Speaker Spotting:
Automatic Annotation of Audio Data with Speaker Identity

by

Patrick Kwon

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science
at the
Massachusetts Institute of Technology

May 21, 1998
[June 1998]

Author _____
Department of Electrical Engineering and Computer Science
May 21, 1998

Certified by _____
Christopher Schmandt
Principal Research Scientist
Thesis Supervisor

Accepted by _____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Speaker Spotting: Automatic Annotation of Audio Data with Speaker Identity

by

Patrick Kwon

Submitted to the
Department of Electrical Engineering and Computer Science

May 21, 1998

in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

Speaker spotting is an application of automatic speaker recognition to multimedia indexing. It aims to find if and when a predetermined individual or member of a predetermined set of individuals is speaking. This project examines the development of a speaker spotter for broadcast television news. In particular, it examines the issue of open-set speaker identification, or differentiating between the individuals being spotted (target) and everyone else (non-target), by testing two different models for non-target speakers, and through implementation of a cohort threshold. Also, issues in segmentation are investigated through the use of speaker transition probabilities and overlapping decision segments. A spotting accuracy of 70% to 80% for both target and non-target speakers is achieved on a test hour of broadcast news.

Thesis Supervisor: Christopher Schmandt
Title: Principal Research Scientist

Acknowledgements

Thanks to:

Chris Schmandt, my MIT thesis advisor, for taking me on and for his patience with me as I finished this thesis.

Brian Eberman, my supervisor at Digital Equipment Corporation's Cambridge Research Lab (DEC CRL), for his help and support in this research.

Mike Sokolov, Bill Goldenthal, and the rest of the folks at CRL for all their help during the past three years as a VI-A intern there.

Mark Roh, Bernie Chang, and Jennifer Choi, as well as my staff team at MIT Korean Christian Fellowship: James Choung, Gary Crichlow, Chinsan Han, and Jean Kang for their friendship and concern for me.

Most of all, I thank God for seeing me through this. Truly, I could not have done this without God having been with me all the way. In the end, that's all that matters...

Soli Deo Gloria

Table of Contents

Chapter 1
Introduction

1.1. Project Overview

This project describes an application of speaker recognition technology to the problem of multimedia data retrieval. More specifically, speaker identification is used to locate if and when certain speakers are speaking in an audio stream, a process called "speaker spotting."

1.2. Project Motivation

In the past, retrieving multimedia data required locating an audio or video recording in an archive, then listening or viewing it to locate the desired portion. Today, technology has made it possible to store multimedia in digital form, allowing for quicker access to the desired data via on-line databases that can serve the data to the user. However, it is not immediately clear how to index audio and video for retrieval, thus making it difficult to make full use of the data.

One way to perform indexing is with text annotations attached to the audio or video media. This is especially useful with recorded speech, as transcripts labeled with what was said, who said it, and when it was said capture much of the information needed to provide a useful index. Currently, these transcriptions must be produced by human transcribers, and the transcription task is time-consuming and laborious. For one project, it took over fifty hours for one person to produce a time-aligned, annotated transcript of one hour of broadcast news (Garofolo 1996).

Such effort is clearly not viable for large amounts of audio data, so automatic annotation tools are called for. This project will examine one particular annotation tool: speaker spotting.

The most obvious tool for creating annotations is automatic speech recognition. However, the current state of the art, used on "found speech," or speech recorded from broadcasts not specifically designed for speech recognition, has a best word error rate of 17.4%, according to results from a 1997 DARPA-sponsored benchmark test (Liggett 1997). This error rate of about one in every six words is clearly not accurate enough to provide good dialogue transcriptions. So, for the time being, transcription of what was said requires human intervention. Other tools may be able to assist in providing annotations as well as in reducing the human effort needed in transcribing. In particular, speaker recognition technology can provide a useful tool for multimedia indexing.

Since the needed human effort is so much, and with the overwhelming amount of audio data that is being produced, there is a question of whether or not something is worth transcribing. For example, for long-term information retrieval purposes, in a news broadcast, it is not very useful to know what the anchor said; it is much more useful to hear what, for example, President Clinton had to say in the audio clip associated with a story. But, a human must listen to the entire news broadcast to figure out if President Clinton spoke before even being able to transcribe his words. An automatic tool to identify speakers would give a user a convenient means to select the parts that might be worth transcribing.

Identity information can also be used in order to augment automatic transcription of speech by machines. Speech recognition systems can be divided into two categories: speaker dependent and speaker independent systems. Speaker dependent systems are trained to recognize the speech of a particular person, capturing that person's speech peculiarities such as pronunciation habits. Speaker independent systems are trained to recognize speech from a generic speaker. Since the speaker dependent system captures the behavior of a single person, it will perform

better for that speaker than a speaker independent system will. However, if the speaker is not identified, a speaker independent system must be used. With an initial identification of the speaker, the speech can be run through a speaker dependent speech recognizer, resulting in better word recognition performance.

1.3. Speaker Spotting

Speaker spotting is a means of providing such identity information. It is an application of automatic speaker recognition aimed at finding if and when a predetermined individual or member of a predetermined set of individuals is speaking. This is in contrast to previous uses of speaker recognition, which have been focused primarily on security and single-user interactive systems--answering the question, "who is speaking?" Speaker spotting attempts to mark the times where one desired speaker is speaking as well as to identify which desired speaker it is—not only asking "who is speaking?" but also "when is he speaking?" For example, a speaker spotter could attempt to find the times where a representative from Massachusetts is speaking given the audio from a session of the House of Representatives. Here, it must be determined if one of these representatives is speaking in the particular session, and if so, at what times one started and stopped speaking. In short, the system attempts to "spot" the speaker in the audio sequence.

This project examines development of a speaker spotter for broadcast television news, implementing a method of speech segmentation via open-set speaker recognition. Broadcast television news is chosen because it demonstrates a practical use for such a system in spotting noteworthy individuals. It also provided a limited domain for development while providing sufficient variety to be interesting. The performance of the system is judged by the accuracy, both in determining who is speaking as well as in finding the boundaries of each speaker's speech.

1.4. Document Overview

The first section gives a general introduction to the project and to the paper. The second section describes some background information to the project, including speaker recognition and related

work. The third section describes the parts of the project, including the speaker identifier as well as the data and performance evaluation. The fourth section presents the actual research done, including procedure, results, and analysis of the results. The fifth section provides final conclusions.

Chapter 2

Background Information

2.1. Speaker Recognition

There are two major tasks in the area of speaker recognition: identification and verification. In speaker identification, the test speaker (applicant) simply submits a speech sample to the system. The system compares a model of this utterance to models of each speaker on which the system has been trained (enrolled), and returns the suspected identity of the speaker. In speaker verification, the applicant identifies herself and submits a speech sample to the system. The system either verifies or denies that the voice matches the given identity.

Speaker recognition technology can be further divided into closed-set and open-set recognition. This refers to whether or not the system must have previously been trained on the applicant. In every system, all speakers that the system will recognize must be enrolled in the system. Closed-set systems impose the additional constraint that every potential applicant must be enrolled in the system. In contrast, open-set systems do not require that every potential applicant be enrolled. The open-set problem thus tends to be a more difficult problem, as the system must do more than match the applicant to the correct enrolled model, but must also reject any applicant that does not match any of the enrolled models.

Most emphasis in speaker recognition has been put on the closed-set problem due to the more complex nature of the problem of identification—identification has more degrees of freedom (namely, in the lack of a proposed speaker identity) than verification does. Speaker verification, on the other hand, focuses more on the open-set problem, since the open-set problem has potentially wider application than the closed set problem, and the simpler nature of verification allows for more complexity in its field of application.

Another axis of classification of speaker recognition systems is text-dependency. Text-dependent recognition systems require that the speakers always speak a predetermined phrase. Thus, the system has to deal with less variation in the utterance and so can potentially do a better job in recognizing the speaker. A less constrained system is a vocabulary-dependent system, which requires that the speaker use a predetermined vocabulary (such as digits) in test phrases. However, these systems are limited in the domains in which they can be used. In contrast, text-independent systems put no limits on the speech used. Such systems have the widest application, though speech recognition technology has not yet reached the point where this can be fully realized. They also may not perform as well as the other systems, as less useful information can be extracted from the unconstrained speech.

Speaker spotting is an open-set text-independent speaker identification problem. Speakers presented to the system may not have been previously enrolled, the speech presented is completely unconstrained, and the speakers do not give a proposed identity to the system. The added difficulties of open-set speaker identification are partially countered by limiting the number of enrolled speakers to a few (under 10).

2.2. Related Work

The following is a summary of previous work related to this project.

2.2.1. Segmentation of Speech with Multiple Speakers

Siu et al. (Siu 1992) have developed a method to separate and group the speech of several speakers, without any previous knowledge of the speakers (i.e., "unsupervised"). The method makes a number of assumptions about the domain of application, namely air traffic control conversations. First, the number of speakers and the entire audio sequence, from start to finish, are available beforehand. Also, intervals of noise separate the utterances of the speakers. Finally, a conversation between two speakers lasts at least 50 utterances, so in any contiguous set of 50 utterances, there are at least two and at most 4 speakers.

The process begins by identifying contiguous portions of speech and of noise; each contiguous portion of speech is considered to be an utterance of a single speaker. Consecutive utterances are grouped into sets of 50, and models of the speakers in each set are produced such that there are two models per set of utterances. These are then clustered into as many sets as there are known speakers.

The assumptions made by Siu render their method inappropriate for speaker spotting in this domain. In speaker spotting, the number of speakers is not necessarily available beforehand. Also, the speaker utterances are not necessarily separated by silences, and the audio does not necessarily consist of conversations of any particular length between two individuals.

2.2.2. Audio Indexing Using Speaker Identification

Wilcox et al. (Wilcox 1994) also have developed a system for identifying the speech of several speakers, with or without previous knowledge of the speakers. This method uses a hidden Markov network with sub-networks for each speaker. If the speakers are known beforehand, these sub-networks are trained with this data. If the speakers are not known beforehand (though the number of target speakers is known), the sub-networks can be trained via Baum-Welch training, an iterative process by which the network parameters are locally optimized, with initial parameters found either randomly or by agglomerative clustering to segment the data (Rabiner

1993). An ideal state sequence is found through Viterbi decoding of the postulated state sequences. In order to improve performance (required when training is unsupervised, but also useful when training is supervised), the speaker sub-networks are retrained using the results of the decoding, and the process is repeated. These iterations are done until the state sequence converges.

This speaker identification method does not seem to be ideal for speaker spotting. It uses as a basic unit of identification decision a 20 ms segment of audio. But, since any utterance with comprehensible speech is of length significantly greater than 20 ms (a single word being on the order of tenths of a second), the short segments provide more resolution than is really necessary. So, it is potentially possible to improve identification performance on each individual segment by using a longer segment. Nonetheless, the modeling by use of a hidden Markov model does appear beneficial. A similar technique is attempted in speaker spotting.

2.2.3. Speaker Indexing

The Speaker Indexing Algorithm developed by Roy (Roy 1995) likewise attempts to separate and group the utterances of many speakers in an unsupervised system. Like Siu, the system has multiple requirements on the nature of the audio. Each utterance must be at least 5 seconds long, with pauses between utterances of different speakers of at least 0.2 seconds long. Also, the entire audio sequence must be available beforehand.

First, the audio is segmented into utterances by locating pauses of longer than 0.2 seconds. Then, the vowels in each utterance are found by locating energy peaks in a spectral representation of the audio waveform. Mel-frequency spectral coefficients of these vowels are passed through a back-propagation neural network for identifying speakers, using positive and negative examples for training.

The actual identification begins by assuming that the first 5 seconds of audio are from the first speaker. The neural network is trained with positive examples from these 5 seconds and negative examples randomly chosen from the rest of the audio. After training, each detected vowel is classified by the neural network as either belonging to the first speaker or not. Then, each utterance is identified as either belonging to the first speaker or not, by a majority vote of the labels of the vowels in the utterance. The process is then repeated, except the audio used for positive examples is now composed of all utterances labeled as belonging to the speaker, and the audio for negative examples comes from the rest. The cycle for the first speaker is repeated until no additional utterances are attributed to this speaker.

After locating the first speaker, each additional speaker is located by the same process. The audio that has already been labeled is removed from the testing set, and the process is repeated until all audio has been labeled.

This method does not appear to be appropriate for speaker spotting. The requirements on the audio constrain the process more than desired, as utterances may not necessarily be 5 seconds long and may not be separated by pauses. Also, only using vowels (or a portion of the vowels) for identifying speakers potentially ignores a great deal of information about the speaker that is captured by the consonants. As a result, identification performance may be weaker than can be achieved.

Chapter 3

Project Components

3.1. System Overview

The speaker spotting system consists of three parts: the audio handler, the speaker identifier, and the score handler. The audio handler takes the incoming audio, divides it into segments of a selected length, and sends these segments to the speaker identifier. The speaker identifier then performs identification (described below) on each segment. For every segment, it creates a set of scores, one score for each model against which the segment has been compared. The score handler then evaluates the scores according to various criteria, and returns a postulated speaker identity. The methodology of dividing the audio into segments (as done by the audio handler) and

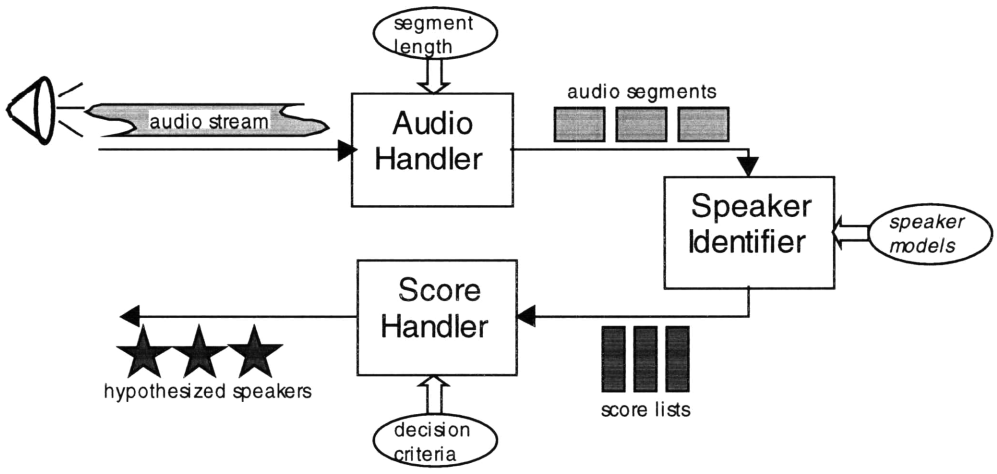


Figure 3.1 : The Speaker Spotting system.

of evaluating the scores (as done by the score handler) are the principal areas of experimentation in this project.

3.2. Speaker Identification

The speaker identifier used in this system begins with an input audio waveform (utterance), which is assumed to contain the speech of a single individual to be identified. The utterance is first scaled so that the variance in energy of the utterance is 1.0. Then, a feature vector is computed for each "frame," every 5 milliseconds using a Hamming window of width 25 milliseconds. The feature vector's 20 components are found by taking the initial 20 coefficients of the inverse Fourier transform of the log magnitude of the mel-scale warped Fourier transform of the waveform. This "mel-cepstral" representation of the audio is used because spectral/cepstral analysis captures characteristics of the vocal tract, and variations in the vocal tract lead to the auditory features that allow for distinguishing between different individuals' voices (Rabiner 1993).

The mel-scale warping modifies the frequency scale such that frequencies under 1 kilohertz are scaled linearly and the frequencies above 1 kilohertz are scaled logarithmically. This warping is done so that the system mimics the behavior of the human auditory system, which puts greater emphasis on the higher frequencies (Rabiner 1993). The coefficients produced by this entire process are referred to as mel-frequency cepstral coefficients (MFCCs).

Research has demonstrated that low-energy frames do not carry much information about inter-speaker variation (Mason 1992). So, after the MFCCs have been computed for each frame, all frames with energy (as represented by MFCC 0) lower than the energy in the frame with the lowest energy plus a predefined threshold are rejected. The threshold is appropriately chosen to balance between the benefit from rejecting lower energy frames and the loss from having less information available to make a judgment.

After the low-energy frames have been removed, the means and covariances of the 20 MFCC components are computed. However, the averages are not used for identification, as they model not only characteristics of the speaker's vocal tract, but also other factors such as background noise, recording equipment peculiarities, and speech effort (i.e., variations in loudness). These other factors tend to obscure the speaker-specific information in the averages (Reynolds 1995). Thus, ignoring the means improves identification performance. This leaves the 20 by 20 covariance matrix, which is the model for the utterance.

If the utterance is used for training, the resulting model becomes the canonical model for the speaker. If the model is used for testing, it is compared against the speaker models using a Wishart model, which models the covariances of a multivariate normal distribution (Anderson 1984). Assuming a Wishart distribution, a score proportional to the log-likelihood of the covariance can be found from the formula

$$\ell(\Sigma_j, S) = -\tfrac{n-1}{2}\log|\Sigma_j| - \tfrac{n}{2} tr\Sigma_j^{-1} S$$

where Σ_j is the model for the jth speaker, S is the model for the test utterance, and n is the number of frames in the utterance (Gish 1990). The score is proportional to the log likelihood because it ignores a term that is constant for a single utterance. The model is identified as the identity of the speaker with the highest scoring model.

3.3. Data

3.3.1. General Description

The data used for this project is from the *News Hour with Jim Lehrer*. This show is chosen because of the variety in speaker interactions in its various sections. These sections include standard newscasts, stories, and discussions.

The typical newscast consists of an anchor speaking most of the time, with occasional recorded clips of other people. The transitions between the anchor and the clips are characterized by a period of silence, and the audio tends to be studio-quality.

The stories consist of newsmagazine-style prerecorded segments of a few minutes in length, narrated by a reporter, with frequent insertions of other speakers. The audio varies from studio-quality voice-overs to noisy on-site interviews, but the speakers tend to be isolated from one another with periods of non-speech (either silence or other sounds).

The discussions consist of one-on-one or roundtable interactions between a reporter and others. These segments have studio-quality audio with speech segments ranging from isolated (with pauses in between) to simultaneous speakers.

The variety of formats in this show provides a challenge in spotting speakers, as many structural cues that may identify speaker change in one format will not work in other formats. For example, in some domains, it may be known that a pause of at least one second occurs between speakers. However, this is not true of this particular domain. As a result, such structural heuristics cannot be used. However, the ability to spot speakers in this environment demonstrates that the solution is robust enough to be used in other domains as well.

Five hours of data (i.e., five one-hour news broadcasts) were collected for this project. Each hour was considered to be a single data set. For reference purposes, each data set is identified by the date of broadcast. For example, "NH611" refers to the *News Hour* broadcast on June 11, 1997.

Data Set Label	Broadcast Date
NH611	June 11, 1997
NH612	June 12, 1997
NH616	June 16, 1997
NH617	June 17, 1997
NH620	June 20, 1997

Table 3.1: Data Sets

3.3.2. Data Processing

Capturing of this data is done by recording the show off of cable television onto professional quality Betamax tape using a Sony UVW-1800 Video Cassette Recorder. From the tape, the audio is digitized using a Digital Celebris FP 590 PC with an Ensoniq Jazz sound card. The audio is stored as 16 bit PCM samples at 16 kilohertz, with some meta-information contained in a NIST sphere-format header. This is the actual audio used for speaker spotting.

3.3.3. Labeling Speakers

For development and evaluation purposes, the speakers in each show are identified manually and their identities are time-aligned to the audio. This is done with a combination of examining a spectral view of the audio waveform, viewing the video associated with the audio, and listening to the audio itself. In addition to labeling the speakers, all non-speech (including silences and noise) of greater than 0.5 seconds is labeled, as well as non-speech of any length that occurs between speakers.

The labeling was done using a program, WinSpec, created by the Speech group at Digital Equipment Corporation's Cambridge Research Lab specifically for the task of creating such labels. WinSpec takes as input an audio file, and displays the file, or a portion of the file, in a

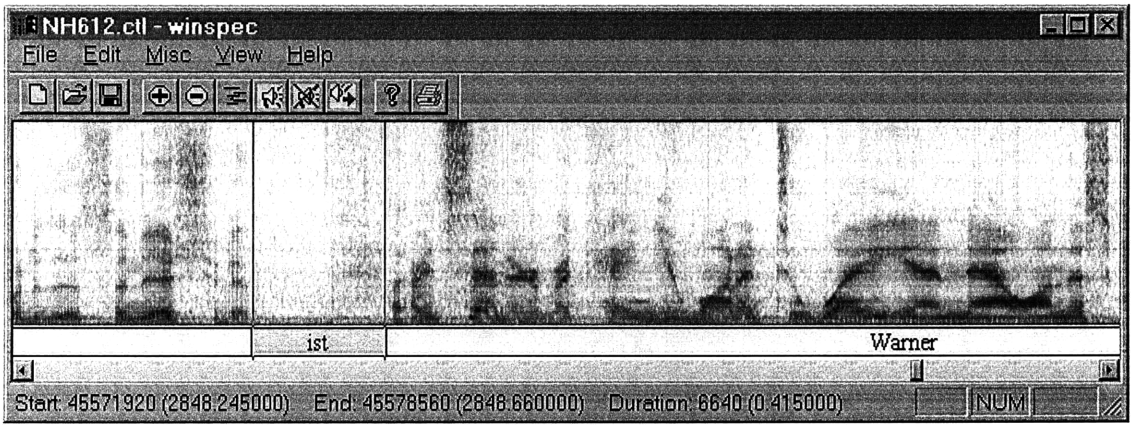


Figure 3.2: Sample WinSpec view.

spectral view. Below this spectrogram is a bar that allows for time-aligned transcription. Within this bar, boundaries between speakers can be marked, and speaker identities can be added. In addition, WinSpec allows for playback of a marked section of audio.

3.3.4. Categorization

3.3.4.1. Speaker Sets

The speakers in the five hours of data are categorized into three sets:

- The "target" set of five individuals consists of speakers who appear at least once in five of the six hours, and who spoke for at least 35 seconds in the two hours chosen to provide training data. These are the speakers that the system attempted to "spot."
- The "trained" set of 19 individuals consists of speakers who spoke for at least 35 seconds in the two hours chosen to provide training data. These speakers helped provide the means to determine whether or not an utterance belonged to a member of the target set.
- The "unseen" set consists of everyone else.

Speaker Set	Speakers	
target	Clinton, Bill Gingrich, Newt Lehrer. Jim	Solman, Paul Warner, Margaret
trained	Abler, Ronald Bearden, Tom Davies, Tim Emanuel, Rahm Haeberli, Martin Hobson, Jenny Holman, Kwame McCain, John McConnell, Mitch	McVeigh, William Meehan, Marty Murray, David Natsios, Andrew Ornstein, Norman Page, Clarence Sobel, David Sullivan, Tim Varney, Christine
unseen	all other speakers	

Table 3.2: Speakers in each speaker set.

For completeness' sake, "non-target" refers to speakers not contained in the target set, or equivalently, contained in the trained or unseen sets.

Table 3.3 shows the amount of time speakers from each set appear in each hour of data.

		Data Set				
		NH611	NH612	NH616	NH617	NH620
Speaker Set	target	10:46	12:26	13:27	10:57	4:06
	trained	36:08	31:42	0:00	4:20	2:02
	unseen	5:12	7:17	35:53	36:51	44:52

Table 3.3 : Time spent in each speaker set (min:sec).

3.3.4.2. Training and Testing Audio Sets

Using these speaker set definitions, training and testing audio sets are defined.

The training data set is created with audio from two of the hours of data (NH611 and NH612). Thirty seconds (+/- 0.5 seconds) of the audio from each speaker in the target and trained speaker sets is designated as training data. The audio is selected by taking the audio from these two hours and cutting it into one utterance per manually labeled speaker segment. Then, sufficient segments are chosen for each speaker such that the total time over all selected segments is within 0.5 seconds of thirty seconds.

The rest of the audio of NH611 and NH612, as well as all of the audio of NH616, NH617, and NH620 are designated as the testing data set.

3.3.4.3. Model Sets

On the basis of the speaker sets, three sets of speaker models are defined:

- The "target" set of speaker models consists of 5 models, each corresponding to a speaker in the target set of speakers.

- The "homogeneous garbage" (or "homogeneous") set of speaker models consists of 19 models, one per speaker in the "trained" set of speakers.
- The "gender garbage" (or "gender") set of speaker models consists of two models, one per gender.

"Garbage" models may refer to either the homogeneous or gender sets of models, or both.

The actual speaker models are created using the audio of the training audio set. In creating the models in the target and homogeneous model sets, for each speaker, each segment used for training is individually passed to the speaker identifier, and a model computed for it. Then, the aggregate results (means and covariances, appropriately weighted according to the length of each utterance) over all of the segments are used in order to create the actual speaker model. This is done instead of concatenating all of the audio together in order to prevent spectral discontinuities at the boundaries, which may adversely affect the model.

For each of the gender speaker models, a total of 30 seconds of audio from the training audio set is used. The first model is trained with a 1.76-second segment of audio from each of the 17 male speakers in the trained speaker set. The second model is trained with a 15-second segment of audio from each of the two female speakers in the trained speaker set. In both cases, the segments are a subset of the audio used to train the gender speaker models. As with the homogeneous models, the segments are individually passed to the speaker identifier, and a speaker model computed over all the segments.

3.4. Evaluation Methodology

The performance of the system is evaluated by calculating the percentage of frames it correctly identifies. A frame is considered to be correctly identified if the identity produced by the speaker spotter matches the identity found during manual transcription. In addition, if the manually labeled identity is silence or noise, the frame is also considered to be correctly identified if the speaker spotter label matches either of the manually produced labels of the adjacent segments that are

not noise or silence and the label is within the length of a segment used for performing speaker identification. Although it is not a significant contributor to performance, this additional condition is necessary due to the fact that manual labeling of endpoints is not precise to the frame, and the focus of the system is to analyze speech, not silence or noise. In addition, frames that have been manually identified as music are discarded in performance evaluation.

Chapter 4

Speaker Spotting Experiments

This chapter describes the development of the speaker spotter. First, the performance of the speaker identification system is analyzed, both in closed-set and open-set identification. Also, a potential improvement to the speaker identification process, a cohort threshold is investigated. After examining the speaker identifier, the focus shifts to speaker spotting. A simple means of spotting via fixed segment lengths is evaluated, followed by testing performance with the addition of speaker transition probabilities, and with the use of overlapping segment lengths.

For all the experiments below, the speaker models described in section 3.3.4.3 are used.

4.1. Speaker Identification

The first set of experiments evaluates the performance of the speaker identification system (described in section 3.2). These experiments consist of closed-set identification evaluation and open-set identification evaluation.

4.1.1. Closed-Set Identification

To provide a baseline performance measure, the closed-set identification performance of the speaker identifier is analyzed according to identification segment length. This experiment uses as its model set the union of the target and homogeneous sets. For evaluation purposes, no distinction is made between these two sets. The testing data used in this experiment consists of

the target and homogeneous speakers' speech from the designated NH611 and NH612 testing data.

In order to provide equal-length testing segments, the segments of each individual speaker's speech from NH611 and NH612 are concatenated together into one long "utterance" for each speaker. This is then divided into the appropriate segment length. Although this method of artificially creating utterances leads to spectral discontinuities at the points of concatenation, the effects of these discontinuities in performance appear to be minimal and are ignored in analysis.

The segment lengths for testing start at 0.5 seconds and proceed in 0.5-second increments to a final length of 5.0 seconds. For each segment length, the concatenated speaker utterance is cut into segments of that length, and each of these segments is used as an utterance for testing the identifier.

For each segment, the speaker identifier returns a single hypothesized speaker name. If this name matches the name of the actual speaker of that segment, the result is considered correct. If the name does not match the name of the actual speaker, the result is considered incorrect.

These identification results are totaled according to the segment length. Figure 4.1 plots the percentage of segments correct for each of the segment lengths.

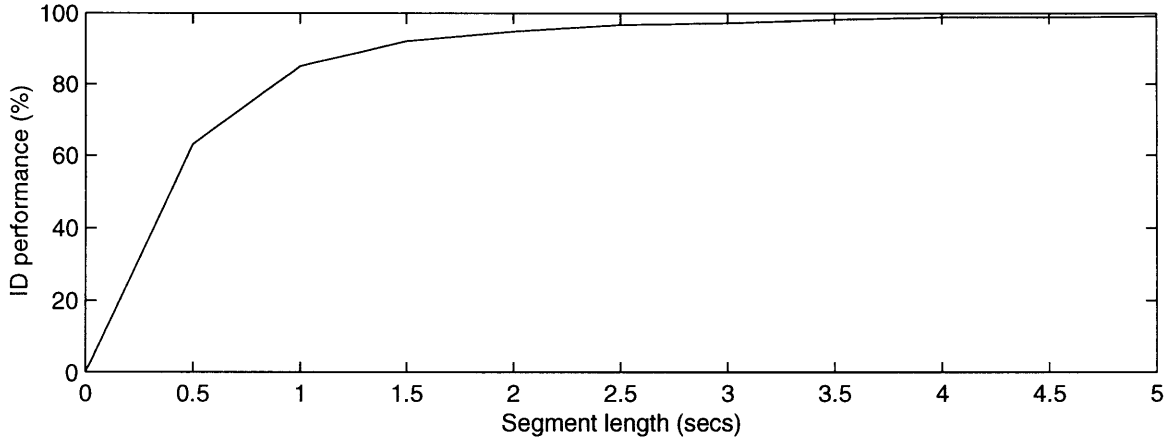


Figure 4.1: Closed-set identification performance.

The plot indicates that the speaker identifier performs relatively well for closed-set identification given an appropriate segment length. With a sufficiently long segment, performance approaches 98% correct. In addition, it demonstrates that a segment length below 2 seconds is probably not appropriate for speaker identification for this system, as below 2 seconds, there is a significant gain in performance for increasing segment size.

4.1.2. Open Set Identification

After verifying the performance of the speaker identifier on closed-set identification, the performance on open-set identification is examined. This experiment utilizes all three model sets for analysis. The testing uses speech from testing data of NH611 and NH612. Unlike the closed-set test, this one uses the speech of the unseen speakers as well as the target and trained sets. Again, the audio for each speaker is concatenated together into one utterance for each speaker, then divided into fixed segment lengths.

The first test analyzes the performance of the homogeneous garbage models in identifying whether or not a speaker belongs to the target set. Here, the relevant model sets are the target set and the homogeneous set. As in the closed-set tests, the speaker identifier returns a single hypothesized speaker name. If the actual speaker is one of the five target speakers, and the hypothesized name matches that of the actual speaker, the identification is considered correct. If the actual speaker is one of the nineteen trained speakers, and the hypothesized name is that of *any* trained speaker, then the identification is also considered correct. Unlike the closed-set test, the function of the homogeneous models is solely to identify a speaker as not being one of the target set. If the actual speaker is an unseen speaker, and the hypothesized name matched that of any trained speaker, it is likewise considered correct. Any other case is considered incorrect.

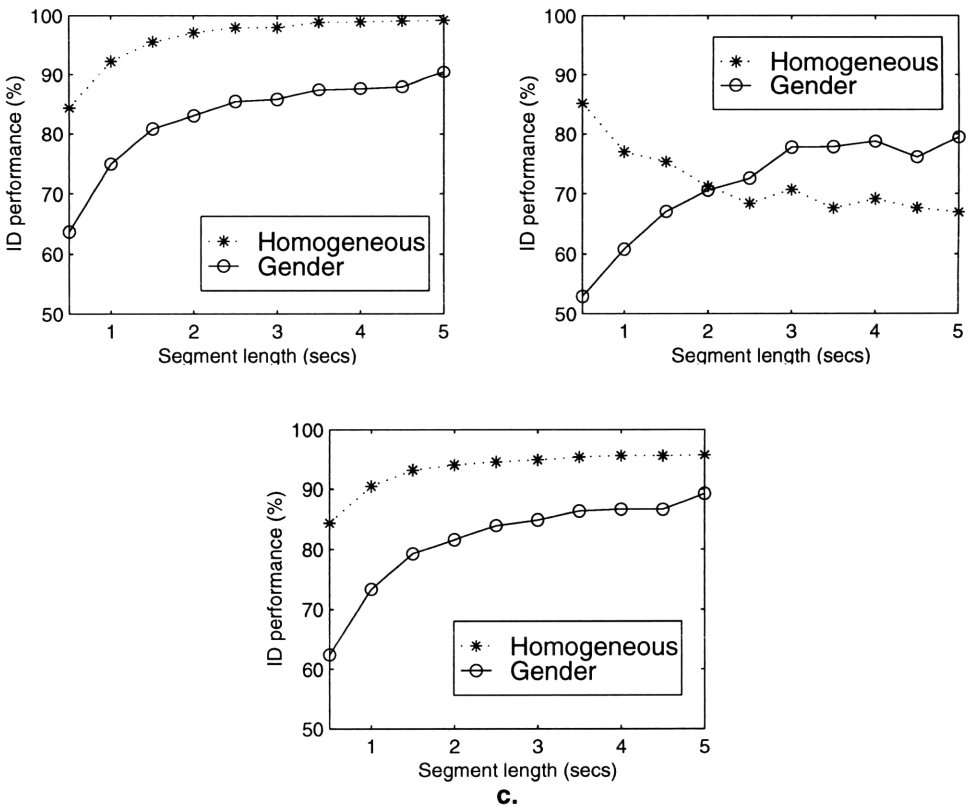


Figure 4.2: Open-set identification performance for (a.) registered speakers, (b.) unregistered speakers, and (c.) all speakers.

Figure 4.2 demonstrates how the performance varies by segment length for each of the speaker sets, as well as overall. It appears that the gender models do an adequate job in separating the target from the non-target speakers for an appropriate segment length of at least 2 seconds. However, Figure 4.2b shows that performance declines for identification of unseen speakers with increasing segment length. Examination of the confusion matrices demonstrates that this is due to the non-uniform distribution of the unseen speakers (or the non-uniform distribution of the garbage models). That is, many of the unseen speakers are clustered near a subset of the speakers of the other sets, and in particular are disproportionately clustered near the speakers of the target set. When the identification segment length is shorter, there is more randomness in the hypothesized identification (since the segment contains less information about the speaker), so performance is approximately near the ratio of the number of garbage models to total number of models. However, as the segment length grows longer, the randomness is lessened, and the

identification converges on a particular speaker. Since the distribution is not uniform, the performance drops.

The second test analyzes the performance of the gender garbage models in identifying whether or not a speaker belonged to the target set. The model sets used for this experiment are the target set and the gender garbage set. The methodology for this test is like that of the first test, except that if the actual speaker is a non-target speaker, the identification is correct if the speaker identifier returns the name of either one of the garbage models.

Figure 4.2 also demonstrates the open-set performance of the speaker identifier using the gender garbage models. The relatively worse performance on the target and trained sets of the gender models versus the homogenous models demonstrates that the multiple-speaker nature of the gender models effectively captures a general characteristic of the speakers as desired. The performance on the unseen speakers also reflects this, as the gender models do not suffer from the convergence problem that the homogenous models do.

4.1.3. Utterance Length Analysis

For speaker spotting, an appropriate segment length must be determined. A longer segment length is desirable, as the previous experiments indicate, in order to increase identification performance. However, a shorter segment length is desirable in order to more precisely determine speaker changes, as with the basic segmentation scheme used here, the precision has an upper bound of the segment length. In addition, a longer segment may miss short utterances, if the segment length is greater than the length of an utterance.

The experiments to this point have established that the minimum segment length should probably not be less than 2.0 seconds, as identification performance drops significantly below this point. Analysis of the length of speaker utterances in NH611 and NH612 help refine this decision.

Using the manual labeling of NH611 and NH612, a list of all segments and their lengths is composed. From this list, all non-speech segments (including silence, music, and background noise) are removed. Additionally, segments which contained overlapping speakers are also removed from the list. Figure 4.3 shows the distribution of the remaining segments.

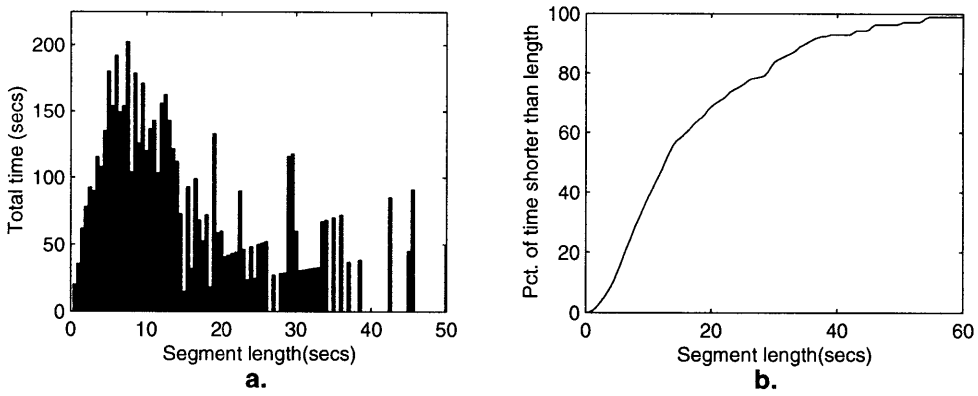


Figure 4.3: (a.) Distribution of amount of time spent in segments of length *n*; (b.) percentage of time spent in segments shorter than length *n*.

Figure 4.3b demonstrates that 13% of the utterances are shorter than a segment length of 5.0 seconds. Utterances significantly shorter than the segment length will tend to be missed by the spotter (though it is likely that utterances near the length of the segment will be caught by the spotter). So, this value seems viable as an upper bound on the segment length, as about 10% is a reasonable limit to the amount of audio that the spotter can be allowed to miss.

Based on the experiments done thus far, a spotting segment length of 2.5 seconds appears to be appropriate. Only 3.9% (by time) of the utterances are shorter than this (and so will be missed by the spotter with high probability), and identification performance seems reasonable, with a performance of 94.6% correct with the homogeneous garbage models and 83.9% correct with the gender garbage models. This segment length will be used for the rest of the experiments.

4.1.4. Threshold

One potential improvement in open-set speaker identification that is examined is implementing a cohort threshold for certifying an identification hypothesis of a target speaker. This is based on the belief that a correctly matched model will score significantly higher than any other model; an incorrectly matched model will thus have a score that is not significantly different from the scores of the other models.

If the top scoring model corresponds to one of the target speakers, a further analysis is made by the score handler. The difference between the top scoring model ("candidate") score and the second highest scoring model ("cohort") score is computed. If the difference is greater than a threshold value, then the identification hypothesis stands. However, if the difference is less than the threshold, then the identification hypothesis is considered to be wrong, and the hypothesis is changed to a garbage speaker.

The choice of an appropriate threshold value is investigated by testing a variety of different threshold values. The threshold value is chosen to balance the tradeoff between falsely identifying a target speaker as a non-target speaker and falsely identifying a non-target speaker as a target speaker. Figure 4.4 shows the performance of the various threshold values with the two garbage model alternatives.

A threshold value of 100 appears to maintain an appropriate level of performance for both garbage model sets; values much higher than 100 tend to become unreasonably poor in their performance in identifying the target speakers.

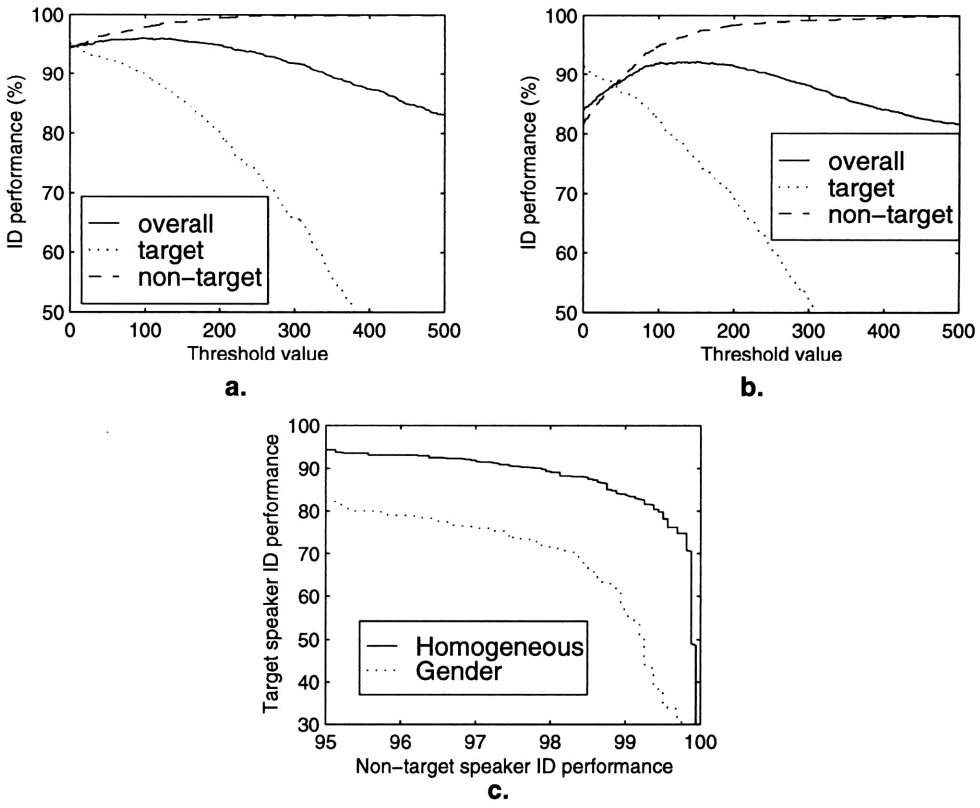


Figure 4.4: Threshold performance using (a.) homogeneous and (b.) gender garbage models; (c.) target vs. non-target identification performance.

4.1.5. System Tests

In the previous experiments, testing is done on the same sets of data as are used for training the speaker models. This means that there is a definite close similarity between the testing and training data. However, the conditions for the other hours of data may not match as closely with the others for a variety of reasons: for example, the *News Hour* may use different microphones for recording, the equalization may differ, etc. Since the similarity between the testing data and the training data is the basis for identification, it can be expected that the results from testing with NH611 and NH612 will act as an upper bound with respect to the results from other data sets.

In order to check the robustness of the speaker identification system to this variation, further testing is done on two more hours of data. The speaker models used in these experiments are

the same ones as used previously (trained from NH611 and NH612). The data sets used for testing were NH616 and NH620.

Two experiments are performed with these two sets of data. The first experiment analyzes the open-set performance of the speaker identifier without a cohort threshold. The second analyzes the performance with the cohort threshold. As done previously, the segments of each individual speaker's speech are concatenated together. However, instead of iteratively testing different segment lengths, these single utterances are divided into 2.5 second-long segments for identification. In the experiment with the cohort threshold, a threshold value of 100 is used.

Table 4.1 shows the results of these experiments. For comparison purposes, the results of the tests on NH611 and NH612 at 2.5 seconds are shown as well.

		Target		Non-target		Overall	
		Hom.	Gen.	Hom.	Gen.	Hom.	Gen.
Without threshold	NH611 and NH612	95.2	91.3	94.3	81.6	94.5	83.9
	NH616	70.9	55.9	42.8	58.4	50.7	57.7
	NH620	84.5	77.3	53.2	72.7	55.8	73.1
With threshold = 100	NH611 and NH612	89.9	82.2	97.9	94.8	96.0	91.8
	NH616	46.5	40.3	66.3	67.7	60.7	60.0
	NH620	73.2	60.8	76.2	82.8	75.9	84.7

Table 4.1: Identification performance (in percentages) for various data sets.

The results from NH616 and NH620 demonstrate the same trends as those of NH611/NH612. However, a marked decrease in performance from NH611/NH612 to NH616 and NH620 can be seen, demonstrating that the system is not significantly robust to changes in the environment of the speech. Additionally, the occasionally large differences between the performance on NH616 and NH620 demonstrate that performance is dependent on the environment as well.

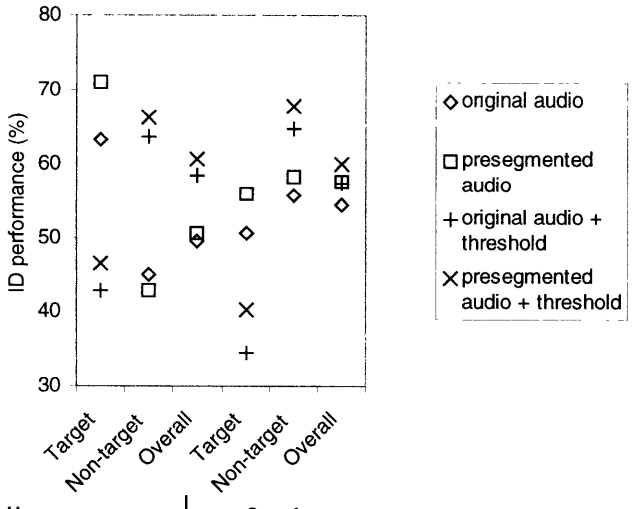
4.2. Speaker Spotting

The second part of the development of the speaker spotter uses the results of the speaker identification tests and applied them to the problem of speaker spotting. Specifically, the next series of experiments examine audio segmentation.

4.2.1. Fixed Segment Length Testing

The first spotting experiment uses a simple fixed-length identification window. The original audio is divided into consecutive 2.5-second segments, and identification is performed on each segment. With this segmentation, each segment may contain the speech of multiple speakers. Performance is measured according to the procedure outlined in section 3.4.

Four segmentation tests of this sort are performed on each data set: using the gender garbage model set, the homogeneous garbage model set, and both of these using a cohort threshold of 100. Figure 4.5 shows the results of this experiment on two different hours of data.



Homogeneous			Gender				Homogeneous			Gender		
Tar.	NT	All	Tar.	NT	All		Tar.	NT	All	Tar.	NT	All
63.3	45.0	49.5	50.5	55.8	54.4	original	76.7	55.1	56.9	68.2	74.5	73.9
70.9	42.8	50.7	55.9	58.4	57.7	presegmented	84.5	53.2	55.8	77.3	72.7	73.1
42.8	63.6	58.5	34.4	64.8	57.4	orig. + thresh	64.4	79.2	77.9	53.2	85.7	82.9
46.5	66.3	60.7	40.3	67.7	60.0	preseg + thresh	73.2	76.2	75.9	60.8	82.8	84.7

a. b.

Figure 4.5: Identification performance using 2.5-second segment length for (a.) NH616 and (b.) NH620.

Comparing these results with the results from the presegmented single-speaker audio, in general, a decline in performance is seen. Specifically, this can be seen with respect to the target model identification. This is as expected, since the fixed segment length does not guarantee that there is only one speaker in each segment. Additionally, there may not be sufficient speech in a segment in order to perform identification reliably (due to the elimination of silences longer than 0.5 seconds in the presegmented audio). The better performance of the fixed-length segments over the presegemented audio in the non-target sets (and therefore in some cases in the overall results) demonstrates that the garbage models capture generic speaker information, and does not necessarily actually reflect a significant "improvement" in system performance. That is, true improvement requires that performance for one of the speaker sets increases without a significant decrease in performance for the other speaker set.

4.2.2. Speaker Transition Probabilities

The utterance length analysis of section 4.1.3 provides additional information that can be useful in improving system performance. Since utterances tend to be longer than the identification segment length of 2.5 seconds, it is reasonable to assume that consecutive segments are likely to have the same speaker. Furthermore, knowing the nature of the data, there is reason to believe that certain speaker transitions are more likely than other speaker transitions. For these reasons, adding speaker transition probabilities to the spotting process may be beneficial.

Borrowing a technique from speech recognition, a forward search algorithm is applied to the speaker spotter. The continuous speaker spotting problem can be modeled as a probabilistic process captured via a hidden Markov model (HMM). Then, the forward algorithm can be used to augment the current speaker probabilities of the speaker identifier with probabilistic information about past speakers.

In modeling the spotting process as an HMM, the possible speakers (as represented by the speaker models) are the set of states

$$\mathbf{s} = \{s_1, s_2, \ldots, s_N\}$$

The current state is

$$q_t \in \mathbf{s}$$

The segments of audio being identified, represented by their respective models, are the observations

$$\mathbf{o} = \{o_1, o_2, \ldots, o_T\}$$

The state transition probabilities

$$\mathbf{A} = \{a_{ij}\}, \text{where } a_{ij} = \mathrm{P}(q_{t+1} = s_j \mid q_t = s_i),\ 1 \le i, j \le N$$

are the probabilities that speaker j follows speaker i. The observation probabilities

$$\mathbf{B} = \{b_j(k)\},\ \text{where } b_j(k) = \mathrm{P}(o_k \text{ at } t \mid q_t = s_t), 1 \le j \le N, 1 \le k \le T$$

are the probability scores (given as likelihoods) found by the speaker identifier. The *a priori* state probabilities

$$\pi = \{\pi_t\},\ \text{where } \pi_i = \mathrm{P}(q_1 = s_t), 1 \le i \le N$$

are the probabilities that the first state is speaker i.

The forward procedure then works as follows: first, define the forward variable $\alpha_t(i)$ as the probability of the partial observation sequence up to time t and also of being in state s_t at time t, or

$$\alpha_t(i) = \mathrm{P}(o_1 o_2 \ldots o_t, q_t = s_i)$$

For the initial step, it is seen that

$$\alpha_1 = \pi_t b_t(o_1), 1 \le i \le N$$

Furthermore, it is observed that

$$\mathrm{P}(\mathbf{o}) = \sum_{i=1}^{N} \alpha_T(i)$$

Therefore, by induction,

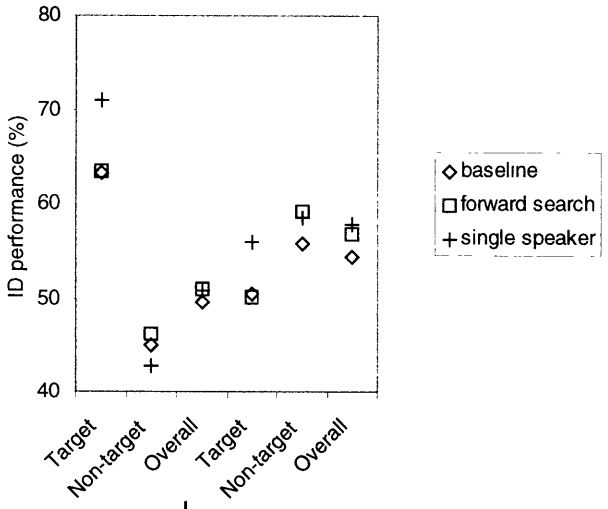
$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i)\alpha_{ij}]b_j(o_{t+1}), \quad \begin{matrix} 1 \le t \le T-1 \\ 1 \le j \le N \end{matrix}$$

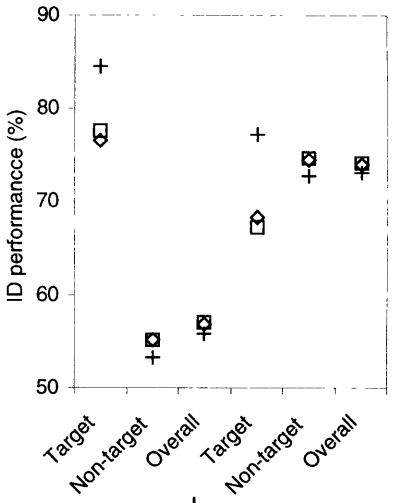
Using this, the new speaker score therefore takes on information both about the current segment (from the scores of the speaker identifier) and about the history of past segments. The hypothesized speaker identity is selected by selecting the speaker with the highest new score.

The *a priori* speaker probabilities are obtained by calculating the relative percentage of time each speaker spoke in the NH611 and NH612 data sets. Only actual speech is included for these probabilities; music, noise, and marked silences were excluded. For the garbage models (both homogeneous and gender), the total time spent in non-target speakers is added and equally distributed among all models.

The transition probabilities are obtained by dividing NH611 and NH612 into consecutive 2.5-second segment lengths and tallying the number of occurrences of each possible speaker transition. As with the *a priori* probabilities, transitions that involved non-target speakers are evenly distributed among all models.

Using these probabilities, the forward search is added to the spotting process. Figure 4.6 shows the results of the experiments with the search.





Homogeneous			Gender				Homogeneous			Gender		
Tar.	NT	All	Tar.	NT	All		Tar.	NT	All	Tar.	NT	All
63.3	45.0	49.5	50.5	55.8	54.4	baseline	76.6	55.1	56.9	68.2	74.5	73.9
63.5	46.1	51.0	50.1	59.2	56.7	forward search	77.6	55.1	57.0	67.2	74.7	74.1
70.9	42.8	50.7	55.9	58.4	57.7	single speaker	84.5	53.2	55.8	77.3	72.7	73.1

a. **b.**

Figure 4.6: Identification performance using forward search for (a.) NH616 and (b.) NH620.

The performance on target speakers using the gender model set appears to decline slightly with application of the search, while using the homogeneous model set it appears to rise slightly. This is most likely due to the fact that since there are only two gender garbage models, and all non-target speakers get mapped to these models, the probability that a garbage model is next, according to the transition probabilities, is artificially high. However, with the homogeneous garbage models, there are more models for the non-target speakers to map to, so their probabilities are not inflated as they are with the gender models. In any case, the difference in performance with application of the forward search is minimal, due to the fact that the acoustic likelihoods (i.e., original scores from the speaker identifier) dominate over the contribution from the transitions.

4.2.3. Overlapping Segments

A major limiting factor in the accuracy of finding speaker changes is the identification segment length. The 2.5-second length implies that when the system is functioning properly, a speaker transition is within 2.5 seconds of the point at which it is noted. This is because the identifier may identify a speaker in a segment if anywhere in the segment, that speaker is speaking (there is no guarantee that the identifier will simply identify the speaker who is speaking the most in the segment). So, a transition may fall anywhere in a segment. However, decreasing this segment length will decrease identification performance, which is clearly undesirable. One possible means of increasing segmentation performance without decreasing identification performance is to use overlapping segments.

Overlapping segments attempt to improve spotting performance by decreasing the effective segment size without decreasing the actual window size. An identification window of constant size and a step size smaller than this window size are used. For each step, identification is performed on the audio contained in the window; then, the window is shifted by the step size.

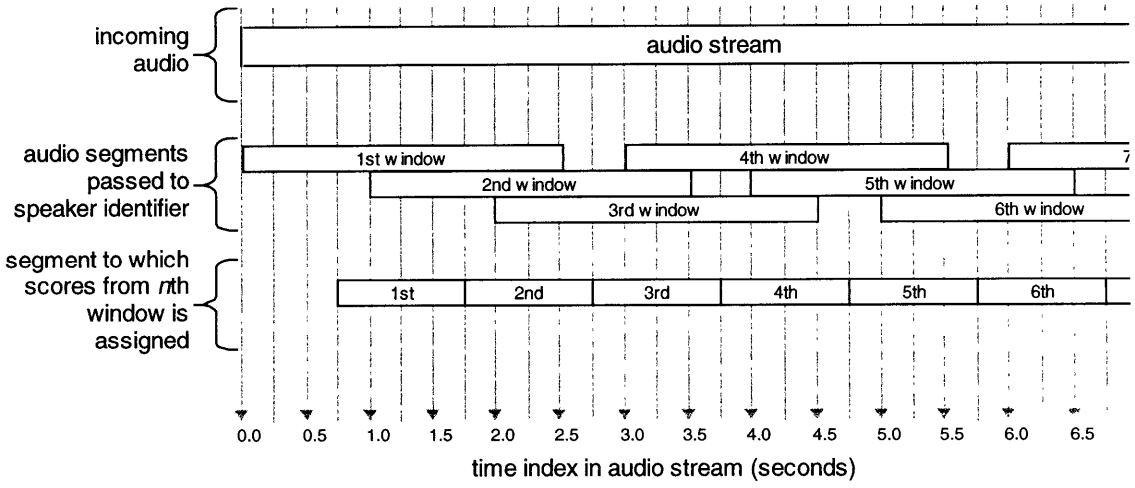
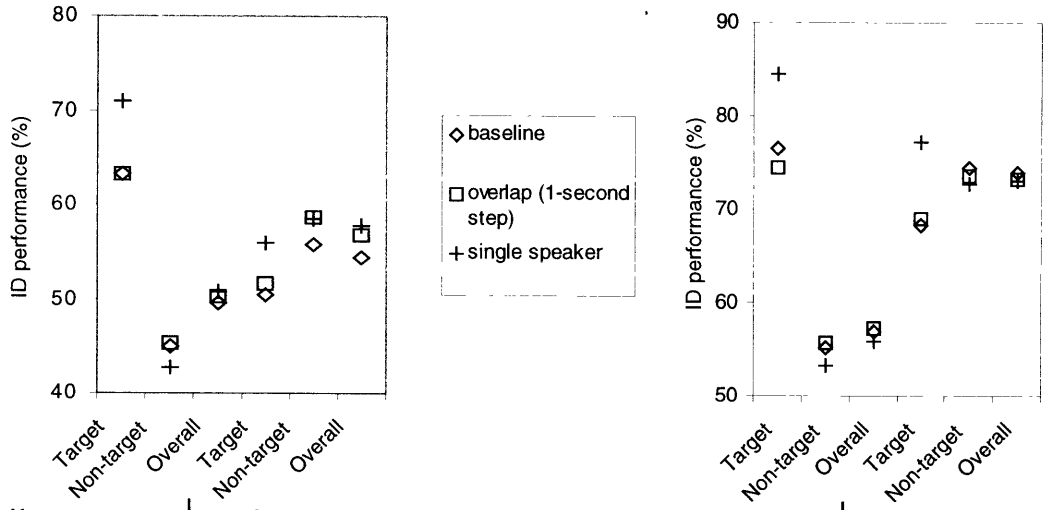


Figure 4.7: 2.5-second overlapping windows with step size of 1 second.

In the first experiment with overlapping windows, a window size of 2.5 seconds and a step size of 1 second were used. At each step, the identification scores reported were assigned to the middle 1 second of the window. This is based on the assumption that utterances are greater than 1



a.

Homogeneous			Gender			
Tar.	NT	All	Tar.	NT	All	
63.3	45.0	49.5	50.5	55.8	54.4	baseline
63.2	45.3	50.3	51.6	58.6	56.7	overlap (1-second step)
70.9	42.8	50.7	55.9	58.4	57.7	single speaker

b.

Homogeneous			Gender		
Tar.	NT	All	Tar.	NT	All
76.6	55.1	56.9	68.2	74.5	73.9
74.4	55.7	57.3	69.0	73.6	73.2
84.5	53.2	55.8	77.3	72.7	73.1

Figure 4.8: Identification performance using overlapping 2.5-second windows with 1-second step for (a.) NH616 and (b.) NH620.

second long, so that the audio surrounding a particular 1 second segment will belong to the same speaker with high probability, an assumption justified in section 4.1.3.

The results demonstrate that for the gender speaker models, performance generally increases for target speakers. For the homogeneous speaker models, performance generally decreases for target speakers. This discrepancy is due to similar reasons as those explained in section 4.1.2: certain homogeneous models are more like the target speakers than other models. As a result, by increasing the number of segments scored, the homogeneous garbage models that are similar to the target models have more opportunities to obscure the correct target models. However, this problem does not exist for the gender speaker models, so the gender speaker models increase performance while the homogeneous models decrease performance.

Having examined the impact of these overlapping segments on identification performance, the impact on segmentation is examined. Examining performance at the speaker changes when there was a difference in the segmentation between the two step sizes reveal ambiguous results. At some boundaries, performance is improved, with a hypothesized speaker change closer to the actual speaker change when the step size is decreased. However, at other boundaries, the segmentation is actually worse. Figure 4.9 demonstrates an instance of each of these circumstances.

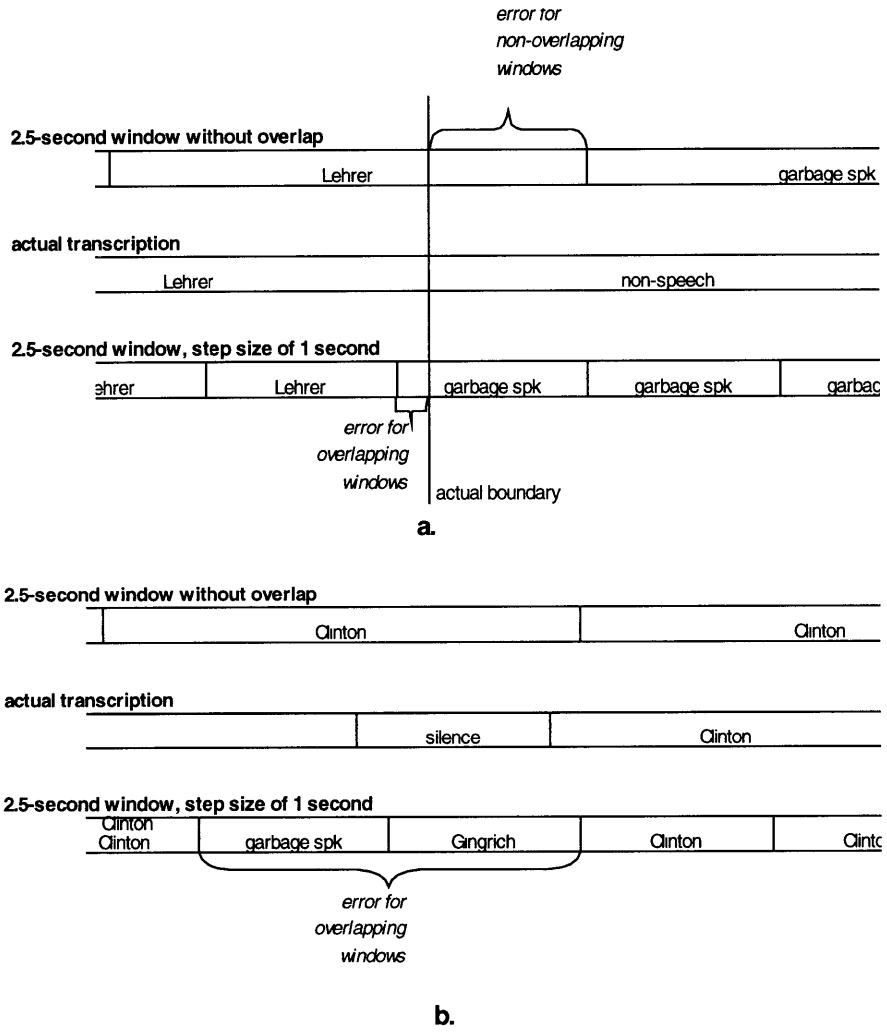


Figure 4.9: Segmentation performance (a.) increase and (b.) decrease with overlapping windows.

The improved performance is due to the increased resolution given by using overlapping windows. The decreased performance is due to the increase in segments, allowing for more segments to make errors. The potential for errors is greatest near speaker transitions, as the presence of multiple speakers and/or silence may confuse the identifier.

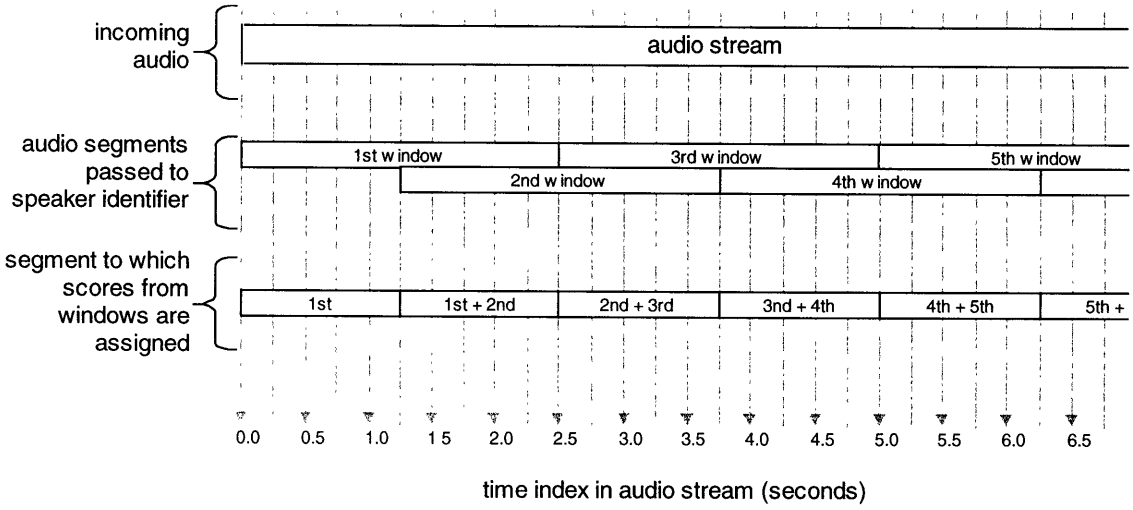
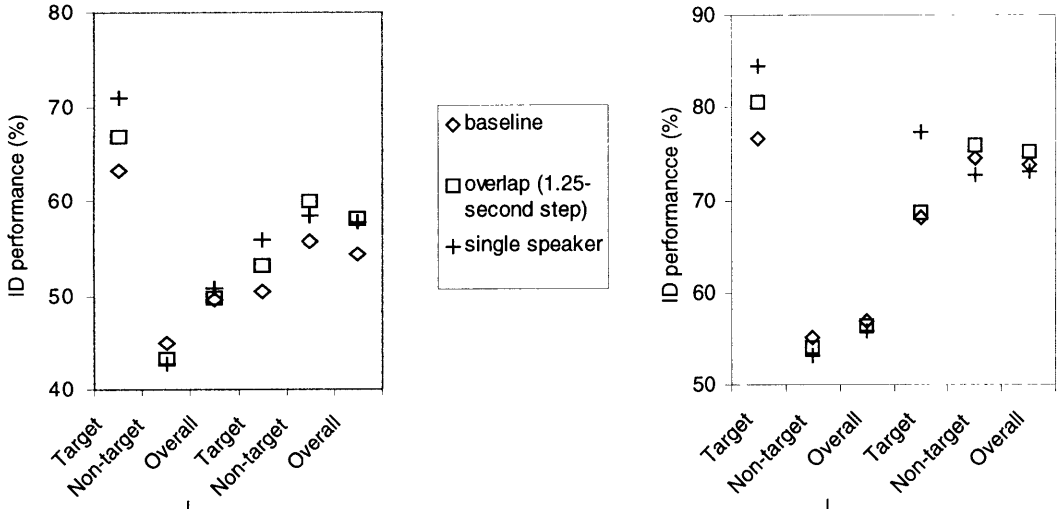


Figure 4.10: 2.5-second overlapping windows with step size of 1.25 seconds.

A second experiment is performed using overlapping windows of length 2.5 seconds and step size of 1.25 seconds, as shown in figure 4.10. A score for a 1.25 second segment is calculated as the sum of the scores for the windows that contained that segment. This attempts to maximize the utilization of the available data--unlike the previous experiment, all the information known about a segment (as contained by the scores) is used for identification in that segment. The results of this experiment appear in figure 4.11.

In each case, target speaker performance increases, while performance increases for the gender garbage models and decreases for the homogeneous garbage models. This is consistent with the results presented in section 4.1.2: the method of computing scores for a segment in this

experiment effectively lengthens the identification segment length. As a result, non-target speaker performance for the gender garbage models is expected to rise, while performance for the homogeneous garbage models is expected to drop.



Homogeneous			Gender				Homogeneous			Gender		
Tar.	NT	All	Tar.	NT	All		Tar.	NT	All	Tar.	NT	All
63.3	45.0	49.5	50.5	55.8	54.4	baseline	76.6	55.1	56.9	68.2	74.5	73.9
63.2	45.3	50.3	51.6	58.6	56.7	overlap (1.25-sec. step)	80.6	54.1	56.4	68.7	75.9	75.3
70.9	42.8	50.7	55.9	58.4	57.7	single speaker	84.5	53.2	55.8	77.3	72.7	73.1

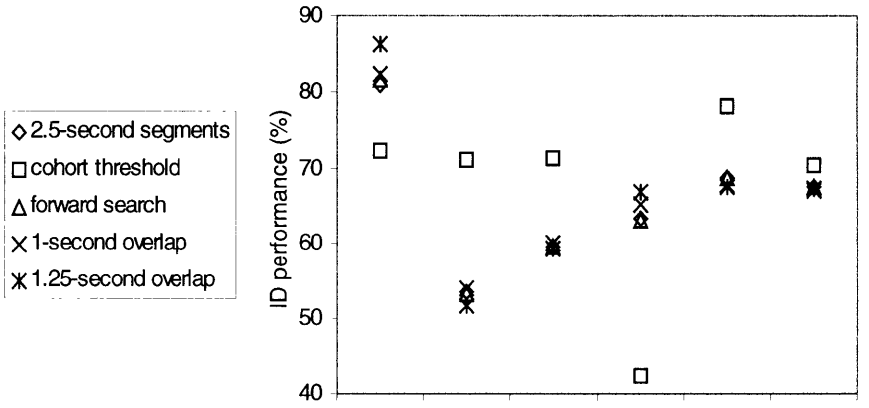
a. b.

Figure 4.11: Identification performance using 1.25-second overlap for (a.) NH616 and (b.) NH620.

4.3. Final System Analysis

A final set of experiments is performed using one hour of data that had not been previously used. These results are here to verify the previously obtained results. The data set used for this test ss NH617. Figure 4.12 exhibits the results of these tests.

These results show similar trends to the previous ones. In particular, there is a small increase in performance for target speakers (with a small decline in performance for non-target speakers) with application of the forward search using the homogeneous garbage models. Also, using the 1.25 second overlapping windows gives a measurable performance boost for the target speakers with a moderate decrease for non-target speakers.



	Homogeneous			Gender		
	Tar.	NT	All	Tar.	NT	All
2.5-second segments	80.9	53.6	59.6	63.4	68.7	67.5
cohort threshold	72.3	71.0	71.2	42.4	78.1	70.3
forward search	81.6	53.4	59.6	63.2	68.8	67.6
1-second overlap	82.2	54.0	60.2	65.3	67.8	67.2
1.25-second overlap	86 3	51.7	59.3	66.9	67.5	67.4

Figure 4.12: Identification performance for NH617.

Chapter 5
Conclusions and Future Work

5.1. Conclusions

This project examined the possibility of applying speaker recognition technology to the problem of speaker spotting. A number of various techniques in open-set identification and segmentation were examined. These utilized a fixed-length identification segment and added other methods as well. Adding a cohort threshold demonstrated how false positive identifications of target speakers could be decreased at the expense of increasing false negative identifications. A forward search was added in hopes of increasing spotting performance by using probabilistic information about speech patterns, but demonstrated little benefit. The use of overlapping decision windows showed how segmentation precision could be increased, but with a cost of accuracy in falsely identifying transitions. Overlapping windows also demonstrated how the tradeoff between longer identification windows increasing identification performance and shorter windows increasing segmentation accuracy and precision could be overcome, thus maintaining accuracy and precision while increasing identification performance.

Overall, these results demonstrated that speaker spotting is a potentially achievable goal. However, performance at this point is not good enough for practical use of speaker spotting. By adding such techniques as these, the performance can be improved to a point; futher improvements depend wholly on the performance of the speaker identification system.

5.2. Future Work

Several possible steps can be taken to improve speaker spotting:

5.2.1. Speaker Identification Improvements

The performance of the speaker identification system provided an upper bound to the performance of the spotter. Simply improving identification performance should thus improve spotting performance as well. Three particular areas of improvement are noteworthy:

- Speaker Modeling. By improving the speaker models, identification performance within a particular environment would be increased. One particular possible alternative to the single Gaussian model used here is a Gaussian mixture model.
- Robustness. Improving the robustness of the speaker identifier to environmental conditions means that performance would be constant across data sets. One method of coping with changing environmental conditions is the use of multiple models for each speaker, each one tuned to a particular environment.
- Window size. It is desirable to reduce the window size needed to perform speaker identification. By decreasing the window size, segmentation performance could be improved.

5.2.2. Non-target Speaker Modeling

The models developed for identifying non-target speakers were not developed with much effort toward maximizing their performance. In particular, for the homogeneous models, the models could be selected such that they are reasonably well distributed across the space of speakers. This would reduce some clustering effects that may have led to decreased identification performance. For the gender models, investigation of the number of speakers used in training the models could increase their coverage in identifying speakers. This is particularly true with respect to the female garbage model—the use of the speech of only two females may have negatively impacted performance. Alternately, using other means of creating the garbage models may be lead to performance increases.

5.2.3. Variable-length Windows

Fixed-length identification windows were used here due to their simplicity. However, a variable-length window based on clustering could be beneficial to improving identification and segmentation performance. For example, if two adjacent windows had similar models, the system could determine with high probability that they shared the same speaker. The two windows could then be clustered in order to identify the speaker. With this longer effective window, identification performance for that speaker could be improved.

Alternately, since short silences are relatively common, many fixed-length windows may not actually contain their full complement of speech. With a variable-length window, if the window contained a portion of silence, the window could be extended so that it captures enough speech to reliably identify the speaker.

References

Anderson, T. E., *An Introduction to Multivariate Statistical Analysis*, New Your, NY, John Wiley and Sons.

Garofolo, J.S., Fiscus,J.G., and Fisher, W.M., February 2-5,1997, "Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora," *DARPA Speech Recognition Workshop*, Chantilly, VA.

Gish, H., and Schmidt, M., 1994, "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, Vol. 11, No. 4, pp. 18-32.

Ligget, W. and Fisher, W., February 8-11,1998, "Insights from the Broadcast News Benchmark Tests," *DARPA Speech Recognition Workshop*, Chantilly, VA.

Mason, J. and Eatock, J., 1992, "Simple Speech Classifiers for Improved Speaker Recognition," *European Signal Processing and Imaging Conference*, Brussels.

Rabiner, L., and Juang, B.H., 1993, *Fundamentals of Speech Recognition*, Engelwood Cliffs, NJ, Prentice Hall.

Reynolds, D.A., and Rose, R.C., 1995, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72-83.

Roy, D.K., 1995, "NewsComm: A Hand-Held Device for Interactive Access to Structured Audio" (S.M. thesis), Massachusetts Institute of Technology, Program in Media Arts and Sciences, School of Architecture and Planning.

Siu, M.H., Yu, G. and Gish, H., March 1992, "An Unsupervised, Sequential Learning Algorithm for the Segmentation of Speech Waveforms with Multiple Speakers," *International Conference on Acoustics, Speech, and Signal Processing,* San Francisco.

Wilcox, L., Chen, F. Kimber, D., and Balasubramanian, V., April 1994, "Segmentation of Speech Using Speaker Identification," *International Conference on Acoustics, Speech and Signal Processing*, Adelaide.