lower out of vocabulary rates. Figure 3-7 does confirm our hypothesis, showing that the out of vocabulary rate for a particular vocabulary size v is lower when word frequencies are accounted for when building the

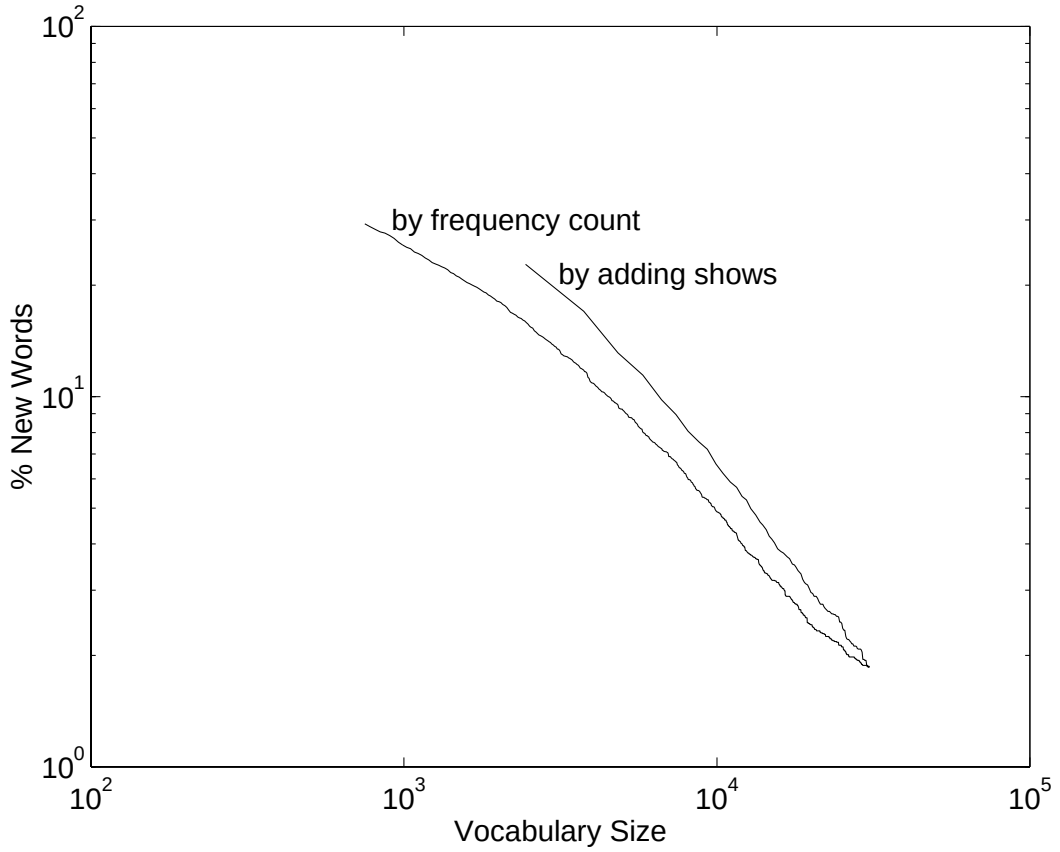

Figure 3-7: NPR-ME out of vocabulary rate for different methods of determining vocabulary size. For the top plot, we varied the training set size and set the vocabulary to include all unique words. For the bottom plot, we use the *entire* training set to compute word frequencies and varied the vocabulary size v by setting the vocabulary to include only the most frequent v words.

vocabulary. This is especially true when v is relatively small. The two curves converge as v approaches its maximum size.

Regardless of what method we use to compute the out of vocabulary rate curves, we see that a large training set (nearly 1,000,000 words, which results in a vocabulary size of over 30,000 words), still leaves us with nearly 2% of out of vocabulary words for the NPR-ME data. This is a potentially serious problem for two reasons. First, as we will see in Section 3.5, it is these words that are most important to recognize if we are interested in transcribing the data for an information retrieval system. Second, the misrecognition of new words has a ripple effect in a speech recognition system, causing other in-vocabulary words to be misrecognized. Hetherington [42] showed

that 1.5 word errors are encountered per every new word, some of which occur in neighboring in-vocabulary words. Therefore, not only are the new words missed by an ASR system, but neighboring words, which may also be high content words, may be affected.

We found that the out of vocabulary rate analysis results on NPR-ME were similar to analysis performed on the Hub4 task. Jain et al. [46] computed the out of vocabulary rate on the 1996 Hub4 development test set as a function of vocabulary size. They found that as the vocabulary size was increased, the out of vocabulary rate decreased, reaching a minimum of 1.1% with a vocabulary size of 60,000 words. A vocabulary size of 30,000 words yielded an out of vocabulary rate of 1.9%, similar to the results found on the NPR-ME corpus.

3.5 Part of Speech Analysis

In this section, we further investigate the properties of the cumulative, common, and new word vocabularies compiled from our NPR-ME corpus. We are interested in understanding the distribution of the words in each vocabulary to determine if the words could be obtained from common dictionary sources, or if a large collection of task-specific training data is required. In addition, we are interested in determining the characteristics of the out of vocabulary word list. We suspect that these words are high content words (i.e., proper nouns), that would be important to recognize for an information retrieval task.

For our analysis, we examine the syntactic part-of-speech tags for each vocabulary. We suspect that each vocabulary will have a very different part-of-speech distribution. The cumulative vocabulary should be similar to that of the general English language, while the new word vocabulary should contain a majority of proper nouns. We also suspect that the common vocabulary will be similar to the most common words in the English language, with the possible addition of a few common corpus-specific words.

For this analysis, we used Brill's part-of-speech tagger [6], and collapsed the large set of 48 tags [12] down to eleven: proper nouns, nouns, adjectives, adverbs, verbs,

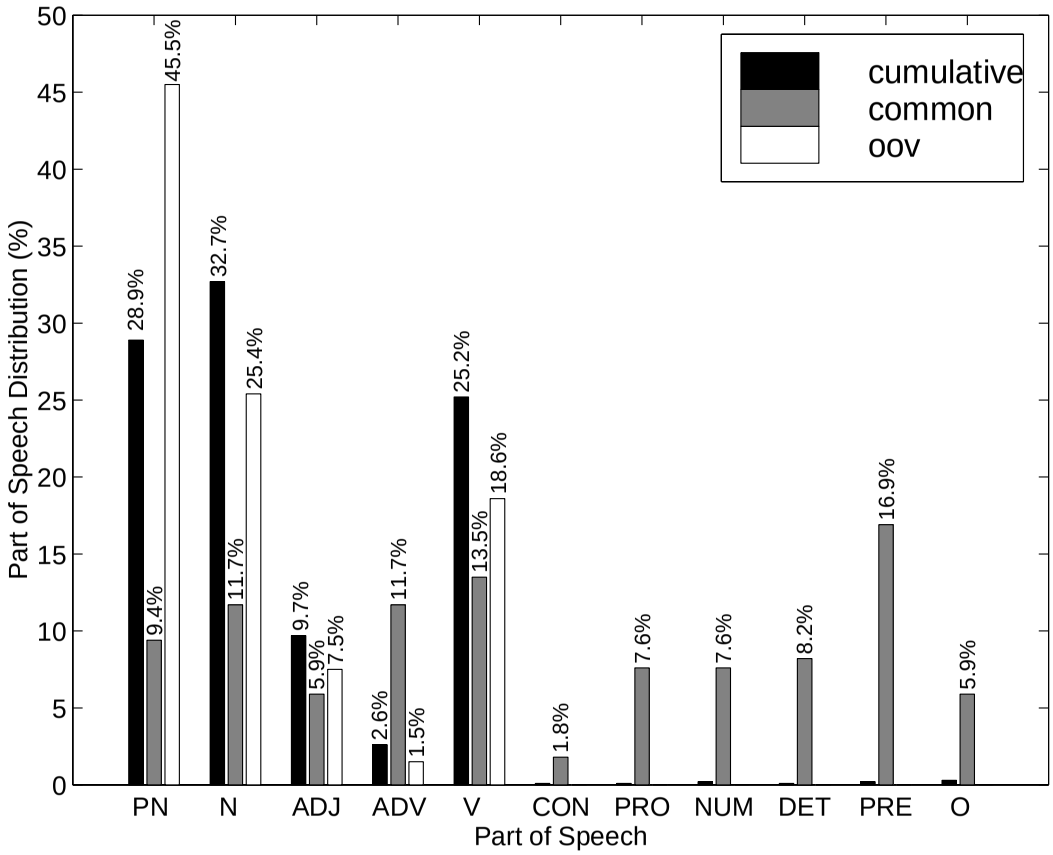

Figure 3-8: Summary of the part-of-speech distributions for the NPR-ME cumulative, common and out of vocabulary word (oov) vocabularies. The distributions are unweighted by word frequency. The part of speech distributions shown are: proper noun (PN), noun (N), adjective (ADJ), adverb (ADV), verb (V), conjunction (CON), pronoun (PRO), number (NUM), determiner (DET), preposition (PRE), and other (O).

conjunctions, pronouns, numbers, determiners, propositions, and "other." We tagged the transcripts of our entire NPR-ME corpus and computed the part-of-speech distributions for the cumulative, common and new word vocabularies. The out of vocabulary list was constructed by finding those words in a held out test show that had not appeared in remainder of the NPR-ME corpus.

Figure 3-8 displays the part-of-speech distributions for the NPR-ME vocabularies. The distributions are unweighted by word frequency. We found that the cumulative vocabulary primarily consists of nouns, verbs, proper nouns and adjectives. We also

A_F_L_C_I_O	Jehovah	Rembrandt
Beardstown	July	Rossner
Belmonte	Kosovos	Serino
Bonne	Latvia	Shane
Brubeck	Laureen	Sienook
Clayton	Lynnhurst	Stewicky
Cuno	Marshaun	T_R_G_I
Degas	Mcgaw	Tritch
Doyle	Michelle	Vaughn
Dracut	Noradom	Venetian
Enos	O_F_C_E	Vermeers
Fillipe	Oaklandvale	Verve
Fitzpatrick	Pacific	Vidrine
Flexon	Pearl	Vulgova
Foxx	Presioso	W_A_V_E
Gilbart	Primakov	Yipgeni
Givadi	Prior	Yuvanovich
Greenwich	Q_U_A_N_T_I_C	
I_R_S	Ranured	

Table 3.3: Proper nouns not found in the NPR-ME vocabulary.

found that, as we had expected, a large percentage of out of vocabulary words are proper nouns, which make up over 45% of the out of vocabulary list. This indicates that it will be very difficult to collect these words when trying to construct a vocabulary for an NPR-ME recognition task. However, it is these words that are most important in a content description task. Table 3.3 lists the proper nouns found in the out of vocabulary list. Among others, we see that "Degas" (mentioned in a story about stolen museum paintings), "Primakov", "Vulgova" and "Yuvanovich" (Russian foreign minister, Kosovo Albanian leader, and Yugoslav foreign minister, respectively, mentioned in a story about the conflict in Kosovo), were not found in the NPR-ME vocabulary. Therefore, we would not be able to properly index the stories these words appear in since we will would not be able to recognize these words when automatically transcribing the data.[3]

We found that the out of vocabulary word list contained a large percentage of verbs (18%). Upon closer examination, we found that over 56% of these words were

[3]A complete list of the out of vocabulary words can be found in Appendix B.

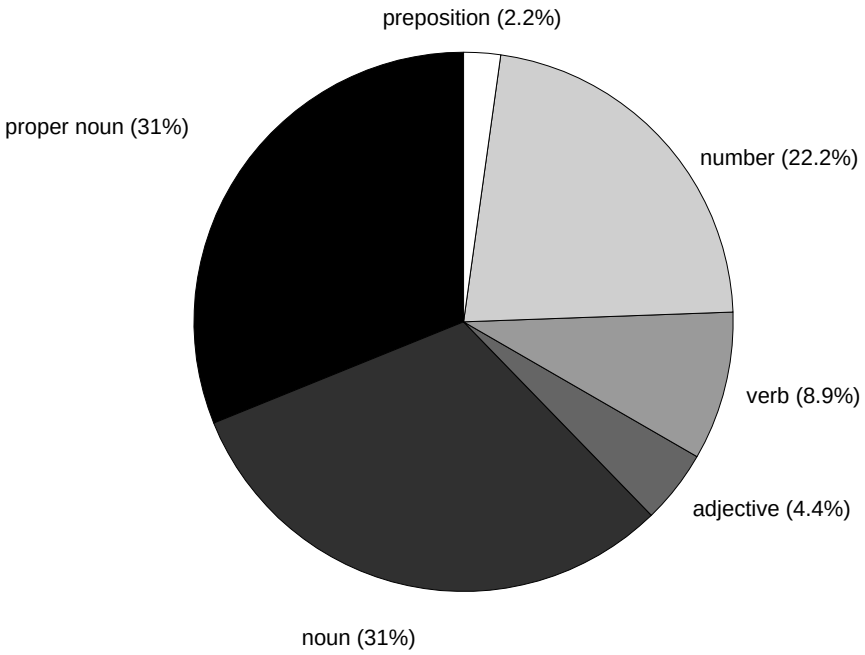

Figure 3-9: Summary of the part of speech distribution for the NPR-ME common vocabulary not found in the 200 most frequent words in the Brown Corpus.

forms of verbs that exist in the cumulative vocabulary. For example, "combing" and "squawking" were found in the new word vocabulary, while "comb" and "squawk" exist in the cumulative vocabulary. While methods to automatically generate all forms of a given verb could be developed, these previously unseen verb forms are indeed new words and would have to be included in a recognizer's vocabulary.

We further investigated the components of the common vocabulary to see if they were similar to the most common words of English. To accomplish this, we compared the NPR-ME common vocabulary to the most frequent 200 words found in the Brown Corpus [52]. The Brown Corpus consists of over one million words of running text from 500 selections of about 2000 words each. We found that just over 73% (123 of the 168 common vocabulary words) of the NPR-ME common vocabulary was present in the top 200 Brown Corpus vocabulary. A part-of-speech breakdown of the words not found in the top 200 Brown Corpus is illustrated in Figure 3-9. We found that the majority of these words were proper nouns, nouns and numbers. The proper nouns not found in the top 200 Brown Corpus are listed in Table 3.4. We can see that these

Bob
Boston
Edition
Massachusetts
Morning
N_P_R
National
News
President
Radio
U_S
University
W_B_U_R
Washington

Table 3.4: Proper nouns in NPR-ME common vocabulary not found in the 200 most frequent words in the Brown Corpus.

words are very task-dependent, such as Boston, WBUR, Massachusetts, etc.[4]

3.6 Cross Corpus Effects

In Section 3.4.2, we investigated the out of vocabulary rate on an NPR-ME test show using NPR-ME training data to build the working vocabulary. This would constitute a best case scenario, since task-specific data are used to build the vocabulary for an ASR system. In this section, we evaluate the out of vocabulary problem when a different corpus is used to develop the working vocabulary. This scenario may very well be a more realistic evaluation of the out of vocabulary problem, as it is often difficult and time consuming to gather a large collection of domain-specific data. However, we often have access to a collection of similar data to use for system development. In these experiments, we used the Hub4 data to construct the vocabularies. The NPR-ME and Hub4 tasks are very similar, as they are both broadcast news, speech-based data. We also found that the vocabulary characteristics of these corpora were very

[4]A complete list of the common words not found in the top 200 Brown Corpus can be found in Appendix B.

similar, so the Hub4 data seems to be a good choice to use for development of an NPR-ME transcription system.

Figure 3-10 shows the effect of using an out-of-domain training set to construct the system vocabulary on the out of vocabulary rate problem. Because we assumed we would have all of the material for the training set ahead of time, we built the vocabularies in decreasing order of word frequency, as we discussed in Section 3.4.2. The NPR-ME out of vocabulary rate curve is also shown for comparison. We can see that the use of training data from a different corpus, even though very similar in content, exacerbates the out of vocabulary rate problem. For small vocabulary sizes (fewer than 1000 words), the Hub4 and NPR-ME curves are very similar. This makes intuitive sense because of the way the curves were generated. Vocabularies of that size primarily consist of high frequency words (e.g., "the," "of," etc.), which would presumably be task-independent. As the size of the vocabulary grows, however, the curves separate substantially. With a vocabulary size of over 25,000 words, the use of out-of-domain data yields an out of vocabulary rate of over 4%. If we compare this with the NPR-ME results, we find that for a similar vocabulary size, the in-domain data yields an out of vocabulary rate of just over 2%.

Others have addressed the out of vocabulary rate problem for GAD by constructing very large vocabularies consisting of both in-domain and out-of-domain data. BBN [51] constructed a 45,000 word vocabulary using both broadcast news and newspaper sources, which yielded an out of vocabulary rate of 0.9% on the 1996 Hub4 evaluation set. Other sites [13, 28, 73, 81, 85] constructed a 65,000 word vocabulary using broadcast news training texts, newswire texts, and additional names that frequently appeared in the broadcast news data, which yielded an out of vocabulary rate of 0.7% on the 1996 Hub4 evaluation set.

While the combination of in-domain and out-of-domain data can reduce the out of vocabulary rate to just under 1% on the Hub4 data set used in the analysis cited above, this does not guarantee that the rate will *remain* at that level for all future collections of broadcast news data. As new topics and names appear in the news, the out of vocabulary rate will in all likelihood increase.

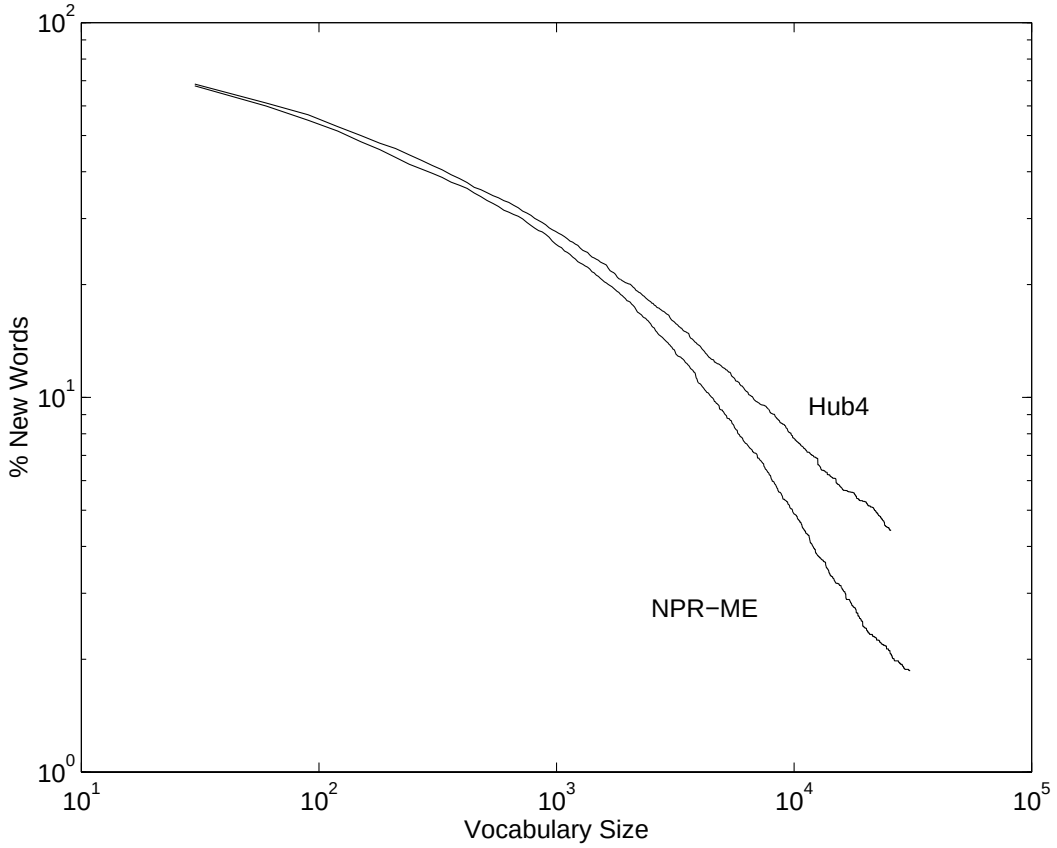

Figure 3-10: Out of vocabulary rate as a function of training data. Vocabularies are built by observing the entire training set and adding words in decreasing order of frequency. The top curve, labeled Hub4, illustrates the out of vocabulary rate as a function of out-of-domain data. The bottom curve, labeled NPR-ME, illustrates the out of vocabulary rate as a function of in-domain data.

3.7 Summary

In this chapter we have examined the lexical aspects of GAD. Our analysis of the transcriptions of the NPR-ME corpus revealed some interesting general characteristics. It contains many speakers and stories, with numerous turn takings. We also discovered that no single speaker dominates the speech data, and that the majority of speakers have never been seen previously.

Our vocabulary analysis found that the vocabulary of a single NPR-ME show was modest (approximately 2600 words). However, we also found that even with a large training set (nearly 1 million words), new words are still encountered. Using task-specific training data with a vocabulary of over 30,000 words, we found that

the out of vocabulary rate was just over 2%. For a strict transcription task, this rate may be acceptable, since new words are typically very low-frequency words. Therefore, the overall word error rate in a transcription task may not be affected too adversely. However, for an information retrieval task, a 2% error rate may not be acceptable because these words are typically important for conveying content. Our part-of-speech analysis found that new words were predominately proper nouns and nouns, which would be very important to recognize if we were describing the content of this data. This problem was magnified when we investigated the more realistic scenario of constructing a training set from a out-of-domain source. In this case, we found that the out of vocabulary rate nearly doubled to 4%.

The analysis completed in this chapter uncovered some potentially serious problems for a word-based approach for the transcription of GAD. An alternative to a large vocabulary continuous speech recognition approach is to use a subword unit representation. The benefit of a subword unit approach is that the vocabulary of the recognizer is constrained while it provides full coverage of the corpus lexicon. Therefore, we are able to overcome the large vocabulary and out of vocabulary problems that we discovered in our analysis of GAD.

Ng [62] recently investigated the feasibility of using subword unit representations for spoken document retrieval. He examined a range of subword units of varying complexity derived from phonetic transcriptions. He found that subword units achieved comparable performance to text-based word units if the underlying phonetic units were recognized correctly. In Chapter 5 we will explore different training and testing methods for the phonetic recognition of GAD.

Chapter 4

Sound Recognition

To fully describe the content of general audio data, we have argued that a complete acoustic description must be created, in addition to transcribing the speech material. This description would indicate regions of speech and non-speech, identify segments spoken by particular speakers, indicate the speaking environment, etc. While developing a complete acoustic description system is beyond the scope of this work, we have chosen to concentrate on the development of a sound recognition system that segments GAD into general sound classes. Not only would such a system contribute to the description of the acoustic content, but it may also be a useful preprocessing step for a speech recognition system. First, segments of audio containing non-speech can be detected, so the speech recognition system's resources won't be wasted on segments of music, silence, or other non-speech audio. Second, segmenting the audio into acoustically homogeneous blocks and using appropriate models for each segment may improve overall speech recognition results, a topic explored in Chapter 5.

In this chapter we asked the following three questions.

1. How many meaningful classes exist in GAD?

2. What are their general acoustic characteristics?

3. Can they be reliably determined using an automatic recognition system?

To answer these question, we first divide the data into several categories through

preliminary investigation, and examine their general characteristics and distributions. Second, we determine how well we can automatically segment the sound stream into these acoustic categories. After we develop a recognition system to accomplish this task, we evaluate the results to determine if our subjectively defined acoustic classes need further refinement.

4.1 Acoustic Analysis

In this section we are interested in discovering the acoustic characteristics of GAD. Specifically, we will determine what general sound classes exist in the data, and how they differ from one another.

4.1.1 Sound Classes

We began our analysis by thoroughly examining two hours of NPR-ME shows. In addition to carefully listening to each show, we also viewed spectrograms of the data. We found that there are a number of categories of speakers in the data, each with varying sets of different general acoustic conditions. First, there are clearly main hosts of the show. Their speech is typically prepared (i.e., apparently read from a prepared script) and is either acoustically very clean, or appears over background theme music. Second, there are a number of reporters that call in their news reports from the field. This speech is often spontaneous, and spectrograms of this data reveal that it is clearly of telephone bandwidth. Third, there are reporters recorded on-location present in the data. This speech is similar to the telephone reporters in that it is spontaneous, but the spectrograms revealed that this data is not bandlimited. This speech contains a variety of background noise, ranging from gunfire, to street noise.

In addition to the above speech conditions, we found a number of non-speech conditions in the NPR-ME data. First, there are a number of music segments. These segments are typically found not only at the beginning and end of the broadcast, but also between news stories. Long stretches of silence were also identified in the data.

These silence segments typically occur at speaker boundaries and between major news stories. Finally, there are some segments of noise, such as sounds of gunfire within a story about warring nations.

From this analysis, we reached the preliminary conclusion that there are seven logical categories into which the NPR-ME data may be classified. We identified four unique speech categories: clean speech, music speech, noisy speech and field speech. Three non-speech categories were identified: music, silence, and miscellaneous. The seven categories are described below.

- **Clean Speech (c_s):** The clean speech class consists of wideband (up to 8 kHz[1]) speech from anchors and reporters. The speech is typically prepared (i.e., read from a script rather than spontaneously generated), and is recorded with high quality acoustic conditions (a high quality microphone is presumably used, potentially sophisticated digital signal processing techniques are used to enhance the vocal appeal of the speakers [15], and there is no audible background noise).

- **Music Speech (m_s):** The music speech class consists of wideband speech in the presence of background music. This is typically found both before and after the theme music which leads the show, and it is also found separating major stories. Music speech also often exists in stories about a particular musical group or music genre.

- **Noisy Speech (n_s):** The noisy speech class consists of wideband speech in the presence of background noise. This type of speech is typically generated from reporters on location. The background noise ranges from office noise, to traffic noise, to cross-talk from other speakers (e.g., a translator speaking over the native speaker).

[1]The data may actually contain energy at higher frequencies. However, as explained in Chapter 2, the data was downsampled to 16 kHz upon transfer to computer disk, and is therefore limited to an upper frequency of 8 kHz.

- **Field Speech (f_s):** The field speech class consists of bandlimited (4 kHz) speech, collected over the telephone. This type of speech is typically generated from reporters calling in from the field, so the overall quality is generally quite poor. Higher quality telephone speech is seen occasionally from individuals calling into the show. For example, the local host might be talking to an "expert" on the current topic over the telephone.

- **Music (m):** In addition to the opening NPR-ME theme music, selections of music are found scattered throughout the program separating major stories. In addition, there are often stories about musical topics which contain selections of music.

- **Silence (sil):** Segments of silence longer than 250 ms were identified as a separate class. Anything shorter than 250 ms could be confused with the closure portion of stop consonants, or the short, natural pauses that occur between words or utterances.

- **Miscellaneous (misc):** The miscellaneous class captures anything that didn't fall into one of the other six classes. Typically, these segments are used for effect in a story (e.g., sounds of gunfire in a story about a military action).

Figures 4-1, 4-2, and 4-3 illustrate some of the spectral differences among the seven identified sound classes. A spectrogram of a segment of music followed by a segment of speech superimposed on the background music (i.e., music speech) is shown in Figure 4-1. We can clearly see the fine harmonic structure in the music segment, indicated by the evenly-spaced horizontal lines in the spectrogram. We can also see that the harmonics carry through to the music speech segment. Figure 4-2 is a spectrogram of a segment of clean speech, followed by field speech[2]. Here, we see the bandlimited nature of the field speech. While the clean speech segment contains energy through 8 kHz, the field speech segment contains no energy at frequencies

[2]This spectrogram was manually created for illustration purposes by splicing together a segment of clean speech, followed by a segment of field speech. The edit point can be seen between the two segments.

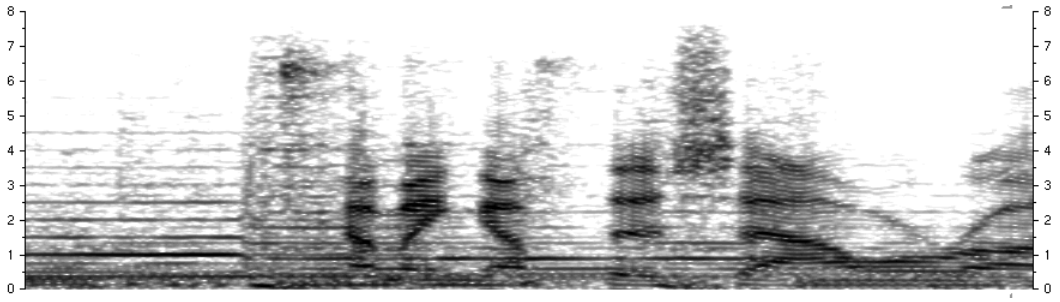

Figure 4-1: Spectrogram of a segment of music followed by speech superimposed on the background music. Note the harmonics in the music segment. The harmonics, indicated by evenly spaced horizontal lines in the spectrogram, also carry through into the music speech segment.

over 4 kHz. A spectrogram of a sample of noisy speech is shown in Figure 4-3. If we compare this spectrogram with the clean speech portion of Figure 4-2 (the initial portion of the figure), we can clearly see the effect of the background noise. The individual speech sounds are much less distinct in the noisy speech sample, making it very difficult to identify the formants in the speech. Although we do not explicitly extract these characteristic frequencies (seen as dark bands of energy in the spectrograms) from the speech signal, they are important characteristics that are known to carry important linguistic information. Noisy speech poses a number of challenges to our speech recognition system, SUMMIT. First, as we described in Chapter 2, SUMMIT proposes phonetic segment boundaries at locations of spectral change. The presence of background noise makes these boundaries much less distinct. Second, the formant information is implicitly captured by the MFCC representation of the speech signal. The presence of background noise makes the differences among the speech sounds (seen clearly in the clean speech portion of Figure 4-2) much less distinct. This would presumably make noisy speech more difficult for an automatic speech recognition system to transcribe.

4.1.2 Corpus Preparation

In order to complete further acoustic analysis on the NPR-ME data, we had to label the data with the seven acoustic classes defined in the previous section. As described

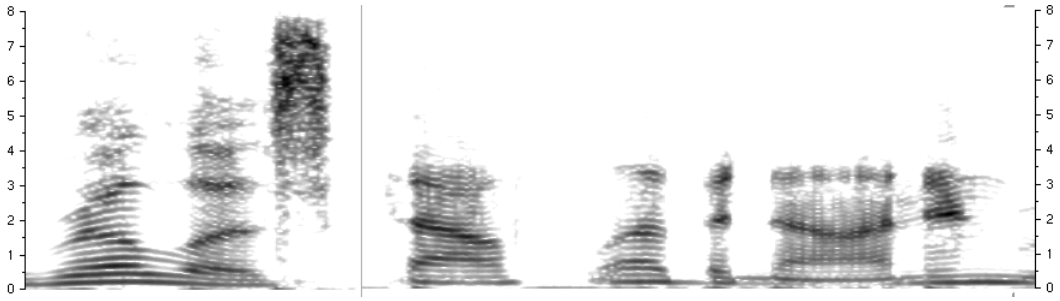

Figure 4-2: Spectrogram of a segment of clean speech followed by field speech. Note the bandlimited nature of the field speech segment, as compared to the segment of clean speech, which contains energy through 8 kHz.

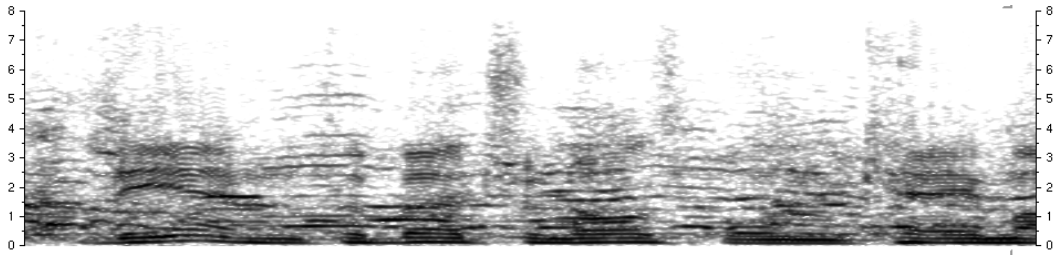

Figure 4-3: Spectrogram of a segment of noisy speech. When comparing this spectrogram to the clean speech portion of Figure 4-2, we can clearly see the background noise throughout the frequency range.

in Chapter 2, ten hours of NPR-ME data was segmented into manageable sized waveform files at silence breaks. The data was then manually labeled with one of the seven acoustic labels once every 10 ms. The labeling was done through visual examination of spectrograms and through critical listening of the data. Eight hours of the tagged data was used for the analyses presented in Section 4.1.3 and for system training in Section 4.2, and the remaining two hours of data was used for system test. Table 4.1 summarizes the amount of training and testing data (in minutes) available in each sound class.

Throughout our investigations in this chapter, we made heavy use of the Transcription View facility in SAPPHIRE [43], which can simultaneously display the waveform, spectrogram and transcription for each file. In addition to using SAPPHIRE as a display tool, we also took advantage of its editing capabilities to manually label the data with the sound class labels.

Sound Class	Training Data	Testing Data
Clean Speech	244.1	53.7
Field Speech	72.4	28.6
Music Speech	59.4	14.5
Noisy Speech	66.9	11.9
Music	21.7	5.4
Silence	25.7	5.3
Miscellaneous	5.3	2.3

Table 4.1: Amount of training and testing data available (in minutes) for each sound environment.

4.1.3 Characteristics of Sound Classes

Many acoustic differences were apparent while viewing the spectrograms of the NPR-ME data. In this section, we study the characteristics of the seven sound classes more closely.

For each waveform in the NPR-ME training set, the discrete Fourier transform (DFT) was computed every 10 ms using a 40 ms analysis window. The magnitude of the DFT was computed, and the average value was then computed for each sound class. Figure 4-4 is a plot of the average spectra for each of the seven sound classes. We see a number of differences in the average spectra for the sound classes. Silence and field speech are visually distinct from other classes both in terms of energy and spectral shape. We see the minimal energy across the entire frequency range in the silence spectrum, while the field speech spectrum drops off substantially at frequencies over 4 kHz. Music differs from speech in its fine harmonic structure, which is seen by the ripples in the spectrum. These quasi-periodic ripples can also be seen in the music speech spectrum. We also see that the miscellaneous class has a significant decrease in spectral energy at frequencies above 6 kHz. Differences in the average spectra of the clean, noisy, and music speech categories are more subtle, suggesting that confusions may result if these sounds were to be classified using purely spectral features.

Figure 4-5 shows the distribution of sound classes in the NPR-ME training data. The distribution was computed based on the total amount of data in minutes in each

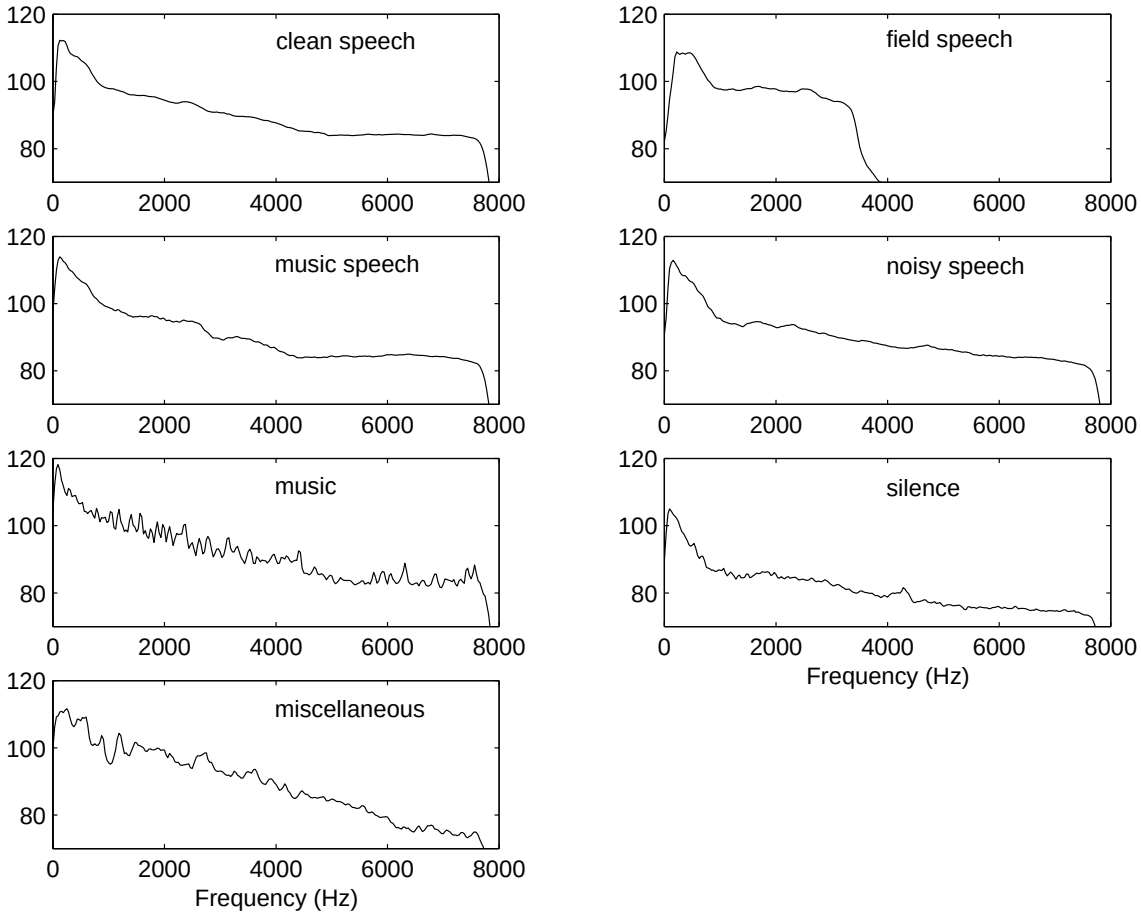

Figure 4-4: Average power spectrum (in dB) for each of the seven sound classes found in GAD.

class. High quality studio speech constitutes only about half of the entire corpus. Another 26% of the data contains speech superimposed with other sounds. Nearly 15% of the data is of telephone bandwidth, and the remaining 10% of the data is non-speech. Among the non-speech classes, we see that there is substantially more silence and music data than miscellaneous data. Closer examination of the data revealed that silences occurred not only between speakers and stories, but also within sentences at natural, syntactic boundaries.

We computed the average segment length to determine if differences existed among the classes. The results are shown in Table 4.2. Values for speech and non-speech were also computed. The speech sound class contains all clean, field, music and noisy speech segments. The non-speech class contains all music, silence, and miscellaneous

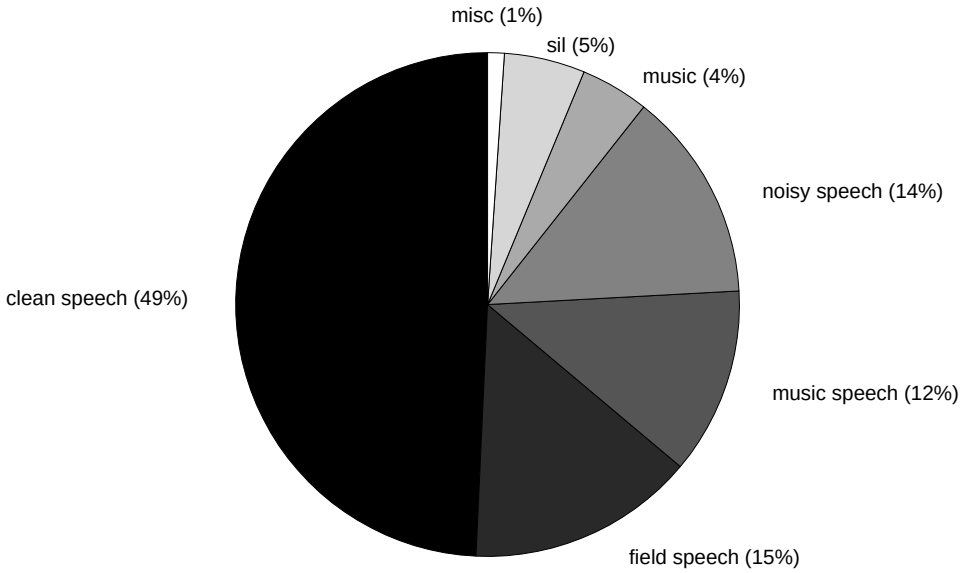

Figure 4-5: Distribution of sound classes in the NPR-ME training data, computed with respect to total time in each class.

segments. We see that speech segments are substantially longer than non-speech segments (4.8 seconds vs. 0.7 seconds). Within the speech classes, field and music speech segments are slightly longer than the clean and noisy speech classes. These results indicate that segment duration may be a useful feature in an automatic segmentation system.

4.2 Automatic Recognition of Sounds

Based on the analysis presented in the previous section, it would seem that GAD can be acoustically characterized by seven general sound classes. We are now interested in determining if these classes can be reliably extracted from the audio signal. Our goal is to segment GAD into homogeneous regions so that regions of different nature can be handled differently. For example, regions of music and noise can be eliminated from future processing by an automatic speech recognition system. Also, different models can be developed for the recognition of each speech environment. The effect of such segmentation on a speech recognition system will be explored in Chapter 5.

The following section describes approaches that have been taken by others for this

Sound Class	Average Segment Length
Clean Speech	4.7
Field Speech	5.2
Music Speech	5.0
Noisy Speech	4.4
Music	4.2
Silence	0.4
Miscellaneous	1.8
Speech	4.8
Non-Speech	0.7

Table 4.2: Average segment length (in seconds) for each sound class. The speech sound class is comprised of all clean, field, music and noisy speech segments. The non-speech class is comprised of all music, silence, and miscellaneous segments.

task. We then describe the approach that we have chosen to take in this work, and present our results. Analysis of the results led us to refine our subjectively determined sound classes.

4.2.1 Related Work

In this section, we review some related work on approaches for segmenting audio data. Three general segmentation algorithms have been proposed in the literature. First, a model-based segmentation approach [69, 70, 80] builds different models for a fixed set of acoustic classes from a training set, and frames of the incoming audio stream are classified with a maximum likelihood approach. Segment boundaries are proposed at locations where there is a change in the assigned acoustic class. Second, a decoder-guided segmentation approach [36, 50] decodes the input audio stream with a speech recognition system trained with a collection of class-dependent models. For example, a gender-dependent phonetic recognizer may be used for this task. During the decoding process, in addition to producing the recognized output string, class assignments are also generated. Segment boundaries are then proposed at locations where there is a change in the class label. Third, a metric-based segmentation approach [11, 30, 74] proposes boundaries at maxima of distances between neighboring windows placed at every sample along the audio stream. Distances such as the Kullback-Leibler dis-

tance [14] and the generalized likelihood ratio distance [3] have been used. In this approach, class labels are not generated for the resulting segments. Examples of each of these approaches are described below.

A number of research sites have been involved with the DARPA sponsored Hub4 [61] task, which has encouraged researchers to focus on the problems of processing speech materials which have not been created specifically for the purpose of speech system development, such as broadcast news. Participating Hub4 sites have used segmentation and classification of the test utterances to improve their recognition results. Dragon Systems took a model-based approach in their 1996 Hub4 recognition system [80]. They segmented the testing data into four classes: non-speech, which consisted of pure music, noise or silence, and three speech channel conditions: full-bandwidth, narrow-bandwidth, or speech with music. First, the data is segmented into manageable pieces based on the overall energy. Second, the resulting segments are reprocessed to exclude pure music or noise, based on a measure of "harmonicity" [45]. Third, channel analysis is performed to classify the data into one of the speech channel conditions. A model of each channel is trained from a small selection of training data. Each model is a single probability distribution, which consists of a mixture of 256 diagonal covariance Gaussians. Each second of the original data stream is classified as one of the three defined classes. The channel changes that are detected from this analysis are used to refine the segmentation obtained in the first two steps. To analyze the performance of their segmenter and channel classification system, they compared the amount of data (in seconds) automatically classified as one of the four classes (non-speech, full-bandwidth, narrow-bandwidth and speech with music) to the true data labels. This system achieved an overall classification accuracy of 82.2% on the Hub4 1996 evaluation test set. They found that the majority of their classification errors resulted from noisy, full-bandwidth data being misclassified as speech with music.

The 1998 HTK Hub4 system [36] used a combination of model-based and decoder-guided approaches to segment the audio stream into homogeneous chunks. The first pass labels each audio frame according to bandwidth and discards the non-speech

segments using a mixture of diagonal Gaussian models and a conventional Viterbi decoder. The second pass uses a phone recognizer that outputs a sequence of phones with male, female or silence tags. Segment boundaries are determined when there is a switch in either bandwidth or gender, or when a silence segment is found. They report a segmentation frame accuracy up to 95%, and after further processing increased their accuracy to 99%.

IBM [11] took a metric-based approach to the Hub4 segmentation problem. They modeled the input audio stream as a Gaussian process in the cepstral space. A maximum likelihood approach was used to detect turns in this Gaussian process based on the Bayesian information criterion [72]. They analyzed their data in terms of insertions and deletions of boundaries. They achieved a very low insertion rate (4.1%), and a 33.4% deletion rate. The majority of their deletions occurred when a segment was less than 2 seconds in length. These most likely occurred because there wasn't sufficient data to adequately develop the Gaussian model for these segments.

There has been research completed in the area of speech/music classification in addition to the Hub4 task. Saunders [69] uses a straightforward model-based approach to the discrimination of speech and music. A simple multivariate Gaussian system is trained using features that were determined to discriminate between music and speech. He found that using statistics computed from the zero crossing rate, he could achieve a classification performance averaging 90%. By including additional information about the energy contour, he improved his accuracy on the training set to 98%. Performance on an independent test set ranged from 95% to 96%. Scheirer and Slaney [70] report similar results on their model-based speech/music discriminator. The discriminator was based on various combinations of 13 features such as 4-Hz modulation energy, zero crossing rate, and spectral centroid. They investigated a number of classification strategies, such as Gaussian mixture models and K-nearest-neighbor classifiers. When looking at long-term windows (2.4 seconds), they achieved a classification rate of 98.6% on FM radio broadcasts.

4.2.2 System Development

We chose to develop a model-based approach to segment GAD into its salient classes. In this approach, we build models for each of the seven sound classes found in GAD from a training set. Our test data is then recognized, frame by frame, by a maximum likelihood process, and the segmentation boundaries are found at locations where there is a change in acoustic class.

The maximum *a-posteriori* probability (MAP) approach [64] was used to recognize each frame as one of the seven sound categories determined in Section 4.1.1. The details of this approach are reviewed here.

In the MAP approach to recognition, we assume a probabilistic model where the sequence of units to be recognized (i.e., the frame-based sequence of sound class labels), C, produces a sequence of acoustic observations, Y, with probability $P(C, Y)$. The goal is to then determine the most probable sound class sequence, $\hat{C}$, based on the acoustic observations, so that the hypothesized sequence has the maximum *a-posteriori* probability:

$$\hat{C} \ni P(\hat{C} \mid Y) = \max_{C} P(C \mid Y) \tag{4.1}$$

Using Bayes' Rule, $P(C \mid Y)$ can be written as:

$$P(C \mid Y) = \frac{P(Y \mid C)P(C)}{P(Y)} \tag{4.2}$$

Since $P(Y)$ is independent of C, it can be eliminated from the computation, and the MAP decoding rule of Equation 4.1 is:

$$\hat{C} = \arg\max_{C} P(Y \mid C)P(C) \tag{4.3}$$

The first term in Equation 4.3 is the acoustic model, which estimates the probability of a sequence of acoustic observations given the sound class sequence. After performing principle components analysis on the feature vectors Y, the acoustic models $P(Y \mid C)$ were represented by mixtures of diagonal Gaussian distributions. The

second term in Equation 4.3 is the language model, which estimates the probability of the hypothesized sequence. In initial experiments, a unigram language model (which simply estimates the *a-priori* sound class probabilities) was used to estimate $P(C)$. The unigram probabilities were estimated from the frequency of occurance of sound classes found in the training set. In subsequent experiments, a bigram language model was developed to better model the sequential constraints of the sound classes. The bigram was trained from the frame-based sound class transcriptions.

For acoustic modeling, fourteen Mel-frequency cepstral coefficients (MFCC) were computed every 10 ms using a 20 ms Hamming window. To capture the longer-term spectral characteristics of each class, the feature vector for each frame was formed by averaging the MFCCs of adjacent frames centered around the frame of interest. This is illustrated in Figure 4-6. Experiments were performed to determine the optimal segment size. The number of frames included in the analysis segment was varied from 15 (7 frames on each side) to 81 (40 on each side). Two hours were selected from the training set for these experiments. One hour was used for system training and the second hour was used for development. In these initial experiments, a simple unigram language model was used.

As shown in Figure 4-7, the recognition accuracy on the development set increased steadily as more context was included in the analysis segment, eventually reaching a peak value of 76.5% (for an analysis segment of 51 frames). The accuracy then began to level off and decrease slightly, as the analysis segment began to include too much data from neighboring classes. Therefore, an analysis segment length of 510 ms (51 frames) was used for all subsequent experiments.

4.2.3 Feature Refinement

Examination of the results from the development set led to further refinement of the feature sets. First, many of the misrecognized frames were found to contain small portions of neighboring classes in their analysis segments. To potentially alleviate this problem, MFCC averages and derivatives across the first and last thirds of the analysis segment were added to the feature vector. Second, examination of the aver-

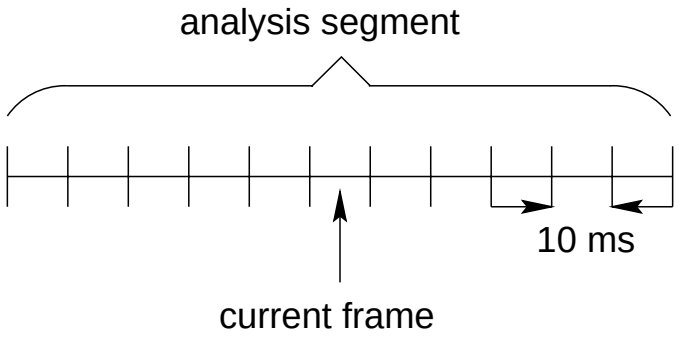

Figure 4-6: Illustration of the measurement window used in the sound recognition system.

age spectra for each sound class indicated that the average spectral energy in a frame may be a distinguishing feature for the music speech, noisy speech and clean speech sound classes. Referring back to Figure 4-4, we can see that music speech has the largest average spectral energy, followed by noisy speech and clean speech, respectively. Adding these measurements into the feature vector increased the development set recognition accuracy to 80.0%. Finally, a bigram language model was added to model the sequential constraints of the sound classes. Adding the bigram language model into the recognition system resulted in a development set recognition accuracy of 82.5%.

The recognition algorithm we have developed was evaluated on the test set (two hours of previously unseen data) after training on the full training set (eight hours of data). Using the optimal analysis segment size and measurement vector previously determined, the system achieved a recognition accuracy of 78.6%.

Details of the results of this experiment are shown numerically in Table 4.3 in the form of a confusion matrix, and are illustrated graphically in the bubble plot of Figure 4-8. The rows of Figure 4-8 are the reference classes, and the columns are the hypothesized classes. The radius of the bubbles in each entry are directly proportional to the likelihood that the reference class is recognized as the hypothesis class. If perfect recognition were achieved (i.e., recognition accuracy of 100%), this plot would display large, equal-sized bubbles along the diagonal and no off-diagonal entries.

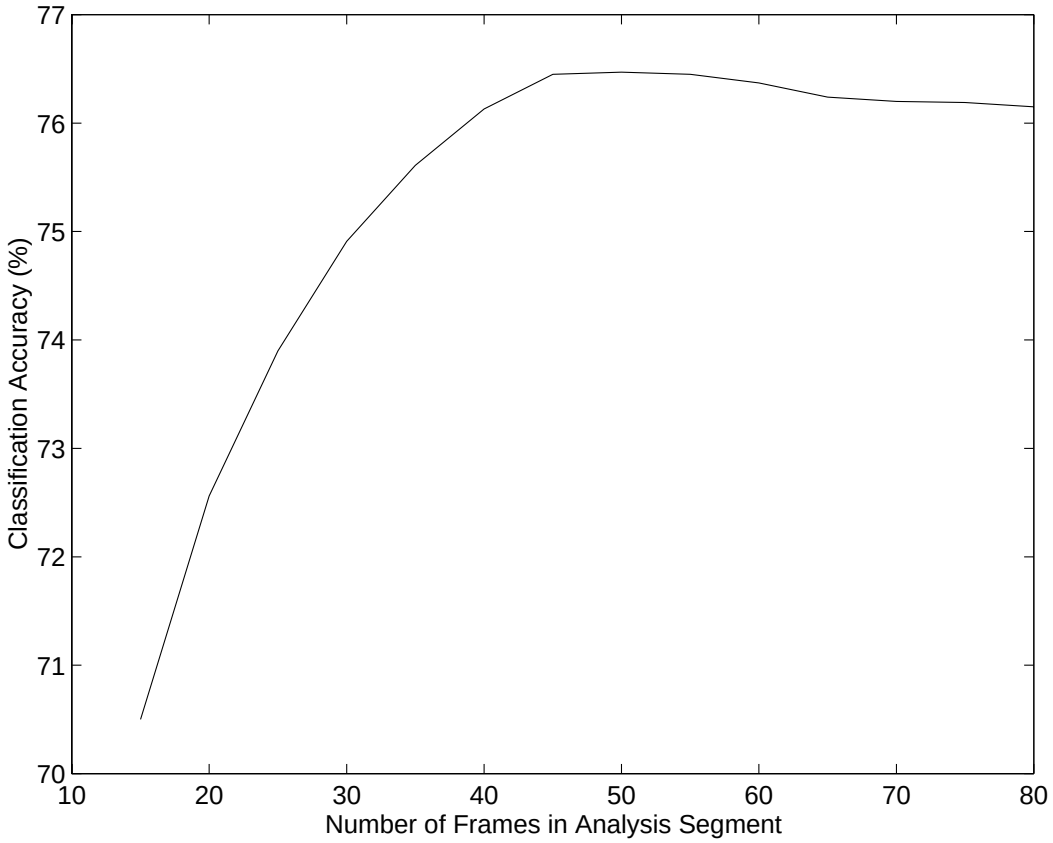

Figure 4-7: Recognition accuracy as a function of the analysis segment size.

We found that the primary confusion in the recognition system was noisy speech with music speech. Since both of these classes consist of wideband speech superimposed on other sounds, this confusion makes intuitive sense. This indicates that the music speech and noisy speech classes are difficult to distinguish. The use of more discriminating features may have to be explored to improve the results on these two classes.

The miscellaneous class is also commonly misrecognized, with most of the misclassified frames being labeled as music. One reason for this may be due to an insufficient amount of training data for this class. In Table 4.1, we see that there is just over 5 minutes of miscellaneous class data available for training. This may not be enough data to adequately train the mean vectors and covariance matrices of the Gaussian models. Since the majority of the misclassified frames are labeled as other non-speech classes, and since there is very little testing data for this class (2.3 minutes), we are

Reference	Hypothesis						
	Clean Speech	Field Speech	Music Speech	Noisy Speech	Music	Silence	Misc
Clean Speech	87.8	0.6	0.9	7.4	0.0	3.1	0.0
Field Speech	1.9	85.7	1.9	0.9	1.1	8.1	0.4
Music Speech	4.9	0.9	79.5	6.9	6.9	0.6	0.3
Noisy Speech	10.4	3.0	50.1	24.3	6.8	1.9	3.5
Music	0.3	0.1	6.2	2.5	85.4	2.4	2.9
Silence	2.6	3.7	0.0	0.8	0.1	92.1	0.7
Misc	0.3	10.7	15.0	2.6	39.2	11.7	20.6

Table 4.3: Confusion matrix for seven class sound recognition system. The overall recognition accuracy for this experiment was 78.6%

not very surprised with this result.

4.2.4 Speech / Non-Speech Recognition

Given the fact that the non-speech sounds are all significantly different from speech sounds, we decided to perform an additional experiment to determine the separability of speech and non-speech frames. The speech class was formed from the union of the clean speech, field speech, music speech and noisy speech classes. The non-speech class consisted of the union of the music, silence and miscellaneous classes. Each frame in the corpus was labeled with the proper speech or non-speech tag, and a new recognition system was developed. Using the measurements that achieved the best recognition performance in the seven class system, a recognition accuracy of 92.9% was achieved on the test set. The confusion matrix for this experiment is shown in Table 4.4. We see that speech (with a recognition accuracy of 92%) is slightly easier to recognize than non-speech (with a recognition accuracy of 89.2%).

Table 4.5 shows the details of the speech / non-speech recognition experiment, broken down by original sound class. Closer inspection of the results revealed that many short segments of silence within a speech utterance are being recognized as non-speech. Since the original transcriptions required a silence segment to be greater than 250 ms for its constituent frames to be labeled as silence, these shorter silence segments were originally transcribed as one of the speech classes. This indicates that

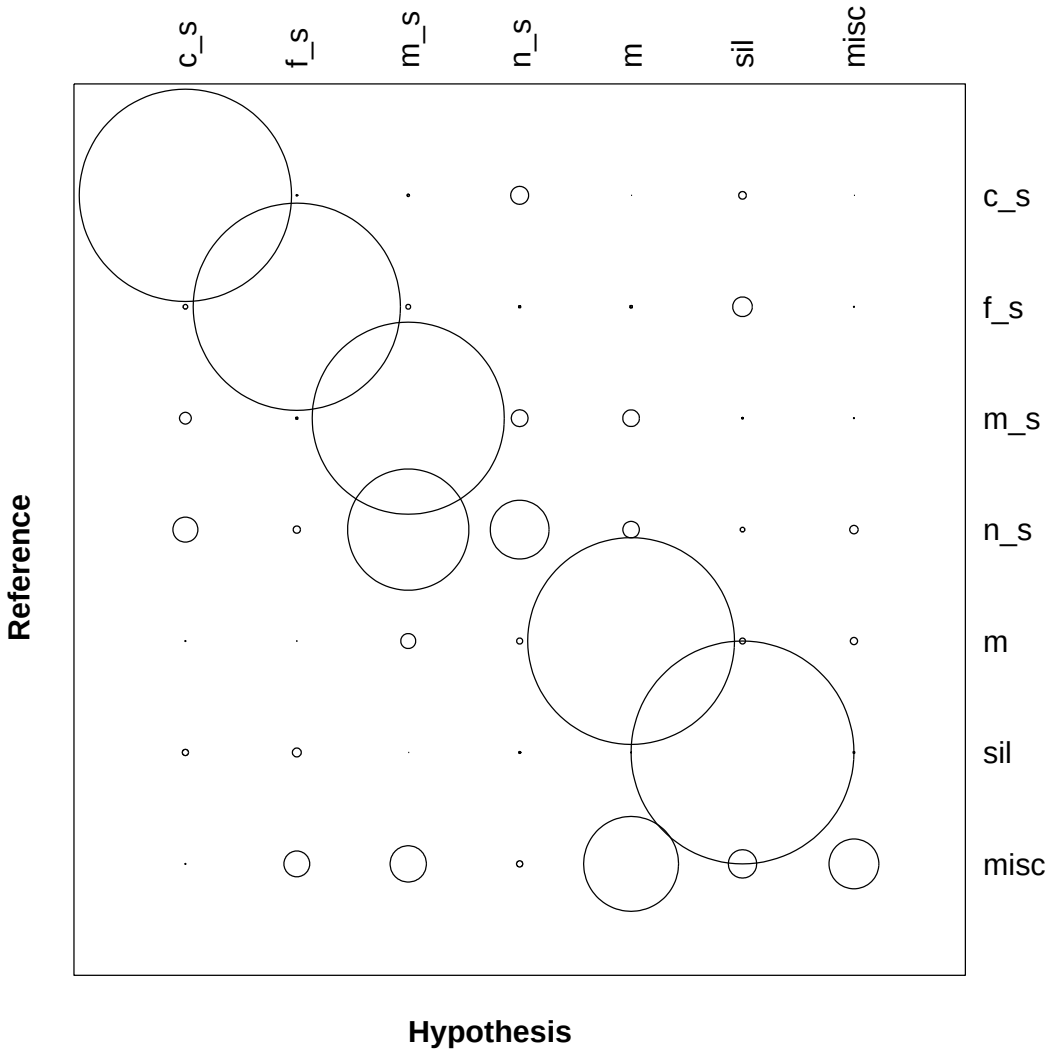

Figure 4-8: Confusion matrix for sound recognition experiment. The radius of the bubbles in each entry are directly proportional to the likelihood that the reference class is recognized as the hypothesis class. The overall recognition accuracy for this experiment was 78.6%.

Reference	Hypothesis	
	Speech	**Non-Speech**
Speech	92.0	8.0
Non-Speech	10.8	89.2

Table 4.4: Confusion matrix for speech / non-speech recognition system. The overall recognition accuracy for this experiment was 91.7%

Reference	Hypothesis	
	Speech	**Non-Speech**
Clean Speech	97.4	2.6
Field Speech	88.3	11.7
Music Speech	85.7	14.3
Noisy Speech	82.0	18.0
Music	5.0	95.0
Silence	7.5	92.5
Miscellaneous	24.0	76.0

Table 4.5: Confusion matrix for speech / non-speech recognition system, broken down by original sound class. The overall recognition accuracy for this experiment was 91.7%.

some of the speech / non-speech distinctions are difficult, if not arbitrary, which makes it hard to be definitive about proper class assignment.

4.2.5 Smoothing Experiments

While we have developed a frame-based recognition system, we have to remember that the overall goal of the system is to generate acoustically homogeneous *segments*. Segment boundaries will be proposed at locations where there is a change in the assigned acoustic class. Therefore, single-frame misrecognition errors will adversely effect the generation of these segments (i.e., many short segments will be erroneously proposed). For example, if the recognizer outputs a hypothesis string such as:

```
m m m m m m m_s m m m m m
```

three segments would be generated (a music segment, consisting of the first six frames, a music speech segment, consisting of one frame, and another music segment, consisting of the remaining four frames). However, it is likely that the single music

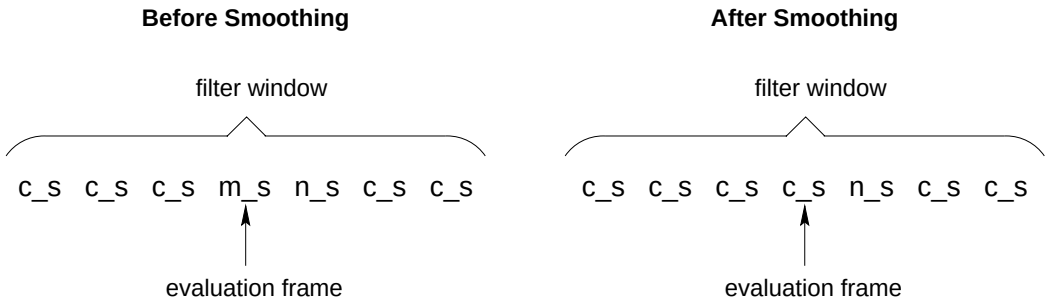

Figure 4-9: Illustration of the median filter used to smooth the output of the recognition system. The evaluation frame (m_s) is changed to the majority class (c_s).

speech frame (labeled m_s) should actually be labeled as music (m), since it is surrounded by a number of music frames. This hypothesis string would generate a single, music segment. In this section, we investigated the use of smoothing techniques to correct these singleton errors.

To accomplish this, a median filter is applied to the output of the recognition system. For each frame, the majority class is computed from a window centered on the evaluation frame. The evaluation frame is then reclassified as this majority value. Figure 4-9 illustrates the performance of the median filter. In this example, the evaluation frame (m_s) is reclassified as the majority value (c_s).

Experiments were performed to determine the optimal median filter size. The number of frames included in the filter was varied from 3 (1 frame on each side) to 49 (24 frames on each side). As shown in Figure 4-10, the recognition accuracy on the development set increases slightly as the filter size increases. After reaching a peak value of 74% (for an analysis segment size of 27 frames), the accuracy began to decrease as the filter began to expand past a true segment boundary.

The optimal filter was used to process the recognition output on the test set for both the seven class and speech / non-speech systems. The use of the filter on the seven class system yielded an increase in recognition accuracy from 78.6% to 79.4%. This slight increase may indicate that there are not many singleton errors in the seven class system. The speech / non-speech system benefitted a bit more from the smoothing process, which increased in accuracy from 92.9% to 94.2%.

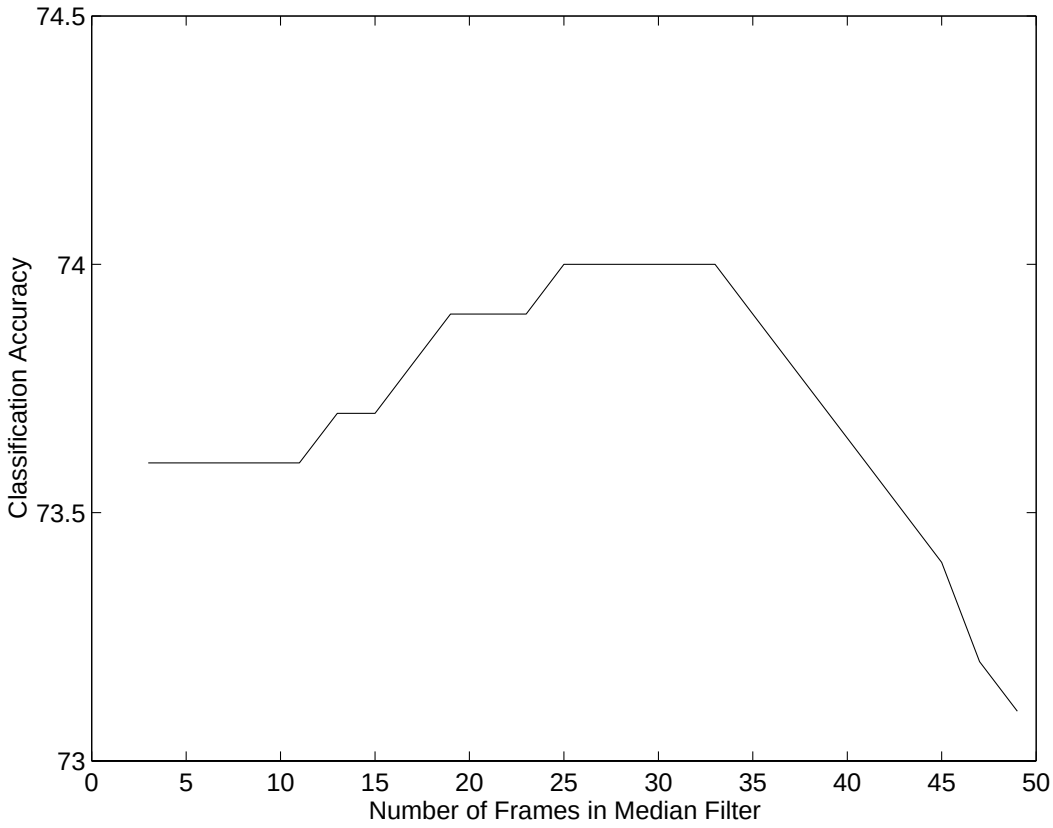

Figure 4-10: Recognition accuracy as a function of the median filter size.

4.3 Clustering Experiments

One of the goals of this chapter is to understand the acoustic nature of GAD. The acoustic classes that we have been experimenting with up to this point were determined subjectively. In this section, we investigate the use of clustering methods to provide us with insight into the similarities among the classes.

The classes were clustered according to the confusion matrix produced by the seven class recognition experiment developed in Section 4.2. The confusion matrix is shown in Table 4.3. Each row in the matrix can be viewed as a probability density. Specifically, the entry in position (a, b) of the matrix represents the probability that a frame will be recognized as b when the correct class label is a. The symmetric *Kullback Leibler* distance [14] was calculated for each pair of classes as shown below:

$$d(a,b) = \sum_{x \in X} p(x|a) \log \frac{p(x|a)}{p(x|b)} + \sum_{x \in X} p(x|b) \log \frac{p(x|b)}{p(x|a)} \tag{4.4}$$

where,

$p(x|a)$: the probability of confusing class x with class a.

$p(x|b)$: the probability of confusing class x with class b.

This distance metric provides a measure of the divergence between the conditional probability distributions of the classes. The new symmetric matrix of between-class "distances" was then used for bottom-up clustering [19] of the classes, resulting in the tree shown in Figure 4-11. The vertical axis is a measure of the distance between classes or clusters.

The clustering experiment produced some interesting results. We see that the first classes to be clustered are the music speech and noisy speech classes. These classes are acoustically very similar in that they both contain wideband speech superimposed on interference sounds. We then see that the wideband speech forms its own cluster, as the music and noisy speech cluster merges with the clean speech class. Wideband non-speech also forms its own cluster, as the music and miscellaneous classes merge. Next, these two wideband clusters merge higher in the tree. The bandlimited speech and silence classes also form a single cluster, presumably due to their low overall energy levels as compared to the wideband classes. Finally, very high in the tree, all of the clusters merge into a single class.

The results of the clustering experiment explain some of the confusions found in our sound recognition experiments. We found that the noisy speech class was most often confused with the music speech class. The results of the clustering experiment show that these two classes merge very low in the tree, and are therefore acoustically very similar. Also, we found that the miscellaneous class was often confused with the music class. These two classes also merge fairly early in the clustering tree.

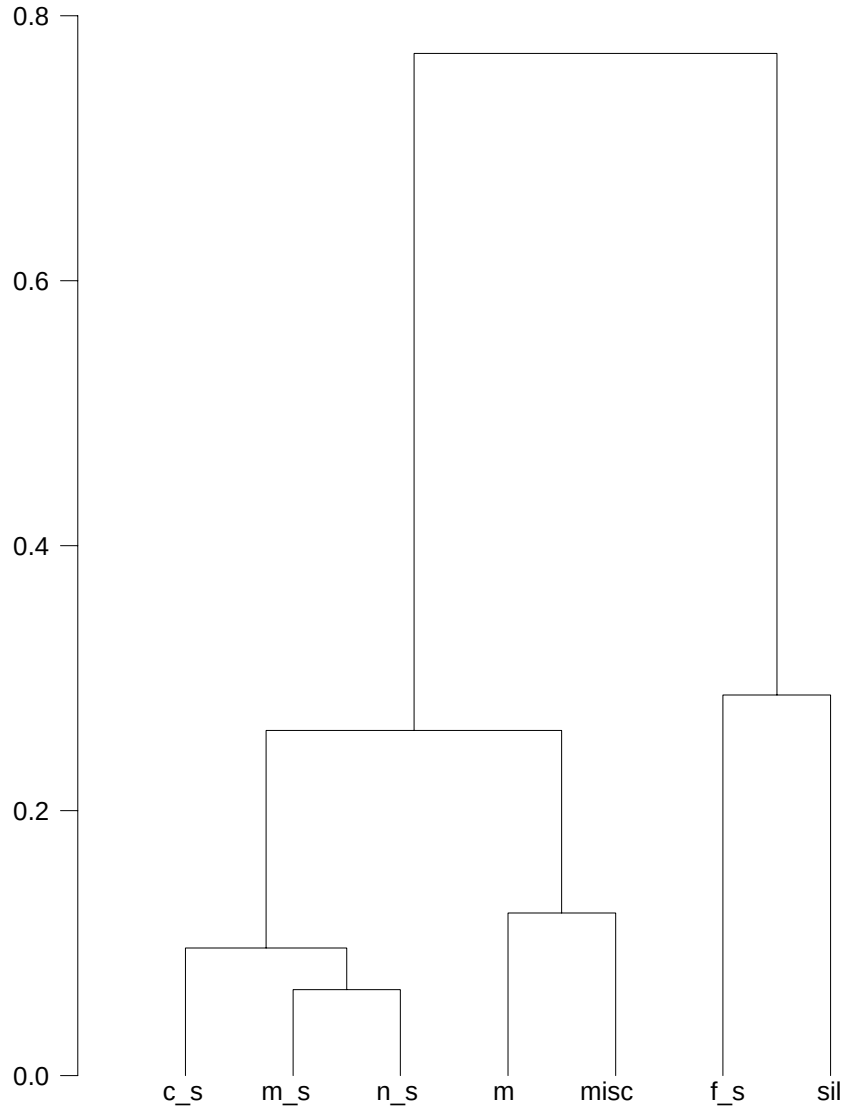

Figure 4-11: Clustering tree based on Kullback-Leibler distance computed from the confusion matrix of Table 4.3.

4.4 Modified Acoustic Classes

Since both the sound recognition system and clustering experiment results indicated that the music speech and noisy speech classes were acoustically very similar, we decided to collapse them into a single class, called interference speech (i_s). Each music speech or noisy speech frame in the NPR-ME corpus was relabeled with the new interference speech tag, and a new recognition system was developed using the resulting six classes. Using the measurements that achieved the best recognition performance in the seven-class system and a bigram trained on the relabeled data, a recognition accuracy of 85.9% was achieved on the test set. The use of the median filter developed in Section 4.2.5 increased the overall accuracy to 86.6%.

Details of the results of this experiment are shown numerically in Table 4.6 in the form of a confusion matrix, and are illustrated graphically in the bubble plot of Figure 4-12. As in Figure 4-8, the radius of the bubbles in each entry are directly proportional to the likelihood that the reference class is recognized as the hypothesis class. While the miscellaneous class is still commonly misrecognized, the majority of the classes are recognized correctly, illustrated by the large bubbles lying along the diagonal. However, we see that a small percentage of the interference speech frames are being misrecognized as clean speech. One possible reason for this confusion is that some of the frames labeled as interference speech actually contain very low levels of background music or noise. While those frames may have been erroneously recognized, such a mistake may not be very detrimental to the goal of providing a segmentation for a subsequent ASR system. First, many of these frames were difficult for the human transcriber to initially label, so perhaps they *should* have been labeled as clean speech. Second, very low level acoustic disturbances may not affect a speech recognizer's performance significantly. This will be investigated further in Chapter 5, when we study the use of our sound recognition system as a preprocessing step for phonetic recognition.

Reference	Hypothesis					
	Clean Speech	Field Speech	Interference Speech	Music	Silence	Misc
Clean Speech	92.8	0.8	4.1	0.0	2.3	0.0
Field Speech	2.0	88.0	1.6	0.8	7.1	0.6
Interference Speech	9.6	2.3	77.0	5.8	1.3	4.2
Music	0.3	0.1	5.6	87.7	2.2	4.1
Silence	4.9	5.5	0.7	0.5	88.0	0.4
Misc	0.8	10.8	16.3	20.9	10.9	40.2

Table 4.6: Confusion matrix for sound recognition system with modified acoustic classes. The overall recognition accuracy for this experiment was 86.6%

4.5 Summary

In this chapter we have examined the acoustic characteristics of GAD, and have developed a sound recognition system to segment the audio into its salient sound classes. For the NPR-ME corpus, we subjectively identified seven acoustically distinct classes based on visual and aural examination of the data. We found that these classes differed in their spectral characteristics, statistical profile, and segment duration. Specifically, we found that high quality, prepared speech constitutes only half of the entire corpus, another 25% of the data contains speech superimposed on other sounds, nearly 15% of the data was of telephone bandwidth, and the remaining 10% of the data was non-speech. We also found that while pure music segments are similar in length to speech segments, other non-speech segments are substantially shorter in length.

We were able to achieve a recognition accuracy of 79.4% for these seven classes on unseen data, using relatively straightforward acoustic measurements and pattern recognition and smoothing techniques. The results of our seven class recognition system and clustering experiments revealed that the noisy speech and music speech classes were acoustically very similar, and perhaps should be combined into a single class. A six class system was developed to investigate the consequences of merging these two classes. The resulting six class system achieved a recognition accuracy of nearly 87%. A speech / non-speech recognizer achieved an accuracy of 94.2%. These

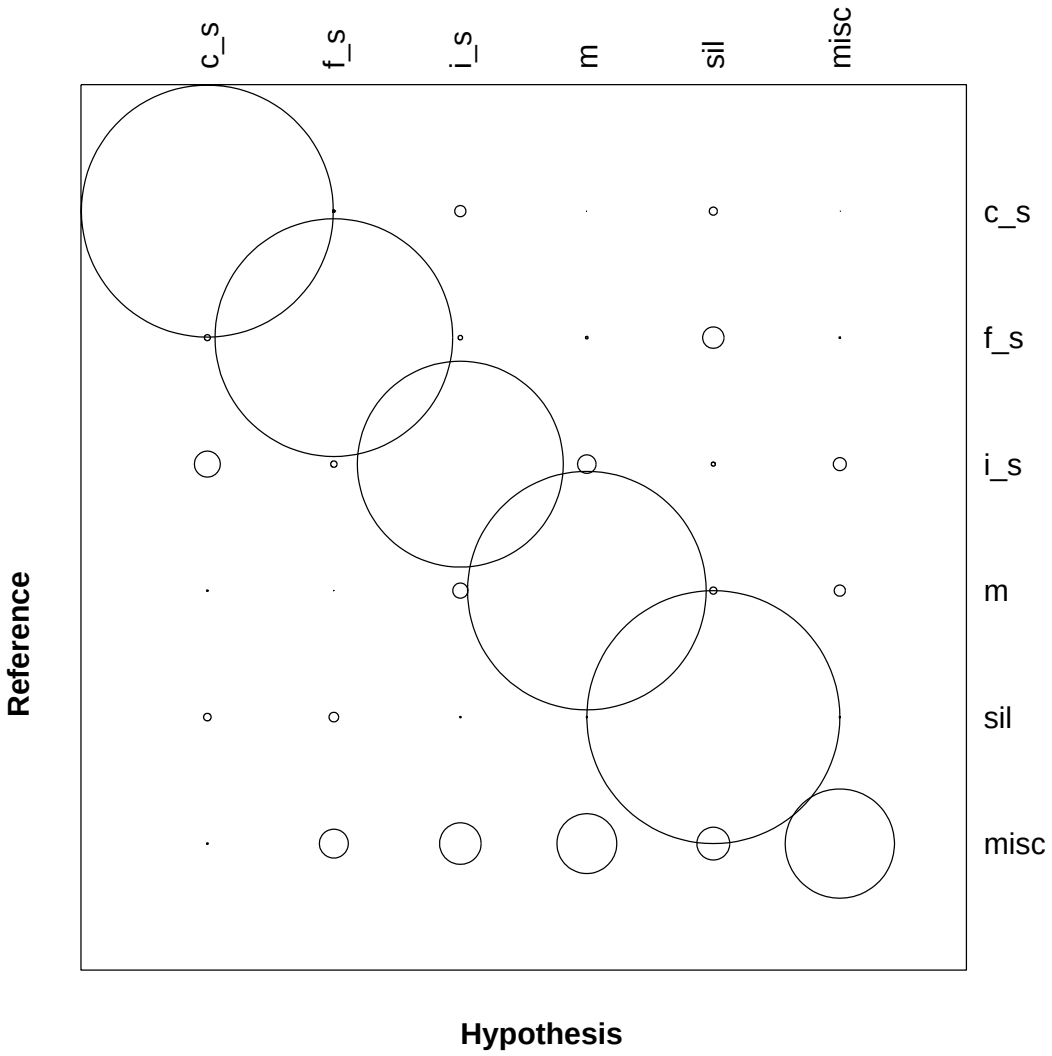

Figure 4-12: Confusion matrix for sound recognition experiment with the music speech and noisy speech classes collapsed to a single, interference speech (i_s) class. The radius of the bubbles in each entry are directly proportional to the likelihood that the reference class is classified as the hypothesis class. The overall recognition accuracy for this experiment was 86.6%.

results are difficult to compare with others found in the literature since different corpora and sound classes were used. However, our results may indicate that the finer distinctions in the sound class definitions are helpful for the recognition of this data since the results of our six-class system were significantly better than the results of the four-class system described by Dragon Systems [80]. Our speech / non-speech

recognition results are slightly worse than the results presented by Saunders [69] (95-96%) and Scheirer and Slaney [70] (98.6%). However, again, it is difficult to make a direct comparison since different corpora were used.

The level of performance needed for a sound recognizer is clearly related to the ways in which it will serve as an intelligent front-end to a speech recognition system. We will investigate this in the following chapter, where we use our seven class system to classify test data in a phonetic recognition task. We will also evaluate how the merging of the noisy and music speech classes impacts the speech recognition system.

Chapter 5

Phonetic Recognition

The lexical analysis completed in Chapter 3 revealed some potential problems with a large vocabulary, continuous speech recognition approach to the transcription of GAD. We found that the NPR-ME corpus, which consists of 102 hours of data, has a vocabulary of over 30,000 words. While this is within the current capabilities of automatic speech recognition systems, the growth of the vocabulary shows no sign of abating as more data is accumulated. Analysis of similar corpora (WSJ and LA-Times), indicates that the vocabulary of NPR-ME could grow to over 100,000 words if a whole year's worth of just this one show was collected. In addition, we found that the majority of new words that are encountered are high content words (i.e., proper nouns), which could not be found in any standard dictionary. However, these words would be very important to recognize if we were describing the content of this data for an information retrieval system.

An alternative approach that has the potential to deal with the above problems is the use of a subword unit for the recognition of GAD. The use of subword units in the recognizer constrains the size of the vocabulary needed and reduces the amount of data required for training. Word-based approaches to the transcription of GAD traditionally use vocabularies on the order of 64,000 words, and hundreds of millions of words from text documents for language model training. Conversely, subword-based approaches have much smaller vocabularies, and therefore require much less data for language model training. For example, a subword phonetic recognizer has a closed

vocabulary of 61 phones and only requires hundreds of thousands of phone occurrences for training the language models. However, the constraining power of phonetic units is much less than that of word units. Larger subword units will presumably be required for acceptable performance.

Ng [62] recently investigated the feasibility of using subword unit representations for spoken document retrieval. He examined a range of subword units of varying complexity derived from phonetic transcriptions. The performance of a spoken document retrieval system was compared when the underlying phonetic transcriptions were perfect and when they contained phonetic recognition errors. He found that some subword units achieved comparable performance to text-based word units if the underlying phonetic units were recognized correctly, and that the information retrieval performance was directly correlated with the phonetic recognition performance. Therefore, it is important to recognize the phonetic units correctly. In this chapter we determine how well phonetic units can be extracted from the speech signal for a subword representation.

In Chapter 4 we found that GAD contained a number of different acoustic speaking environments. Since the performance of ASR systems can vary a great deal depending on speaker, microphone, recording conditions and transmission channel, we have proposed that the transcription of GAD would benefit from a preprocessing step that first segments the signal into acoustically homogeneous blocks so that appropriate models could be used during test. We explore this hypothesis in this chapter and try to determine the best training and testing approach for the phonetic recognition of GAD.

5.1 Corpus Preparation

The experiments presented in this chapter required additional processing of the NPR-ME corpus. As described in Chapter 2, ten hours of NPR-ME data was segmented into manageable sized waveform files at silence breaks. Each 10 ms frame was then manually labeled with one of the seven sound classes defined in Chapter 4. If any

of the waveform files contained multiple acoustic classes (e.g., a segment of music followed by a segment of clean speech) they were manually split at these boundaries. Therefore, each file was homogeneous with respect to general sound class.

To complete phonetic recognition experiments, we must be able to train acoustic models for each phone to be recognized. This requires phonetically aligned data. Files that contained speech materials required further processing to generate these alignments. First, word-based orthographies were manually generated for each waveform file from the complete show transcript. This was done by listening to each waveform file and tagging the show transcript to indicate the word strings spoken in each individual file. Second, a word-based lexicon was developed for each show. The configuration of SUMMIT that was used in this work typically operates with vocabulary sizes on the order of 3,000 words. Therefore, the development of the lexicon was done on a show by show basis, rather than on the entire NPR-ME corpus as a whole (which has a vocabulary size of over 30,000 words). The function of the lexicon is to list the unique words present in the NPR-ME data and to represent their alternate pronunciations. The lexicon was first created by generating a unique vocabulary list for each show. For each word in the vocabulary a *baseform* pronunciation was then specified. For example, the baseform for "Boston" might be /b ao s t ax n/. The baseform pronunciations were initially generated from the Carnegie Mellon Pronouncing Dictionary [7]. Any pronunciation that was not found in the dictionary was generated manually. Pronunciation rules were then applied to generate alternative pronunciations within and across words.

SUMMIT was used with a forced Viterbi search to generate the phonetic alignments. The Viterbi search matches the constraints provided by: 1) the word orthographies, which specifies what words are spoken in each waveform file, 2) the lexicon, which specifies allowable sequences of phones corresponding to the word pronunciations, and 3) the acoustic phonetic network, which generates likely phonetic segments and the scores of each segment. Given these constraints, the search finds the best scoring pronunciations and phonetic alignments for each waveform file. This was completed with an iterative process. First, acoustic models trained with the

Environment	Training Data	Testing Data
Clean Speech	59.2%	56.4%
Music Speech	11.6%	4.9%
Noisy Speech	13.8%	9.2%
Field Speech	15.4%	29.6%

Table 5.1: Distribution of training and testing data for each speaking environment.

TIMIT corpus (described in Chapter 2) were used to generate initial alignments for the NPR-ME clean speech data. The acoustic models were then retrained using the newly aligned clean speech data. This alignment and retraining procedure was repeated until the phonetic recognition performance on this data reached a local minimum. The resulting models were then used to generate phonetic alignments for the remainder of the NPR-ME speech data.

From this collection of data, eight shows were used for system training, containing 6.5 hours of speech data. The remaining two shows were used for system test, containing a total of 1.5 hours of speech data. Table 5.1 summarizes the distribution of training and testing data in each speaking environment.

5.2 Experimental Set-up

The phonetic recognition system used in this work is based on the SUMMIT recognizer described in Chapter 2. Some of the details of the system are reviewed here. Our implementation of SUMMIT uses 61 context-independent segment acoustic models corresponding to the TIMIT phone labels (a list of the 61 phones in the TIMIT corpus with their IPA symbols, TIMIT labels, and example occurrences is shown in Table 2.4). The feature vector used in the segment models has 77 measurements consisting of three sets of 14 Mel-frequency cepstral coefficient (MFCC) and energy averages computed over segment thirds, two sets of MFCC and energy derivatives computed over a time window of 40 ms centered at the segment beginning and end, log duration, and a count of the number of internal boundaries proposed in the

segment. Context-dependent boundary models, which try to model the transitions between two adjacent segments, are used in conjunction with the segment models. The boundary model feature vector has 112 dimensions and is made up of eight sets of MFCC averages computed over time windows of 10, 20, and 40 ms at various offsets ($\pm 5, \pm 15$, and ± 35 ms) around the segment boundary. For the 6.5 hours of acoustic training data in the NPR-ME corpus, there is a total of 3160 unique transitions. Since many of these transitions occur infrequently, models are not trained for all 3160 transitions found in the training data. Transitions that occur fewer than 50 times are combined with acoustically similar transitions. Boundary models are then trained for the resulting 790 boundary classes. Cepstral mean normalization [1] and principle components analysis [78] are performed on the acoustic feature vectors.

Since the EM algorithm [16] used to train the acoustic models makes use of random initializations of the parameter values and only converges to a *local* optimum, different sets of models can result from different training runs using the same training data. It has been shown that combining the different models into a single, larger model results in better performance than using just the set of models that yield the best performance on a development set [41]. We used this approach and combined five separate acoustic models trained using different random initializations. The models are combined using a simple linear combination with equal weights for each model. In initial experiments, we verified that this approach did improve our baseline results, by up to 5% in some cases.

A statistical bigram language model was used to constrain the forward Viterbi search during decoding. The bigram model was trained using the phonetic transcriptions of the complete training set, which consisted of approximately 340,000 phone occurrences.

Results, expressed as phonetic recognition error rates, are collapsed down to the 39 labels typically used by others to report recognition results [9, 32, 37, 55]. The mapping between the 61 phones used for acoustic modeling and the 39 classes used for computation of results is shown in Table 2.5.

All of the experiments completed in this chapter use only those segments that con-

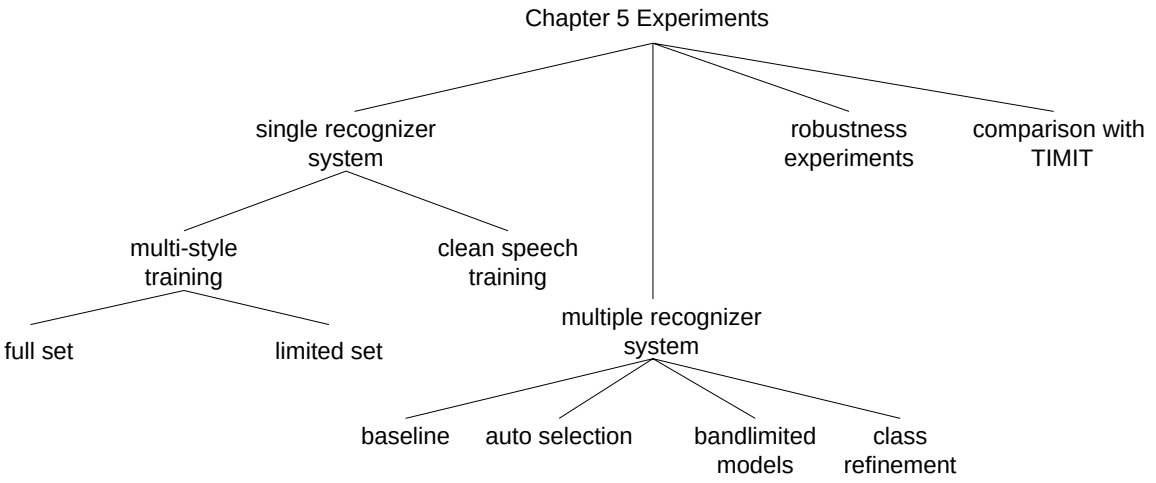

Figure 5-1: Road map of the phonetic recognition experiments presented in this chapter.

tain speech material. All non-speech segments (e.g., music, silence and miscellaneous) were not considered for the work described in this chapter.

A number of recognition systems and experiments are presented in this chapter. Figure 5-1 illustrates how they are related to one another and presents a road map for the remainder of the chapter. In Section 5.3 we investigate the use of a single recognizer system for the phonetic recognition of GAD. Two training methods are explored, namely, a multi-style approach, and a clean speech approach. To compare these two methods fairly, the multi-style approach is developed with two different amounts of training data (labeled full set and limited set in the figure). Section 5.4 investigates the use of a multiple recognizer system for the phonetic recognition of GAD. First, baseline performance is determined. Next, the sound recognition system developed in Chapter 4 is used to automatically select the models used in the multiple recognizer system (labeled as auto selection in the figure). Next, a bandlimited system is developed to determine if we could improve the performance on the field speech data. We then refine our sound classes and determine how they affect the phonetic recognition results. Finally, in Sections 5.6 and 5.7, we examine robustness issues and compare our results with those found in the literature on the TIMIT corpus.

5.3 Single Recognizer System

One approach to the design of a recognition system is to use a single recognizer for all testing conditions. If a single recognizer is to be used for all four different types of speech material present in the NPR-ME corpus, we can utilize the training data in two different ways. First, we can use a large quantity of mixed quality data. This would attempt to capture all of the acoustic differences that exist among the four speaking environments in a single model, while utilizing all of the training data available. Second, we can use a smaller amount of acoustically clean data. This would allow us to model only the speech units themselves, without any possible corruption from background noise. In this section, we explore both approaches, and try to determine the trade-offs between them.

5.3.1 Multi-Style Training

Multi-style training [58] was originally used to train a speech recognition system on multiple speaking styles in order to increase robustness to mismatched conditions. The system was trained on normal speaking styles and tested on abnormal speaking styles, such as speaking under stress. In general, these experiments showed that incorporating a variety of styles in training improves performance under different styles in testing, thereby reducing the potential mismatch between training and testing conditions. In addition to having a more diverse training set, this approach enables us to utilize a large amount of data to train our acoustic models, which may also make them more robust. To accomplish this, acoustic models were trained using all of the available training data in all four speaking environments (clean speech, music speech, noisy speech, and field speech). This amounted to a total of 6.5 hours of speech training data for the multi-style system.

The first row of Table 5.2, labeled multi-style1, shows the phonetic recognition error rates for each speaking environment in the test set. We see that the performance on the test set varied widely across speaking environments, with the lowest error rate arising from clean speech (28.3%) and the highest error rate arising from field speech

(48.3%). The phonetic recognition error rate on the entire test set was 35.8%.

5.3.2 Clean Speech Training

Although multi-style training reduces the mismatch between the training and testing data, this approach may produce models that are too general for some of the testing environments. To model the wide range of acoustic conditions present in the training data, the multi-style models may have a larger variance than required by some of the individual testing environments. An alternative is to train models using only the clean, wideband speech material found in the training set. This amounted to a total of 3.9 hours of training data. The second row of Table 5.2 shows the results of this experiment. Again, we found that the phonetic recognition performance on the test set varied across speaking environments. The overall phonetic recognition error rate was 37.9% for this experiment.

While the clean speech recognizer produced results on the clean speech test data that were comparable to the multi-style approach, the results on the other speaking environments were significantly worse. However, one must keep in mind that the multi-style approach utilized nearly 1.7 times the amount of data for training the acoustic models. This additional data allows the multi-style models to be more accurately trained. Also, the multi-style system has more modeling parameters. In SUMMIT, the number of Gaussian mixtures is determined automatically based on the amount of available training data. To perform a fair comparison between these two approaches, we trained a multi-style system with an amount of training data equivalent to that of the clean speech system, keeping the distribution of the speaking environments equivalent to that of the entire training set. As shown in the third row of Table 5.2, labeled multi-style2, we found this system degraded the full multi-style results to an overall error rate of 36.7%, a relative increase in error rate of nearly 3%. When we compare these results to the clean speech system, we see that the clean speech system performed significantly better on the clean speech testing data. However, for the remaining speaking environments, the limited multi-style system outperformed the clean speech system. This indicates that the superior performance

Training Data	Testing Data				
	Clean Speech	Music Speech	Noisy Speech	Field Speech	Over All
Multi-Style1	28.3	35.3	45.8	48.3	35.8
Clean Speech	28.2	44.8	50.5	53.2	37.9
Multi-Style2	29.4	36.9	46.8	48.8	36.7

Table 5.2: Summary of phonetic recognition error rates for the multi-style and clean speech training systems. The multi-style1 system uses all of the available training data, while the multi-style2 system uses an amount of training data comparable to the clean speech system.

achieved by the multi-style system was not due to the increased amount of training data or the increased number of modeling parameters, but because the models were more robust to acoustic differences presented by the music, noisy and field speech data.

5.4 Multiple Recognizer System

A second approach to the recognition of data with a variety of acoustic conditions is to use multiple recognizers, each trained on one of the unique conditions. A multiple recognizer system involves training separate models to match each speaking environment present in the NPR-ME corpus. Like multi-style training, this approach decreases the mismatch between training and testing conditions by incorporating different data in training. Unlike multi-style training, this approach provides a separate model to directly match each condition rather than pooling the data and averaging the model parameters over different conditions. As a result, a multiple recognizer system may be able to model and represent more diverse testing conditions.

An added complexity in the multiple recognizer system is that the appropriate recognition system must be determined during testing. There are two approaches to this problem. First, each test utterance is run through each recognizer, and the highest scoring output is selected. This method has the advantage that a separate environment classification system is not required. However, each test utterance must

be run through every recognizer, which may not be computationally efficient. The second approach, and the one taken in this work, is to utilize a sound classification system that determines the speaking environment of each test utterance. The matching recognizer is then used to process the utterance.

In this section we explore the use of a multiple recognizer system for the phonetic recognition of NPR-ME, one for each type of speech material. These results will be compared to the single recognizer systems described in the previous section.

5.4.1 Environment Specific Baseline

The environment-specific approach involves training a separate set of models for each speaking environment, and using the appropriate models for testing. In this section, we establish baseline performance by using the manually assigned labels for training and testing. This is equivalent to assuming that the environment classification has been done without error. Table 5.3 details the results in the form of a confusion matrix. Each entry in the table indicates the phonetic recognition error rate for each training and testing condition. The overall recognition error rate for the complete test set is shown in the last column. The matched condition results are shown in boldface type on the diagonal of the table.

In addition to seeing how well each matched condition performs, Table 5.3 illustrates the consequences of testing with a mismatched system. We can see that in the clean, music and field speech testing cases, the matched condition yields significantly better results than any of the mismatched conditions. However, the matched noisy speech condition performs only slightly better than the mismatched noisy/clean and noisy/music speech conditions. This may indicate that the noisy speech class is not significantly different than the music speech or clean speech classes. This result was also found in Chapter 4, where we saw that noisy speech was often confused with music speech. We will investigate the consequence of combining these two classes into a single class on our phonetic recognition system later in this chapter.

The multiple recognizer system achieved an overall error rate of 36.7%, computed from the weighted diagonal entries of the confusion matrix. This result represents a

Training Data	Testing Data				
	Clean Speech	Music Speech	Noisy Speech	Field Speech	Over All
Clean Speech	**28.2**	44.8	50.5	53.2	37.9
Music Speech	45.3	**33.5**	50.7	57.8	48.7
Noisy Speech	39.2	48.1	**50.0**	55.4	45.1
Field Speech	55.6	55.7	57.9	**49.3**	54.1

Table 5.3: Summary of phonetic recognition error rates for the environment-specific training system, in the form of a confusion matrix. The overall error rate for this experiment, computed from the weighted diagonal entries, was 36.7%.

2.5% *increase* in error rate when compared to the multi-style training system. Table 5.4 summarizes the results of the multi-style system for comparison with the multiple recognizer system. If we look at the first row of Table 5.4, we see that for all but the clean speech class, the multi-style system outperforms each of the matched condition cases in the environment specific system. However, we again have to keep in mind that the multi-style system utilizes significantly more training data to develop its acoustic models than any of the individual environment-specific systems. Therefore, the improvement in error rate may be due simply to the increased amount of training data, rather than the inclusion of a variety of speaking conditions. To provide a fair comparison, we retrained the multi-style system for each test case to match the amount of environment-specific training data for that case. For example, the music speech class has 42.8 minutes of training data available. Therefore, to compute multi-style results on the music speech test data, we retrained the multi-style system using an equivalent amount of mixed class data. The distribution of classes in the new multi-style training set was similar to the distribution of classes in the entire training set. These results are shown in the second row of Table 5.4, labeled multi-style3. We see that the results for all of the speaking environments are degraded, resulting in an overall error rate of 38.7%. This represents a relative increase of 8.1% from the full multi-style system. If we compare the multi-style3 results with the environment-specific system, we find that in all but the noisy speech case, the environment-specific matched case outperforms the multi-style3 system. This indi-

Training Data	Testing Data				
	Clean Speech	Music Speech	Noisy Speech	Field Speech	Over All
Multi-Style1	28.3	35.3	45.8	48.3	35.8
Multi-Style3	29.4	43.0	50.0	52.2	38.7

Table 5.4: Summary of phonetic recognition error rates for the multi-style training systems. The multi-style1 system uses all of the available training data while multi-style3 uses an amount of training data comparable to each of the test speaking environment systems (i.e., the multi-style3 system uses a comparable amount of training data to the music speech system when the results on the music speech test data are computed).

cates that the improvement in performance shown by the multi-style1 system was primarily due to the increased amount of training data available.

When compared to the clean speech single recognizer system, the multiple recognizer system results achieve a relative decrease in phonetic error rate of 3.2%. We can see that for every testing condition, the matched condition case in the multiple recognizer system outperformed the clean speech system.

5.4.2 Integrated System

The experiments described thus far have assumed that test utterances have been classified perfectly. We now use the sound recognition system described in Chapter 4 as a preprocessor to classify each test utterance as one of the seven predefined sound classes. This was accomplished by using the sound recognition system to label each frame in the test utterance. The class containing the majority of frames was used to label the test utterance. The environment-specific model chosen by the automatic classifier was then used to perform the phonetic recognition. We should point out that this approach could not be directly applied in practice. Recall (cf. page 96) that we performed a preprocessing step that segmented the NPR-ME test data into acoustically homogeneous files. While here we are using our sound recognition system to determine the class label for each test utterance, in practice the sound recognition system would have to be used to also segment the data, which would possibly degrade

Auto Class	Error Rate	% of Tokens
Clean Speech	28.3	56.2
Music Speech	44.4	9.2
Noisy Speech	40.0	6.9
Field Speech	49.3	27.8

Table 5.5: Auto-class selection phonetic recognition results. The overall error rate for this experiment, computed from the weighted entries, was 36.5%.

the results.

The results for this experiment are shown in Table 5.5. The first column in Table 5.5 shows the phonetic error rate for each of the automatically chosen classes. The second column shows the percent of tokens in each automatically chosen class. The overall error rate for this experiment was 36.5%, which is slightly better than the baseline environment-specific system.

When we compare these results in more detail, we see that the automatically classified noisy speech data performed significantly better than the baseline system's noisy speech data, while the music speech class performed worse. In addition, we see that 9.2% of the test data was automatically classified as music speech, while this class actually makes up only 4.9% of the test data. We also see that only 6.9% of the data is automatically classified as noisy speech, while this class makes up 9.2% of the test data. These misclassified noisy speech utterances appear to have degraded the overall music speech performance.

5.4.3 Bandlimited Field Speech Models

In all of the experiments conducted thus far, the full bandwidth range (upper limit of 8 kHz) has been used in developing our acoustic models. However, as we discovered in Chapter 4, the field speech data is bandlimited to 4 kHz. The MFCCs used to represent the speech signal are computed from Mel-scaled spectral coefficients (MFSCs), using a cosine transformation. Since the higher-order MFSCs (which represent frequencies over 4 kHz) do not contain any energy for the field speech data, they

Bandwidth	Training Data				
	Clean Speech	Music Speech	Noisy Speech	Field Speech	Multi Style
Full-Bandwidth (8 kHz)	53.2	57.8	55.4	49.3	48.3
Bandlimited (4 kHz)	48.4	55.1	51.4	48.9	46.0

Table 5.6: Summary of field speech phonetic recognition error rates on field speech data for bandlimited training system.

do not contribute anything to this transformation. In essence, these coefficients are being wasted in the field speech case. By bandlimiting the analysis to 4 kHz for the field speech data, we are fully utilizing all of the spectral coefficients in the MFCC computation. This may improve our recognition performance on the field speech data. Table 5.6 shows the results of testing the field speech data on full-bandwidth (8 kHz) and bandlimited (4 kHz) models for all training conditions. The first row of the table shows the results for the full-bandwidth system, while the second row shows the results for the bandlimited system.

We can see that for all training cases, the bandlimited models significantly reduce the error rate on the field speech test data. If bandlimited field speech models are used in the environment-specific system, the overall error rate is slightly reduced to 36.3% (from 36.5%). The overall error rate of the multi-style system is also reduced to 35.5% (from 35.8%) if bandlimited models are used to test the field speech data.

5.4.4 Refinement of Sound Classes

We have found that the noisy speech class is often misrecognized as music speech, and this misrecognition significantly increases the error rate on the music speech test data. In Chapter 4, we developed a recognition system that collapsed the music and noisy speech classes into a single class. In this experiment, we investigate how this would effect the environment-specific recognition system.

First, we establish baseline performance by using the manually assigned labels for training and testing. The results of this baseline experiment are shown in Table 5.7,

Training Data	Testing Data			
	Clean Speech	Interference Speech	Field Speech	Over All
Clean Speech	**28.2**	48.6	53.2	37.9
Interference Speech	37.4	**43.1**	54.3	42.8
Field Speech	55.6	57.2	**49.3**	54.1

Table 5.7: Environment-specific training system phonetic recognition results, using the collapsed class set. The overall error rate for this experiment, computed from the weighted diagonal entries, was 36.6%

in the form of a confusion matrix. Each entry in the table indicates the phonetic recognition error rate for each training and testing condition. The overall recognition error rate for the complete test set is shown in the last column. The matched condition results are shown in boldface type on the diagonal of the table.

We can see that the matched condition for the interference speech class performed significantly better than any of the mismatched conditions. The overall error rate for this experiment, computed from the weighted diagonal entries of the confusion matrix, was 36.6%, which is essentially equivalent to the full environment-specific results. If we look at the interference speech results more closely, we find that the constituent noisy speech data benefitted from the new interference speech model, decreasing in error rate from 50.0% to 47.3%. However, the music speech test data was degraded by the interference speech model, *increasing* in error rate from 33.5% to 34.5%, which effectively canceled the gains made on the noisy speech data.

We now use the modified sound recognition system described in Chapter 4 as a preprocessor to classify each test utterance as one of the six defined sound classes. The environment-specific model chosen for each utterance by the automatic classifier was then used to perform the phonetic recognition. The results for this experiment are shown in Table 5.8. The first column in Table 5.8 shows the phonetic error rate for each of the automatically chosen classes. The second column shows the percent of tokens in each automatically chosen class. The overall error rate for this experiment was 35.9%, which represents a relative decrease of 2.2% from the original

Auto Class	Error Rate	% of Total
Clean Speech	28.6	60.4
Interference Speech	42.2	11.9
Field Speech	49.2	27.7

Table 5.8: New auto-class selection phonetic recognition results. The overall error rate for this experiment, computed from the weighted entries, was 35.9%.

environment-specific system results.

When we examine these results in more detail, we see that 60.4% of the test data was automatically classified as clean speech, while this class actually makes up only 56.4% of the test data. We also see that only 11.9% of the data is classified as interference speech, while this class makes up 14.1% of the test data. These misclassified interference speech utterances appear to have only slightly degraded the overall clean speech performance, while increasing the overall interference speech performance. As we discussed in Chapter 4, we found that many of the noisy and music speech utterances contained very low-levels of background music or noise. The result found here indicates that these utterances may be better modeled with clean speech.

5.5 Statistical Significance

The results of the phonetic recognition experiments conducted in this chapter are summarized in Figure 5-2. We found that if a single recognizer system is to be used, training on all of the available data which contains a variety of speaking environments is more effective than using a smaller amount of homogeneous, clean data. We also found that a multiple recognizer system achieved performance similar to a single multi-style recognizer. However, when comparing the differences which exist between the performance of different systems, it is important to consider whether these differences are statistically significant. If the difference between the performance of two systems is not statistically significant then it is possible that a test result indicating

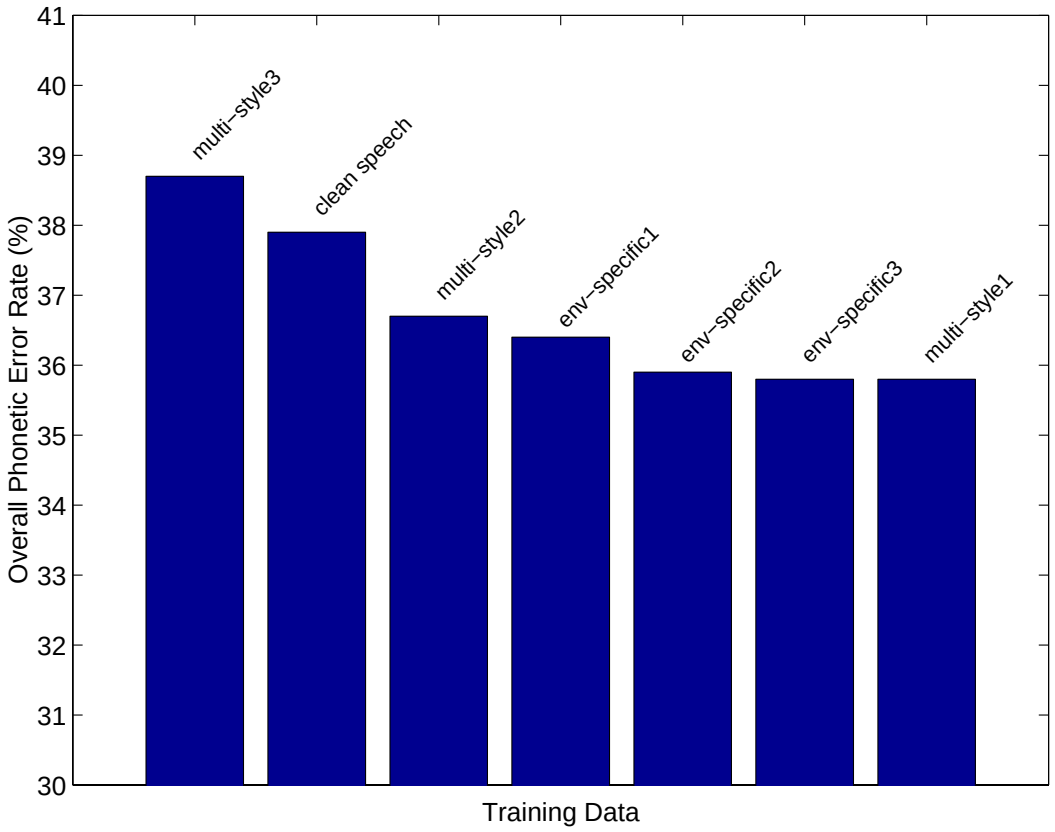

Figure 5-2: Summary of phonetic error rate results for different training methods. The multi-style1 system uses all of the available training data, the multi-style2 system uses an amount of training data comparable to the clean speech system and the multi-style3 uses an amount of training data comparable to each of the test speaking environment systems. Each of the environment-specific systems used the sound recognition system as a preprocessor to select the appropriate models for testing. The env-specific1 system uses the four original speaking classes, the env-specific2 system collapses music and noisy speech into a single class, and the env-specific3 system adds bandlimited models for the field speech data.

	clean speech	multi-style2	env-specific1	env-specific2	env-specific3	multi-style1
multi-style3	*.001*	*.001*	*.001*	*.001*	*.001*	*.001*
clean speech	—	*.001*	*.001*	*.001*	*.001*	*.001*
multi-style2	—	—	*.001*	*.001*	*.001*	*.001*
env-specific1	—	—	—	*.001*	*.001*	*.001*
env-specific2	—	—	—	—	**.131**	*.016*
env-specific3	—	—	—	—	—	**.139**

Table 5.9: Measure of statistical significance of differences between different phonetic recognition systems. Significant differences are shown in *italics* while insignificant differences are shown in **boldface**. All results with a significance level less than .001 are simply listed as having a significance level of .001.

that one system outperforms another is simply the result of chance and not an indicator of true superiority. To evaluate the significance of the results presented in the previous sections, the *matched pairs sentence segment word error test* is utilized [29]. This test measures the likelihood that the differences present during an evaluation between two systems are a result of chance as opposed to genuine differences in the performance of the systems.

Table 5.9 shows the significance values for the comparison of different pairs of recognition systems presented in Sections 5.3 and 5.4. The differences are considered significant if the likelihood of the differences occurring due to chance is estimated to be .05 or less. In other words, the results are considered significant if there is a 95% chance or better that the difference in performance between the two systems is a result of genuine differences in the recognition systems. In the table, significant differences are indicated with *italics* while insignificant differences are indicated with **boldface**. Also, all results with a significance level less than .001 are simply listed as having a significance level of .001 in the table.

The results shown in Table 5.9 indicate that there is a significant difference between the performance of the majority of recognition systems presented in this chapter. The environment-specific systems (listed as env-specific1, env-specific2 and env-specific3 in the table) performed significantly better than the clean speech and limited multi-style systems (listed as multi-style2 and multi-style3 in the table). The table

also indicates that the performance of the environment-specific system with bandlimited field speech models is statistically equivalent to the full multi-style system (listed as multi-style1 in the table).

5.6 Robustness Experiments

In this section we explore issues related to training set size. We found that much of the gain achieved by the multi-style system over the environment-specific system was due to the increased amount of training data available for the multi-style system development. Here, we investigate the effect of training set size on the environment-specific system. First, we attempt to determine the source of the large discrepancy that exists among the environment-specific results. We found that the clean speech data performed significantly better under matched training and testing conditions than any of the other speech classes. The discrepancy between the clean speech data results and the music, noisy and field speech data results could be due to the increased amount of training data available for the clean speech class. The discrepancy could also be due to the more difficult acoustic conditions presented by the music, noisy and field speech classes. Second, we will determine whether more training data could potentially improve the clean speech system.

5.6.1 Comparison of Results Among NPR-ME Classes

In all of the environment-specific experiments presented in this chapter, we found that the music, noisy and field speech data performed significantly worse than the clean speech data under matched training and testing conditions. There are two possible explanations for this result. First, if we examine Table 5.1, we find that the clean speech class has significantly more data available to train the acoustic models. Therefore, the clean speech models may simply be more accurately trained. Second, the acoustic conditions presented by the music, noisy and field speech classes may make this data more difficult to recognize. In an attempt to discern between these two explanations, we trained the clean speech recognizer with amounts of data

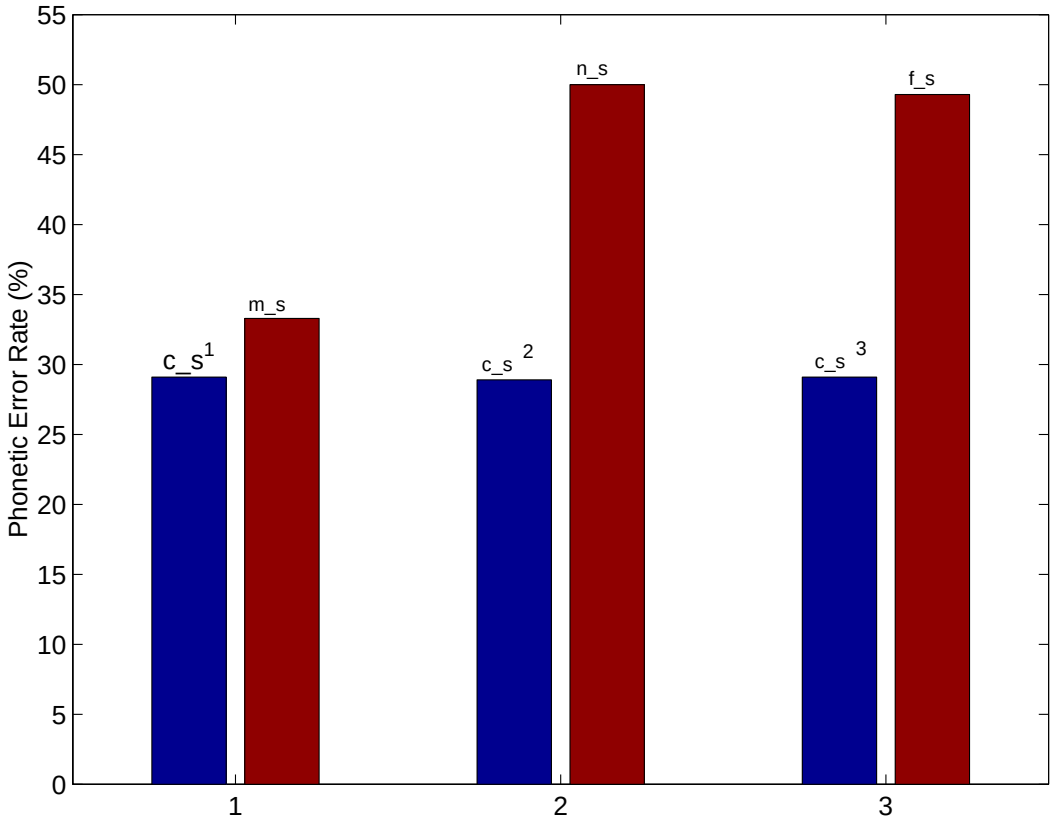

Figure 5-3: Illustration of the effects of limiting the amount of clean speech (c_s) data on phonetic error rate. The c_s1 uses an amount of data equivalent to the music speech (m_s) system, c_s2 uses an amount of data equivalent to the noisy speech (n_s) system, and c_s3 uses an amount of data equivalent to the field speech (f_s) system.

equivalent to each of the other speaking environment cases. We then computed the phonetic error rate on the clean speech test data, and compared each result to the corresponding speaking environment result. The results of these experiments are shown in Figure 5-3.

While the clean speech system results are degraded slightly as the amount of training data is reduced to match each of the music, noisy, and field speech conditions, the error rates are still substantially better than the remaining speech conditions. This indicates that the degraded results are primarily due to the difficult acoustic conditions presented by the music, noisy and field speech classes.

In all of our recognition experiments we found that the noisy speech data performed significantly worse than the music speech data. This result was initially

surprising since both classes consist of speech superimposed on interfering sounds. However, upon closer examination of the music and noisy speech data we found that these classes were very different. The music speech data typically contained moderate to low levels of background music, which made the speech very easy to distinguish. However, much of the noisy speech data contained high levels of background noise, which at times made the speech very difficult to understand. This could explain the large discrepancy in phonetic recognition error rate that exists between the music and noisy speech classes.

5.6.2 Training Set Size vs. Error Rate

We have determined that one of the reasons that the full multi-style system outperforms the environment-specific system is the increased amount of data available to train its acoustic models. Here, we investigate the performance of the environment-specific clean speech system, to study its behavior as the amount of training data is varied. We are interested in determining if more training data would be helpful, and if so, what gains in recognition performance can be expected with the inclusion of more training data.

Figure 5-4 illustrates the performance of the environment-specific clean speech system as a function of the amount of training data used (in minutes) to develop the acoustic models. The top curve shows the performance on the clean speech test set. The bottom curve shows the performance on the clean speech training set. We see that with a small amount of training data (33 minutes), the phonetic error rate on the training set is merely 11.6%, while the error rate on the test set is 38.1%. Here, the system is able to accurately model the details of the training set, but is unable to generalize to unseen data. As the amount of training data increases, we see that the performance on the training set degrades, while the performance on the test set improves. By the time that we have utilized all of the clean speech training data, the test set curve seems to have leveled off. It appears that more training data may help to close the gap that remains between the training and testing curves as the performance on the training set continues to degrade. However, more importantly, it

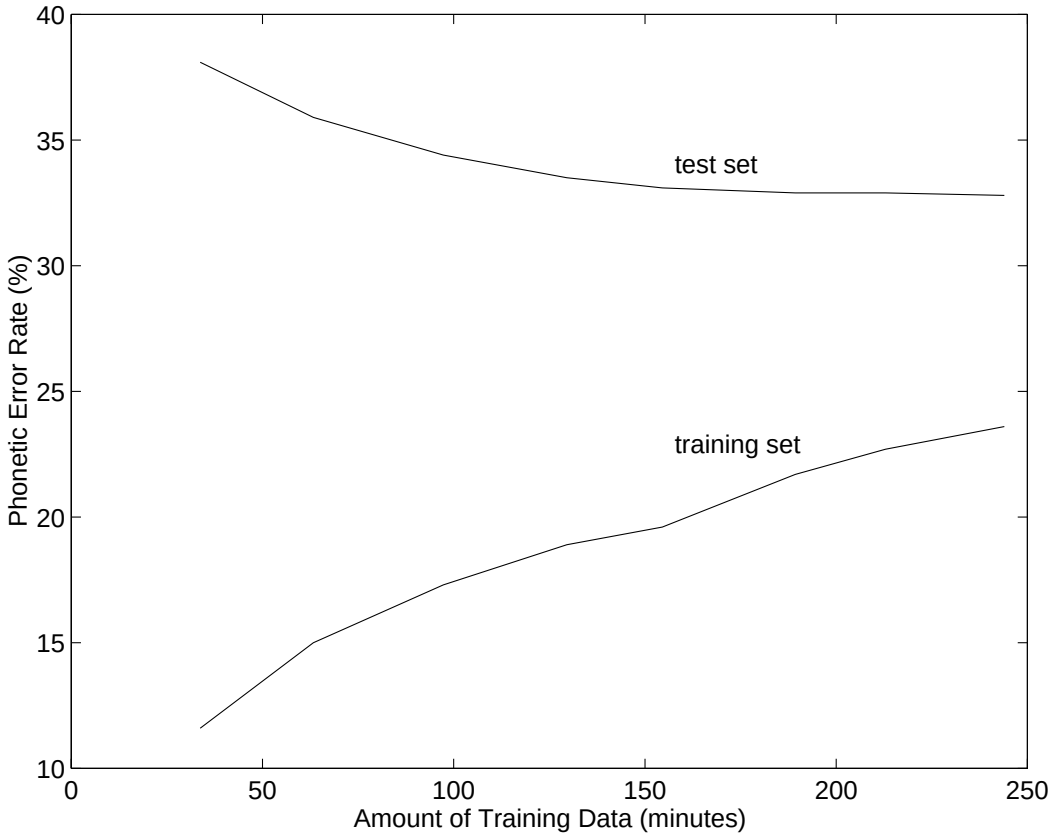

Figure 5-4: Phonetic error rate of the NPR-ME environment-specific clean speech system as a function of amount of training data (in minutes). The top curve illustrates the performance of the test set. The bottom curve illustrates the performance of the training set.

may take a *substantial* amount of additional data to get any significant gains in error rate on the test set.

5.7 Comparison with TIMIT

In all of the experiments presented in this chapter, we found that the phonetic recognition error rates varied widely across the testing conditions, from 28.2% for clean, wideband speech, to 49.3% for telephone bandwidth speech. However, we don't know how these results compare with those for other corpora or other system designs. In this section we compare the phonetic recognizer used in this work to other, state-of-the-art systems. To facilitate a comparison with other systems, we developed a

recognition system for the TIMIT corpus, which is typically used by the speech community for phonetic recognition experiments [8, 32, 35, 37, 53, 67]. We compare the results on TIMIT with our current configuration of SUMMIT (referred to as the anti-phone system) with the current state-of-the-art systems described briefly below.

- **Anti-phone Modeling System:** As described in Chapter 2, the anti-phone system [32] uses an acoustic segmentation algorithm to generate segment graphs. It then uses context-independent segment-based models and context-dependent boundary models, and a phone bigram language model within an anti-phone modeling framework. Both segment and boundary models use mixture of diagonal Gaussian distributions.

- **HMM Modeling System:** The HMM (Hidden Markov Model) system [53] uses two gender-dependent recognizers. Each recognizer uses triphone context-dependent acoustic models and a phone bigram language model within an HMM framework. The HMMs have three states and use mixture of diagonal Gaussian distributions. The higher scoring recognizer output is chosen for each utterance.

- **Near-miss Modeling System:** The near-miss modeling system [8] has two passes. The first pass uses diphone context-dependent frame-based acoustic models and a phone bigram language model within a frame-based framework to generate accurate segment graphs. The second pass reuses both models from the first pass and also adds context-independent segment-based acoustic models within a near-miss modeling framework. Both frame and segment based acoustic models use mixture of diagonal Gaussian distributions.

- **RNN Modeling System:** The RNN (Recursive Neural Network) system [67] uses context-dependent frame-based acoustic models and a phone bigram language model within an HMM framework. The HMMs have one state. The acoustic models use recursive neural networks.

Table 5.10 shows the results on the TIMIT core test set over the commonly used 39 classes. The anti-phone system used in this work achieves an error rate of 30.2%.

System	Error (%)
Anti-phone [32]	30.2
HMM [53]	30.9
Near-miss [8]	25.5
RNN [67]	26.1

Table 5.10: Phonetic recognition error rates on TIMIT's core test set over 39 classes. The anti-phone system was used in the experiments in this work.

This result does not represent the best SUMMIT system. Using the SUMMIT system with heterogeneous measurements and multiple classifiers, Halberstadt [37] achieved a phonetic recognition error rate of 24.4% on the TIMIT core test set. While our TIMIT results do not represent the overall state-of-the-art (achieved by Halberstadt [37]), they do compare favorably with the current state-of-the-art HMM recognizer.

The error rate of 30.2% achieved on the TIMIT core test set is slightly worse than the performance we obtained on the clean speech data (28.2%). However, the NPR-ME corpus contains nearly 1.5 times the amount of clean speech training data, which may account for the discrepancy. If we retrain the NPR-ME clean speech system with an equivalent amount of data, we obtain an error rate of 29.0%. Even when we account for training data size differences, the results on the NPR-ME clean speech data are slightly better than the results on the TIMIT core test set. However, we can not directly compare these results for a variety of reasons. First, the TIMIT corpus is comprised of clean, studio-quality, phonetically balanced read speech, while the NPR-ME clean speech data contains read and spontaneous speech. Second, the TIMIT task was designed to be speaker-independent (i.e., there is no overlap between speakers in the training and test sets), while the NPR-ME data is multi-speaker (i.e., individual speakers appear in both the training and test sets). Third, the TIMIT core test set contains an even balance of the eight major dialects of American English, while NPR-ME does not have such a balance of speaking styles. Finally, the phonetic alignments for the TIMIT data were generated manually, while the NPR-ME alignments were generated automatically by the SUMMIT recognizer with an iterative procedure that optimized the phonetic recognition performance on the training set. This could

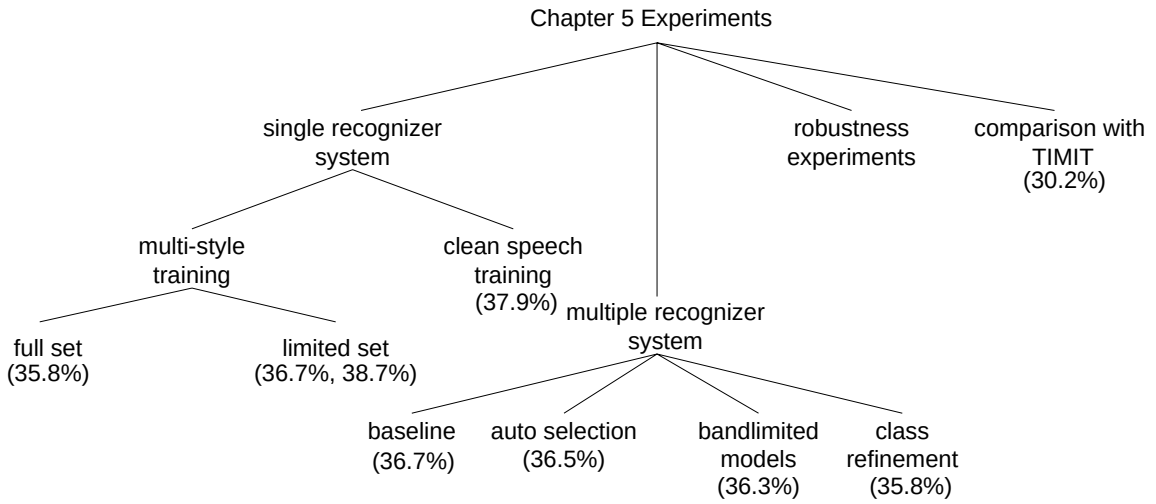

Figure 5-5: Road map of the phonetic recognition experiments presented in this chapter, with recognition results where appropriate.

effectively bias the alignments toward overall better recognition performance.

While we can't directly compare the results of the NPR-ME phonetic recognition performance to the TIMIT results, the experiments conducted in this section suggest that the recognizer used in this work was comparable to those used by others, and that the clean speech data performed similarly to the TIMIT data.

5.8 Summary

In this chapter, we described experiments that we have conducted concerning the phonetic recognition of NPR-ME. We found that for all of the training techniques that we investigated the phonetic error rates varied widely across the NPR-ME speaking environments. By systematically exploring different system designs (one recognizer vs. multiple recognizers) and different training techniques, we were able to discover how each technique affected each environment.

We investigated the use of single and multiple recognizer systems, and different training techniques associated with each. The results of our experiments are summarized in Figure 5-2. We also revisit our road map in Figure 5-5, appended with results where appropriate, to review how all of our experiments are related.

If a single recognizer system is to be used, we found that training on all of the

available data which contains a variety of speaking environments was more effective than using a smaller amount of homogeneous, clean data. This result held true even when we accounted for the discrepancy in the amount of training data available for the two approaches.

While we originally felt that the transcription of GAD would benefit from a pre-processing step that first segments the data into acoustically homogeneous blocks so that appropriate models could be used during test, overall, we found that such a multiple recognizer system achieved performance similar to a single multi-style recognizer. However, upon closer inspection of the results we found that the multi-style system primarily benefitted from the increased amount of data available for training. We may be able to utilize the strengths of both the multi-style and environment-specific approaches by developing interpolated models. Interpolation of models refers to the weighted averaging of the density functions of several models to produce a single model. Model interpolation has been used successfully to combine well-trained context-independent models with less well-trained but more detailed context-dependent models [44]. By interpolating between the well-trained multi-style models and the more detailed environment-specific models, we may be able to improve our phonetic recognition results.

Bandlimiting the training data proved to be an effective method for significantly improving the recognition results for the field speech environment. We also found that automatically selecting the environment-specific models did not degrade our results, and in fact improved them slightly. This indicated that low-levels of background noise or music may be better modeled with clean speech. However, we should again mention that the experiments presented in this chapter assumed that segmentation was performed without error. Our sound recognition system was only used to classify the test utterances. In practice, the results may be degraded if segmentation errors are present.

In this chapter we were primarily interested in the effect of speaking environment on phonetic recognition results. We could have improved our overall results by using more complex language modeling techniques. Ng [62] showed that the use of higher

order n-gram language models improved phonetic recognition results on the NPR-ME corpus by as much as 6%. In addition, more detailed segmentation of the data may improve the recognition results. In this work, we segmented the speech data based only on acoustic environment. It may be helpful to also segment the data based on speaker identity, or gender, to prevent utterances from having a mix of speakers.

Chapter 6

Summary and Future Work

6.1 Summary

General audio data (GAD) from sources such as television, radio, movies, meeting recordings, etc., are fast becoming important sources of data for information retrieval (IR) systems. To incorporate GAD into an IR system, the content of this data must be described and indexed. The main goal of this research was to understand the issues posed in describing the content of GAD. We were interested in understanding the general nature of GAD, both lexically and acoustically, and in discovering how our findings would impact an automatic indexing system. Specifically, three research issues were addressed:

1. What are the lexical characteristics of GAD, and how do they impact an automatic speech recognition system?

2. What general sound classes exist in GAD, and how well can they be distinguished automatically?

3. How can we best utilize the training data to develop a GAD transcription system?

In this thesis, we made the following contributions to research in the area of GAD analysis and transcription:

- **Development of GAD Corpus:** To complete this work, over 100 hours of data were collected and orthographically transcribed for lexical analysis. Ten hours were additionally transcribed for acoustic analysis and recognition experiments. This corpus will be valuable to others working on research issues in GAD.

- **Lexical Analysis of GAD:** We performed a lexical analysis to understand the general lexical characteristics of GAD. This analysis discovered some potential problems for a general LVCSR approach to the transcription of GAD. We found that even for large training set sizes and vocabularies, new words are still regularly encountered, and that these words are primarily high content words (i.e., proper nouns) and therefore would need to be correctly recognized to describe the linguistic content of GAD. We proposed a sub-word based approach to the recognition of GAD.

- **Acoustic Analysis and Development of Sound Recognition System:** We performed an acoustic analysis to determine what sound classes exist in GAD, and discovered the characteristics of the classes. A sound recognition system was developed which would benefit both an acoustic description system, and a recognition system.

- **Discovery of Optimal Recognition Strategies for GAD:** We investigated a number of different training and testing strategies for the phonetic recognition of GAD. We found that knowledge of the speaking environment is useful for phonetic recognition.

In the following sections, we give a brief summary of the main chapters in this thesis and finally close by mentioning some possible directions for future work.

6.1.1 Lexical Analysis

In Chapter 3 we examined the lexical aspects of GAD. Our analysis of the transcriptions of the NPR-ME corpus revealed some interesting general characteristics. It

contains many speakers and stories, with numerous turn takings. We also discovered that no single speaker dominates the speech data, and that the majority of speakers have never been seen previously. Our vocabulary analysis found that the vocabulary of a single NPR-ME show was modest (approximately 2600 unique words out of over 9700 total words). However, we also discovered that new words are still encountered even with a very large training set. With a training set of nearly one million words (resulting in over 30,000 unique vocabulary words), the out of vocabulary rate was just over 2%. Our part-of-speech analysis suggest that new words were predominately proper nouns and nouns, which would be very important to recognize if we were describing the content of this data. This problem was magnified when we investigated the more realistic scenario of constructing a training set from an out-of-domain source. In this case, the out of vocabulary rate nearly doubled to 4%.

This analysis uncovered some potentially serious problems for a word-based approach for the transcription of GAD. An alternative to a large vocabulary continuous speech recognition approach is to use a subword unit representation. This was explored in Chapter 5.

6.1.2 Sound Recognition

In Chapter 4 we examined the acoustic characteristics of GAD and developed a sound recognition system to segment the audio into its salient sound classes. For the NPR-ME corpus we subjectively identified seven acoustically distinct classes based on visual and aural examination of the data. We found that these classes differed in their spectral characteristics, statistical profile, and segment duration. Specifically, we found that high quality, prepared speech constitutes only half of the entire corpus. Another 25% of the data contains speech superimposed on other sounds, nearly 15% of the data was of telephone bandwidth, and the remaining 10% of the data was non-speech. We also found that while pure music segments are similar in length to speech segments, other non-speech segments are substantially shorter in length.

We were able to achieve a 79.4% recognition accuracy for these seven classes on unseen data, using relatively straightforward acoustic measurements and pattern

recognition and smoothing techniques. A speech / non-speech recognizer achieved an accuracy of over 94.2%. These results compare favorably with similar systems found in the literature. The results of our seven class recognition system and clustering experiments revealed that the noisy speech and music speech classes were acoustically very similar. A six class system was developed to investigate the consequences of merging these two classes. The resulting six class system achieved a recognition accuracy of nearly 87%. We used our seven class system to classify test data in the phonetic recognition experiments completed in Chapter 5. We also evaluated how the merging of the noisy and music speech classes impacted the speech recognition system.

6.1.3 Phonetic Recognition

In Chapter 5 we conducted experiments concerning the phonetic recognition of NPR-ME. We found that the phonetic error rates varied widely across the NPR-ME speaking environments for all of the training techniques that we investigated. By systematically exploring different system designs (one recognizer vs. multiple recognizers) and different training techniques, we were able to discover how each technique affected each environment.

We investigated the use of single and multiple recognizer systems and different training techniques associated with each. If a single recognizer system is to be used, we found that training on all of the available data which contains a variety of speaking environments was more effective than using a smaller amount of homogeneous, clean data. This result held true even when we accounted for the discrepancy in the amount of training data available for the two approaches.

Overall, we found that a multiple recognizer system achieved performance similar to the single multi-style recognizer. Upon closer inspection of the results, we found that the multi-style system primarily benefitted from the increased amount of data available for training. Bandlimiting the training data proved to be an effective method for significantly improving the recognition results for the field speech environment. We also found that automatically selecting the environment-specific models did not

degrade our results, but rather improved them slightly.

6.2 Future Directions

The use of GAD as a data source for IR systems is a new and exciting area of research, and there are a large number of areas for extension of this study of GAD. In this section, we mention some of these possible directions for future work.

6.2.1 Lexical Analysis

In this work, we studied the general lexical characteristics of GAD. One potentially interesting area of study is the investigation of the lexical characteristics of topic-specific subsets of the NPR-ME corpus. This analysis may indicate that knowledge of the particular nature of the speech material may help limit the active vocabulary. For example, if we could determine that a portion of a news broadcast concerns the traffic report, we may be able to reduce the recognizer vocabulary to only those words relevant to the subject matter, which may be substantially smaller than the full vocabulary.

Our lexical analysis found that as additional data is added, new words are regularly encountered. We do not know, however, if words are also "retired" with time. For example, words associated with news topics that appeared months ago may not appear in future broadcasts. An interesting area of study would be to understand the life-cycle of the GAD vocabulary.

6.2.2 Sound Segmentation

In the development of our sound recognition system, we found that the music speech and noisy speech classes were highly confusable. One area of future research that may be useful for this task is the discovery of more discriminating features. For example, features that capture the harmonic characteristics of music and music speech may be helpful in discriminating between these classes and noisy speech.

Another interesting area of research is the use of a graph-based approach to the segmentation of GAD. Two general approaches have traditionally been used for the segmentation of GAD. First, a model-based approach, such as we described in Chapter 4 and that has been used by others [5, 36], builds different models for a fixed set of acoustic classes from a training corpus. Data to be processed is then classified by a maximum likelihood selection, and the segmentation boundaries are found at locations where there is a change in acoustic class. Second, a purely data-driven approach proposes segment boundaries at maxima of the distances between neighboring windows placed at samples along the incoming data stream [11].

There are limitations to each of these approaches. First, the model-based approach does not generalize to unseen acoustic conditions. Second, as we have shown, there might not be sufficient training data for some acoustic classes (e.g., individual speakers) to build robust models. Purely data-driven approaches also have limitations. First, many of these require thresholds, which are not very robust. Second, these thresholds would have to be tuned to detect changes between very acoustically different sounds such as silence and music and between very acoustically similar sounds such as speech from different speakers of the same gender. Third, it has been shown that shorter segments are difficult to detect due to the lack of data available to develop the models for these segments [11]. In addition to the limitations described above, the single, linear segmentation that is produced by these methods does not capture all of the possible scales of segmentation that may be desired. For example, we have shown that a segmentation that indicates boundaries between different speaking environments is useful for improving phonetic recognition accuracies. A segmentation that indicates boundaries between speakers is also useful for describing the non-linguistic content of GAD. A complete representation of GAD should indicate both of these scales of segmentation. A graph-based representation that would provide a multi-level acoustic description of GAD in a single framework would be an interesting approach to this problem. A hierarchical clustering algorithm that incorporates temporal constraints has been used successfully to describe speech at a phonetic level [31]. Using such a representation to produce a multi-level acoustic

description of GAD should be explored.

6.2.3 Phonetic Recognition

In Chapter 5 we were primarily interested in the effect of speaking environment on phonetic recognition results, but much additional work can be done to improve the overall recognition results. First, more training data can be used to improve model robustness and more detailed and complex models can be used to try to capture more information from the speech signal. Second, more complex language modeling techniques can be used. Ng [62] showed that the use of higher order n-gram language models improved phonetic recognition results on the NPR-ME corpus by as much as 6%. Third, more detailed segmentation of the data may improve the recognition results. In this work, we segmented the speech data based only on acoustic environment. It may be helpful to also segment the data based on speaker identity or gender to prevent utterances from having a mix of speakers or genders. Finally, rather than using separate models for each speaking condition, which requires a large amount of training data for each condition, the use of general adaptation techniques could be explored. The goal of adaptation is to adjust the density functions (i.e., mixture Gaussians) used by general acoustic models to match the current speaking condition (or speaker, if speaker segmentation has been performed) as closely as possible using whatever adaptation data is available. A popular approach in the speech recognition literature that could be explored here is the maximum likelihood linear regression (MLLR) technique [56].

Finally, in the work presented in this thesis, a single phonetic hypothesis string was proposed for each test utterance. However, this final hypothesis is only the most probable decoding of the acoustic signal, out of a large number of hypotheses that are considered during the recognition process. Including the complete list of hypotheses in the form of an N-best list may be more useful to an information retrieval system. This would offer the hope of including terms that would otherwise be missed if only the single best hypothesis were used. Ng [62] has shown that including the top five phonetic recognition hypotheses slightly improves the performance of an information

retrieval system.

Appendix A

NPR Transcription Conventions

A.1 General Comments

- Each tape will be marked with the NPR show title, the time of the broadcast, and the air date. Each tape will contain two broadcasts - one full broadcast on each side of the tape.

- NPR transcription files will be named according to the following convention:

  show_idMMDDYY.txt

  where the show_id is ME for Morning Edition, FA for Fresh Air, and AC for All Things Considered, and MMDDYY is the date of the broadcast.

- The basic transcription file will be for an entire broadcast. Markers of internal segments like "story" will be included in the transcription file to facilitate later break-outs for testing, etc.

- All number sequences should be spelled out and all letter sequences should be rendered with capital letters with underbars between each letter. For example:

  - C_D_C (acronym for the Centers for Disease Control in Atlanta)
  - four hundred twenty three
  - nineteen ninety four

- thirty three percent

- If an acronym is not pronounced as a letter sequence but as a word (e.g., ARPA), do not place underbars between letters.

- Each turn in the conversation should be preceded by a double-spaced line in the transcription. Speech within one turn in the dialogue should be single-spaced.

A.2 Markings

1. Internal segment markers will be used to segment the transcribed speech and/or specify attributes of the segments:

   - "broadcast", delimiting a broadcast, including an i.d. and revision date, e.g.:
     <broadcast id="morning-edition.072795" rev="080195">
     ...
     </broadcast>

   - "story", delimiting stories, including an id, and topic label, e.g.:
     <story id=1 topic="headline news">
     ...
     </story>

     The id will be an integer number which indicates the order of the story in the broadcast. Note that "credits" and self identification by the anchorperson should be excluded from adjoining stories. Self identification by correspondents or commentators should be included within their stories.

   - "language", delimiting foreign language passages, e.g.:
     <language Spanish>
     ...
     </language>

- "sung", delimiting sung lyrics, e.g.:

  <sung>

  ...

  </sung>

2. Each speaker within a broadcast file will be identified by letters A, B, C, ..., AA, AB, ..., ZZ. When a transcriber is in doubt about whether a new speaker is one they've heard before, they should assume the speaker is new, use the next letter, and flag it with a comment for later verification.

   A separate file will be created to hold speaker information for the broadcast. The broadcast speaker information file will have the same basename as the transcription file, but will have a ".spk" extension. Lines in this file will give as much information about each speaker as can be gleaned from the recording:

   - name, e.g.:
     speaker_a_name: John Smith
   - sex (male, female, unknown), e.g.:
     speaker_a_sex: male
   - dialect (optional, default is native speaker of American English), e.g.:
     speaker_a_dialect: Hispanic
   - age (optional, default is adult [child, adult, elderly]), e.g.:
     speaker_a_age: adult
   - role (if known), e.g.:
     speaker_a_role: high school teacher

3. Each speaker's turn in the broadcast will be prefixed by the letter i.d. of the speaker in uppercase, a colon, and a space, and transcription of turns will be separated by a blank line, e.g.:

   A: And now, here's a report from Madrid.

   B: This is Michael Jones, reporting from...

4. Musical segments and any other non-speech segments should be separately noted as follows:

   A: The time is 19 minutes past the hour.

   [musical_interlude]

   A: Mexican rebels under heavy fire...

   Use [musical_interlude] to note musical segments, and descriptive phrases for any other non-speech segments, e.g. [shelling_gunfire].

5. Stretches of reduced audio quality of the broadcast, which typically occurs with field reporters or telephone guests, will be tagged with a [field] marker. Indicate the beginning, followed by a slash, in brackets [ ], and the end, preceded by a slash, in brackets, e.g.:

   A: We now have Ed Smith, a medical student who has been researching this question for...

   B: [field/] Even if you give to everybody... [/field]

   A: Smith's father, Steven, a doctor and ten year member...

   Field quality speech sounds a bit "muffled" - reporters in the field, sounds clips from speeches or phone-in guests typically fall into category.

6. All transcriptions should be done in real-time verbatim, including filler phrases, (e.g., [um], [er]) and non-speech events (e.g., [sneeze], [phone-ring]). All extraneous sounds should be typed within square brackets, to make it easier to distinguish these from real speech. Please try to use items from the following list for this:

   - [er]
   - [mm]
   - [uh]

- [um]
- [cough]
- [door_slam]
- [phone_ring]
- [grunt]
- [laughter]
- [throat_clear]
- [sneeze]

Please try to be as consistent as possible in using these labels. If you wish to add to the list, please make a note of the addition you made and let Michelle know what it is.

7. Any word or phrase that is difficult to understand for any reason should be surrounded by double parentheses, i.e., (()). If it's possible to hear what was said, put the word or phrase inside the double parentheses. If not, leave one blank space inside the double parentheses, i.e., (( )), to indicate that speech has not been transcribed because it was unintelligible.

8. Continuous background noise should be noted using one of the tags listed below. Note that you can always use [noise] to transcribe something that isn't described by any of the other tags.

   - [noise]
   - [music]
   - [phone_ringing]
   - [paper_rustling]

   Indicate the beginning, followed by a slash, in brackets [ ], and the end, preceded by a slash, in brackets, e.g.:

- A: [music/] This is the Morning Edition, I'm Bob Edwards. [/music] Here is a report from David Welna on the uprising in Cristobal de las Casas.
- B: [noise/] Army planes fired rockets during two twenty minute raids on areas south of the city. [/noise]

9. Simultaneous talking, where the speech of two speakers overlaps in time, should be marked by tagging the beginning and ending of the overlapping sections with a pound sign (#). The speech of both the talkers should be marked this way, e.g.:

   A: I never heard such nonsense, you know, # as I heard that #

   B: # Yeah, I know. #

   A: day when I went to the ...

10. If a word or words is clearly heard and understood, but the proper spelling cannot be determined, an "@" should be prepended to the word or words in question. This may occur frequently with proper names. ALL occurrences of questioned words should contain this notation, not just the first, e.g.: "... Israeli prime minister @Yitzhak @Rabin today and ...".

Appendix B

Vocabulary Lists

B.1 NPR-ME Out of Vocabulary Words

Table B.1 lists the NPR-ME out of vocabulary words.

B.2 Common Words Not in Brown Corpus

Table B.2 lists the words found in the common NPR-ME vocabulary that were not found in the 200 most frequent words of the Brown Corpus.

A_F_L_C_I_O	disclaimer	Marshaun	seascape
absentia	Doyle	Mcgaw	selfadministration
aggressor	Dracut	Michelle	selfhealing
allot	encompass	militaries	Serino
alluded	Enos	misinformation	Shane
apprentice	enthusiast	misrepresented	sickly
arranging	esophogloscopy	monologues	Sienook
artificially	Fillipe	mum	smutts
artworks	fireproofing	newtonbased	sociological
aspire	Fitzpatrick	nonstops	squawking
ass	Flexon	Noradom	Stewicky
assesses	fourparty	O_F_C_E	substantiate
Beardstown	Foxx	Oaklandvale	T_R_G_I
Belmonte	gaslight	octet	totaling
bilk	Gilbart	Pacific	trem
Bonne	Givadi	paintings	Tritch
bronchoscopy	grandmotherly	parlayed	underhanded
Brubeck	Greenwich	parliaments	underperformed
cabling	homage	patented	unhelpful
calculating	I_R_S	Pearl	Vaughn
capitals	il	pilfered	vehement
celebrities	inflaming	Presioso	Venetian
Clayton	innuendo	Primakov	Vermeers
combing	inspects	Prior	Verve
comedians	institution	processed	Vidrine
condolence	internationalize	pullers	Vulgova
councilman	Jehovah	purposely	W_A_V_E
crayons	July	Q_U_A_N_T_I_C	Yipgeni
Cuno	Kosovos	Ranured	Yuvanovich
decontaminated	Latvia	rectify	
Degas	laughable	Rembrandt	
demeanor	Laureen	revelation	

Table B.1: List of all NPR-ME out of vocabulary words.

a_m	going	nine	today
Bob	hour	nineteen	U_S
Boston	hundred	ninety	University
business	listeners	past	W_B_U_R
degrees	making	point	Washington
dollar	Massachusetts	President	week
dollars	morning	radio	whether
edition	N_P_R	reports	yesterday
eight	national	says	
fifty	news	six	
five	next	station	

Table B.2: List of all common NPR-ME vocabulary words that were not found in the 200 most frequent words of the Brown Corpus.

Bibliography

[1] A. Acero and R. Stern. Environmental robustness in automatic speech recognition. In *Proc. ICASSP '90*, pages 849–852, Albuquerque, NM, 1990.

[2] T. Anastasakos, J. McDonough, and R. Schwartz. A compact model for speaker-adaptive training. In *Proc. ICSLP '96*, Philadelphia, PA, 1996.

[3] T. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, New York, 1958.

[4] Association for Computational Linguistics Data Collection Initiative. CD-ROM I, September 1991.

[5] R. Bakis et al. Transcription of broadcast news shows with the IBM large vocabulary speech recognition system. In *Proc. DARPA Speech Recognition Workshop*, 1997.

[6] E. Brill. *A Corpus-Based Approach to Language Learning.* PhD thesis, University of Pennsylvania, May 1993.

[7] Carnegie Mellon University. The CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[8] J. W. Chang. *Near-Miss Modeling: A Segment-Based Approach to Speech Recognition.* PhD thesis, Massachusetts Institute of Technology, May 1998.

[9] J. W. Chang and J. R. Glass. Segmentation and modeling in segment-based recognition. In *Proc. Eurospeech '97*, pages 1199–1202, Rhodes, Greece, September 1997.

[10] S. S. Chen et al. IBM's LVCSR system for transcription of broadcast news used in the 1997 Hub4 English evaluation. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, Landsdowne, VA, 1998.

[11] S. S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, Landsdowne, VA, 1998.

[12] K. W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proc. 2nd Conf. on Applied Nat. Lang. Processing*, pages 136–143, Austin, TX, February 1988.

[13] G. D. Cook and A. J. Robinson. The 1997 ABBOTT system for the transcription of broadcast news. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, February 1998.

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley and Sons, New York, 1991.

[15] G. Davis and R. Jones. Signal processing equipment. In *The Sound Reinforcement Handbook.* Hal Leonard Corp., Milwaukee, WI, 1989.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39:1–38, 1977.

[17] S. Dharanipragada, M. Franz, and S. Roukos. Audio indexing for broadcast news. In *Seventh Text REtrieval Conference (TREC-7)*, Gaithersburg, MD, 1998. NIST-SP 500-242.

[18] B. P. Douglas and J. M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc. DARPA Speech and Nat. Lang. Workshop*, pages 357–362, Harriman, NY, February 1992.

[19] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, NY, 1973.

[20] W. Fisher, G. Doddington, and Goudie-Marshall. The DARPA speech recognition database: specification and status. In *Proc. DARPA Speech Recognition Workshop*, pages 93–96, 1986.

[21] J. Foote. An overview of audio information retrieval. *Multimedia Systems*, 7:2–10, 1999.

[22] J. T. Foote. Content-based retrieval of music and audio. In Kuo C. et al., editors, *Multimedia Storage and Archiving Systems II.* SPIE, 1997.

[23] J. T. Foote. Rapid speaker identification using discrete mmi feature quantisation. *Expert System Applications*, 13(4):283–289, 1998.

[24] J. T. Foote and H. F. Silverman. A model distance measure for talker clustering and identification. In *Proc. ICASSP '94*, pages 17–32, Adelaide, Australia, April 1994.

[25] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dalgren. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. Technical report, National Institute of Standards and Technology, 1993. Report No. NISTIR 4930.

[26] J. Garofolo, E. Voorhees, C. Auzanne, V. Stanford, and B. Lund. 1998 trec-7 spoken document retrieval track overview and results. In *Seventh Text REtrieval Conference (TREC-7)*, Gaithersburg, MD, 1997. National Institute for Standards and Technology. NIST-SP 500-242.

[27] J. Garofolo, E. Voorhees, and V. Stanford. 1997 trec-6 spoken document retrieval track overview and results. In *Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, MD, 1997. National Institute for Standards and Technology. NIST-SP 500-240.

[28] J. L. Gauvain, L. Lamel, and G. Adda. The LIMSI 1997 Hub-4E transcription system. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, February 1998.

[29] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. ICASSP '89*, pages 532–535, Glasgow, Scotland, 1989.

[30] H. Gish and N. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, pages 18–21, October 1994.

[31] J. Glass. *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition.* PhD thesis, Massachusetts Institute of Technology, May 1988.

[32] J. Glass, J. Chang, and M. McCandless. A probabilistic framework for feature-based speech recognition. In *Proc. ICSLP '96*, volume 4, pages 2277–2280, Philadelphia, PA, October 1996.

[33] J. R. Glass, T. J. Hazen, and L. Hetherington. Real-time telephone-based speech recognition in the Jupiter domain. In *Proc. ICASSP '99*, Phoenix, AZ, March 1998.

[34] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP '92*, volume 1, pages 517–520, San Francisco, March 1992.

[35] W. Goldenthal. Statistical trajectory models for phonetic recognition. Technical report MIT/LCS/TR-642, MIT Lab. for Computer Science, August 1994.

[36] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland, and S. J. Young. Segment generation and clustering in the HTK broadcast news transcription system. In *Proc. DARPA Speech Recognition Workshop*, 1998.

[37] A. Halberstadt. *Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition.* PhD thesis, Massachusetts Institute of Technology, November 1999.

[38] D. K. Harman, editor. *Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, MD, 1997. National Institute for Standards and Technology. NIST-SP 500-240.

[39] A. G. Hauptmann and H. D. Wactlar. Indexing and search of multimodel information. In *Proc. ICASSP '97*, volume 1, pages 195–198, 1997.

[40] A. G. Hauptmann and H. D. Wactlar. Indexing and search of multimodel information. In *Proc. ICASSP '97*, volume 1, pages 195–198, 1997.

[41] T. J. Hazen and A. K. Halberstadt. Using aggregation to improve the performance of mixture Gaussian acoustic models. In *Proc. ICASSP '98*, pages 653–657, Seattle, WA, 1998.

[42] I. L. Hetherington. New words: Effect on recognition performance and incorporation issues. In *Proc. Eurospeech '95*, Madrid, Spain, September 1995.

[43] L. Hetherington and M. McCandless. SAPPHIRE. In *Annual Research Summary: Spoken Language Systems*. Laboratory for Computer Science, Massachusetts Institute of Technology, 1996.

[44] X. Huang, M. Y. Hwang, L. Jiang, and M. Mahajan. Deleted interpolation and density sharing for continuous Hidden Markov Models. In *Proc. ICASSP '96*, pages 885–888, Atlanta, GA, May 1996.

[45] M. J. Hunt. A robust method of detecting the presence of voiced speech. In *Proc. 15th Intl. Congress on Acoustics*, Trondheim, Norway, June 1995.

[46] U. Jain et al. Recognition of continuous broadcast news with multiple unknown speakers and environments. In *Proc. DARPA Speech Recognition Workshop*, Harriman, NY, February 1996.

[47] D. A. James. *The Application of Classical Information Retrieval Techniques to Spoken Documents*. PhD thesis, University of Cambridge, February 1995.

[48] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland. Spoken document retrieval for TREC-7 at Cambridge University. In *Seventh Text REtrieval Conference (TREC-7)*, Gaithersburg, MD, 1997. NIST-SP 500-242.

[49] D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. In *Proc. Interface Conference*, Sydney, Australia, 1996.

[50] F. Kubala et al. The 1996 BBN Byblos Hub-4 transcription system. In *Proc. DARPA Speech Recognition Workshop*, Harriman, NY, February 1996.

[51] F. Kubala et al. The 1997 BBN Byblos system applied to broadcast news transcription. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 35–40, Landsdowne, VA, February 1998.

[52] H. Kucera and W. N. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, 1967.

[53] L. Lamel and J. L. Gauvain. High performance speaker-independent phone recognition using CDHMM. In *Proc. Eurospeech '93*, pages 121–124, Berlin, Germany, September 1993.

[54] L. Lamel, R. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proc. DARPA Speech Recognition Workshop*, Palo Alto, CA, February 1986. Report No. SAIC-86/1546.

[55] K. F. Lee and H. W. Hon. Speaker-independent phone recognition using Hidden Markov Models. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(11):1641–1648, November 1989.

[56] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. *Computer Speech & Language*, 9:171–185, 1995.

[57] Linguistic Data Consortium. 1997 English Broadcast News Transcripts (Hub-4). LDC Catalog Number: LDC98T28.

[58] R. Lippmann, E. Martin, and D. Paul. Multi-style training for robust isolated-word speech recognition. In *Proc. ICASSP '87*, pages 705–708, Dallas, TX, 1987.

[59] S. Matsoukas, R. Schwartz, H. Jin, and L. Nguyen. Practical implementations of speaker-adaptive training. In *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, 1997.

[60] Muscle Fish. http://www.musclefish.com.

[61] National Institute for Standards and Technology. *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, February 1998.

[62] K. Ng. *Subword-based Approaches for Spoken Document Retrieval.* PhD thesis, Massachusetts Institute of Technology, February 2000.

[63] D. S. Pallett and J. G. Fiscus. 1996 preliminary broadcast news benchmark tests. In *Proc. DARPA Speech Recognition Workshop*, 1997.

[64] L. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[65] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[66] C. V. Rijsbergen. *Information Retrieval.* Butterworths, London, 1979.

[67] T. Robinson, M. Hochberg, and S. Renals. IPA: Improved phone modeling with recurrent neural networks. In *Proc. ICASSP '94*, pages 37–40, Adelaide, Australia, April 1994.

[68] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

[69] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proc. ICASSP '96*, pages 993–996, Atlanta, GA, May 1996.

[70] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature music/speech discriminator. In *Proc. ICASSP '97*, pages 1331–1334, April 1996.

[71] C. Schmandt. *Voice Communication with Computers (Conversational Systems).* Van Nostrand Reinhold, New York, 1994.

[72] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[73] K. Seymore et al. The 1997 CMU Sphinx-3 English broadcast news transcription system. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 55–59, Landsdowne, VA, February 1998.

[74] M. Siegler, U. Jain, B. Ray, and R. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA Speech Recognition Workshop*, pages 97–99, Chantilly, VA, February 1997.

[75] M. S. Spina and V. W. Zue. Automatic transcription of general audio data: Preliminary analysis. In *Proc. ICSLP '96*, volume 2, pages 594–597, Philadelphia, PA, October 1996.

[76] M. S. Spina and V. W. Zue. Automatic transcription of general audio data: Effect of environment segmentation on phonetic recognition. In *Proc. Eurospeech '97*, pages 1547–1550, Rhodes, Greece, September 1997.

[77] Townshend Computer Tools. Dat-link information. http://www.tc.com.

[78] H. L. Van Trees. *Detection, Estimation and Modulation Theory, Part I.* Wiley, New York, 1968.

[79] A. Viterbi. Error bounds for convolutional codes and an asymptotic optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, April 1967.

[80] S. Wegmann et al. Marketplace recognition using Dragon's continuous speech recognition system. In *Proc. DARPA Speech Recognition Workshop*, Harriman, NY, February 1996.

[81] S. Wegmann, F. Scattone, I. Carp, L. Gillick, R. Roth, and J. Yamron. The 1997 Dragon Systems' broadcast news transcription system. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 60–65, Landsdowne, VA, February 1998.

[82] L. Wilcox, F. Chen, and V. Balasubramanian. Segmentation of speech using speaker identification. In *Proc. ICASSP '94*, volume S1, pages 161–163, April 1994.

[83] M. J. Witbrock and A. G. Hauptmann. Speech recognition and information retrieval: Experiments in retrieving spoken documents. In *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, 1997.

[84] E. Wold, T. Blum, D. Keslar, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[85] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, E. W. D Whittaker, and S. J. Young. The 1997 HTK broadcast news transcription system. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 41–48, Landsdowne, VA, February 1998.

[86] L. Wyse and S. Smoliar. Toward content-based audio indexing and retrieval and a new speaker discrimination technique. In D. F. Rosenthal and H. G. Okuno, editors, *Readings in computational auditory scene analysis*. Lawrence Erlbaum, New York, NY, 1998.

[87] V. Zue et al. JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1), January 2000.