SEQUENTIAL ORGANIZATION IN COMPUTATIONAL AUDITORY SCENE ANALYSIS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Yang Shao, M.S.

*****

The Ohio State University

2007

Dissertation Committee:

Professor DeLiang Wang, Advisor

Professor Eric Fosler-Lussier

Professor Phil Schniter

Approved by

_____

Advisor
Graduate Program in
Computer Science and Engineering

ABSTRACT

A human listener has the ability to follow a speaker's voice while others are speaking simultaneously. In particular, the listener can organize the time-frequency (T-F) energy of the same speaker into a single stream. This aspect of auditory perception is termed auditory scene analysis (ASA). ASA comprises two organization processes: segmentation and grouping. Segmentation decomposes the auditory scene into T-F segments. Grouping combines the segments from the same source into a single perceptual stream. Within the grouping process, simultaneous organization integrates segments that overlap in time, and sequential organization groups segments across time.

Inspired by ASA research, computational auditory scene analysis (CASA) aims to organize sound based on ASA principles. CASA systems seek to segregate target speech from a complex auditory scene. However, almost all the existing systems focus on simultaneous organization. This dissertation presents a systematic effort on sequential organization. The goal is to organize T-F segments from the same speaker that are separated in time into a single stream. This study proposes to employ speaker characteristics for sequential organization.

This study first explores bottom-up methods for sequential grouping. Subsequently, a speaker-model-based sequential organization framework is proposed and shown to yield

better grouping performance than feature-based methods. Specifically, a computational objective is derived for sequential grouping in the context of cochannel speaker recognition. Cochannel speech occurs when two utterances are transmitted in a single communication channel. This formulation leads to a grouping system that searches for the optimal grouping of separated speech segments. To reduce search space and computation time, a hypothesis pruning method is then proposed and it achieves performance close to that of exhaustive search. Systematic evaluations show that the proposed system improves not only grouping performance but also speech recognition accuracy.

The model-based grouping system is then extended to handle multi-talker as well as non-speech intrusions using generic models. This generalization is shown to function well regardless of interference types and the number of interfering sources. The grouping system is further extended to deal with noisy inputs from unknown speakers. Specifically, it employs a speaker quantization method that extracts representative speakers from a large speaker space and performs sequential grouping using obtained generic models. The resulting grouping performance is only moderately lower than that with known speaker models.

In addition to sequential grouping, this dissertation presents a systematic effort in robust speaker recognition. A novel usable speech extraction method is proposed that significantly improves recognition performance. Then, missing-data recognition is

combined with the use of CASA as a front-end processor. Substantial performance improvements are achieved in speaker recognition evaluations under various noisy conditions. Finally, a general solution is proposed for robust speaker recognition in the presence of additive noise. Novel speaker features are derived from auditory filtering and cepstral analysis, and are used in conjunction with an uncertainty decoder that accounts for mismatch introduced in front-end processing. Systematic evaluations show that the proposed system achieves significant performance improvement over the use of typical speaker features and a state-of-the-art robust front-end processor for noisy speech.

Dedicated to my parents Yongwei Zhou and Gengnian Shao, and my wife Chen Chen.

ACKNOWLEDGMENTS

First and foremost, I wish to thank my adviser, Dr. DeLiang Wang, for his intellectual support and guidance. He has taught me what it entails to be a researcher and shown me by his example many of the qualities I needed to cultivate.

Many thanks are due to Dr. Eric Fosler-Lussier, whose open door policy and seminars enabled me to stay up-to-date with the theory and practice of large vocabulary speech recognition. I also thank Dr. Phil Schniter for serving on my candidacy and dissertation committees.

I am grateful to all the friends in the Perception and Neurodynamics Laboratory at The Ohio State University. I will remember many pleasant days when we discussed problems, exchanged opinions, or even argued with each other. In particular, I wish to thank lab alumni, Drs. Mingyang Wu, Nicoleta Roman, Guoning Hu and Soundararajan Srinivasan, who have set great examples of being seniors and leaders of the lab for the rest of us. I also want to thank Yipeng Li and Zhaozhang Jin. They are great fellow graduate students and have made my Ph.D study enjoyable.

Special thanks to my family and friends. They have given me great support all these years. Many thanks go to my wife, Dr. Chen Chen. Her love, care, and kindness have given me tremendous assistance and support.

Finally, I wish to acknowledge the financial support provided by an AFOSR grant, an AFRL grant and an NSF grant.

VITA

April 1, 1976 ……..	Born – Zhejiang, China
June, 1998 ……	B.E., Computer Science, Nanjing University of Aeronautics and Astronautics, Nanjing, China
June, 2001 ……	M.S., Computer Science, Fudan University, Shanghai, China

PUBLICATIONS

Journal Articles

Y. Shao, S. Srinivasan, Z. Jin and D.L. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," submitted to *Computer Speech and Language*, 2007.

Y. Shao and D.L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 289-298, 2006.

Y. Shao and Z.G. Li, "A speaker recognition system using MFCC features and weighted vector quantization," *Computer Engineering and Applications*, China, 2001.

Y. Shao and Z.G. Li, "Home banking service system using HMM framework," *Computer Engineering*, China, 2001.

Conference Papers

Y. Shao, S. Srinivasan and D.L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* vol. IV, pp. 277-280, Honolulu, USA, 2007.

S. Srinivasan, Y. Shao, Z. Jin and D.L. Wang, "A computational auditory scene analysis system for robust speech recognition," In: *Proceedings of Interspeech*, pp. 73-76, Pittsburgh, USA, 2006.

Y. Shao and D.L. Wang, "Robust speaker recognition using binary time-frequency masks," In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. I, pp. 645-648, Toulouse, France, 2006.

Y. Shao and D.L. Wang, "Model-based sequential organization for cochannel speaker identification," In: *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Jeju, Korea, 2004.

Y. Shao and D.L. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. II, pp. 205-208, Hong Kong, China, 2003.

Others

Y. Shao, S. Srinivasan, Z. Jin and D.L. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Technical Report, OSU-CISRC-8/07-TR62*, Dept. of Computer Science and Engineering, The Ohio State University, 2007.

Y. Shao and D.L. Wang, "Model-based sequential organization in cochannel speech," *Technical Report*, *OSU-CISRC-05/04-TR30*, Dept. of Computer Science and Engineering, The Ohio State University, 2004.

FIELDS OF STUDY

Major Field: Computer Science and Engineering

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

One thing I do on my drive to school is to turn on the radio and tune to 'Morning Zoo at WNCI' FM 97.9. This is almost a daily routine for me to laugh a bit and wake up by listening to this humorous program. Dave and Jimmy, the hosts, usually tease with each other while a third host Kelsey and sometimes other invited hosts jump in from time to time to engage in arguments about topics from Washington to Hollywood. I never gave it a second thought until one day it occurred to me that this listening environment was fairly complex with active talkers and other sounds from the background. I realized that I had somehow associated some images with the hosts in my head based on their distinct voices. It is not hard for me to choose to listen to one of them even when there are two or three voices talking at the same time. I had long taken this listening ability for granted.

Like the above experience, a daily auditory scene typically comprises multiple sounds from different sources. Usually there is a target source that one is listening to,

such as a radio host, a piece of music being played etc. Meanwhile, there are acoustic events from other sound sources that are of little interest to the listener, such as an unscrupulous couple talking to each other, a ventilation fan in an office, a projector in a meeting room, or cars on the street etc. In cochannel speech, for example, a combination of utterances from two speakers is transmitted over a single communication channel. Unlike conversations, talkers from different channels are usually not aware of each other under cochannel conditions. In the radio example above, cochannel speech exists when two hosts purposefully talk over each other's voice. Consequently, speech from both channels has large overlap, which presents a considerable challenge to applications such as automatic speaker and speech recognition.

On the other hand, for a cochannel signal that has comparable energies from both talkers, human listeners can readily select and follow one speaker's voice (Brungart, 2001). Even in worse scenarios such as a cocktail party, listeners can select and follow the voice of a particular talker as long as the signal-to-noise ratio (SNR) is not exceedingly low (Helmholtz, 1863; Cherry, 1953; Bregman, 1990). This phenomenon is termed as the 'cocktail party problem' (Cherry, 1953). The human ability to function well in everyday, complex acoustic environments is due to a perceptual process termed auditory scene analysis (ASA), which produces a subjective representation of different sources in an acoustic mixture (Bregman, 1990). In other words, listeners organize the auditory scene into streams that correspond to different sound sources in the input.

According to Bregman (1990), organization in ASA takes place in two main processes: segmentation and grouping (Wang and Brown, 2006). Segmentation

decomposes the auditory scene into groups of contiguous time-frequency (T-F) units or segments, each of which primarily originates from a single sound source. A T-F unit denotes the signal at a particular time and frequency. Grouping involves combining the segments that are likely to arise from the same source together into a single stream. Thus, each of the formed streams gives a perceptual representation of a sound source in the input mixture. Grouping itself is composed of simultaneous and sequential organization. Simultaneous organization involves integration of segments that overlap in time, and sequential organization refers to grouping across time.

A computational auditory scene analysis (CASA) system that segregates target speech from a complex auditory scene is desirable for many applications. For example, CASA is used to separate voice mixture so that a speaker's speech is automatically transcribed in the presence of another speaker (Cooke and Lee, 2006). However, almost all the existing CASA systems (Divenyi, 2005; Wang and Brown, 2006) address only segmentation and simultaneous organization. On the other hand, sequential organization is crucial for building a complete CASA system. This dissertation focuses on the sequential organization aspect of CASA. In other words, our goal is to link together the segments from the same speaker that are separated in time.

CASA studies (Weintraub, 1985; Brown and Cooke, 1994a; Wang and Brown, 1999; Hu, 2006; Wang and Brown, 2006) employ periodicity and temporal continuity for speech segregation. Nevertheless, these systems deal with only voiced speech and they do not lend themselves easily to handling real-world speech inputs. Binaural cues such as interaural time difference and interaural intensity difference have been applied for sound

segregation and speaker tracking (e.g. Roman *et al.* 2003; Stern *et al.* 2006). Additionally, microphone arrays have utilized spatial separation of sources to track a speaker (Ward *et al.*, 2003). Nonetheless, the human auditory system can select and follow a speaker's voice when sound sources originate from the same direction (Brungart *et al.*, 2006). Here, we focus our sequential grouping study on monaural or single-microphone conditions.

Previous model-based systems utilize models from automatic speech recognition for speech organization (Ellis, 1996; Barker *et al.*, 2005; Ellis, 2006). However, by listening to a cochannel mixture, one can follow the voice of either speaker even when they are speaking in a language unknown to the listener (Wang, 2006). Apparently, the auditory system does not require language-specific knowledge for sequential organization; rather it appears to rely on speaker characteristics contained in the speech signal for the grouping purpose.

It is well known that human listeners use speaker characteristics such as excitation and vocal tract information to recognize a speaker's voice (Schmidt-Nielsen and Crystal, 1998; Furui, 2001) and such characteristics have been incorporated in models of automatic speaker recognition (Atal, 1972; Matsui and Furui, 1990; Naik, 1990; Furui, 1994; Reynolds, 1995; Campbell, 1997; Furui, 2001; Reynolds *et al.*, 2003; Bimbot *et al.*, 2004). In this dissertation, we propose to utilize speaker characteristics for sequential organization.

1.2 Objectives and Research Issues

The objective of this dissertation is to design effective algorithms for sequential organization in CASA research. We intend to employ speaker modeling and classification methods for this purpose. As a byproduct, we also intend to improve speaker recognition under noisy conditions. Specifically, this dissertation will address the following research issues.

- *The Goal of Sequential Organization*. As described earlier, in the ASA account, the goal of sequential organization is to link speech from the same speaker that is separated in time and put it into a single auditory stream. In order to translate this goal into a computational objective, we have to first answer the following question: What is the separated speech? From the CASA perspective, the separated speech refers to disjoint homogeneous segments, each of which is composed of contiguous T-F units that primarily originate from a single source. These segments are extracted from a speech input by segmentation and simultaneous grouping. Hence, they are termed as simultaneous streams. The goal of sequential organization in the context of CASA is to organize these streams into their corresponding source streams in computer memory. In other words, it amounts to finding the best assignment of simultaneous streams to source (speaker) streams.
- *Robust features for speaker recognition and sequential grouping*. As described earlier, existing CASA systems (Weintraub, 1985; Brown and Cooke, 1994a; Wang and Brown, 1999; Hu, 2006) utilize periodicity and temporal continuity cues for

sequential grouping. However, periodicity cues such as pitch have difficulty in real-world segregation tasks because such cues have reduced discriminative power with a large number of speakers. In addition, the temporal continuity cue is not able to group speech segments that are separated by silence. Thus, this cue is not applicable as well. As we propose to exploit speaker characteristics for grouping, features from speaker recognition (Atal, 1972; Matsui and Furui, 1990; Naik, 1990; Furui, 1994; Campbell, 1997; Furui, 2001; Bimbot *et al.*, 2004) may be good candidates for speech organization. However, such features, when used directly, do not perform well under noisy conditions (Barger and Sridharan, 1997; Ortega-Garcia and Gonzalez-Rodriguez, 1997; Drygajlo and El-Maliki, 1998; Schmidt-Nielsen and Crystal, 1998; Sivakumaran and Ariyaeeinia, 2000; Lovekin *et al.*, 2001; Yoshida *et al.*, 2001; Shao and Wang, 2003; Shao and Wang, 2006b). Thus, it remains a challenge to find robust features for speaker recognition and sequential grouping.

- *Speaker modeling and scoring methods*. CASA systems usually segregate speech based on the time-frequency decomposition of an input (Wang and Brown, 2006). Speaker features in the T-F domain are typically vectors composed of individual frequency components. Under noisy conditions, some of these components are corrupted while others are relatively intact due to different degrees of spectral overlap between speech and noise. Conventional speaker modeling and scoring methods assume clean feature vectors (Matsui and Furui, 1990; Naik, 1990; Furui, 1994; Campbell, 1997; Furui, 2001; Bimbot *et al.*, 2004), and they are not applicable in the presence of noise.

- *Feature-based or model-based sequential organization*. Bregman (1990) recognizes two kinds of grouping, namely primitive grouping and schema-based grouping. The former relies on innate sound attributes (features) and is thus regarded as a bottom-up process. The latter utilizes acquired schemas (models) and is thus considered as a top-down process. A human listener is able to group speech in a foreign language (Wang and Brown, 2006) and follow a stranger's voice at a party (Cherry, 1953). Apparently, under such conditions, humans rely on general speaker characteristics for speech organization. On the other hand, humans also seem to utilize prior knowledge of a speaker for grouping. For example, a familiar voice helps a listener to better follow that voice. Therefore, we need to address both primitive grouping and schema-based grouping. In the CASA account, we need to study grouping approaches based on innate features and those based on acquired models.
- *Known or unknown speakers.* As implied in the aforementioned radio example, it seems that one's familiarity with the talkers plays a role in the organization of different speakers. It is reasonable to assume, under some conditions, that the speakers in the auditory scene are known *a priori*. In other words, the models of all the speakers that can appear in an input are available in advance. On the other hand, the cocktail party problem shows one condition where one is familiar only with the voice that he or she is paying attention to. This presents an open-set condition where only the target speaker model is available. Furthermore, the listening experience of foreign language mixtures (Wang, 2006) indicates that sometimes one does not need any knowledge of the talkers at all. This is a completely open-set condition where

none of the speaker models in the input are available. From a security application point of view, these three conditions each have their own applicable domains while the last is clearly the hardest to tackle.

As no previous study has used speaker characteristics for sequential organization, there are many questions to be answered as described above. Our general goal is to study sequential organization in the CASA context. Specifically, we seek to design grouping algorithms where inputs are simultaneous streams obtained by segmentation and simultaneous grouping. The system determines the best assignment of these streams to corresponding speakers using speaker characteristics. The system produces organized speaker streams as output.

Specifically, we intend to explore bottom-up methods that directly employ speaker features for speech organization. On the other hand, we also aim to study how to apply acquired schemas in the form of statistical models from speaker recognition to sequential grouping. Sometimes, it is infeasible to assume prior models of the sources that may appear in an auditory scene. Under such conditions, we intend to explore alternative modeling methods to extract generic models that account for a group of sources. We then seek to study how to use the obtained models for sequential organization. In addition, we intend to improve robust speaker recognition by applying novel speaker features and incorporating CASA as a front-end processor.

1.3 Organization of Dissertation

This dissertation presents a systematic effort in developing a sequential organization system. Our study aims to address the issues and trying to achieve the goals as described in Section 1.2. The remainder of the dissertation is organized as follows.

In Chapter 2, we first survey perceptual studies on sequential organization, seeking to identify cues and methods that are applicable for our computational study. Then we describe previous CASA studies on sequential grouping. Finally, we present several previous CASA studies, including multipitch tracking, binary T-F masks and voiced speech segregation. Outputs of these studies lay foundations for our work on robust speaker recognition and sequential grouping. We will also describe binary T-F masks and ideal binary T-F mask as the goal of CASA.

In Chapter 3, we first describe a statistical framework for speaker recognition, including decision rules for speaker identification and verification tasks. The robustness issues are discussed next. We then present a usable speech extraction method that captures minimally corrupted speech segments at the frame level in order to improve identification performance under cochannel conditions. We employ a voiced speech segregation system that produces binary T-F masks as output. The masks indicate reliable (clean) or unreliable (noisy) units on a T-F representation of an input. A missing-data recognition method is employed to utilize such masks. Substantial performance improvements are achieved in both identification and verification evaluations under various noisy conditions. Finally, we propose a general solution to robust speaker recognition in the presence of additive noise. Novel speaker features are derived from

auditory filtering and cepstral analysis. In addition to feature derivation, we apply, as a novel speaker scoring method, an uncertainty decoder that accounts for front-end processing errors in conjunction with estimated binary masks. Our evaluation shows that the proposed system achieves substantial performance improvement over not only typical speaker features but also a state-of-the-art robust front-end processor for noisy speech.

In Chapter 4, we first explore several feature-based grouping methods using features such as pitch, spectrum and vocal tract length. Then we present a speaker-model-based sequential organization system in detail. Specifically, we derive a computational objective for joint speaker identification and sequential grouping under cochannel conditions. Our formulation leads to an exhaustive search that finds the optimal hypothesis in the joint speaker and grouping space. A hypothesis pruning method is introduced to reduce the search space and computation time while achieving a performance level close to that of exhaustive search. Lastly, we present the proposed system as part of a complete CASA system and systematically evaluate its performance on a speech separation and speech recognition task. This system achieves a significant improvement over the baseline speech recognition performance across all the SNR conditions.

In Chapter 5, we extend the model-based sequential grouping system from cochannel speech to conditions where there are more than two talkers in the auditory scene. Subsequently, the system is generalized to deal with non-speech as well as speech interference. Both generalizations incorporate generic models that account for known or unknown interferences. We show that the system is able to function well when only

target speaker models are available, regardless of interference types and the number of interfering sources. Finally we present a system that does not require *a priori* knowledge of the speakers in the auditory scene. Specifically, a generic speaker modeling method is employed to quantize a large speaker space, and the obtained generic models are used for sequential organization of mixtures of unknown speakers. We show that the grouping performance is only moderately lower than that with known speaker models.

Chapter 6 summarizes the contributions of the dissertation. It also discusses insights gained and future research directions.

CHAPTER 2

BACKGROUND

This chapter first surveys perceptual studies that are related to sequential organization. We intend to utilize the cues and methods that are applicable to perceptual grouping for our computational study. Then we describe previous CASA studies that implicitly or explicitly address sequential grouping. Finally, we present three aspects of CASA: multipitch tracking, binary T-F masks and voiced speech segregation. These aspects lay foundations for our work on robust speaker recognition and sequential grouping.

2.1 Perceptual Studies on Sequential Organization

In this section, we first survey how the human auditory system uses various cues to organize speech in general. We then discuss speech organization studies that focus on multi-talker conditions where sequential organization is a crucial process. Since there are not many studies that directly evaluate human sequential organization, we survey such experiments in some detail. In addition, we describe several speech segregation studies

on subjects with hearing loss. Some of these studies employ sine-wave speech as input. Since Gestalt grouping principles play a universal role in perception, we touch upon the roles of such principles in sequential organization in the end.

As described earlier, sequential organization of speech refers to the process that organizes the speech of a talker separated in time into a perceptual stream. This process operates on a longer time scale than simultaneous grouping. Speech signals are composed of different types of acoustic events and carry specific information from a talker to listeners, and it differs from other sounds such as tones. The questions then arise whether humans rely on this linguistic information and knowledge about speech for perceptual grouping. It is shown that a listener is generally better at organizing speech sounds than non-speech sounds, and one is better at connected speech of a natural order than that of a distorted one (Bregman, 1990; Warren *et al.*, 1996). It is also shown that a speech-like signal is likely being perceived as speech (Rand, 1974; Liberman, 1982; Warren, *et al*., 1990; Remez, *et al*., 1994; Warren, *et al*., 1996). These observations indicate that humans do, on the one hand, utilize the life-long acquired linguistic knowledge for organization.

On the other hand, a listener performs grouping regardless whether the utterances are spoken in a native or foreign language (Wang, 2006). In this case, the auditory system has to rely on something other than specific linguistic information for sequential organization. It is suggested that a listener uses perceived pitch trajectories to organize speech when speech is separated by silence (Bregman, 1990). This indicates the use of one type of speaker characteristics for grouping.

Moore (2003) outlines a set of cues for perceptual grouping of sounds. The cues include fundamental frequency, onset disparities, contrasts with previous sounds, changes in frequency or intensity, and sound location. Furthermore, in natural acoustic environments, the auditory system is confronted with a mixture of sounds from multiple active sources. Humans can only organize a limited number of stimuli using a single feature such as pitch or fundamental frequency. Moore states that combining different physical cues provides a good base for parsing the acoustic input.

Speech organization under multi-talker conditions

Under multi-talker conditions, a mixture input is decomposed into small speech segments which are then simultaneously and sequentially grouped into auditory streams. Between the two streams, a listener usually pays attention to one of them, which is deemed as target. The other speaker in the pair is regarded as a masker. Here, we can infer the performance of sequential organization by evaluating how well the listener follows the target speech. This type of evaluation is typically quantified by a speech intelligibility test, which measures the speech reception threshold as the required SNR in decibels (dB) for a 50% intelligibility score.

Darwin *et al.* (2003) report that the most powerful cues for monaural speech segregation relate to vocal variations of the underlying speakers in a two-talker mixture. The authors systematically examine the influence of features such as fundamental frequency (*F*0) and vocal-tract length (VTL) on segregation. In the experiments, an mixture input consists of two utterances that are produced by the same speaker and the

utterances are selected from the Coordinate Response Measure speech corpus (Bolia *et al.*, 2000). A target utterance always contains an anchor word 'Baron' followed by a color and a number and a masker utterance is randomly selected from the corpus. The masker utterance has different contents than that of the target. These utterances are electronically modified by a pitch-synchronous overlap and add (PSOLA) algorithm according to controlled *F*0 contours. Spectral envelopes of the signal are scaled to reflect VTL changes. Note that the utterances are modified in such a way that they can be considered as from different speakers even though the original ones are produced by the same speaker. Subjects are asked to identify the color and number in the target utterance that contains the word 'Baron'. One intelligibility experiment varies *F*0 contours and it shows systematic improvement when *F*0 values between the two utterances differ by greater than two semitones. In another experiment that evaluates VTL, systematic improvement is recorded if the ratio of target VTL and masker VTL is no less than 8%. The third experiment modifies both *F*0 and VTL of one of the utterances and this modification actually reflects a gender change. The resulting performance improvements are as great as those obtained by real recordings of different-gender speakers. Furthermore, the improvements are much greater than those by varying *F*0 or VTL alone. In addition, the authors also find that large intonation variations of a speaker lead to minimal performance improvement.

Besides the performance difference caused by the gender change in the above study, Brungart (2001) finds that speech intelligibility is the worst when both target and interference utterances are produced by the same speaker. Subjects' performance is better

when the utterances originate from different speakers and it is even better when the speakers are of different gender. This finding suggests that the cues that reflect gender contrasts in voices are useful for sequential grouping. Apparently, *F*0 and vocal-tract shape are two of such cues.

Drullman and Bronkhorst (2004) also show the effectiveness of pitch differences in speech intelligibility tasks that use two-talker mixtures. Its experimental setup is similar to the one above. Specifically, interference speech is systematically modified to create pitch differences between interference and target by 2, 4, 8, and 12 semitones. The intelligibility results gradually improve with increasing difference. Meanwhile, it is also found that the higher the energy ratio of target to interference, the better the intelligibility performance. Similarly, earlier studies (Assmann and Summerfield, 1990; Assmann and Summerfield, 1994) demonstrate that increasing the pitch difference between a pair of vowels that are simultaneously presented to the subject improve vowel identification performance.

Culling and Darwin (1993) suggest that when *F*0 contours of two speakers intersect and cross, human listeners rely on timbre continuity to determine whether the contours cross or not. Here, the stimuli are two simultaneous diphthong-like sounds that contain *F*0 contours either diverging or crossing at the intersection. The results show that the subjects are able to discriminate between a cross pattern and a divergence pattern with different timbres of the two sounds. However, the listeners are not able to tell the difference when the timbres are the same. In other words, the cues other than pitch are important for tracking a speaker.

Darwin and Hukin (2000) find that intonation and vocal-tract size cues override spatial cues in attending to a target utterance within a pair of speakers. In this study, listeners are asked to determine a word from a target utterance after hearing two simultaneous utterances. The experiment is configured in such a way as to evaluate intonation, vocal-tract size and spatial cues for tracking a particular speaker over time, instead of measuring how well listeners can recognize a word. A natural intonation that exhibits stress and varying *F*0 is found more effective than monotonous *F*0 contours. Authors also find that vocal-tract size is an effective cue for selective attention to one of the speakers.

Listeners' ability to follow a target in the presence of an interference talker can be attributed to a process termed glimpsing (Miller and Licklider, 1950; Assmann and Summerfield, 2004). The hypothesis is that listeners are able to exploit speech glimpses, defined as time-frequency regions where the target signal is much more energetic than the interference, for understanding speech in noise. Apparently, with an increasing number of interfering sources, there are fewer glimpses to be exploited. Speech intelligibility degrades sharply when the number of interference talkers increases from one to two, and the degradation slows when there are four or more maskers.

In summary, *F*0 or its perceptual counterpart, pitch, is an effective cue for speech segregation under multi-talker conditions. However, the above studies also find that the pitch contours of two speakers require at least a two-semitone difference in order to have a meaningful performance improvement in intelligibility tests. An increase of two semitones corresponds to 16 Hz for a 130 Hz tone or 32 Hz for a 260 Hz tone. Thus,

using pitch for speech organization becomes insufficient in the case of same-gender mixtures. Vocal-tract size appears to be a good candidate for grouping according to the studies described earlier. However, since the values of vocal-tract size are confined within a small range, it may also become insufficient with a number of speakers.

Grouping cues inferred from related studies

Sequential stream segregation tasks require subjects to segregate sequential pure-tone patterns. Grose and Hall (1996) suggest that listeners with cochlear hearing loss are able to perform perceptual organization of sequential stimuli but their performance is poorer than those with normal hearing. Specifically, listeners suffering from cochlear hearing loss require a greater frequency separation between presumed auditory streams than normal-hearing listeners do. In addition, Mackersie *et al*. (2001) and Mackersie (2003) suggest that speech segregation abilities under two-talker conditions are similar to those in sequential stream segregation. Hence, the cues that are deemed useful for sequential stream segregation likely also work for speech organization (Moore and Gockel, 2002). These cues include phase, temporal envelope, etc.

Numerous studies (Bregman, 1990; Hartmann, 1996; Vliegen and Oxenham, 1998; Vliegen and Moore, 1999; Yabe *et al.*, 1999; Moore and Gockel, 2002; Atienza *et al.*, 2003; Moore, 2003; Carlyon, 2004) on auditory scene analysis or auditory streaming employ non-speech inputs such as alternating low-high tones. These studies on tonal inputs find that frequency and time differences between tones are important cues for auditory streaming. However, such cues may not be directly applicable to speech

organization since speech signals are more complex than tonal signals, and conclusions made from tonal inputs do not lend themselves easily to speech inputs (Moore, 2003).

Sign-wave speech (SWS) is typically synthesized by using three time-varying sinusoids which reproduce frequency and amplitude variations of the first three speech formants from natural speech (Remez *et al.*, 1994). Listeners are able to transcribe the SWS to a certain extent even if they have not been exposed to such audio inputs before (Barker and Cooke, 1998). Under a condition where two SWS signals are combined, listeners can extract words from the mixture, but they are not able to group these words into two separate streams. This observation indicates that the acoustic features that are missing from SWS may be important for sequential grouping. Such features include *F*0, the complete spectral structure, and excitation characteristics.

Telephone communication does not transmit signals under 300 Hz because low-frequency sound components are typically deemed as redundant. Although the low-frequency components are unintelligible when presented alone, they greatly improve speech recognition in noise for normal-hearing subjects when presented through a cochlear implant simulation (Chang *et al.*, 2006). This low-frequency enhancement effect is not due to a linear addition of intelligibility between low- and high-frequency components or an increase in the physical SNR. It is suggested that an auditory-based process uses the pitch cues from low-frequency sounds to first segregate the target voice from the competing voices and then group appropriate temporal envelope cues in the target voice for robust speech recognition under realistic listening situations (Chang *et al.*, 2006).

Perceptual grouping principles

Generally speaking, the principles of perceptual grouping described by Gestalt psychologists apply to the perceptual grouping of sounds (Bregman, 1990; Moore, 2003). Wang and Brown (2006) summarize applications of the principles in ASA (Bregman, 1990) as: proximity in frequency and time, periodicity, continuous or smooth transition, onset and offset, amplitude and frequency modulation, rhythm and common spatial location. Some of them are applicable to sequential grouping, such as periodicity; some are applicable to simultaneous grouping only such as onset/offset.

It is interesting to see how to apply these principles to our study of sequential organization. In a computational system, the similarity principle requires quantifying similarity measures. In our account, this principle leads to finding the underlying features that exhibit similarity regardless of what has been said and how it has been said as long as the speech to be organized is produced by a single speaker (source). A speaker has his or her own way of producing sounds because of individual vocal organs and acquired speaking styles. Hence, in our view, the similarity principle leads to identifying and exploiting the speaker characteristics from input signals.

The principle of disjoint allocation suggests that a single component in a sound shall be regarded as originating from one source. This principle implies that sequential grouping decisions shall be exclusive in assigning segments into streams. The principle of good continuation suggests that sounds with continuous frequency trajectories shall be organized into the same stream. Since an auditory scene is typically composed of various

acoustic events, application of the continuity principle may be constrained. On the other hand, the principle of closure in audition suggests that a sound shall be perceived as continuous though it is actually partially removed or masked by another sound (Bregman, 1990). This illusion of continuity indicates that a discontinuous but smooth trajectory may be regarded as the continuation of an existing sound. The principle of common fate states that acoustic components changing in the same way at the same time belong together. This principle applies to simultaneous organization. For example, sound components with a common spacing in frequency that constitute harmonics of the same fundamental frequency are grouped together.

2.2 Computational Studies on Sequential Organization

As the first CASA model, the study by Weintraub (1985) seeks to segregate and reconstruct two simultaneous speakers. This model uses an autocorrelation function to capture periodicity as well as amplitude modulation. He then uses the coincidence function to track pitch contours of two simultaneous utterances. Sounds from different speakers are separated by using iterative spectral estimation according to pitch and temporal continuity. Weintraub's model requires that two speakers in an input mixture belong to opposite genders and they have different pitch ranges. Sequential grouping is performed by grouping separated speeches into speaker streams according to their pitch estimates.

Cooke (1993) proposes a CASA system based on a T-F representation called synchrony strands, which is obtained by evaluating local similarity and temporal

continuity of outputs from a cochlea model. In a sense, these strands are similar to our definition of T-F segments. Strands are merged into groups based on common harmonicity and common amplitude modulation. Pitch contours are then obtained for each group, and the groups with similar contours are organized into the same stream. However, this system requires input of continuously voiced speech and is not able to group speech that is separated in time.

Brown and Cooke (1994a) propose to generate discrete T-F elements based on cross-frequency correlation of filter responses and frequency transition across time. These elements are similar to Cooke's synchrony strands and our definition of segments. Elements are grouped by common periodicity and common onset and offset. Specifically, the elements with similar $F0$ contours are grouped into the same stream. Similar to the study by Cooke (1993), this system also requires continuously voiced speech as input.

Wang and Brown (1999) employ a two-layer oscillator network for speech segregation. In the first layer, segments are formed based on cross-channel correlation and temporal continuity. In the second layer, segments are grouped into two streams, one for the target and the other for its background according to a dominant pitch estimate in each time frame. Similar to the previous two studies, this system does not address sequential grouping of segments that are separated in time.

The above systems are mainly bottom-up approaches, predominantly using periodicity cues. Ellis (1996) develops a prediction-driven system which produces predictions from a world model and compares the predictions against the input. The world model comprises three types of sound elements, noise cloud, transient click, and

harmonic sound. This system is essentially a top-down process. A recent speech decoding system by Barker *et al*. (2005) implicitly addresses sequential grouping and its formulation is derived from a statistical framework of automatic speech recognition. This decoder searches for the most likely word sequence and produces a group of signal fragments that are determined to be speech. The rest of the input signal is regarded as non-speech background.

Timbre-based sequential grouping methods have been proposed for musical sound separation (Brown and Cooke, 1994b; Godsmark and Brown, 1999). The American Standards Association defines timbre as "... that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar" (American Standards Association, 1960). This definition is hard to quantify. Timbre-based methods extract features that relate to timbre such as brightness, onset asynchrony or dynamic changes of spectral shape from music. Such features represent some form of spectral energy distribution which has been widely used as cepstral features and dynamic features in speech processing (Huang *et al.*, 2001).

A recent blind source separation system (Bach and Jordan, 2004) segregates two simultaneous voices by spectral clustering of the spectrogram using auditory features such as pitch and timbre. To obtain parameters for the clustering algorithm, a training session is applied on artificial mixtures. This constrains itself to specific target-to-interference ratio (TIR) conditions and does not generalize easily. In addition, given an arbitrarily long utterance, the clustering operation on the complete T-F decomposition of the input requires a large amount of computation.

Besides the above CASA studies, there are studies in speech processing that relate to sequential grouping. Research has been carried out for decades to extract one of the speakers from cochannel speech by either enhancing target speech or suppressing interfering speech (Quatieri and Danisewicz, 1990; Morgan *et al.*, 1997). Zissman and Seward (1992) examine pitch continuity in cochannel speech and assign pitch contours to a corresponding talker by polynomial contour fitting when pitch contours from two speakers cross. Their results suggest that a method based purely on pitch information is not sufficient. Morgan *et al*. (1997) estimate the dominant pitch and then reconstruct the speech components of both stronger and weaker talker frame by frame using frequency-domain filtering according to the estimated pitch; speech signals are further enhanced by the formants estimated for the stronger talker. Afterwards, a speaker assignment algorithm using a maximum likelihood criterion is applied to group recovered signals into two speaker streams, one for the target and the other for the interferer. The assignment algorithm groups the individual frames by examining the pitch and spectral continuity for consecutive voiced frames, and comparing the spectral similarity of the onset frame of a voiced segment with recently assigned frames using a divergence measure (Carlson and Clements, 1991), which is the symmetrized Kullback-Leibler divergence (Kullback, 1968). Because of such short-time processing, the spectral comparison is biased towards the comparison of phonetic information contained in a frame instead of speaker characteristics, thus degrading the grouping performance.

Studies have also been conducted on speaker detection and tracking tasks in multi-speaker environments such as conversational speech or broadcast news (see e.g. Yu and

Gish, 1993; Dunn *et al.*, 2000). Speakers under such conditions are usually aware of each other, resulting in long single-speaker segments. Various methods, supervised or unsupervised, have been explored for speaker detection and tracking tasks. A typical method (Dunn *et al.*, 2000) exploits log-likelihood ratio scores, calculated from trained feature distribution for speakers and a universal background model, to partition a recording into homogeneous segments, which are subsequently clustered for tracking purposes. However, the segments that CASA deals with typically last 30 ms to 300 ms (Shao and Wang, 2006a), far shorter than the optimal segment length of around 2.5 *sec* and the typical minimum length of 1 *sec* for reliable speaker clustering (Dunn *et al.*, 2000). In addition, as pointed out by Lovekin *et al.* (2001), the discriminative ability to differentiate speakers is sharply reduced if a segment is shorter than 500 *ms*. Thus, even though tracking tasks are analogous to sequential grouping because they both answer the questions of who is talking and when, the former deal with long inputs with little speaker overlap while the latter faces the opposite. This distinction makes the use of the methods in speaker detection and tracking difficult to apply to sequential grouping.

2.3 Aspects of Monaural Speech Segregation

2.3.1 Multipitch tracking

As discussed in the preceding sections, pitch is an important cue for auditory organization. However, when an auditory scene comprises multiple speakers, conventional pitch estimation algorithms are unable to track pitch contours for all the

speakers simultaneously. Here, we introduce a multipitch tracking system proposed by Wu *et al*. (2003) to deal with the multi-talker condition. This system outputs pitch contours for up to two speakers for an input utterance. We will exploit pitch outputs for our sequential grouping study.

The multipitch tracking system first obtains a T-F decomposition of an input signal by passing it through a bank of Gammatone filters (Wang and Brown, 2006). The system then calculates envelopes in high-frequency channels (center frequency greater than 800 Hz) and also computes normalized correlograms (autocorrelations) for each frequency channel (Wang and Brown, 2006). Peaks of a correlogram indicate the periodicity of the signal in the corresponding frequency channel. Some peaks are inconsistent with the pitch because of pitch dynamics and the fact that harmonics are unresolved in high frequency channels. Additionally, in a noisy input the peaks do not always agree with the pitch. In order to minimize the effects introduced by these false peaks, channels deemed corrupted are removed and the peaks are further screened in the retained channels (Wu *et al.*, 2003).

The system uses a statistical method to capture the relationship between true pitch periods and the observed peaks. Specifically, a mixture of a Laplacian distribution and a uniform distribution is employed to model the distribution of time-lag differences between a true pitch period and the closest peak in a selected channel (Wu *et al.*, 2003). The distribution parameters are trained from clean speech by a maximum likelihood estimator. Thus, the system formulates a way to calculate the probability of a frequency channel supporting a pitch hypothesis. An integration method is then used to produce the

conditional probability of observing the selected peaks in all selected channels in a time frame given a hypothesized pitch period. Finally, a continuous hidden Markov model (HMM) is used to model pitch dynamics. More specifically, the HMM states represent all the possible hypotheses in a time frame and the above mixture distribution is used as the state observation density. Transitions between the HMM states represent the probabilistic pitch dynamics, and not only model pitch changes across time but also the shifts between zero-pitch, single-pitch and two-pitch spaces. These spaces correspond to pitch periods of pauses, single speaker and two speakers respectively.

Figure 2.1 shows an example of multipitch tracking results. The input is a two-talker mixture created from two female utterances. Prior pitch values are obtained from premixing utterances using the Snack toolkit (Sjolander and Beskow, 2000). It is an open source version of the popular ESPS/waves+ toolkit. Even though the two female speakers have the same pitch range, the multipitch tracking system is able to produce pitch contours that fit the true pitch values well. It is evident from the figure that in the mixture, there are portions that contain only one speaker's voiced speech, and portions that contain both speakers' voiced speech. There are also portions considered by the algorithm to contain one speaker's voiced speech but they actually contain both speakers' voiced speech. A typical reason for this mistake is that one speaker's voiced energy is much lower than that of the other.

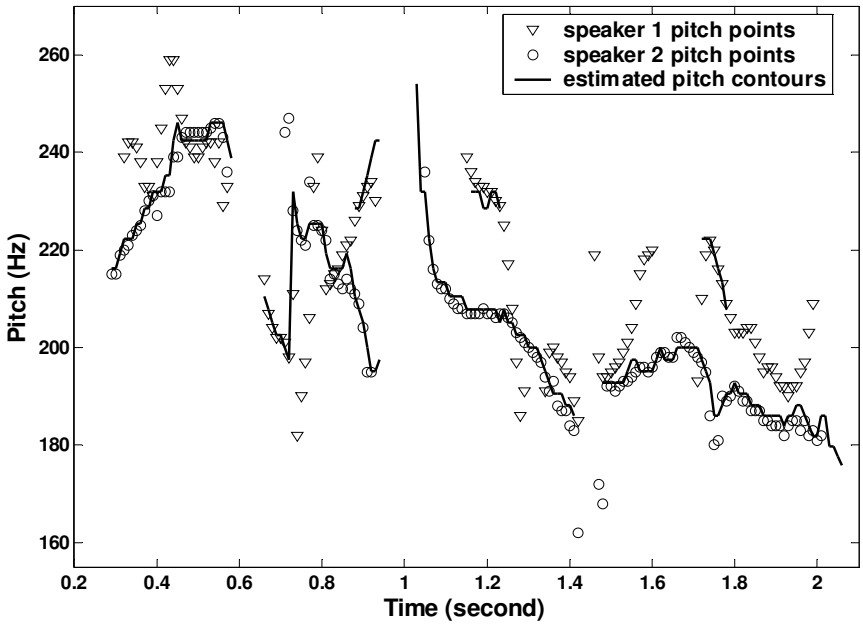

Figure 2.1: Estimated pitch contours from multipitch tracking compared with single-speaker pitch points. The solid lines represent the pitch contours obtained from a two-talker mixture using the multipitch tracking algorithm. The triangles and circles represent the pitch values obtained from the premixing utterances.

2.3.2 Binary time-frequency masks

Two-dimensional time-frequency decompositions of signals are widely used in speech processing. The well-known spectrogram and cochleagram (Wang and Brown, 2006) are good examples of such representations. Within this representation, a binary T-F mask furnishes the information about whether a T-F unit is reliable or not when the input signal is a combination of target and interference signals. The notion of an ideal binary T-

F mask has been proposed as the computational goal of CASA (Wang, 2005). The ideal binary mask is a binary matrix, defined as follows:

$$M(f,t) = \begin{cases} 1, & if\ S(f,t) > N(f,t) \\ 0, & otherwise \end{cases}. \tag{2.1}$$

$M(f,t)$ is the T-F mask indexed by frequency f and time t. $S(f,t)$ refers to energy from the target source in the frequency channel centered at f and in time frame t; $N(f,t)$ is the corresponding energy from the interference source. If a T-F unit contains stronger energy from target than interference, the corresponding mask element is labeled 1; it is assigned 0 otherwise. This implies a local SNR criterion of 0 dB. Given premixing target and interference signals, the ideal binary mask can be readily constructed.

The ideal binary mask is motivated by the human auditory masking phenomenon (Moore, 2003), and it has many desirable properties. The ideal binary mask provides the maximum SNR gain of all the binary masks (Hu and Wang, 2004; Ellis, 2006). Furthermore, such masks have been applied to robust speech recognition and shown to be highly effective as front-ends (Cooke *et al*., 2001; Roman *et al*., 2003). Besides, depending on which one is the target among multiple sources, the ideal binary mask can be constructed accordingly. For example, the binary values of a T-F unit in the mask correspond to the two underlying speakers in a cochannel mixture. If one speaker is of interest to the user, it is designated as target, and the other speaker is regarded as interference. If both speakers are desired, after selecting one as the target, the other speaker corresponds to the complement mask. Under non-speech noisy conditions, the speech signal is regarded as target, and the ideal binary mask can be defined accordingly.

2.3.3 Voiced speech segregation

Construction of the ideal binary mask requires premixing recordings of target and interference. To estimate the ideal mask from an input, we employ a pitch-based speech segregation system (Hu and Wang, 2004; Hu and Wang, 2006; Hu, 2006). This system makes minimal assumptions about the underlying noise sources and significantly improves the SNR of segregated speech under various noisy conditions. The system performs voiced segmentation and simultaneous grouping, and produces binary T-F masks of segregated simultaneous streams together with their pitch contours. A simultaneous stream refers to a group of segments that have been simultaneously organized.

The periodic nature of voiced speech provides useful cues for segmentation. For example, a harmonic usually activates a number of adjacent auditory channels because the pass-bands of adjacent Gammatone filters have significant overlaps, resulting in high cross-channel correlation. In addition, the periodic signal usually lasts for some time, within which it has good temporal continuity. Thus, the speech segregation system performs segmentation of voiced speech by merging T-F units using cross-channel correlation and temporal continuity (Hu and Wang, 2004). Specifically, neighboring T-F units with sufficiently high cross-channel correlation in a correlogram response are merged to form segments in the low frequency range. A correlogram is a periodicity representation, consisting of autocorrelations of filter responses across all the filter channels (Wang and Brown, 2006). In the high frequency range, where a Gammatone filter responds to multiple harmonics, the system merges the T-F units on the basis of

cross-channel correlation of response envelopes. Along the time dimension, temporal continuity is employed to merge neighboring units if they show high cross-channel correlations (Hu and Wang, 2004).

Since pitch is a useful cue for grouping (Bregman, 1990; Wang and Brown, 2006), the system estimates pitch contours for up to two sources for the entire utterance based on the aforementioned correlogram and use them for simultaneous grouping. T-F units are labeled according to their consistency of periodicity with the pitch estimates. Specifically, for low-frequency channels where harmonics are resolved, if a unit shows similar response at an estimated pitch period, the corresponding T-F unit is labeled consistent with the pitch estimate; it is labeled inconsistent otherwise. For high-frequency channels that respond to several harmonics, an amplitude modulation model is used to determine whether a unit response shows beating at the pitch period and thus pitch-consistent (Hu and Wang, 2006). Subsequently, a voiced segment is grouped into a stream that corresponds to the pitch period if more than half of its units are labeled consistent with the pitch estimate. The stream is further expanded by absorbing neighboring units that have the same label. The two estimation processes for pitch and simultaneous streams are repeated to improve the estimates. This iterative algorithm does not terminate until both estimates converge (Hu, 2006).

A simultaneous stream is represented by a binary mask with the T-F units labeled as foreground (target-dominant or 1) if they are consistent with the pitch estimate and others as background (interference-dominant or 0). Figure 2.2 (b) shows a collection of simultaneous streams obtained from the mixture in Figure 2.2 (a) by the Hu (2006)

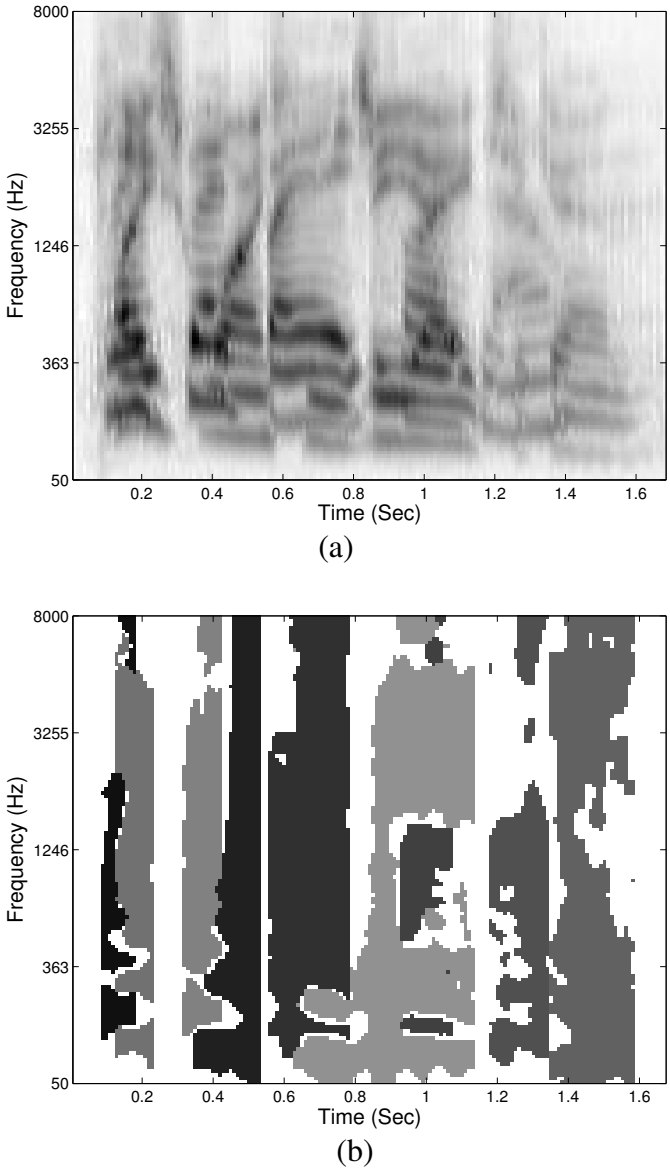

Figure 2.2: Illustrations of noisy speech and estimated simultaneous streams. Plot (a) shows a cochleagram of a two-talker utterance mixed at 0 dB SNR. Darker color indicates stronger energy within the corresponding time-frequency unit. Plot (b) presents derived simultaneous streams from utterance in (a). White color shows the background. Different gray-colored regions indicate that the streams have been grouped across frequency but not across time.

system. The background is shown in white, and the different gray regions represent different simultaneous streams. These segregated streams have been grouped across frequency, but they are still separated in time. In our sequential grouping study, we will develop methods to organize these simultaneous streams into complete streams.

CHAPTER 3

ROBUST SPEAKER RECOGNITION

Speaker recognition studies in recent years have achieved substantial improvement under clean conditions (Martin and Przybocki, 2001; Przybocki and Martin, 2004; Przybocki *et al.*, 2006). However, robust speaker recognition under noisy conditions remains a challenging problem. A speaker recognition system is typically trained on clean speech from a group of registered speakers and faces the mismatch problem when tested in the presence of interference. In this chapter, we first describe how a speaker recognition system works and its robustness issues. Then, we present our system that recognizes a target speaker in the presence of another speaker. This system extracts minimally corrupted speech segments termed *usable speech* and significantly improves recognition performance. Unlike a T-F segment introduced in CASA, a usable speech segment is composed of contiguous time frames of speech that are deemed speaker homogeneous. The former is defined in terms of both time and frequency while the latter only in time. We further improve the recognition performance by employing binary T-F masks that are generated by voiced speech segregation. Specifically, a missing-data

method recognizes speakers based on information of reliable (clean) or unreliable (noisy) T-F units as indicated by the masks. In the last section, we propose a general solution for robust recognition under noisy conditions. This solution extracts novel speaker features through auditory filtering and cepstral analysis, and employs an uncertainty decoder to account for errors from front-end processing. This system substantially improves recognition performance compared to conventional speaker features as well as a state-of-the-art robust feature.

3.1 Speaker Recognition

There are a large number of applications for speaker recognition in both military and civilian areas, such as security control, identity authentication, and forensic applications. Due to these applications, research on automatic speaker recognition has been conducted for more than four decades (Furui, 2005). Generally speaking, speaker recognition utilizes human voices for classification or verification of speaker identities. According to how a recognition decision is made, speaker recognition can be categorized into two tasks: speaker identification (SID) and speaker verification (SV). SID seeks to determine which speaker produces an input utterance among a group of registered speakers. Unlike SID, SV makes a binary decision that either accepts or rejects an identity claim of an input utterance. In other words, this task verifies whether a claimed speaker actually utters the input. SID is typically a closed-set task while SV deals with open-set conditions. Additionally, a speaker recognition system is regarded as text-dependent if it knows linguistic content prior to recognition, and it is considered text-independent if otherwise.

The latter is more challenging than the former (Furui, 2005) since the system not only has to differentiate speakers but also faces linguistic variations in speech. In this dissertation, we deal with text-independent conditions.

A speaker recognition system typically comprises three processing stages (Naik, 1990; Furui, 1994; Campbell, 1997; Furui, 2001; Bimbot *et al.*, 2004). The first stage extracts features that characterize speakers from input signals. Then at the scoring stage, these features are compared with registered speaker templates or models to calculate a measure of similarity between the input and the model. Finally, the decision stage produces either a speaker identity or a binary output of acceptance/rejection based on the obtained similarity measures.

Since the system relies on speaker characteristics for classification, the derived speaker features play an important role in the recognition process. Acoustic features such as pitch and various versions of cepstral coefficients have been explored (Atal, 1972; Furui, 1989; Naik, 1990; Furui, 1991; Campbell, 1997; Furui, 2001; Bimbot *et al.*, 2004). Widely used features include Mel-frequency cepstral coefficients (MFCC), perceptual linear predictive (PLP) coefficients. Additionally, cepstral mean normalization (CMN) has been applied to enhance features by removing distortions from telephone transmission (Furui, 1981). Robust features such as RASTA (Hermansky and Morgan, 1994) have also been investigated for speaker recognition (Reynolds, 1994). These features are designed to capture short-term low-level information about human excitation or vocal tract shape, and they are also widely used in automatic speech recognition (Huang *et al.*, 2001). On the other hand, high-level features such as word idiolect,

prosody, etc., have recently been explored to complement conventional speaker features (Reynolds *et al.*, 2003).

There are many speaker modeling approaches (Furui, 2005). Among them, Gaussian mixture model (GMM) is predominantly employed to simulate the distribution of speaker features. Parameters of a GMM are usually trained using an EM algorithm (Reynolds, 1995). As an alternative approach, models can be adapted from a speaker-independent background model using Bayesian adaptation (Reynolds, 1997). Modeling methods such as polynomial based classifiers (Campbell *et al.*, 2002; Wan and Renals, 2002) and decision tree classifier (Foo and Lim, 2002), are found to yield comparable recognition performance as GMM. In text-dependent tasks, template matching (Furui, 1981) and hidden Markov model (HMM) have been applied to speaker recognition (Naik *et al.*, 1989; Rosenberg *et al.*, 1990; Savic and Gupta, 1990; Furui, 2005). Artificial neural networks (ANN) have also been employed for speaker recognition (Yegnanarayana *et al.*, 2001; Baker and Sridharan, 2006).

3.1.1 Decision framework

In a statistical framework, a speaker identification task can be derived as maximum-likelihood classification (Reynolds, 1995) using Bayesian analysis (Rice, 1995). Assuming that a set of speakers are registered as $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$ and that a sequence of M feature frames $O = \{X_1, X_2, \ldots, X_M\}$ has been observed from an input utterance, the optimal goal of SID is to find a speaker $\hat{\lambda}$ that maximizes the posterior probability of a model given the observations. Mathematically, the SID decision rule is

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} P(\lambda \mid O). \tag{3.1}$$

Applying the Bayesian rule, we have

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} \frac{P(O \mid \lambda) P(\lambda)}{P(O)}. \tag{3.2}$$

Without prior knowledge about test utterances, the prior probabilities of speakers are typically assumed to be uniformly distributed. In addition, the maximization over λ is not affected by $P(O)$. Therefore, $P(\lambda)$ and $P(O)$ are dropped from the equation, resulting in a decision based on likelihood maximization. Assuming observation independence and taking a log transformation, (3.2) can be written as

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} P(O \mid \lambda) = \arg\max_{\lambda \in \Lambda} \sum_{m=1}^{M} \log p(X_m \mid \lambda) . \tag{3.3}$$

Here m is a time frame index of the feature sequence. The right-hand-side (RHS) of this equation contains summation of log likelihood scores of a feature frame given a speaker model. This score represents a similarity measure between the input and the registered speakers.

A speaker verification task determines whether the input O is produced by a claimed speaker λ_c. We can evaluate how similar the input is to the claimant using a log likelihood normalization as

$$\log L(O) = \log P(O \mid \lambda_c) - \log P(O \mid \lambda \neq \lambda_c) . \tag{3.4}$$

Here, $L(O)$ refers to a verification score and is calculated as the likelihood ratio of the conditional probability of the feature sequence given the claimant model to the conditional probability of the features given an imposter model. A positive value of (3.4)

indicates a valid claim, thus an output of acceptance, while a negative value indicates otherwise. This output entails two types of decision errors, a *miss* detection of a target speaker or a *false alarm* of an imposter. Instead of the fixed value of 0, a threshold can be applied on (3.4) and adjusted to artificially suppress one error type while increasing the other. Therefore, SV results are usually reported using receiver operating characteristic (ROC) curves (Przybocki *et al.*, 2006). From a statistical analysis point of view, (3.4) is actually a form of the generalized hypothesis test (Rice, 1995). In other words, this equation performs a test to select either the hypothesis that the claimant utters the input utterance or the hypothesis that some other speaker does.

However, (3.4) can not be literally implemented because it is impossible for a system to enumerate all the imposters in the world given a claimant. One approach to simulate the imposter model is to construct a cohort set that is composed of speakers that are acoustically close to the claimant (Furui, 2001). This is based on the assumption that the density of a specific observation for speakers other than the claimant is dominated by the density for the nearest speakers. The other approach constructs a universal background model (UBM) that incorporates all the speakers in the world (Reynolds, 1997),

$$\log L(O) = \log P(O \mid \lambda_c) - \log P(O \mid \lambda_{UBM}) . \tag{3.5}$$

In a specific task, UBM is usually trained on the pooled training data that excludes the claimant or a set of potential claimants in the corpus.

A set of normalization methods have recently been developed to reduce the distribution variances of claimant scores and imposter scores (Bimbot *et al.*, 2004). Likelihood scores are normalized by subtracting their mean and then dividing their

standard deviation. Both quantities are estimated from pseudo-imposter distributions since it is much easier to obtain imposter scores than the claimant scores. Different normalization methods (Bimbot *et al.*, 2004) have been proposed depending on how this imposter distribution is obtained.

3.1.2 Robust speaker recognition

The National Institute of Standards and Technology (NIST) has coordinated a series of annual speaker recognition evaluation since 1996 (Martin and Przybocki, 2001; Przybocki and Martin, 2004; Przybocki *et al.*, 2006). This series mainly evaluate speaker verification systems. The major challenge in the evaluations is the convolutive distortions (Huang *et al.*, 2001) introduced by different telephone handsets and transmission channels, ranging from landline to wireless. The mismatch between training and testing conditions leads to performance degradation. Robust recognition methods such as cepstral mean normalization, feature variance normalization, feature warping, model adaptation, score normalizations, etc., have been proposed to account for the channel distortions and improve the recognition performance (Furui, 2005).

Apart from channel distortions, speaker recognition also faces additive noise from channel or background. Modern telephones and microphones can reduce noise and thus improve SNR of the target signal. However, unlike a quiet office environment, an auditory scene typically comprises multiple sound sources in the background, such as car and human voice, especially during the use of mobile phones. A study has shown that speaker recognition performance degrades sharply with input SNR below 30 dB

(Yantorno, 1999). The degradation also occurs when noise source is a competing speaker (Lovekin *et al.*, 2001; Shao and Wang, 2003). To tackle this robustness problem, speech enhancement methods that are widely used in speech recognition, such as spectral subtraction, subband modeling, have been explored for robust speaker recognition (Barger and Sridharan, 1997; Ortega-Garcia and Gonzalez-Rodriguez, 1997; Drygajlo and El-Maliki, 1998; Sivakumaran and Ariyaeeinia, 2000; Yoshida *et al.*, 2001), but they are ineffective when noise is nonstationary (Shao and Wang, 2006b). RASTA filtering (Hermansky and Morgan, 1994) and CMN have also been widely used but they are mainly intended for convolutive noise. Nevertheless, recent studies of robust speech recognition on Aurora (Parihar and Picone, 2003) have yielded an advanced front-end feature extraction algorithm (AFE) (STQ-AURORA, 2005-11), standardized by the European Telecommunication Standards Institute (ETSI). ETSI AFE derives robust MFCC features using a set of state-of-the-art front-end processes, including Wiener filtering.

An alternative approach to feature enhancement seeks to model the noise and combines it with the clean speaker models (Rose *et al.*, 1994; Matsui and Furui, 1996; Wong and Russell, 2001; Gong, 2002; Yoma and Villar, 2002). However, such systems are limited when applied to novel interference types because they rely heavily on the use of *a priori* information of noise sources.

On the other hand, similar to speech recognition, humans are found to perform better than machines in speaker recognition tasks (Schmidt-Nielsen and Crystal, 1998). Human performance is comparable with that of the best computer system in the matched handset

condition. When different handsets are used, human performance degrades much less than computer performance. Additionally, humans are more robust when input signals are corrupted by noise in the background such as crosstalk and poor channel conditions. The superiority of the auditory system motivates us to explore computational auditory scene analysis for robust speaker recognition.

3.2 Usable Speech for Cochannel Speaker Recognition

3.2.1 Cochannel speaker identification

One type of noisy condition is cochannel speech where an input signal is a combination of speech utterances from two talkers. This condition usually occurs when two speech signals are transmitted over a single communication channel, or when two speakers are talking at the same time and they are unaware of each other. Cochannel speech is more challenging than telephone speech such as that used in the NIST evaluations because the former has a much higher proportion of speech overlap while the latter is conversation speech with relatively little speech overlap.

Research has been carried out for decades aiming to extract one of the speakers from cochannel speech by enhancing target speech or suppressing interfering speech. However, in automatic speaker recognition, as pointed out by Lovekin *et al.* (2001), the intelligibility and quality of extracted speech are not important. What the system needs are portions of the speech that contain speaker characteristics unique to an individual speaker, classifiable and long enough for the system to make identification or verification

decisions. These portions of speech, or segments, are defined as consecutive frames of speech that are minimally corrupted by interfering speech, and are thus called usable speech (Lovekin *et al.*, 2001).

Previous studies find that voiced segments contain most of the information for SID, and different criteria such as frame-level TIR or spectral autocorrelation ratio have been developed to extract usable speech in cochannel speech (Krishnamachari *et al.*, 2000; Lovekin *et al.*, 2001). According to these criteria, a significant amount of cochannel speech can be considered usable for SID. Frame TIRs are easily calculated with premixing speech utterances, and usable speech extracted based on a TIR threshold retains frames where target speaker is much stronger in terms of overall energy than the other. Spectral autocorrelation ratio estimates the ratio between dominant peak and valley in autocorrelation of a spectral frame. This ratio is used to determine whether a frame is usable, meaning the spectrum is well structured (single-speaker speech), or not. This approach is simple and effective and shows a substantial improvement in SID performance. However, the authors use *a priori* frame TIRs and they are hard to estimate from mixture directly. A further study explores a maximum likelihood decision in an attempt to determine the speakers that generate usable speech segments (Smolenski *et al.*, 2002).

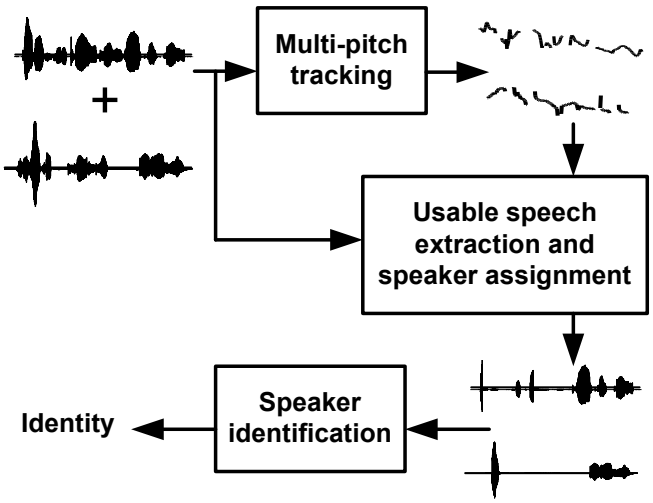

Figure 3.1: Diagram of usable speech extraction and speaker identification. First, pitch tracks are obtained using a multi-pitch tracking algorithm. Then usable speech segments are extracted and assigned accordingly. Finally, speaker identity is determined using a speaker identification method.

We propose a novel method to extract usable speech for speaker identification. This approach is based on a robust multipitch tracking algorithm (Wu *et al.*, 2003) that estimates pitch contours of up to two speakers (see Section 2.3.1). As shown in Figure 3.1, the proposed method is composed of three stages. First, the multipitch tracking algorithm is employed to produce pitch contours from a cochannel input. Then, the usable speech extraction method removes the segments with concurrent pitch contours. Silence and unvoiced segments are removed as well. Thus, only the segments that have single pitch contours are retained. Subsequently, the usable segments are assigned to two

speaker groups, corresponding to the two speakers in the mixture. Finally, speakers are identified using the assigned segments based on Equation (3.3).

3.2.2 Usable speech extraction and assignment

The multipitch tracking system outputs multiple pitch tracks and Figure 2.1 presents an illustration of the outputs. It is evident from the figure that in a cochannel mixture, there are portions that contain only one speaker's voiced speech, and portions that contain both speakers' voiced speeches. There are also portions considered by the multipitch tracking algorithm to contain one speaker's voiced speech but they actually contain both speakers' voiced speeches. A typical reason for this mistake is that one speaker's voiced energy is much lower than that of the other. This kind of mistake, however, is rather benign as far as usable speech extraction is concerned.

Usable speech extraction seeks to determine segments or sequences of frames that comprise a single speaker's voice, and are thus usable for SID. Segments with concurrent pitch contours are not usable for SID because both talkers have strong energy in the segments. This distorts the acoustic features used in SID. More specifically, the harmonics and formants from two talkers are added together in the power spectrum, which ruins the second frequency analysis for deriving cepstral features (Huang *et al.*, 2001). Speech enhancement methods such as spectral subtraction (Berouti *et al.*, 1979) are not applicable here because the interfering speech is highly nonstationary. Therefore, we remove segments with concurrent pitch contours from cochannel speech.

For the segments with a single pitch track, the competing voice is either silent or others' unvoiced speech. In the former case, the power spectrum is intact; in the latter case, the power spectrum is minimally contaminated because unvoiced speech is usually much weaker than voiced speech. Thus, we consider the segments with a single pitch track as usable speech. The remaining signals are considered unusable and removed. To ensure the homogeneity of a usable speech segment, if neighboring pitch values changes as much as 10 Hz, the segment is split into two shorter segments.

In cochannel speech, either speaker can randomly appear as the stronger speaker or the weaker one at a time. Hence, the extracted segments are separated in time and need to be sequentially organized into speaker streams for SID. Here, we leave the discussion of sequential grouping to Chapter 4 and assume that segments are ideally grouped into streams based on *a priori* pitch information. A segment is grouped if the majority of its frames contain pitch estimates that are consistent with the *a priori* pitch values.

3.2.3 Evaluations

As in Lovekin *et al*. (2001), we employ the evaluation data from the TIMIT speech corpus. The speaker set is composed of 38 speakers from the "DR1" dialect region, 14 of which are female and the rest are male. Each speaker has 10 utterance files, ranging from about 1.5 *sec* to 6.2 *sec* in length. For each speaker, 5 out of 10 files are used for training and the remaining 5 files are used to create cochannel mixtures for testing. For each speaker deemed as the target speaker, 1 out of 5 test files is randomly selected and mixed with randomly selected files of every other speaker, which are regarded as interfering

utterances. For each pair the TIR is calculated as the energy ratio of the target speech over the interference speech,

$$TIR = 10\log_{10}\left(\sum_n s_T^2(n) \Big/ \sum_n s_I^2(n)\right), \tag{3.6}$$

in which s_T and s_I are the speech samples of target and interference speakers in the time domain. The interference utterance is either cropped or concatenated with itself to match the length of the target utterance. Speech signals are scaled to create the mixtures at different TIRs: -20 dB, -10 dB, -5 dB, 0 dB, 5 dB, 10 dB and 20 dB. For example, 0 dB TIR means that the overall energy of target is equal to that of interference. Thus, for each TIR, a total of 1406 cochannel mixture files are created for the testing purpose.

SID is performed based on (3.3). We employ the widely used 12 MFCCs and their first-order dynamic coefficients as speaker features. The resulting feature vector contains 24 coefficient components. Speakers are modeled using 16-mixture GMMs, which are trained using the EM algorithm (Reynolds, 1995) from the training samples.

To demonstrate the usefulness of our extraction method, usable speech is recognized after the segments are ideally assigned into streams. If the target speaker is of interest, then the speech signal from the other speaker is considered noise. Here, we choose the target speaker SID as our evaluation criterion. Figure 3.2 gives the target speaker recognition accuracy. As a baseline, a conventional SID system is applied to the cochannel speech to recognize the target speaker. The baseline performance documents the top two identified speakers. The accuracy degrades sharply when TIR decreases because the target speech is increasingly corrupted. Comparable results are obtained by

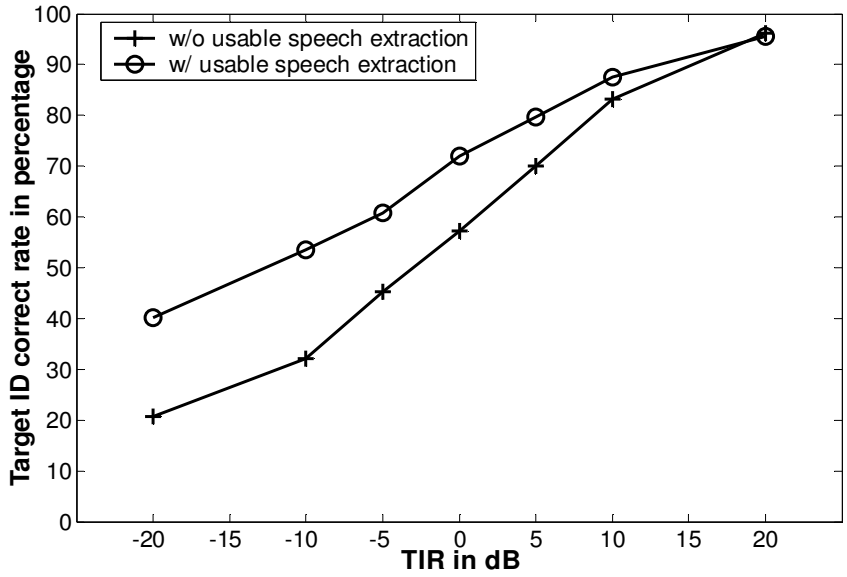

Figure 3.2: Target SID accuracy before and after usable speech extraction. SID is considered correct when the target speaker is identified from cochannel speech. Sequential grouping is performed using prior pitch information.

Yantorno *et al.* (2001) that seeks to understand how cochannel speech impacts SID performance.

The first observation from the figure is that, under cochannel situations, usable speech extraction significantly improves SID performances; the average improvement is about 12% absolute. Secondly, the improvements are consistent across all TIR levels. Performance improvement decreases at higher TIRs because target speaker dominates the mixture. However, target speaker is dominated by interference at lower TIRs, resulting in better performance after usable speech extraction.

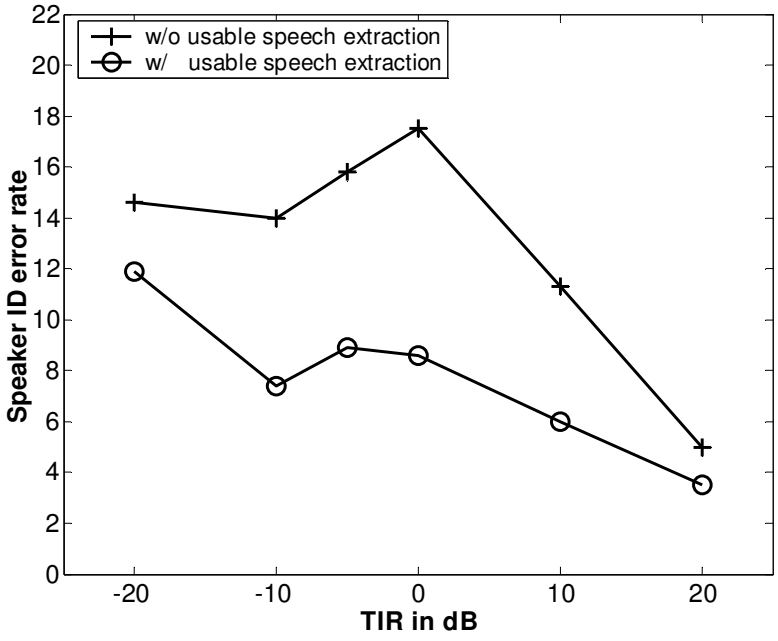

Figure 3.3: SID error rate before and after usable speech extraction. SID is regarded correct when cochannel speech is identified as either target or interfering speaker of a cochannel mixture.

When either of the speakers in a cochannel mixture is of interest, we choose a different performance criterion as shown in Figure 3.3. Here a test is regarded correct if the input is identified as either the target or the interferer in the mixture. From the figure, it can be observed that usable speech substantially improves the identification performance. At 0 dB TIR, the error rate is almost cut in half. In addition, performance improvements occur across all TIR mixture levels. One might expect the result curves to be symmetrical around 0 dB because of the evaluation criterion. The asymmetry is due to the fact that interference signals are scaled to create cochannel mixtures.

3.3 Binary Time-Frequency Masks for Robust Speaker Recognition

A usable speech segment comprises a sequence of frames which are deemed to be speaker homogeneous. However, these frames may contain some speech energy from the other speaker because of the strong overlap in the cochannel condition. Instead of extracting usable speech at the frame level, it is desirable to identify usable speech at the level of time-frequency (T-F) units so that recognition performance could be further improved. Here, we employ a binary T-F mask representation for such purposes. This mask labels a T-F unit reliable when it contains more energy from target than interference, and labels unreliable if otherwise (see also Section 2.3.2). Since the ideal binary mask provides the maximum SNR gain of all the binary masks (Hu and Wang, 2004; Ellis, 2006), we also evaluate the ideal binary mask for robust speaker recognition as a performance upper-bound.

Spectral subtraction has been proposed for binary mask estimation (Drygajlo and El-Maliki, 1998; Drygajlo and El-Maliki, 2001). This method works well when noise is stationary, but its performance degrades sharply under nonstationary noise conditions. We employ the monaural speech segregation system (Hu and Wang, 2004) as described in Section 2.3.3 to obtain the mask. This system produces an estimate of the ideal binary T-F mask without making assumptions of the underlying noise source. Our evaluations also compare with those using binary masks estimated by spectral subtraction. The evaluations are conducted when speech is corrupted by cocktail party noise or rock music.

3.3.1 Missing-data recognition

To utilize the binary masks for robust speaker recognition, we employ a missing data method (Drygajlo and El-Maliki, 1998; Drygajlo and El-Maliki, 2001). The basic idea is to treat the noise-dominant (unreliable) T-F units as missing data during recognition. In a typical speaker recognition system, the probability distribution of an extracted feature vector X, produced by a speaker λ, is modeled as a GMM (Reynolds, 1995). GMM is a weighted linear combination of K unimodal multivariate Gaussian densities, usually parameterized with diagonal covariance matrices (Cooke *et al.*, 2001). Given a binary mask showing whether a feature component is reliable or missing, the feature vector can be split into reliable components X_r or unreliable ones X_u and its probability distribution becomes,

$$p(X \mid \lambda) = \sum_{k=1}^{K} w_k \prod_{X_i \in X_r} p(X_i \mid \mu_{i,k}, \sigma_{i,k}^2) \prod_{X_j \in X_u} p(X_j \mid \mu_{j,k}, \sigma_{j,k}^2). \tag{3.7}$$

w_k is the weight of the kth Gaussian mixture. X_i and X_j refer to a reliable and unreliable feature component in X respectively; μ_k and σ_k^2 are their corresponding means and variances in the kth Gaussian mixture.

The first likelihood term on the right-hand-side of (3.7) can be easily obtained from training since the features are considered reliable (clean). However, the second likelihood term is hard to compute because the feature component is regarded as missing and its distribution is unknown. There are two methods to deal with this unknown distribution, marginalization and imputation (Cooke *et al.*, 2001). The latter seeks to impute the

missing features and replace them with estimates. The former reduces the distribution by integrating over the missing components. Imputation increases computational complexity but does not necessarily produce better verification results (Drygajlo and El-Maliki, 2001). Thus, we use marginalization and compute the overall likelihood as,

$$p(X \mid \lambda) = \sum_{k=1}^{K} w_k \prod_{X_i \in X_r} p(X_i \mid \mu_{i,k}, \sigma_{i,k}^2). \tag{3.8}$$

The likelihood of a noisy utterance given a specific speaker model is computed as the likelihood product of feature vectors of individual frames. For a SID task, the speaker model that gives the maximum likelihood value is selected as the identified speaker. For SV tasks, we use a universal background model (UBM) for score normalization. Specifically, corresponding dimensions of the UBM distribution are also marginalized to calculate the log likelihood ratio in (3.5).

3.3.2 SID evaluation under cochannel conditions

This experiment demonstrates the SID performance using the missing-data method and the binary masks when the noise source is a speaker. To have a consistent comparison with previous studies (Lovekin *et al.*, 2001; Shao and Wang, 2003; Shao and Wang, 2006a), we use the same cochannel corpus as described in Section 3.2.3. The training data for a speaker has an average of 10 *sec* of clean speech, and 16-mixture GMMs are trained using the EM algorithm (Reynolds, 1995).

Figure 3.4 presents the results of this experiment. As a baseline, we extract 12 MFCCs and their first-order derivatives as the feature vector. To compare with usable

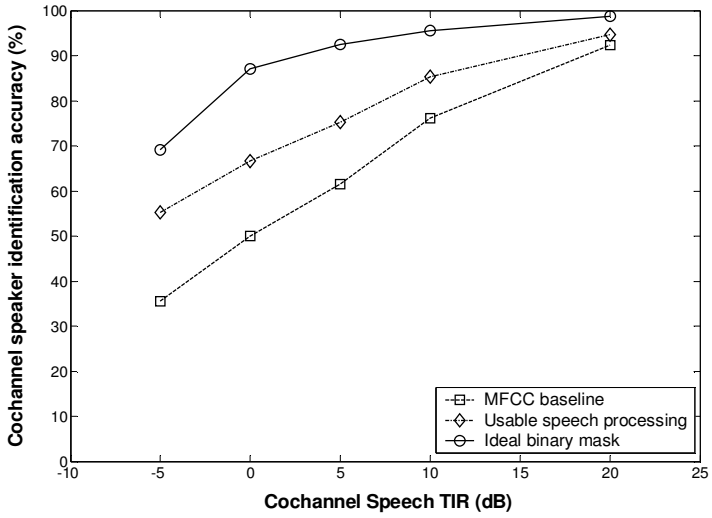

Figure 3.4: Speaker identification performance under cochannel conditions. The square line shows the performance when MFCCs are used. The diamond line shows the results of extracted usable speech segments after they are *a priori* assigned. The circle line gives performance achieved by the ideal binary mask using the missing data method.

speech processing, we apply the usable speech extraction method and ideally assign the extracted segments into the target stream using *a priori* pitch information as described in the preceding section. The same type of MFCCs is derived and identification is performed on the target stream. To evaluate the binary masks, we implement the missing data recognizer with 255-coefficient DFT feature vectors. Specifically, vectors are extracted from the log-compressed power spectrum of 20 ms frames with 10 ms overlap. The frames are extracted by applying a running Hamming window.

It can be observed from the figure that the ideal binary mask performs significantly better than the usable speech method, which in turn is much better than the baseline

performance. This is to be expected since the ideal binary mask provides reliable/unreliable information at a finer level than usable speech, and the T-F redundancy facilitates identification when features are partially missing. Evaluation of the estimated binary mask is not performed in this task because it is hard to determine target pitch contours for the speech segregation system under cochannel conditions.

3.3.3 SID evaluation under non-speech noisy conditions

In this experiment, we demonstrate the effectiveness of binary masks in adverse environments when the intrusion source is not a speaker. Two types of noise are selected from a noise database collected by Cooke (1993): cocktail party noise and rock music. The spectrogram illustrations of these two types of noise are presented in Figure 3.5. Both are wide-band and non-stationary, containing significant energy below 2 kHz. It is also observed that both noise types have some harmonic structure because the cocktail party noise contains speech-like sounds and the rock music contains musical instruments. The noisy speech utterances are simulated by mixing all the test files with the selected noises at -5 dB, 0 dB, 5 dB, 10 dB and 20 dB SNRs. For each pair of SNR and noise, 190 mixtures are created for testing.

Figure 3.6 shows the SID results for both noise conditions at various SNRs. The baseline system uses MFCCs and their first-order derivatives. CMN is applied for robustness. We also employ spectral subtraction to estimate binary mask for the missing data recognizer as proposed by Drygajlo and El-Maliki (1998). Specifically, the average

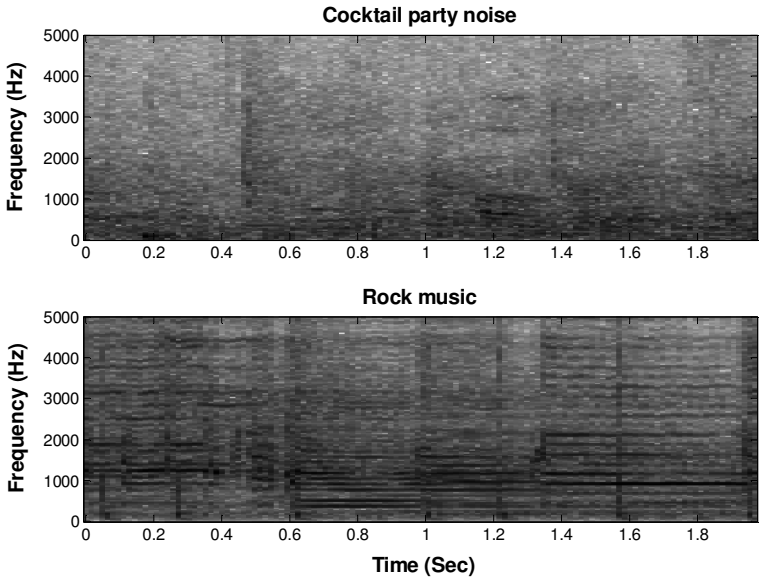

Figure 3.5: Spectrograms of cocktail party noise and rock music, selected from the noise database collected by Cooke (1993).

noise spectrum is estimated from the initial 10 frames of the mixture, and subtracted from each subsequent mixture spectrum. If the resulting component is greater than the noise estimate in energy, the corresponding mask element is labeled 1 and 0 otherwise. The implied 0 dB SNR criterion is preferred over the negative energy criterion because it produces better results (Cooke *et al.*, 2001).

To estimate the ideal binary mask, target pitch contours are determined by applying the widely-used Praat toolkit (Boersma, 2001) on the noisy speech. Note that an estimated mask is obtained using the auditory filterbank that models human's auditory response and it has large overlaps between neighboring filters. Directly using the filterbank energy gives SID accuracy of 94.2% on clean speech, which is significantly

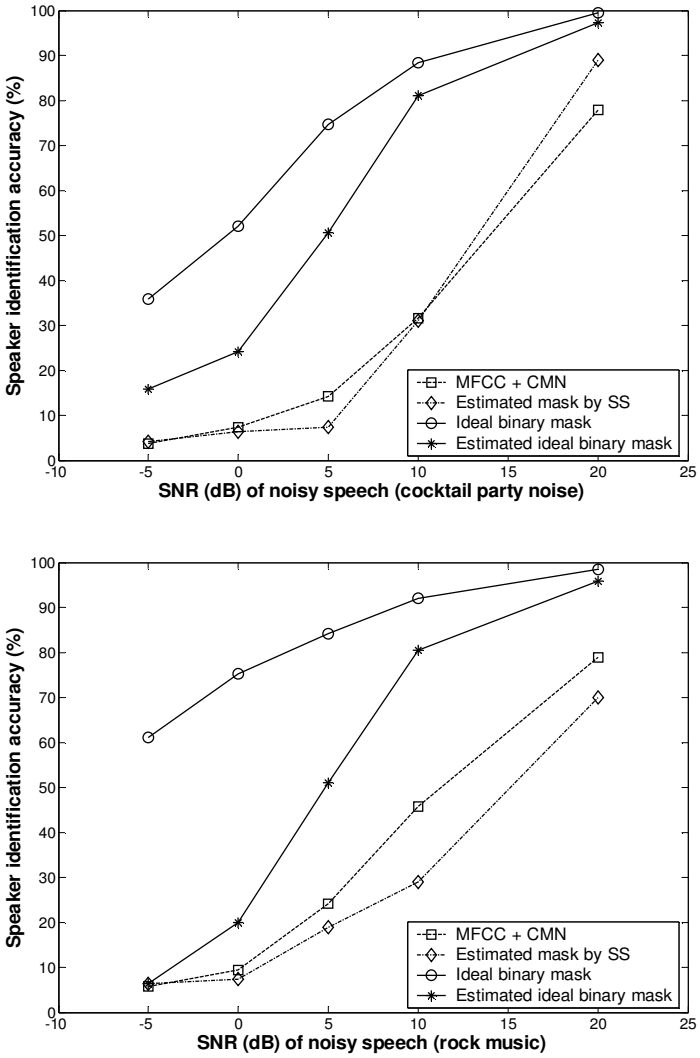

Figure 3.6: Speaker identification performance under noisy conditions. The top plot shows the results for cocktail party noise, and the bottom one for rock music. The square line represents baseline results of the GMM recognizer using cepstral mean normalized (CMN) MFCCs. The diamond line shows the missing data recognition results using binary masks estimated by spectral subtraction (SS). The circle line gives performance achieved by the ideal binary mask. The star line shows the results of the estimated ideal binary mask.

lower than that using the DFT coefficients, 99.5%. Thus, we transform the estimated mask from Gammatone frequency bands into DFT domain by labeling the corresponding frequency bins. Subsequently, the same missing data recognizer is used as in the previous experiment.

It can be observed from the figure that the estimated binary mask performs significantly better than the baseline system using MFCC-CMN. As both noises are non-stationary, spectral subtraction is unable to provide a good mask estimate, and its performance degrades sharply with decreasing SNR. The ideal binary mask produces best performance. The performance gap between the ideal binary mask and estimated mask leaves much room for improvement by adopting more effective mask estimation approaches.

3.3.4 Speaker verification evaluations

In a similar configuration as the preceding experiment, we evaluate binary masks for speaker verification tasks. Here, only the mixtures with the cocktail-party noise are tested on the 38-speaker set. One mixture file contributes 1 true score for the target speaker and 37 imposter scores for the other speakers in the set. For each SNR, there are 190 true scores and 7030 imposter scores. The scores are normalized using a UBM of 4096 mixtures, which is trained from the entire TIMIT training set, excluding the above 38 speakers.

Evaluation results are reported in Figure 3.7 using the decision error tradeoff (DET) curves provided by NIST (Martin *et al.*, 1997). DET curve is a variant of the widely used

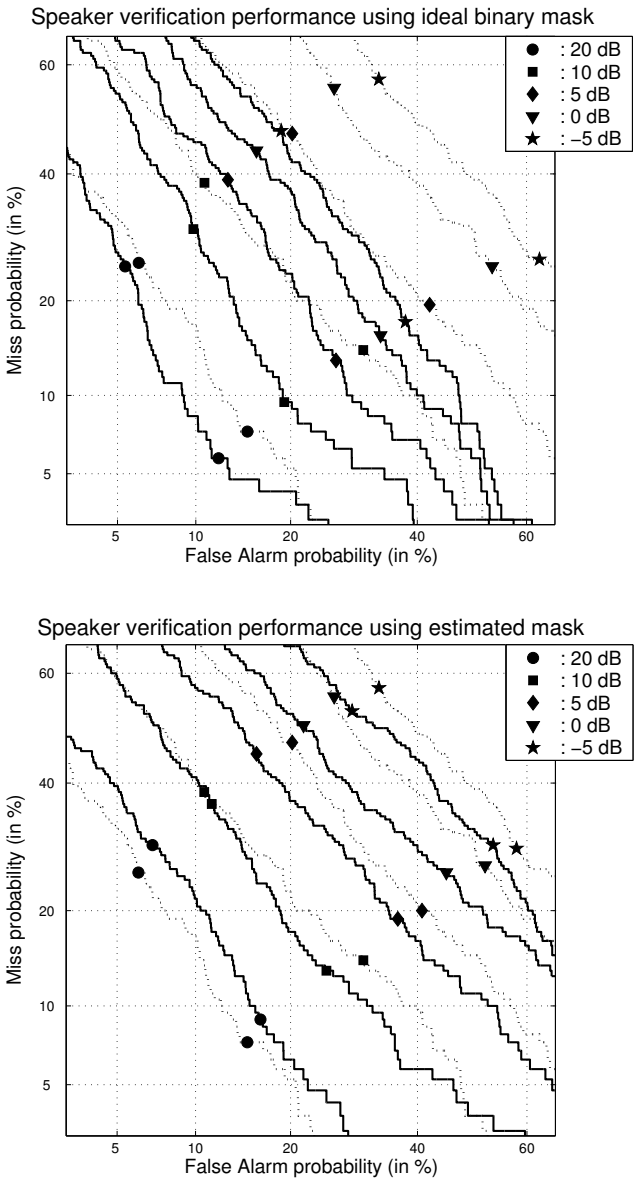

Figure 3.7: Speaker verification performance under cocktail party noise. The top plot shows the results for the ideal binary mask, plotted in solid curves against MFCC baseline in dotted curves. The bottom one shows performance of the estimated binary mask in solid curves against the same baseline in dotted curves.

ROC curve. Unlike the latter, DET plots the two error rates on the *x* and *y* axes on a normal deviate scale. We adopt this metric because NIST provides a DET toolkit for the annual speaker recognition evaluations (Martin and Przybocki, 2001; Przybocki and Martin, 2004; Przybocki *et al.*, 2006). This tool standardizes the performance evaluations by taking verification scores as input and producing DET plots as output. The ideal binary mask yields substantial performance gains over the baseline in the entire range of SNR levels. The estimated mask achieves significant improvement from 10 dB to -5 dB. It under performs only at the 20 dB condition, largely due to the segregation strategy that attempts to reconstruct the target signal by grouping harmonic components. Consequently, inharmonic target components are removed even when interference is very weak.

3.4 A Complete CASA-based Speaker Recognition System

In this section, we present a complete CASA-based robust speaker recognition system. We first propose two novel speaker features based on an auditory periphery model (Patterson *et al.*, 1992). We find that these features achieve comparable SID performance to ETSI-AFE features under both clean and noisy conditions. To account for the deviations of noisy features from clean ones, we employ an uncertainty decoder (Srinivasan and Wang, 2007) that is based on binary T-F masks estimated by a speech segregation system (Hu, 2006), which has been briefly described in Section 2.3.3. Our system achieves substantial improvement over ETSI-AFE features in a wide range of SNR conditions. Conceptually, our system improves noise robustness in two stages of a

speaker identification system; novel robust auditory features in the feature extraction stage, and feature uncertainty estimation and decoding in the scoring stage.

Figure 3.8 presents a diagram of the overall system. Input speech is decomposed using a Gammatone filterbank (Wang and Brown, 2006) to generate a time sequence of auditory features. In addition, we feed the input signal to the speech segregation system (Hu, 2006) that iteratively estimates the ideal binary mask and produces better estimates than earlier systems (Hu and Wang 2004, Hu and Wang 2006). A T-F unit of this mask indicates whether the corresponding Gammatone feature component is reliable or corrupted within a time frame. The corrupted components within a frame vector are then reconstructed using a speech prior (Raj *et al.*, 2004) and the reconstruction uncertainties are also estimated. Subsequently, enhanced Gammatone features and their uncertainties are transformed into the cepstral domain by a discrete cosine transform (DCT) (Oppenheim *et al.*, 1999). Finally, an uncertainty decoder (Deng *et al.*, 2005) identifies speaker using the enhanced cepstral features and the transformed uncertainty estimates. Note that this system can be directly generalized to speaker verification tasks by calculating the UBM likelihood using uncertainty decoding and applying the verification decision rule of (3.5).

3.4.1 Auditory feature extraction

Our system first models auditory filtering by decomposing an input signal into the time-frequency domain using a bank of Gammatone filters (Wang and Brown, 2006). Gammatone filters are derived from psychophysical observations of the auditory

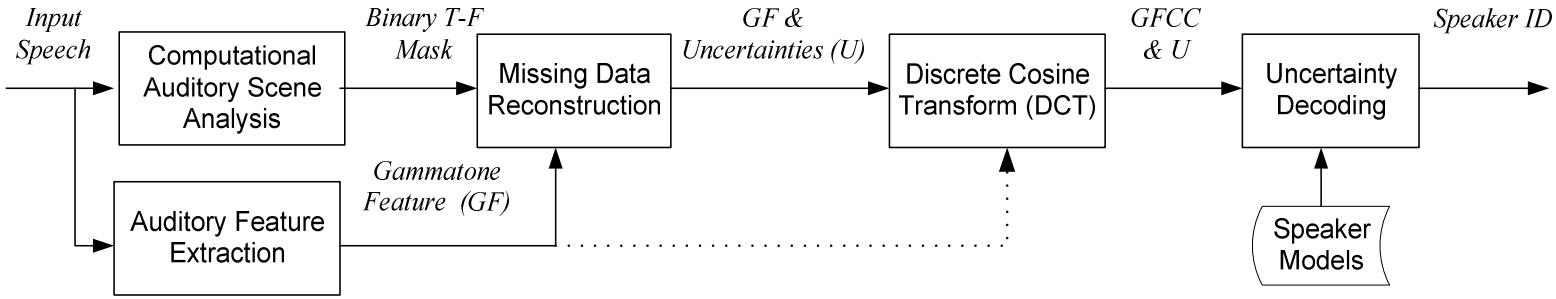

Figure 3.8: Schematic diagram of a complete CASA-based speaker identification system. Input speech is passed through a computational auditory scene analysis system to produce a binary time-frequency (T-F) mask. Then, extracted Gammatone features (GF) are used in conjunction with the binary mask to reconstruct missing T-F units from a speech prior. GF uncertainties are also estimated in the reconstruction process. GFs and their uncertainties are then transformed into "cepstrum" by the discrete cosine transform (DCT). Finally, uncertainty decoding searches for the best-matched speaker model given the resulting Gammatone frequency cepstral coefficients (GFCC) and uncertainties. The dotted path denotes how GFCCs are extracted from clean speech for the purpose of speaker model training.

periphery and this filterbank is a standard model of cochlear filtering (Patterson *et al.*, 1992). The impulse response of a Gammatone filter centered at frequency *f* is:

$$g(f,t) = \begin{cases} t^{a-1}e^{-2\pi bt}\cos(2\pi ft), & t \geq 0 \\ 0, & else \end{cases} \tag{3.9}$$

t refers to time; *a*=4 is the order of the filter; *b* is the rectangular bandwidth which increases with the center frequency *f*. We use a bank of 128 filters whose center frequency *f* ranges from 50 Hz to 8000 Hz. These center frequencies are equally distributed on the ERB scale (Moore, 2003) and the filters with higher center frequencies have wider frequency ranges.

Since the filter output retains original sampling frequency, we down-sample the 128 channel outputs to 100 Hz along the time dimension. This yields a corresponding frame rate of 10 ms, which is used in many short-time speech feature extraction algorithms (Huang *et al.*, 2001). The magnitudes of the down-sampled outputs are then loudness-compressed by a cubic root operation. The resulting responses form a matrix, representing a T-F decomposition of the input. This T-F representation is a variant of cochleagram (Wang and Brown, 2006), which is analogous to the widely used spectrogram. Note that unlike the linear frequency resolution of a spectrogram, a cochleagram provides a much higher frequency resolution at low frequencies than at high frequencies. Figure 3.9 shows a cochleagram and a spectrogram of an utterance. Cochleagram retains higher frequency resolution at low frequency range for the same number of frequency components. We base our subsequent processing on this T-F representation.

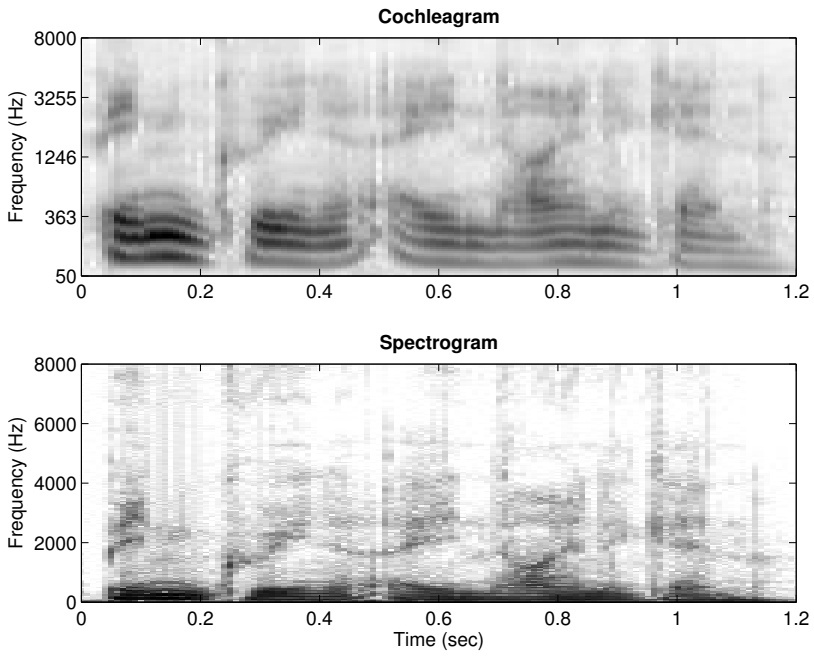

Figure 3.9: Illustrations of a cochleagram (top) and a spectrogram (bottom) of a clean speech utterance. Note the asymmetric frequency resolution at low and high frequencies in the cochleagram.

We call a time frame of the above cochleagram a Gammatone feature (GF). Hence, a GF vector comprises 128 frequency components. Note that the dimension of a GF vector is much larger than that of feature vectors used in a typical speaker recognition system. Additionally, because of overlap among neighboring filter channels, the Gammatone features are largely correlated with each other. Here, we apply a discrete cosine transform (Oppenheim *et al.*, 1999) to a GF in order to reduce its dimensionality and de-correlate its components. We call the resulting coefficients *Gammatone frequency cepstral coefficients* (GFCC) (Shao *et al.*, 2007).

The components of a GF vector G are indexed by variable i that ranges from 1 to N. Here, N=128, referring to 128 Gammatone frequency channels. Cepstral coefficients, $C[j]$ j=0…N-1, are obtained as follows,

$$C[j] = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} G[i] \cos\left(\frac{j\pi}{2N} (2i+1) \right), \; j=0\ldots N\text{-}1 \tag{3.10}$$

Note that the 0^{th} order coefficient is summed using all the GF components. Thus, it relates to the energy of a GF vector. In implementation, one can group the transform factors in (3.10) into a matrix, and multiply it with a GF matrix that is composed of a sequence of GF vectors to obtain GFCCs for an entire cochleagram.

Rigorously speaking, the newly derived features are not cepstral coefficients because a cepstral analysis requires a log operation between the first and the second frequency analysis for the deconvolution purpose (Oppenheim *et al.*, 1999). Here we regard these features as cepstral coefficients because of the functional similarities between the above transformation and that of a typical cepstral analysis.

3.4.2 Feature reconstruction and uncertainty decoding

As described before, the probability distribution of an extracted feature vector X, produced by a speaker λ, is modeled as a GMM, typically parameterized by diagonal covariance matrices (Reynolds, 1995). Under noisy conditions, the aforementioned monaural speech segregation system produces a binary T-F mask that indicates whether a GF feature component is reliable or corrupted (missing). Thus, the feature vector can be partitioned into reliable components $X_{\boldsymbol{r}}$, and unreliable ones $X_{\boldsymbol{u}}$,

$$X = \begin{bmatrix} X_{\boldsymbol{r}} \\ X_{\boldsymbol{u}} \end{bmatrix} \tag{3.11}$$

We propose to use the auditory cepstral feature GFCC in conjunction with the binary mask. In order to apply the DCT transform to a corrupted GF *X*, we first reconstruct the missing GF components from a speech prior model, which is similar to the universal background model (UBM) in a typical speaker verification system. Specifically, the speech prior $p(X)$ is modeled as a GMM (Raj *et al.*, 2004), and obtained from pooled training data:

$$p(X) = \sum_{k=1}^{K} p(k) p(X \mid k), \tag{3.12}$$

where *K* is the number of mixtures, *k* is the mixture index, and $p(k)$ gives the prior of a mixture, or in other words the mixture weight. $p(X|k)$ is the *k*th Gaussian distribution with a mean vector μ_k and a diagonal covariance σ_k^2. Given a binary mask, the components of the mean and variances of each Gaussian can be split into reliable and unreliable ones. We then calculate the *a posteriori* probability of the *k*th mixture given reliable GF components as in

$$p(k \mid X_{\boldsymbol{r}}) = \frac{p(k) p(X_{\boldsymbol{r}} \mid k)}{\sum_{k=1}^{K} p(k) p(X_{\boldsymbol{r}} \mid k)}. \tag{3.13}$$

As shown in Cooke *et al.* (2001) and Srinivasan and Wang (2007), the unreliable components are estimated as the expected value or the mean conditioned on $X_{\boldsymbol{r}}$.

$$\hat{X}_{\boldsymbol{u}} = \sum_{k=1}^{K} p(k \mid X_{\boldsymbol{r}}) \mu_{\boldsymbol{u},k}, \tag{3.14}$$

$\mu_{\boldsymbol{u},k}$ refers to the mean vector of the unreliable components of the *k*th mixture in the speech prior. The reliable components are retained in the reconstruction.

Although (3.14) gives a good estimate of the unreliable GF components, errors in reconstruction will cause degradation of recognition performance. Estimates of the reconstruction uncertainties likely mitigate such degradations by accounting for the reconstruction errors in the speaker likelihood calculation. Specifically, the uncertainties are estimated as in Srinivasan and Wang (2007),

$$\hat{\sigma}^2 = \sum_{k=1}^{M} p(k \mid X_{\boldsymbol{r}}) \left\{ \left(\begin{bmatrix} X_{\boldsymbol{r}} \\ \hat{X}_{\boldsymbol{u}} \end{bmatrix} - \mu_k \right)^2 + \begin{bmatrix} 0 \\ \sigma^2_{\boldsymbol{u},k} \end{bmatrix} \right\}. \tag{3.15}$$

$\sigma^2_{\boldsymbol{u},k}$ is the variances of the unreliable components of the *k*th mixture in the prior model. Thus, we have obtained the reconstructed GF and its variances, which are then transformed into the cepstral domain through DCT.

A registered speaker is modeled using a GMM. Therefore,

$$p(Z \mid k) = N(Z; \mu_{Z,k}, \sigma^2_{Z,k}) \tag{3.16}$$

calculates the likelihood of observing a GFCC frame, *Z*, given mixture component *k*; $\mu_{Z,k}$ and $\sigma^2_{Z,k}$ are the mean and the variances of the *k*th Gaussian mixture components. If noisy speech is processed by an unbiased speech enhancement algorithm, it is shown by Deng *et al.* (2005) that the observation likelihood shall be computed as

$$\int_{-\infty}^{\infty} p(Z \mid k) p(\hat{Z} \mid Z) = N(\hat{Z}; \mu_{Z,k}, \sigma_{Z,k}^2 + \hat{\sigma}_Z^2) . \quad (3.17)$$

Here, $\hat{Z}$ is the enhanced GFCC and $\hat{\sigma}_Z^2$ refers to the diagonal covariances of the cepstral transformed $\hat{\sigma}^2$ in (3.15). The non-diagonal covariance coefficients are numerically small because of the de-correlation of DCT (Oppenheim *et al.*, 1999) and dropped from computation. It can be seen that the uncertainty decoder increases the variances of individual Gaussian mixture components to account for the mask estimation errors (Deng *et al.*, 2005; Srinivasan and Wang, 2007).

3.4.3 Speaker identification evaluations

We evaluate the noise robustness of our proposed auditory features and the uncertainty estimation method in a SID task. The standard MFCC features are used to obtain the baseline performance. We also compare the performance of our system with the state-of-the-art robust front-end ETSI-AFE (STQ-AURORA, 2005-11).

We employ the speech materials from a recent speech separation challenge (SSC) (Cooke and Lee, 2006). The training data is drawn from a closed set of 34 talkers, 18 males and 16 females, and consists of 17,000 utterances. We use the speech-shaped noise (SSN) portion of the test set for our SID evaluation. The SSN data was generated by mixing clean utterances with speech-shaped noise at 4 SNRs: −12, −6, 0 and 6 dB. The test set contains 600 utterances in each SNR condition. The speakers are modeled using 64-mixture GMMs and trained on the training set of SSC directly. The speech prior

model comprises 2048 Gaussian mixtures, and is constructed from the pooled training utterances of all speakers. SID scores are only calculated on the voiced speech frames.

Figure 3.10 presents the SID evaluation results. 'MFCC_D_Z' denotes the baseline SID performance obtained using 24 MFCC features including deltas and after cepstral mean normalization. They are extracted using the HTK toolkit (Young *et al.*, 2000). 'ETSI-AFE' represents the enhanced 24 MFCC features including deltas, derived from the ETSI-AFE front-end feature extraction algorithm. 'ETSI-AFE_Z' denotes the cepstral mean normalized ETSI-AFE feature.

GF and 'GFCC_C0' are the auditory features described earlier with 128 and 23 dimensions respectively. 'GFCC' is the GFCC feature but with the first cepstral coefficient C_0 removed. 'GF_MD' stands for the missing data recognition method using the GF features and estimated binary T-F masks as described in Section 3.3.1.

'GFCC_C0_U' denotes SID performance by the uncertainty decoder using GF reconstruction and estimated uncertainties in the GFCC feature. 'GFCC_U' shows the same feature configuration but without C_0. 'GF_U' shows the SID performance when the uncertainty decoder is directly applied in the GF domain, before the DCT transform.

It is observed from the figure that the proposed GF feature performs significantly better than the baseline MFCC feature at low SNR conditions. More importantly, the GFCC features, especially the GFCC without C_0, not only achieve substantial improvement over the baseline feature, but also obtain comparable identification results with the robust features extracted by ETSI-AFE. Since C_0 relates to the overall energy of a feature frame, it is very susceptible to noise degradation. Thus, removing C_0 is

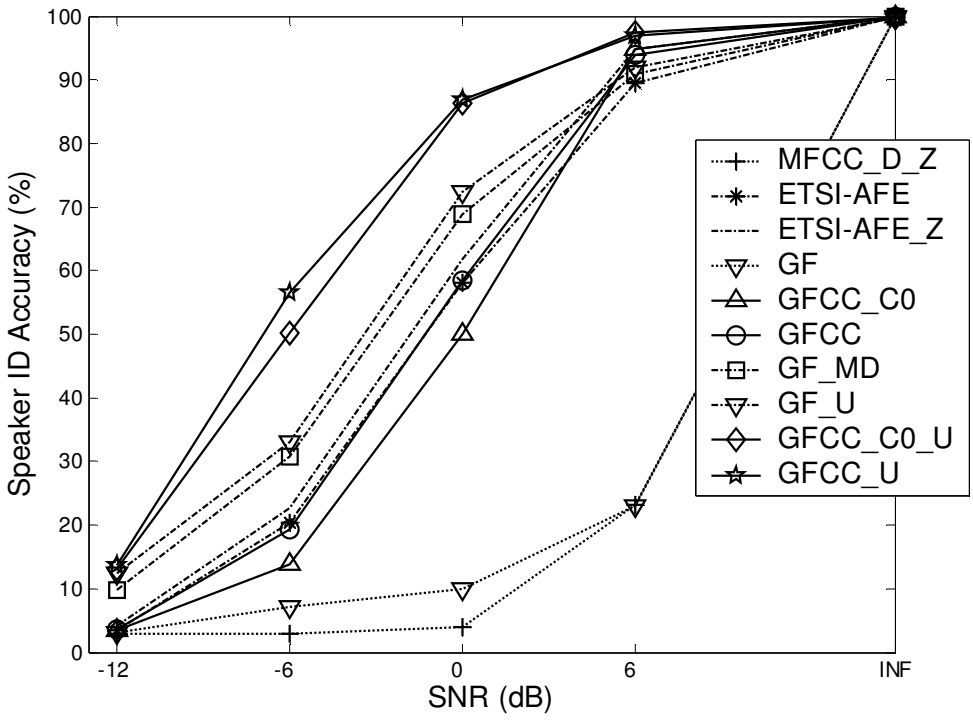

Figure 3.10: Accuracies of speaker identification in the presence of speech-shaped noise.

Legend	*Description*
_D	delta feature
_Z	cepstral mean normalization
_C0	the 0th order cepstral coefficient
_MD	missing data recognition
_U	uncertainty decoding
MFCC	Mel-frequency cepstral coefficients
ETSI-AFE	robust features by ETSI-AFE, including delta features
GFCC	Gammatone frequency cepstral coefficients

beneficial at low SNR conditions. Note that C_0 has been removed from MFCC and ETSI-AFE features.

The missing data method using marginalization performs significantly better than ETSI-AFE. GF reconstruction and uncertainty decoding in the GF domain further improve SID accuracies. Substantial improvement over ETSI-AFE is obtained after the GF feature and the uncertainty are transformed into the GFCC domain. In summary, GFCC features provide a substantial contribution to the robustness of the system.

3.4.5 Feature dimensions and dynamic features

In the experiment above, the lower 23-order GFCC coefficients are used as speaker feature vectors. We chose 23 coefficients because they are observed to be compact and retain the majority information of a GF frame. In addition, this number relates to the number of typical Mel-frequency filters in MFCC feature extraction (Huang *et al.*, 2001).

After performing cepstral transformation of GFCC, we find that the lower 30-order coefficients capture most of the GF feature information while the coefficients above 30th are close to 0 numerically, which means that they provide minimal information. Figure 3.11 illustrates a GFCC transformed GF and a cochleagram using 30 GFCCs. The top plot shows a cochleagram of an utterance. The middle plot shows a comparison of a GF frame at 1 *sec* of the top plot and the resynthesized GF from its 30 GFCCs. The bottom plot presents the resynthesized cochleagram from the top plot using 30 GFCCs. As can be seen from the figure, the lowest 30-order GFCCs retain the majority information in a 128-dimensional GF. This is due to the "energy compaction" property of DCT

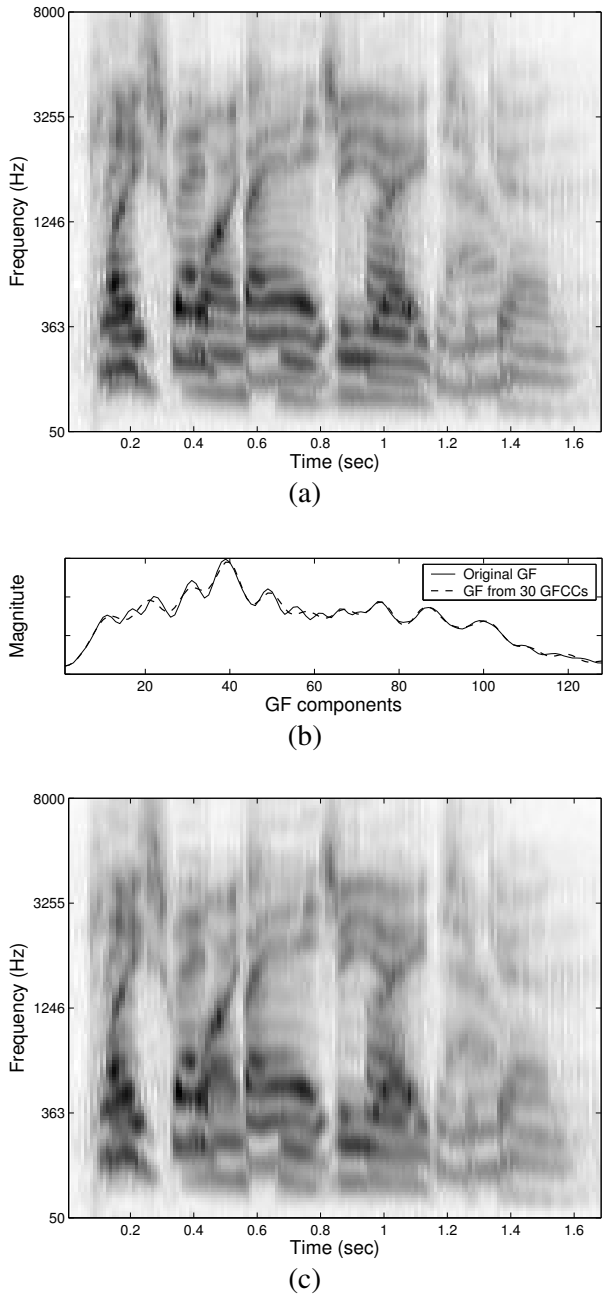

Figure 3.11: Illustrations of energy compaction by GFCCs. Plot (a) shows a cochleagram of an utterance. Darker color indicates stronger energy within the corresponding time-frequency unit. Plot (b) shows a GF frame at time 1 *sec* of (a). The original GF is plotted as the solid line and the resynthesized GF by 30 GFCCs is plotted as the dashed line. Plot (c) presents the resynthesized Cochleagram from (a) using 30 GFCCs.

(Oppenheim *et al.*, 1999). Hence, we switch to using the 30-dimensional GFCCs as feature vectors from now on.

Since a typical speaker recognition system uses MFCCs and their first-order (delta) dynamic coefficients. Thus, it is desirable to study how GFCC dynamic features fare for recognition. The delta feature Z_D at time t is calculated from a set of neighboring GFCC vectors Z around time t.

$$Z_D(t) = \frac{\sum_{w=1}^{W} w \cdot \left(Z(t+w) - Z(t-w) \right)}{2 \sum_{w=1}^{W} w^2} \tag{3.18}$$

w is a neighboring window index; W refers to the half-window length and it is set to 2 here. In other words, the delta-window is of length 5. Then, the obtained delta coefficients are appended to the 30-dimensional GFCCs, resulting in a 60-dimensional feature vector.

According to the uncertainty decoder described earlier, the enhanced feature vector Z assumes a Gaussian distribution. Thus, given the linear Equation of (3.18), the delta feature uncertainties are derived from GFCC uncertainties $\hat{\sigma}_Z^2$ as

$$\hat{\sigma}_D^2(t) = \frac{\sum_{w=1}^{W} w^2 \cdot \left(\hat{\sigma}_Z^2(t+w) + \hat{\sigma}_Z^2(t-w) \right)}{\left(2 \sum_{w=1}^{W} w^2 \right)^2}. \tag{3.19}$$

Second-order dynamic coefficients, known as the acceleration feature, can be calculated by replacing the GFCCs and their uncertainties (3.18-3.19) with the delta coefficients and their uncertainties.

We evaluate the new set of GFCC features using the same recognition system and evaluation configurations as reported in the preceding section. The results are shown in Table 3.1 and the symbols used in the table are explained in Table 3.2. Note that the default experiment configuration includes missing-data reconstruction from binary T-F masks. Increasing the number of GFCCs from 23 to 30 improves identification performance under low SNR conditions. Similar improvements are also observed when C_0 is removed. Therefore, we use as default 30-dimensional GFCC as feature vectors in the rest of this dissertation. Note that the first cepstral coefficient C_0 relates to the overall energy of a feature frame, thus it is susceptible to noise degradation. Therefore, it is beneficial to keep it under high SNR conditions and remove it under low SNRs.

It can be seen from the table that the delta-augmented GFCCs achieve significantly better performance than GFCC alone except under -12 dB condition, where the missing-data reconstruction does not perform well with few reliable T-F units. Using the uncertainty decoder improves the identification accuracies under -6 dB and 0 dB where the performance has not saturated. Unlike other features in the table, 'GFCC(30)_C_0_D' denotes that delta coefficients are appended but uncertainty decoding is not applied to the deltas. Controlled comparison of this feature set with others shows that delta feature alone improves identification accuracy and that applying uncertainty decoding further improves performance. However, we find that including the acceleration feature rather

Feature	-12 dB	-6 dB	0 dB	6 dB	Clean
GFCC(23)_C_0_U	13.33	51.17	87	97.33	99.67
GFCC(22)_U	13.67	56.5	86.83	96.83	99.67
GFCC(30)_C_0_U	14.83	54.67	89	97.67	99.67
GFCC(29)_U	13.33	58.5	88.5	97.33	99.67
GFCC(30)_C_0_D	9.67	56.5	90.5	98.5	99.67
GFCC(30)_C_0_D_U	9.83	58.83	92.17	98.67	99.67
GFCC(30)_C_0_D_A_U	7.67	37.83	81	97.33	99.67

Table 3.1: Accuracy (%) of robust speaker identification using GFCCs, dynamic features and uncertainty decoding. Symbols are explained in Table 3.2.

Experiment Symbol Descriptions	
GFCC	Gammatone frequency cepstral coefficients
(23) (30)	total number of GFCCs
_C_0	the 0th order cepstral coefficient
_D	first-order dynamic feature, delta feature
_A	second-order dynamic feature, acceleration feature
_U	uncertainty decoding
_Z	cepstral mean normalization
ETSI-AFE	robust features by ETSI-AFE

Table 3.2: Symbol notations.

hurts the system performance. This is because the acceleration window requires 9 frames while the binary masks with the speech-shaped-noise do not contain consecutively reconstructed frames that can provide reliable acceleration feature estimates.

3.4.6 SID evaluations under other non-stationary noise conditions

The experiments of the preceding sections are conducted under cochannel speech and speech-shaped-noise conditions. In this section, we evaluate our system under four other non-stationary noisy conditions. Specifically, the four noise types are selected from the Noisex 92 corpus (Varga and Steeneken, 1993) which is widely used for robust speech recognition studies: speech babble noise, destroyer (a navy ship) operation room noise, F-16 cockpit noise and factory noise. The first two types contain a noisy background with many people speaking at the same time. We will use the simplified notations of "Babble", "Destroyer", "F16" and "Factory" to refer to the four noise types respectively.

Each noise type was stored as a single recording of approximately 320 *sec*. We create the noisy utterances for test purposes by mixing the clean utterances of the SSN task with the four noise types. The mixtures are created at -12 dB, -6 dB, 0 dB, 6 dB, 12 dB and 18 dB SNRs. In order to incorporate all the noise statistics, we randomly select a portion of the noise recording that has the same duration as the clean utterance for each mixture pair. This guarantees that the noise signals are different for each utterance but they have the same type within each SNR condition. Thus, there are 34 speakers and 600 test utterances for each of the 6 SNR conditions of the 4 noise types.

This expanded test set is evaluated using GFCC, delta feature and uncertainty decoding, and their results are present in Table 3.3. Specifically, we use the 30-dimensional GFCCs and their delta coefficients because they achieved the best overall performance in the preceding experiments. In addition, MFCC features extracted by

Babble	**-12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**18 dB**
GFCC_C_0_D	3.83	25	83.83	97.5	99.17	99.5
GFCC_D	4.33	30.67	82.83	96.67	98.67	99.5
GFCC_D_C_0	3.83	28	83.17	97.33	99.17	99.5
ETSI-AFE_D	3.17	19	69.83	96.5	99.67	99.83
ETSI-AFE_D_Z	2.16	15	61.5	93.33	99.67	99.83
Destroyer	**-12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**18 dB**
GFCC_C_0_D	4	16.5	76.83	97	98.67	99.17
GFCC_D	3.67	14.67	76.83	94.33	98.5	99.17
GFCC_D_C_0	3.33	13.17	73.67	94.83	98.5	99.17
ETSI-AFE_D	3.33	12.83	44.5	76.17	95	99.33
ETSI-AFE_D_Z	3	11.67	48.67	88.33	98.17	99.5
F16	**-12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**18 dB**
GFCC_C_0_D	6.83	41.67	83.5	96.5	99.17	99.33
GFCC_D	6.83	45.17	84.33	95.5	99.17	99.5
GFCC_D_C_0	6.17	42.17	84.5	95.67	99	99.5
ETSI-AFE_D	3.33	3.83	37.83	77.5	96.5	99.67
ETSI-AFE_D_Z	2.17	7.5	35.33	81.17	97	99.5
Factory	**-12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**18 dB**
GFCC_C_0_D	8.17	46.17	87.83	97.83	99.33	99.33
GFCC_D	8.67	46.17	87.67	97.17	99.33	99.5
GFCC_D_C_0	10	43.17	86.67	97.33	99	99.5
ETSI-AFE_D	3.67	9.5	43.5	79.17	95.67	99.5
ETSI-AFE_D_Z	3.33	9	38	82.17	96.17	99.33

Table 3.3: Accuracy (%) of robust speaker identification using GFCCs, dynamic features and uncertainty decoding. Performance by using ETSI-AFE is presented for comparison. Notations are explained in Table 3.2. Note that symbol '_U' is dropped because all the configurations with GFCCs employ the uncertainty decoder.

ETSI-AFE with their delta features and CMN are evaluated for comparison purposes since they provide the best SID performance without using GFCC features.

The results in Table 3.3 corroborate the conclusions in the SSN experiments that GFCC features significantly outperform ETSI-AFE features except at SNRs of 12 dB and 18 dB where identification performance saturates. Additionally, the first cepstral coefficient C_0 is susceptible to noise degradation since it relates to the overall energy of a feature frame, and removing it improves performance under low SNR conditions.

We have also evaluated the GFCCs with C_0 removed from the static feature but with C_0 added for the delta feature. This type of feature configuration has shown robustness in speech recognition studies since it is believed that delta C_0 is robust against noise. However, according to the results in the Table, this feature combination does not consistently improve accuracies. This is mainly due to poor GFCC reconstruction from sparse binary T-F masks at low SNRs.

3.4.7 Speaker verification evaluations

All the preceding experiments are conducted on speaker identification tasks. Our system can be easily generalized to handle speaker verification tasks by adopting the decision rule in (3.5). Since the verification process requires a UBM for score normalization purpose, the UBM likelihood is calculated by the uncertainty decoder in the same way as the speaker likelihoods. We evaluate the verification system on the same test set used in the preceding section. Specifically, one noisy utterance contributes 1 true score for the corresponding target speaker and 33 imposter scores for the others in the

speaker set. Therefore, there are 600 true scores and 19800 imposter scores for every SNR condition. The UBM is a 2096-mixture GMM, obtained from the training data pooled from all the 34 speakers.

The verification results are shown as DET curves (Martin *et al.*, 1997) in Figures 3.12-3.16 at the end of this section. Note that all the -12 dB conditions are not included because the recognition errors are too high for the DET toolkit to plot meaningful curves. From the figures, it can be observed that GFCC features significantly outperform ETSI-AFE features under most of the conditions, except at 12 dB of Factory and F16 noise types, 6 dB of Babble and Destroyer noise, where results are comparable. However, GFCC underperforms ETSI at 18 dB of all the noise types. We suspect that there are two reasons. The first is that reconstruction is not as effective as at high SNR conditions as using ETSI features. In the speaker identification evaluations, identification accuracies are improved from above 90% at 6 dB to close to 100% at 12 dB and 18 dB. The second reason is that the estimated uncertainties flatten the Gaussian mixture distribution, leading to less discriminative scores. We conduct several control studies to understand why the GFCC performance lags behind at high SNRs. The results are also shown in Figure 3.12-3.16. First, the delta feature uncertainties are removed to see whether delta uncertainties distort the likelihood estimates. Then, we remove the uncertainty decoder and use the enhanced GFCC features directly for verification. From the 18 dB plots, it is evident that GFCC features, both baseline ones (extracted directly from noisy speech) and reconstructed ones, substantially outperform baseline MFCC features. Babble noise is an exception, which yields comparable speaker verification results at 18 dB. However,

since the improvement margin of speaker identification experiments is the smallest among all the four noise types, this exception is probably due to the noise itself. Invariably, the enhanced GFCC features, both with and without uncertainty decoding, perform significantly lower than GFCC baseline features. This finding supports our first hypothesis that feature reconstruction is not as effective in higher SNR conditions as in lower SNR conditions. For the enhanced GFCC features, removing uncertainties actually improves performance numerically even though the resulting differences may not be statistically significant. This supports our second hypothesis that uncertainties reduce the discriminative power of reconstructed features.

ETSI-AFE features are derived using sophisticated noise reduction processing (STQ-AURORA, 2005-11). It contains a two-stage Mel-warped Wiener filtering, voice activity detection for noise estimation, gain normalization and several other advanced front-end signal processing modules. On the one hand, it is not surprising for ETSI features to outperform GFCCs under mildly noisy conditions such as 18 dB since GFCC extraction does not undergo any signal-level noise reduction processing. On the other hand, the model-based GFCC reconstruction, which estimates global means for missing values given reliable T-F units, outperforms the ETSI-AFE under the remaining noisy conditions.

In summary, considering the facts that baseline GFCCs achieve significantly better performance than MFCCs under high SNR conditions and that enhanced GFCCs with uncertainty decoding outperform ETSI-AFE features under most of the other SNR conditions, our GFCC extraction method provides a robust feature set for speech processing.

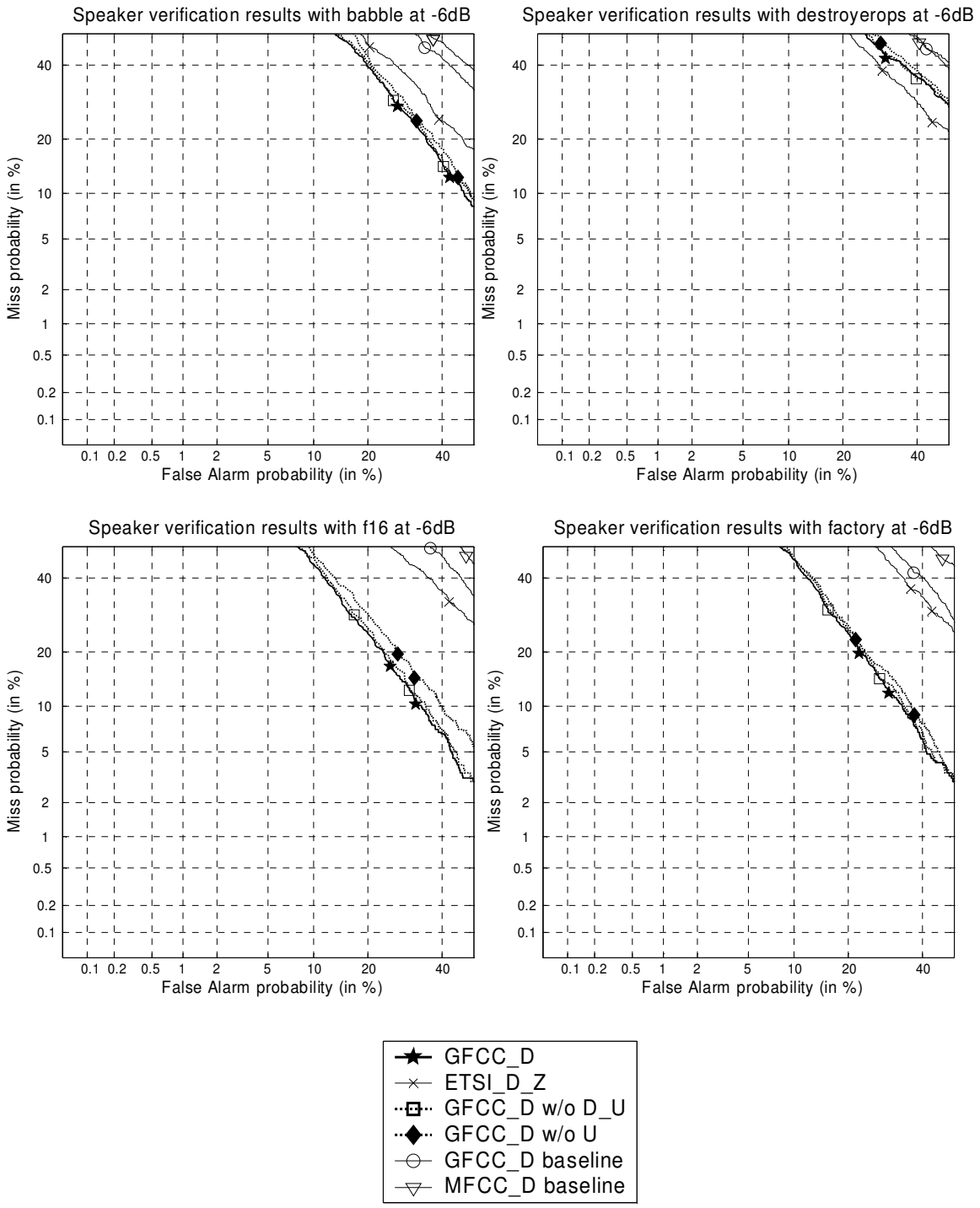

Figure 3.12: Speaker verification evaluation under -6 dB of Babble, Destroyer, F16 and Factory noise conditions. _D means delta feature, _U refers to uncertainty decoding, _Z means cepstral mean normalization.

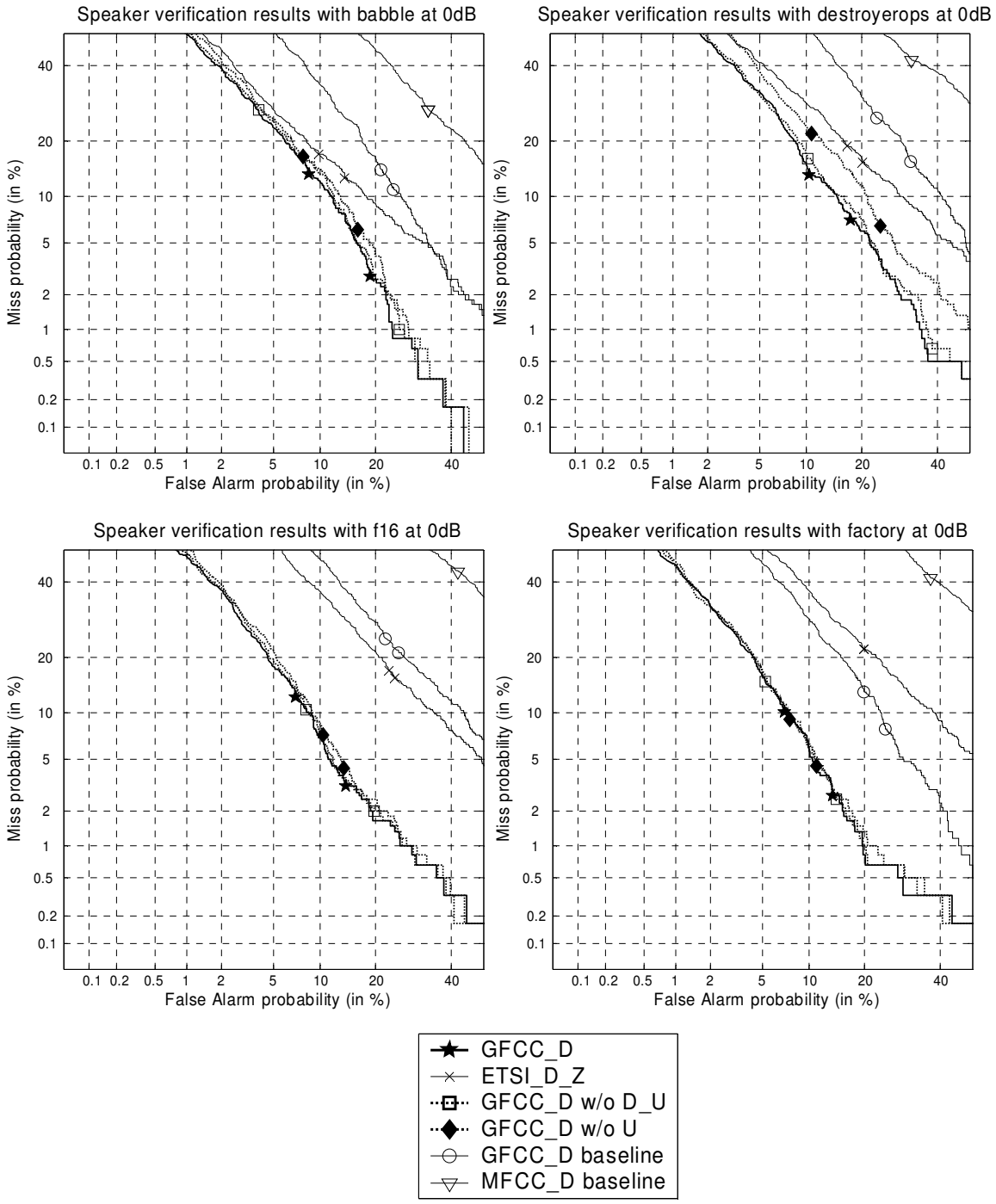

Figure 3.13: Speaker verification evaluation under 0 dB of Babble, Destroyer, F16 and Factory noise conditions. _D means delta feature, _U refers to uncertainty decoding, _Z means cepstral mean normalization.

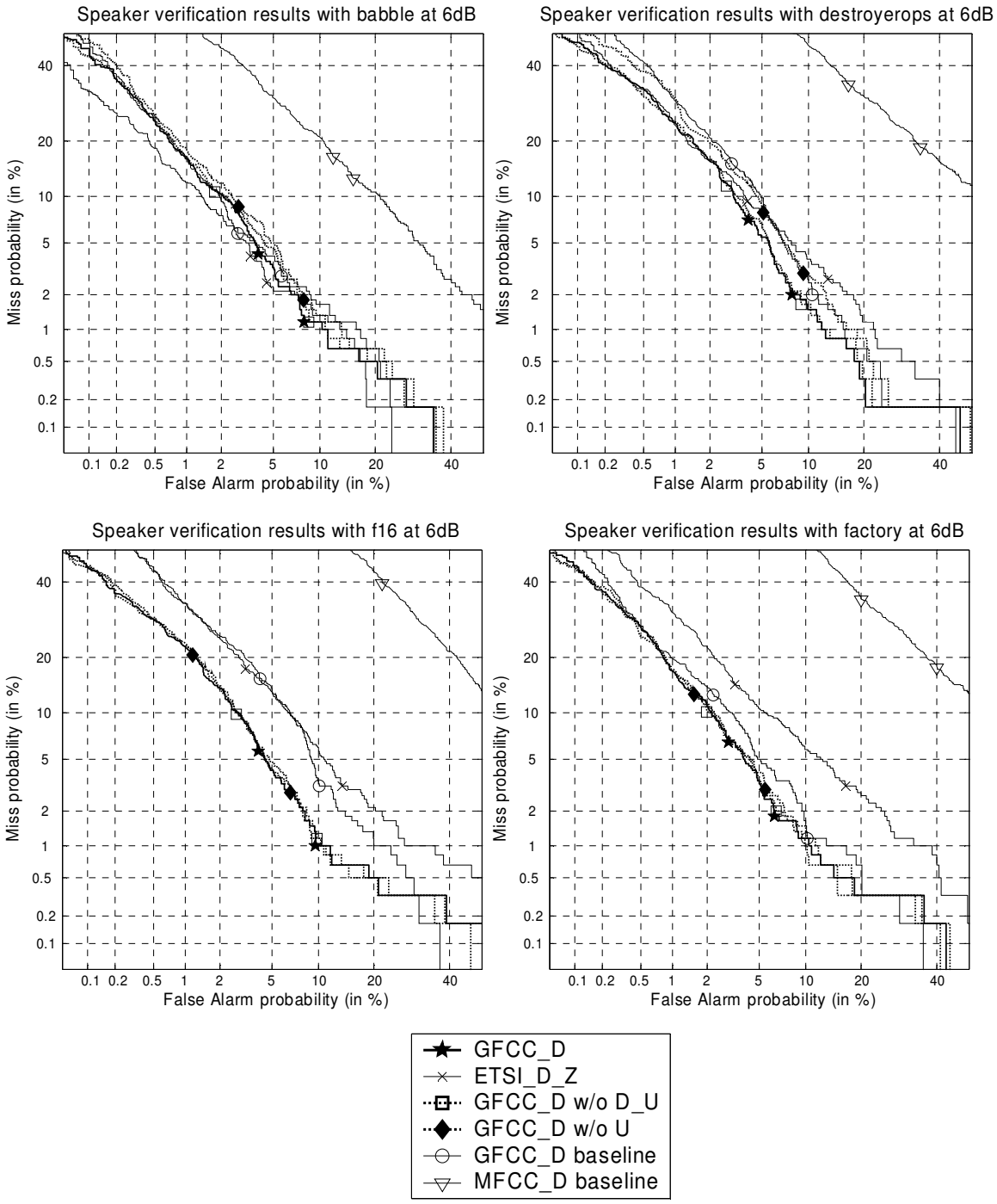

Figure 3.14: Speaker verification evaluation under 6 dB of Babble, Destroyer, F16 and Factory noise conditions. _D means delta feature, _U refers to uncertainty decoding, _Z means cepstral mean normalization.

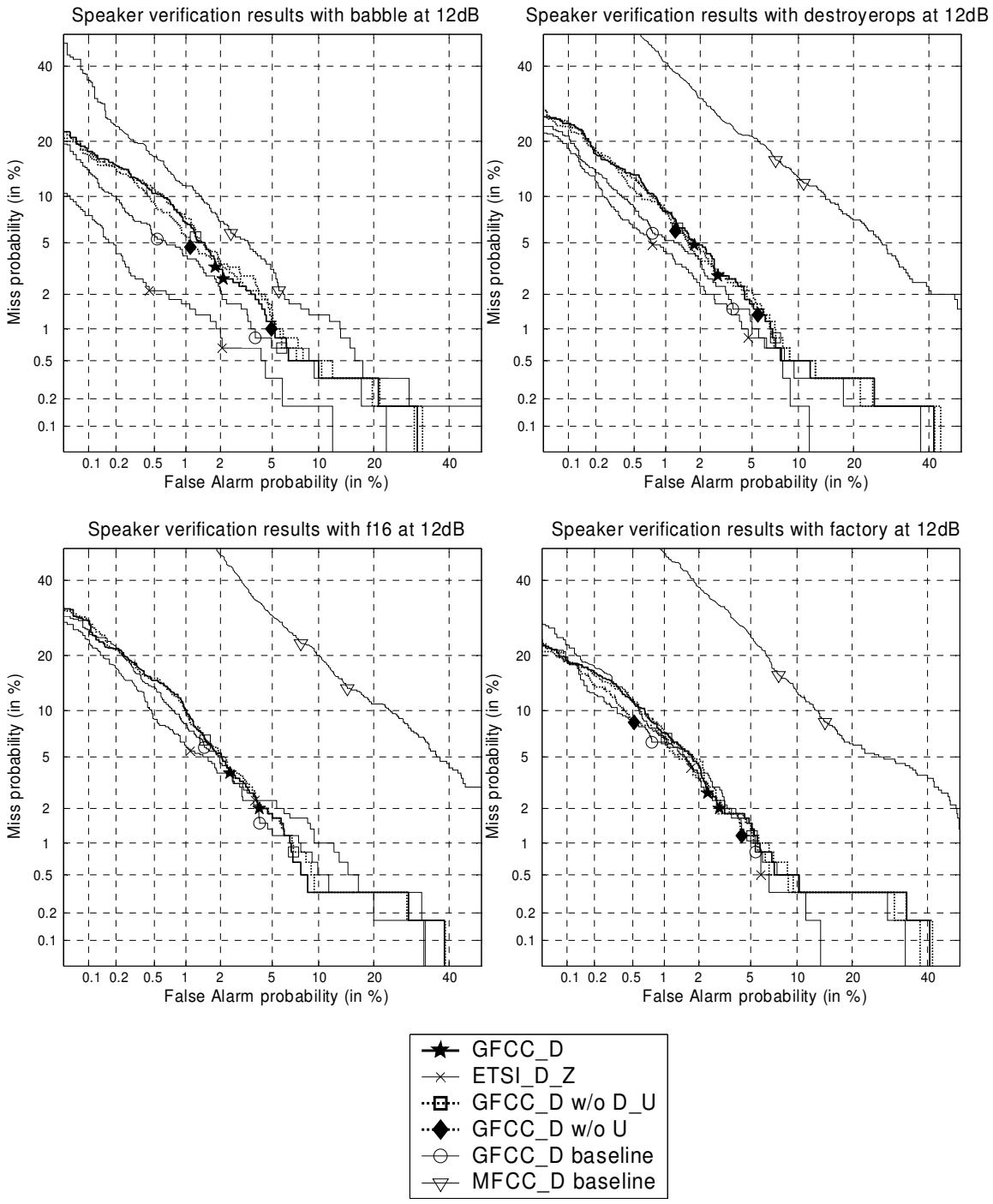

Figure 3.15: Speaker verification evaluation under 12 dB of Babble, Destroyer, F16 and Factory noise conditions. _D means delta feature, _U refers to uncertainty decoding, _Z means cepstral mean normalization.

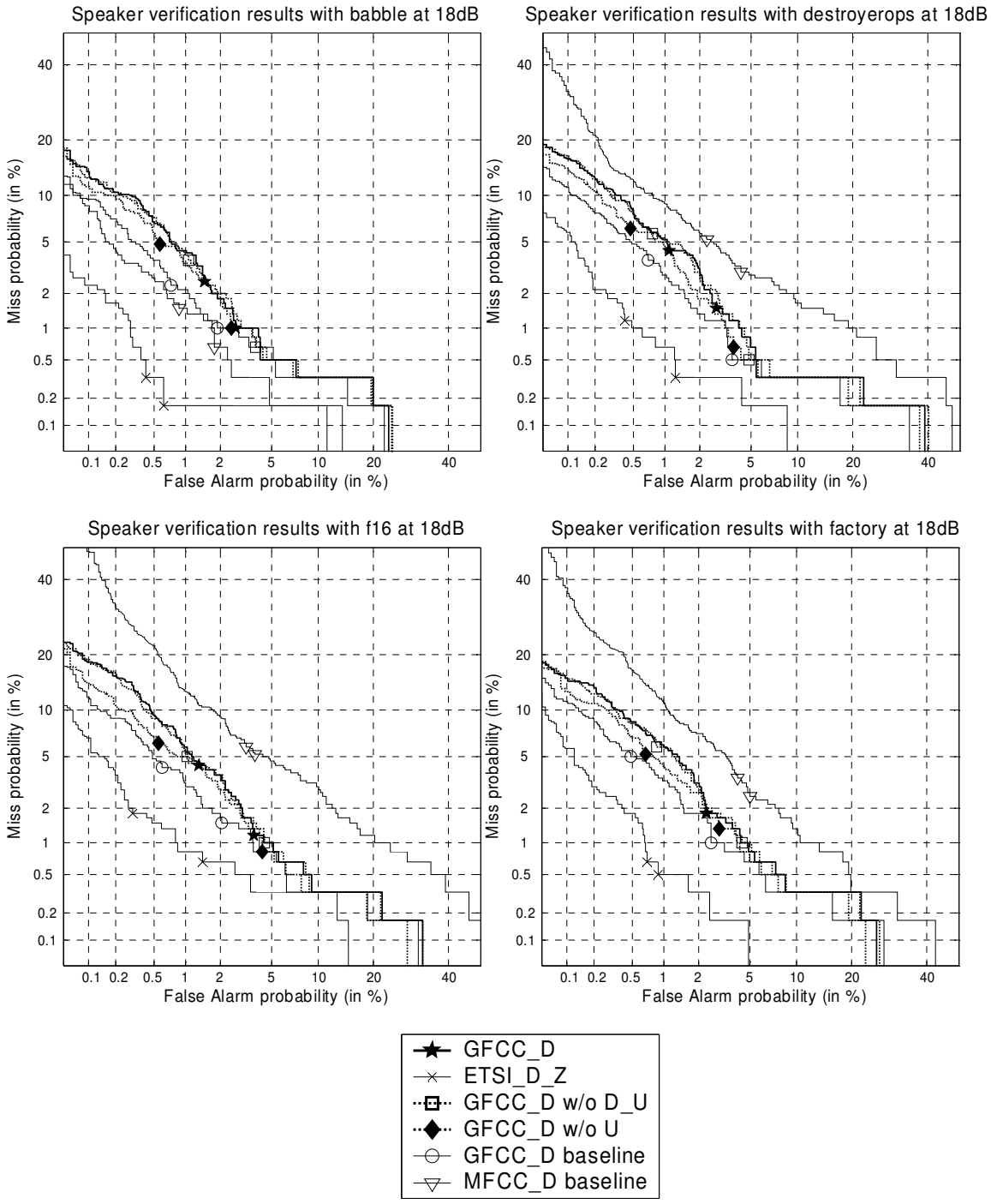

Figure 3.16: Speaker verification evaluation under 18 dB of Babble, Destroyer, F16 and Factory noise conditions. _D means delta feature, _U refers to uncertainty decoding, _Z means cepstral mean normalization.

CHAPTER 4

FEATURE AND MODEL BASED SEQUENTIAL GROUPING

As described earlier, the segmentation stage of a CASA system decomposes an input signal into groups of contiguous time-frequency units or segments. Each of these segments originates from a single speaker. Subsequently, the simultaneous organization process groups these segments across frequency, resulting in simultaneous streams. The goal of sequential organization is to further group these streams, which are still unrelated in time, into speaker streams.

In this chapter, we first explore feature-based grouping methods using features discussed in Chapter 2, including pitch, spectrum, timbre and vocal-tract length. Since the organizational goal entails classification and assignment of the segments according to their inherent speaker identities, we propose to base sequential grouping on the speaker feature and modeling methods described in Chapter 3. More specifically, we derive a computational objective for joint speaker recognition and sequential grouping based on speaker models. The derivation of the objective leads to an algorithm that searches for the optimal hypothesis in the joint speaker and grouping space. This model-based algorithm

groups the usable speech segments and its performance is then compared with the feature-based methods under cochannel conditions.

Subsequently, we extend our grouping algorithm to incorporate a finer level representation of the segments using binary time-frequency masks and novel auditory features. To handle the missing T-F units denoted by the binary masks, this extension employs the reconstruction method and the uncertainty decoder from our robust speaker recognition study. The grouping algorithm is evaluated on a speech separation and recognition task.

4.1 Feature-based Sequential Grouping

4.1.1 Pitch-based sequential grouping

Previous studies have demonstrated the importance of pitch information for speaker recognition; see e.g. Atal (1972). Perceptual studies have also shown the importance of pitch in speech grouping, e.g. Darwin *et al*. (2003). Pitch can help differentiate speakers and thus could be very useful for sequential grouping. However, pitch alone is inadequate for speech grouping because a speaker's pitch may vary considerably and different speakers can have substantial pitch range overlap. On the hand, dynamic aspects of pitch is more discriminative (Atal, 1972). In this section, we propose a pitch dynamic feature for the sequential grouping purpose.

The dynamic feature is extracted based on multipitch tracking (see Section 2.3.1) and usable speech segments (see Section 3.2.2). First, the time gap between two pitch

segments and the difference between the ending pitch of the preceding segment and the beginning pitch of the following segment are collected. We then multiply the two obtained quantities together. The resulting product reflects a pitch change between two segments. Basically, the larger the product, the less likely these segments belong to the same speaker. Hence, it is considered as a pitch dynamic feature. Since a Gaussian-like peak is observed from the histogram of training samples, we employ a mixture of a Gaussian distribution and a uniform distribution to model the feature distribution. Maximum likelihood estimators of the distribution parameters are obtained from the training samples. For sequential grouping, a binary decision is made regarding whether to group the current segment with the preceding segment by thresholding the likelihood of the pitch dynamic feature given its distribution. This pitch-based grouping method is compared with other grouping methods in Section 4.2.3.

4.1.2 Timbre features

As described in Section 2.2, the definition of timbre is general and does not prescribe what are the attributes that constitute timbre. In order to employ timbre for sequential organization, as a first step, the timbre attributes need to be defined.

Brown and Cooke (1994b) propose a timbre-based approach for music sound segregation. The authors suggest two features, brightness and onset asynchrony, obtained from two dimensions of a timbre space. This space is derived from studies on perceptual music grouping. One dimension of the space describes the distribution of spectral energy. Within this dimension, the brightness feature is defined as a spectral centroid. In exact

terms, it is the mean spectral amplitude weighted by frequency indices. The other dimension relates to harmonic synchronicity at the onset of a tone. The onset asynchrony feature quantifies the quality of this synchronicity by measuring the slope of onset times of all the spectral components. Since samples of music instruments exhibit well-separated clusters in the feature space, sequential grouping is easily done by comparing the timbre features. A subsequent study models dynamics of the timbre features by tracking the features across time (Godsmark and Brown, 1999). More specifically, the features are estimated from continuous time frames, and the estimates are further smoothed in time to form timbre tracks. Since a musical instrument has a unique track, sequential grouping is performed by inspecting whether an input follows the timbre track.

The success of applying timbre in music grouping is mainly due to the fact that music instruments tend to have invariant timbre features. However, timbre features do not exhibit the same type of invariance for speech. For example, Figure 4.1 shows histograms of spectral centroid estimates for six speakers from the TIMIT corpus. We randomly select two utterances for each speaker and calculate spectral centroid within 20 ms time frames with 10 ms frame shift. Here, only voiced frames are shown in the figure. Similar to the study by Brown and Cooke (1994b), the spectral centroid is calculated by weighting GF components with their channel indices. It is evident from the figure that male speakers tend to have smaller spectral centroid than females, but there is no obvious pattern to distinguish different speakers. Therefore, we will not use the timbre features for speech grouping. On the other hand, acoustic features such as MFCC and GFCC have already incorporated the signal properties described by the timbre features and the timbre

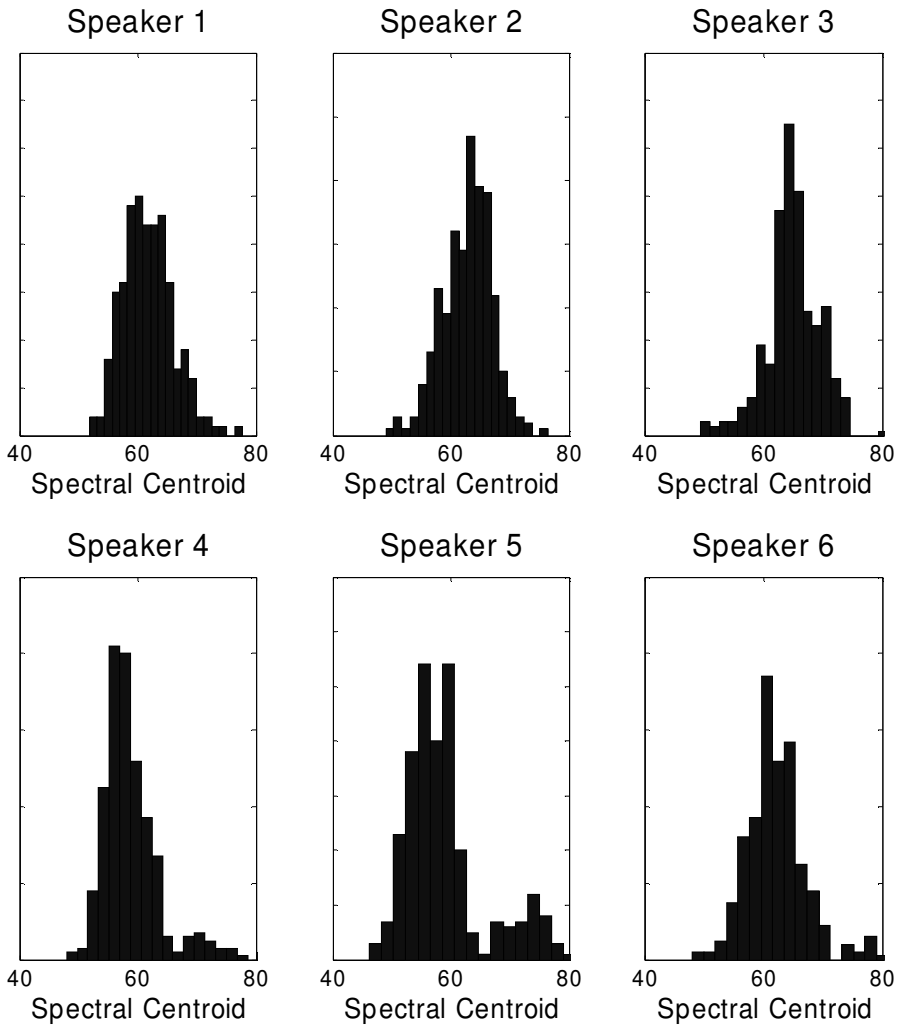

Figure 4.1: Histograms of spectral centroid estimates for six speakers from the TIMIT corpus. The top three speakers are female and the bottom three are male.

track is a simplified variant of the dynamic features presented in speaker recognition studies. Hence, we will employ the MFCC, GFCC and dynamic features for sequential grouping.

4.1.3 Vocal tract length

Speech production is typically modeled as a source-filter process (Furui, 2001). The source in the model refers to a sequence of pulses that simulates the airflow passing through the glottis when voiced sound is produced; it is typically simulated as white noise when unvoiced sound is produced. Vocal tract is usually modeled as concatenated acoustic tubes, which filter the source signal into the sounds being produced. Given this model, the frequency responses of the vocal tract can be estimated from spectral envelope of the signal. In turn, the shape of the vocal tract can be estimated from acoustical analysis of voiced speech (Schroeder, 1966; Paige and Zue, 1969; Wakita, 1973; Wakita, 1977; Ladefoged *et al.*, 1978; Necioglu *et al.*, 2000; Dang and Honda, 2002). Vocal tract length (VTL) is one prominent factor that defines this shape. Intuitively speaking, the longer the VTL, the lower the fundamental frequency and formant frequencies, and vice versa. In addition, VTL estimation and normalization have been employed to remove inter-speaker variability and yield performance gains for automatic speech recognition (Huang *et al.*, 2001). For sequential organization, the key is to group separated simultaneous streams according to speaker identities. Since VTL measures an anatomical difference between speakers, we explore how to use VTL for speech grouping in this section.

Here, we employ the algorithm proposed by Wakita (1977) for VTL estimation. In this algorithm, vocal tract shape is modeled by an acoustic tube with concatenated cylindrical sections of equal length. By applying the linear prediction method in a frame of input signal, the area function of each section can be determined from the frequencies and bandwidths of the detected formants. Since there is an infinite number of VTLs for a given set of formants, the length that gives the most uniform shape of the tube is selected as the estimate from a VTL range.

Figure 4.2 shows histograms of VTL estimates for six speakers from the TIMIT corpus. Two utterances are randomly selected for each speaker and VTL is estimated within 20 ms time frames with 10 ms frame shift. Estimates from voiced frames are shown in the figure as done in Figure 4.1. It is evident that all the speakers have large overlaps in the VTL range and that there is no distinction among different speakers. This is mainly due to the fact that formant frequencies and bandwidths relate to phonemes and that the same phoneme uttered by different speakers tends to exhibit close formants (Furui, 2001). Since VTL estimation depends on formants, variations of VTLs from the same speaker likely originate from different phonemes, which exhibit different formants. More importantly, under noisy conditions, robust formant estimation remains a challenge problem (Huang *et al.*, 2001). Thus, we do not further explore VTL or related features for sequential grouping.

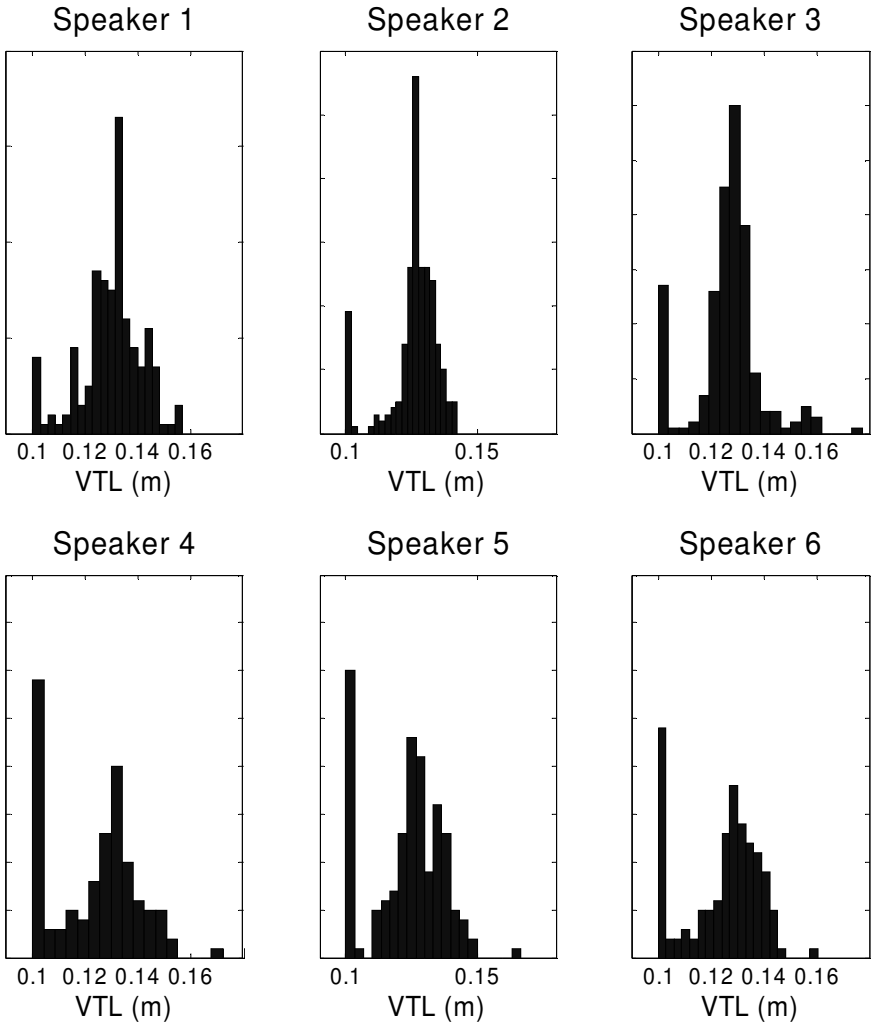

Figure 4.2: Histograms of VTL estimates for six speakers from the TIMIT corpus. The top three speakers are female and the bottom three are male.

4.1.4 Spectrum-based sequential grouping

Speech spectrum carries speaker characteristics (Furui, 2001). Assuming that spectral features of the same speaker are more similar than those of different speakers, spectral similarity can be used for sequential grouping. The symmetrized spectral divergence measure proposed by Carlson and Clement (1991) estimates such similarity using the linear predictive coding method. This measure is employed by Morgan *et al*. (1997) to assign separated and enhanced speech to two speaker streams in a cochannel mixture. Specifically, the assignment relies on frame-level spectral comparison of an unassigned frame with recently assigned frames using the divergence measure.

Given the assignment of initial frames in a cochannel mixture, we calculate the minimum divergence measure between an unassigned frame and the assigned ones for either speaker. The frame is assigned to the speaker that yields the smaller divergence. This spectrum-based grouping method is compared with other grouping methods in Section 4.2.3. Since assignment of the initial frames greatly influences the following grouping decisions, we manually assign the initial fifty frames of either speaker streams. Hence, the reported results represent the best performance that the spectrum-based method can achieve.

4.2 Model-based Sequential Organization

Unlike the preceding section, we focus on using speaker models for sequential organization in this section. Specifically, we propose a model-based grouping system that

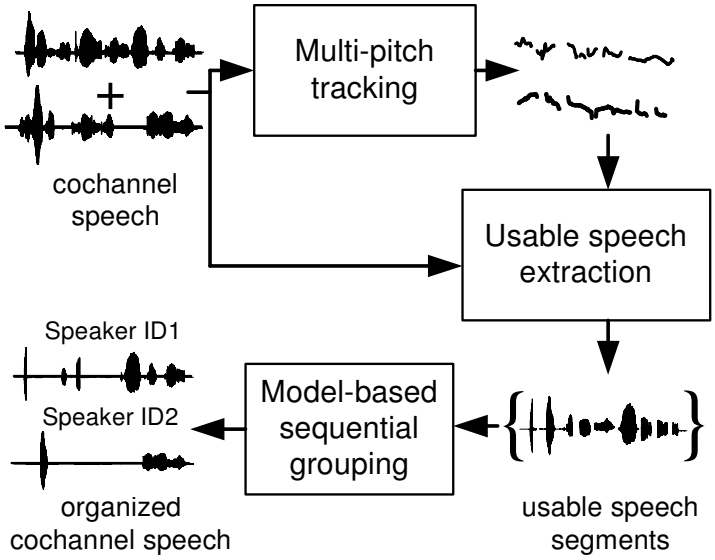

Figure 4.3. Schematic diagram of the proposed model-based sequential grouping system. First, cochannel speech is passed through a multipitch tracking algorithm and pitch contours are obtained. Then usable speech segments are extracted based on the pitch information. Finally, a model-based sequential grouping algorithm organizes segments into two streams and corresponding speaker identities are also produced.

organizes usable speech segments under cochannel conditions. In other words, we regard the usable segments as simultaneous streams. Figure 4.3 presents a diagram of the system. First, a multipitch tracking algorithm processes a cochannel input and produces pitch contours (Wu *et al.*, 2003). Then usable speech segments are extracted from the input based on the pitch contours. Finally, the model-based sequential grouping algorithm organizes these usable speech segments into two speaker streams (refer to Section 2.3.1

for detailed description of multipitch tracking, and Section 3.2.2 for usable speech extraction.)

4.2.1 Derivations

Here, we seek to construct a model-based sequential organization framework. Since we propose using speaker characteristics for the grouping purpose, we derive our computational goal from the speaker recognition framework as described in Section 3.1.1. Given a set of K registered speaker models $\Lambda=\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$, a speaker identification system seeks to find the speaker model that maximizes the posterior probability for a feature sequence $O = \{X_1, X_2, \ldots, X_M\}$ that has been extracted from an input speech utterance. By applying Bayesian analysis, the SID decision rule becomes Equation (3.3). For the sake of completeness, we include the same equation as follows,

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} P(O \mid \lambda) = \arg\max_{\lambda \in \Lambda} \sum_{m=1}^{M} \log p(X_m \mid \lambda) . \tag{4.1}$$

m is the time frame index of the feature sequence. This maximum-likelihood classification has been well established (Reynolds, 1995). However, in order to organize speakers in cochannel speech, this probability framework for a single speaker needs to be extended to multiple speakers.

Given a cochannel input, the usable speech extraction method extracts N speech segments, $Y=\{S_1, S_2, \ldots, S_i, \ldots, S_N\}$, each of which is a usable speech segment that is composed of frames X; in other words, $S_i=\{X\}$. Given Y, (4.1) can be extended as follows

$$\hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}} = \arg\max_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda} P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \mid Y). \tag{4.2}$$

This decision rule finds a pair of speaker models $\hat{\lambda}_{\mathrm{I}}$ and $\hat{\lambda}_{\mathrm{II}}$ from the speaker set Λ that maximize the posterior probability given usable speech segments. As mentioned earlier, the single-pitch segments must be organized into two speaker streams because in cochannel speech one speaker can dominate in some portions and be dominated in other portions. For example, a possible segment assignment (grouping) may look like $\{S_1^0, S_2^1, ..., S_i^1, ..., S_N^0\}$, where superscripts, 0 and 1, do not represent the speaker identities but only denote that the segments marked with the same label are from the same speaker. Therefore, the joint computational objective of sequential grouping and SID may be stated as finding a pair of speaker models $\hat{\lambda}_{\mathrm{I}}$ and $\hat{\lambda}_{\mathrm{II}}$ together with a segment assignment $\hat{g}$ that jointly maximize the posterior probability:

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}}, = \underset{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, g \in G}{\operatorname{argmax}} P(g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \mid Y), \tag{4.3}$$

where G is the assignment space, which includes all possible assignments (label sequences) of the segments.

The posterior probability in (4.3) can be written as

$$P(g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \mid Y) = \frac{P(g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, Y)}{P(Y)} = P(Y \mid g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) P(g \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \frac{P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})}{P(Y)}. \tag{4.4}$$

Since the assignment is independent of specific models, $P(g \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$ becomes $P(g)$ which, without prior knowledge on segment assignment, we assume to be uniformly distributed. Assuming the independence of speaker models and using the same assumption from traditional speaker identification that prior probabilities of speaker

models are the same, we insert Equation (4.4) into (4.3) and remove the constant terms. The objective then becomes finding two speakers and an assignment that have the maximum probability of assigned usable speech segments given the corresponding speaker models as follows.

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}} = \underset{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, g \in G}{\arg\max} P(Y \mid g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}). \tag{4.5}$$

Note, the conditional probability is essentially the joint SID score of assigned segments. Given an assignment g, we denote Y^0 as the subset of usable speech segments labeled 0, and Y^1 the subset labeled 1. Since Y^0 and Y^1 are complementary, the probability term in (4.5) can be written as follows,

$$P(Y \mid g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) = P(Y^0, Y^1 \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}). \tag{4.6}$$

The g term is dropped from the above equation because the two subsets already incorporate the labeling information.

Assuming that any two segments, S_i and S_j, are independent of each other given the speaker models and that segments with different labels are produced by different speakers, the conditional probability in (4.6) can be written as

$$\begin{aligned} P(Y^0, Y^1 \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) &= P(Y^0 \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) P(Y^1 \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \\ &= \prod_{S_i \in Y^0} P(S_i \mid \lambda_{\mathrm{I}}) \prod_{S_j \in Y^1} P(S_j \mid \lambda_{\mathrm{II}}). \end{aligned} \tag{4.7}$$

The probability of having a segment S from a pre-trained speaker model λ is the product of likelihoods of that speaker model generating each individual observation frame X of the segment, assuming the observations are independent of each other. In other words,

$$P(S \mid \lambda) = \prod_{X \in S} p(X \mid \lambda). \tag{4.8}$$

4.2.2 Computational methods

The computational objective in (4.5) is to find the optimal hypothesis of two speakers and one assignment that yield the maximal probability using (4.6)-(4.8). Given the extracted usable speech segments and individual speaker models trained from clean speech, the likelihood maximization amounts to a search for the globally optimal hypothesis in the joint speaker and assignment space, Λ and G.

A. Exhaustive search

The brute-force way to find the maximum is exhaustive search. For a cochannel mixture file, this involves calculating the probability of the assigned segments given a pair of speaker models, $P(Y \mid g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$, for every possible pair out of K speakers in Λ and every assignment in G. Let the calculation of $P(Y \mid g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$ take a unit time, then total computation time is on the order of $O(K^2 \cdot 2^N)$. However, according to (4.6)-(4.7), once an assignment is given, the likelihood maximization is simply finding the best speaker for each segment subset, and corresponding likelihood values are then multiplied, resulting in a complexity of $O(K \cdot 2^N)$.

Similarly, for a given pair of speakers, the likelihood maximization amounts to finding the best assignment for each segment, and the overall probability is the product of these segment likelihood values. The speaker pair with the highest probability gives the

search result together with its associated segment assignment. This way the search complexity is reduced to $O(K^2 \cdot N)$. In implementation, the real computation time can be further reduced by storing all the likelihood scores of a segment given a model in the memory as a table and looking up a score from the table when needed. This implementation avoids repetitive computations of the same score and swaps some memory space for computation time.

B. Hypothesis pruning

In the search space, some hypotheses have very low probabilities. Therefore, if these hypotheses could be identified and pruned from further consideration, the computation time could be greatly reduced. The results of exhaustive search indicate peaky distributions with each peak occupied by several assignment hypotheses in the search space. Thus, keeping a small number of hypotheses could be sufficient. If we associate two states with each segment, representing the hypotheses that the segment is labeled as 0 or 1, a trellis is formed from the first segment to the last one, whose paths represent all the possible assignments of the segments. This way, the search amounts to finding the best path in the trellis, and the hypotheses with low probabilities can then be pruned. We propose an iterative hypothesis pruning algorithm to keep only the two best hypotheses in each iteration. More specifically, the first segment is arbitrarily labeled and starting from the second segment, only two hypothesis states are retained corresponding to the current segment being labeled as either 0 or 1. The better path (out of the two) leading to each state is selected, and path selection is based on SID scores given the partial assignment.

After the last segment is labeled, the best out of the two hypothesis states is then chosen; the best path from the first segment to the last is constructed from the chosen paths at all preceding iterations. This algorithm can be viewed as finding the best path via Viterbi decoding. The evaluation results in the next section show that the proposed algorithm achieves a level of performance close to that of exhaustive search.

For each unlabeled segment, it retains two hypotheses, each of which calculates $P(Y \mid g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$ twice in the worst case, resulting in the polynomial time complexity on the order of $O(K \cdot N)$. The computation time could be further reduced by skipping the pairs of speakers whose partial scores are below a threshold or much lower than others. We give the detailed algorithm as follows.

Hypothesis pruning algorithm:

Step 0. Order the segments in $Y = \{S_1, S_2, \ldots, S_N\}$ sequentially in time.

Step 1. Label S_1 in Y with 0 (assign it to Y^0). This initial assignment is arbitrary.

Step 2. For S_2 in Y, form two hypotheses: H_0, H_1, and create a label path for each of them. H_0 assumes that the current segment belongs to set Y^0, and H_1 assumes that the current segment belongs to Y^1. The label paths are

$$Path[2][H_0] = (0,0),\ Path[2][H_1] = (0,1).$$

$Path[n][\bullet]$ records assignment labels for the past n-1 segments and the hypothesized assignment of the current segment.

Step 3. For an unprocessed segment S_n, n>2, form H_0 and H_1. Then expand the label path for H_0 and H_1 as follows,

$$Path[n][H_0] = \left(Path[n-1]\left[\arg\max_{H \in \{H_0, H_1\}} L\big(Path[n-1][H], 0\big) \right], 0 \right)$$

$$Path[n][H_1] = \left(Path[n-1]\left[\arg\max_{H \in \{H_0, H_1\}} L\big(Path[n-1][H], 1\big) \right], 1 \right)$$

where the L function, as defined below, estimates the joint SID score by considering the best partial segment assignment from 1 to n.

$$L\big(Path[n-1][H], l\big) = \max_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda} P(Y \mid (Path[n-1][H], l), \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}), \tag{4.9}$$

l = 0 or 1, refers to the hypothesized labeling for the current segment.

Step 4. Repeat Step 3 until the last segment S_N is processed. For S_N, compare the likelihood values returned by L for H_0 and H_1. The final winning hypothesis is the one with the higher likelihood. Obtain the corresponding two speaker identities that maximize (4.9) and the segment assignment for this hypothesis.

The L function in (4.9) is the same as (4.6) except that L only considers the partial segment assignment from S_1 to S_n. Figure 4.4 gives an illustration of this iterative algorithm. Since every usable segment could be produced by either of two speakers in the mixture, it is hypothesized as either H_0 or H_1 and labeled with 0 or 1 respectively (S_1 is initialized to hypothesis H_0). The two hypothesis states bifurcate iteratively and our

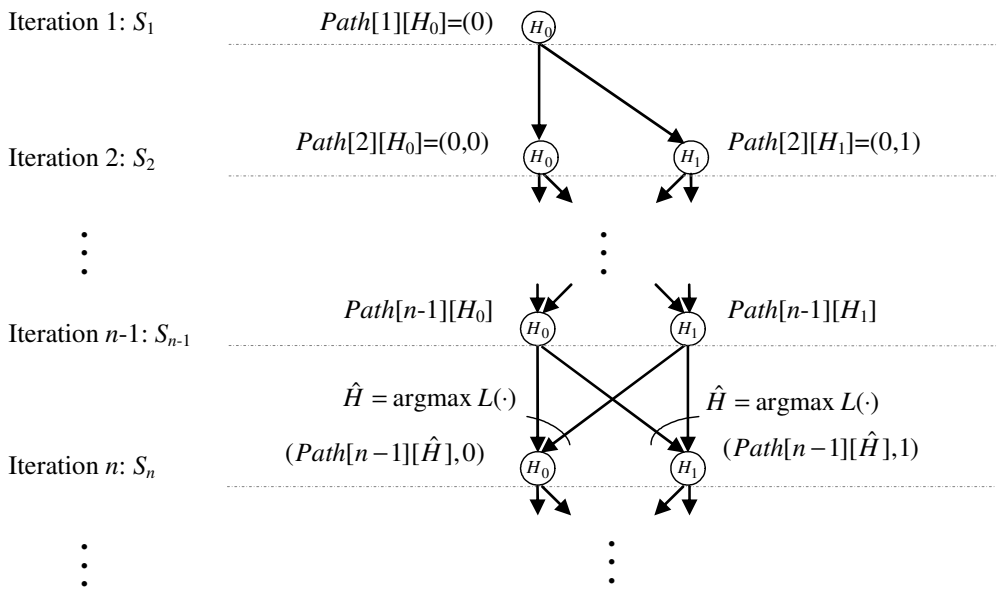

Figure 4.4: Illustration of the hypothesis pruning algorithm. The algorithm is executed segment by segment. Every segment is hypothesized to be either H_0 or H_1 and labeled with 0 or 1 respectively, except that S_1 is identified with hypothesis H_0. *Path* records the best label path. For either hypothesis of the segment to be considered, the better label path from the preceding iteration is chosen by comparing $L(\cdot)$ defined in (4.9), and its label path is copied to the current path. The algorithm repeats until the last segment is processed.

pruning algorithm always retains the best path to a state and is recorded in *Path*. For each state, we compare the partial SID scores, considering the label paths recorded with the preceding hypothesis states. The SID score is defined by the L function in (4.9). The better path is then chosen. The algorithm repeats until the last segment is processed.

The joint maximization for sequential grouping amounts to maximization over speaker pairs and maximization of segment assignments for a specific speaker pair. In (4.7), we assume segment independence given a speaker pair. With this assumption, making a local decision to assign a segment to either speaker in a pair guarantees optimality within a speaker pair. However, it does not guarantee global optimality among all the speaker pairs. Since the pruning algorithm prunes assignments of a subset of segments, the hypothesis pruning algorithm may not find the optimal hypothesis.

Here an example is presented to illustrate the difference between the exhaustive search algorithm and the hypothesis pruning algorithm. Assume that there are 4 segments to be grouped $\{S_1, S_2, S_3, S_4\}$ and 3 speakers {A, B, C}. The likelihoods of a segment produced by a speaker are calculated and stored in a table as follows.

	A	B	C
S_1	0.7	0.6	0.65
S_2	0.5	0.8	0.5
S_3	0.1	0.4	0.5
S_4	0.8	0.5	0.6

Using the exhaustive search algorithm, we first evaluate the best segment assignment for each pair of speakers.

Exhaustive Search

	A-B: L=0.1792		A-C: L=0.14		B-C: L=0.156	
Segment	Label	Speaker	Label	Speaker	Label	Speaker
S_1	0	A	0	A	1	C
S_2	1	B	0	A	0	B
S_3	1	B	1	C	1	C
S_4	0	A	0	A	1	C

Apparently, speaker pair A-B yields the highest overall likelihood and its segment grouping, (0 1 1 0), is selected as output.

The hypothesis pruning algorithm initializes the first two segments as follows,

Hypothesis Pruning:

L=0.48	
Label	Speaker
0	B
0	B

S_2 - H_0

L=0.56	
Label	Speaker
0	A
1	B

S_2 - H_1

Note that the first segment is labeled 0 and hypotheses, H_0 and H_1, are constructed for the second segment, meaning that this segment is labeled 0 and 1 respectively. For H_0 and H_1 of the third segment, we evaluate their hypotheses using (4.9).

L=0.192		**L=0.26**	
Label	Speaker	Label	Speaker
0	B	0	C
0	B	1	B
0	B	0	C

S_3 - H_0

L=0.24		L=0.224	
Label	Speaker	Label	Speaker
0	B	0	A
0	B	1	B
1	C	1	B

S_3 - H_1

For H_0 of S_3, we select the better assignment (0 1 0). For H_1, (0 0 1) is the chosen one. Note that the assignment (0 1 1) has been pruned here. Thus, the best grouping (0 1 1 0) will not be produced by the hypothesis pruning algorithm in the end. For the fourth segment, it is straightforward to show,

***L*=0.156**		*L*=0.14	
Label	Speaker	Label	Speaker
0	C	0	A
1	B	0	A
0	C	1	C
0	C	0	A

S_4 - H_0

L=0.13		***L*=0.144**	
Label	Speaker	Label	Speaker
0	C	0	B
1	A	0	B
0	C	1	C
1	A	1	C

S_4 - H_1

We select (0 1 0 0) for H_0 of S_4 and (0 0 1 1) for H_1. Finally, the algorithm outputs: speaker pair B-C and segment assignment (0 1 0 0). Its overall likelihood is 0.156, close to likelihood, 0.1792, of the optimal hypothesis by exhaustive search. Its segment assignment (0 1 0 0) differs from the optimal hypothesis (0 1 1 0) for the label of the third segment.

C. Alternative methods

We have also explored a number of variations of the hypothesis pruning algorithm. Because the algorithm prunes certain paths, it resembles beam search (Russell and Norvig, 2003). In the evaluation section, we also examine a beam search algorithm with width of 1 or 2 for performance comparison with preceding methods.

If the main objective is cochannel speaker identification, rather than sequential organization, a comprehensive approach is to directly identify speaker pairs from a closed set. One way of formulating the problem is to omit the assignment variable from the computational goal and replace usable speech segments by mixture itself. This may be viewed as integrating over the speaker assignment variable, and hence can produce the maximum SID performance. To reduce the computational complexity associated with

training speaker-pair models, one approximation is to model a speaker-pair model by simply merging two corresponding single-speaker models: $p(O \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) = 0.5\big(p(O \mid \lambda_{\mathrm{I}}) + p(O \mid \lambda_{\mathrm{II}})\big)$. In other words, the joint likelihood of a mixture utterance is taken to be the average of the likelihoods given by each constituent model. This method is denoted as combined GMM and we also evaluate its SID performance.

4.2.3 Evaluations

In this section, we present the evaluation results of the described sequential organization methods together with some alternative approaches by performing a SID task under cochannel conditions.

We employ the same evaluation data as used in cochannel speaker recognition in Section 3.2. Specifically, the speaker set consists of 38 speakers from the "DR1" dialect region, 14 of which are females and others are males. Each speaker has 10 utterance files, ranging from about 1.5 *sec* to 6.2 *sec* in length. For each speaker, 5 out of 10 files are used for training and the remaining 5 files are used to create cochannel mixtures for testing. For each speaker deemed as the target speaker, 1 out of 5 test files is randomly selected and mixed with randomly selected files of every other speaker, which are regarded as interfering utterances. The interfering utterance is either cropped or concatenated with itself to match the length of the target utterance and it is scaled to create the mixtures at different TIRs. For example, 0 dB TIR means that the target speech overall energy is equal to that of the interfering speech. Thus, for each TIR, a total of

1406 cochannel mixture files are created for the testing purpose. In the experiments, speakers are modeled as 16-mixture GMMs, which are tested to be sufficient for the data, and the observations or features used are MFCCs and their first-order dynamic coefficients (Young *et al.*, 2000). Note that no background model is used.

This experiment evaluates the performance of our model-based sequential organization approach. For this evaluation we only consider cochannel mixtures with overall TIR equal to 0 dB to simulate real cochannel situations. To facilitate a better understanding and comparison, we combine the evaluation results into a single table, Table 4.1, including the results from the alternative methods and the feature-based grouping methods.

The 2nd column in Table 4.1 shows the correct rate of speaker assignment by counting correctly assigned frames. To calculate the ratio, the denominator is the total number of extracted usable speech frames. To find the numerator, the two sets of usable frames labeled by the system as 0 and 1 are compared with the two ideal sets labeled with single-speaker pitch points derived from premixing utterances. There are two possible correspondences between the two system-labeled sets and the two ideal sets, and for each correspondence the number of matching frames is recorded. The larger number out of the two correspondences is used as the numerator. Note that the SID performance does not impact the speaker assignment results.

The 3rd and 4th column in the table present the SID performances with two different criteria. Like the evaluation in the preceding section, the speaker from a specified channel – target speaker – can be of interest. Thus, the first criterion measures target identification

Method	Frame Assignment Accuracy (%)	Speaker Identification Accuracy (%)	
		Target	Target & Interferer
Random Assignment	50.0	N/A*	N/A
Ideal Assignment by Prior Pitch	94.1	72.0	43.3
Exhaustive Search	**77.4**	**70.4**	**40.2**
Hypothesis Pruning	**76.2**	**68.8**	**37.5**
Conventional SID	N/A	57.2	13.1
Hypothesis Pruning (open set)	73.0	68.4	N/A
Beam Search (beam = 1)	66.0	51.5	21.0
Beam Search (beam = 2)	76.0	68.1	37.2
Combined GMM	68.2	76.9	48.7
Pitch Dynamics	68.2	52.5	22.3
Spectral Divergence	66.2	N/A	N/A

*N/A: unavailable.

Table 4.1: Grouping accuracy for sequential organization and cochannel speaker identification accuracy.

correct rate. The second criterion records the percentage of mixtures where both speakers are correctly identified; this is the more stringent criterion.

In the table, the baseline rate of correct grouping corresponding to random labeling of each usable frame is 50.0%. The 2nd row shows that ideal assignment by prior pitch achieves 94.1% correct rate. Note that ideal assignment is applied at the segment level: A segment takes the label of a majority of the frames in the segment, where each frame is labeled by comparing the detected pitch with the prior pitch before mixing. The less-than-

perfect result reflects that a single-pitch segment does not always contain frames from the same speaker, which is expected considering the nature of cochannel speech.

Exhaustive search achieves 77.4% correct assignment rate. It reflects the effectiveness of using speaker characteristics for sequential organization. From the derivation it is evident that exhaustive search places an upper limit on the performance of model-based sequential grouping. Our proposed hypothesis pruning method achieves 76.2% correct rate, approaching the upper limit set by exhaustive search.

Table 4.1 also gives the evaluation results for pitch-based and spectrum-based grouping as described earlier in the chapter. The method that uses pitch dynamics clearly performs worse than the pruning algorithm, but produces a significant improvement over the baseline case without usable speech processing. The grouping method based on spectral divergence yields 66.2% correct rate for grouping, comparable in performance to the pitch dynamics method, but it is less effective than our proposed method. As a result the SID results are not shown.

When the beam search algorithm is applied with a beam width of 1, it yields assignment accuracy of 66.0%, which are significantly worse than the pruning algorithm. In the case where the beam width is 2, this method produces results close to those obtained by the hypothesis pruning algorithm.

The combined GMM methods ignores sequential grouping and identifies underlying speaker by a combined speaker pair. Its SID performance is higher than the proposed method that considers speaker assignment. Part of the reason for the better performance is that usable, or single-pitch, frames may still contain energy from both speakers and

forcing a decision of one speaker may degrade identification performance. Of course, correct recognition of a speaker pair does not lend itself to sequential organization directly. However, with the recognized speaker pair, each usable speech frame can be classified into the two speaker sets by comparing its likelihood values given the speaker models. This way, the combined GMM method achieves 68.2% correct assignment rate, lower than that of the hypothesis pruning method.

In terms of SID accuracy, the baseline performance is taken to be identification accuracy by recognizing individual speakers directly. In this case, the two SID criteria document the top two identified speakers. Ideal assignment produces much higher SID performance though it is not 100% correct because of imperfect assignment and limited segment lengths. For the model-based approach, exhaustive search approaches the ceiling SID performance with ideal assignment, and the hypothesis pruning method performs almost as well as exhaustive search, while cutting the overall computation time from an average of 0.491 seconds per file to 0.037 seconds on a Pentium III workstation (The computation time for the exponential version of exhaustive search is on average 7 minutes per file.) Since the search is based on SID scores, the performance gap between the model-based method and ideal assignment is smaller than that of sequential grouping performance.

In the formulation of sequential organization and SID, we assume both speaker models are available – a closed-set situation. To test how the algorithm functions in an open-set situation, we apply the hypothesis pruning algorithm on cochannel speech where one speaker is not registered. This is a task of identifying a familiar speaker in cochannel

mixtures where no model is available for the interfering speaker. For this experiment, the same mixture files are used as in previous evaluations. Specifically, for each test mixture, we remove the corresponding interferer model from the speaker set. In this case, only the SID criterion for target speaker is applicable. The corresponding results are 73.0% for correct assignment and 68.4% for target speaker identity (see also Table 4.1). These results are not much worse than in the closed-set situation. We suspect that the coherence of speaker features in an utterance enables the selection of a speaker model from the registered speakers that is closest to the unregistered speaker. Of course, when none of the two speaker models are known, it would not make sense to use a model-based approach and other methods such as pitch-based organization introduced earlier should be explored instead. We will discuss organization under such conditions in the next chapter.

While comparing average results of different methods, it is useful to note statistical significance. With 1406 test utterances a one-tailed test for the recognition accuracy at around, say, 68.8% requires about 2.9% difference for statistical significance at 5% level (Gillick and Cox, 1989). This suggests, for example, that the performance difference in target speaker recognition between the hypothesis pruning algorithm and exhaustive search is not statistically significant. For speaker assignment performance it is more difficult to construct a statistic for the hypothesis test because frame-level decisions are not independent within segments.

4.3 Incorporating Binary T-F Masks

As described in Chapter 1, the main goal of this dissertation is develop algorithms for sequential grouping in the CASA framework. The outputs of sequential grouping are streams represented by binary time-frequency masks. The inputs are the simultaneous streams obtained by segmentation and simultaneous grouping. The usable speech segment used in the preceding sections comprises a sequence of frames which are deemed to be speaker homogeneous, which reflects one form of the simultaneous stream. In this chapter, we employ the monaural speech segregation system as described in Section 2.3.3 to generate simultaneous streams (Hu, 2006). This system estimates ideal binary T-F masks of the streams. We modify the model-based sequential grouping system to replace the usable segments with the binary T-F masks. In addition, besides voiced speech, unvoiced segments are also extracted based on an onset/offset segmentation method (Hu and Wang, 2007) for sequential grouping. We evaluate the system using a recent speech separation and recognition task (Cooke and Lee, 2006). The objective of this task is to segregate two-talker mixtures and recognize keywords in the separated target.

4.3.1 Extended sequential grouping algorithm

In sequential grouping of usable speech segments, the algorithm groups segments based on aggregated likelihood scores of a speech frame given a model as $p(X \mid \lambda)$ in Equation (4.8). On the other hand, a binary T-F mask that represents a simultaneous stream indicates reliable and unreliable T-F units. To incorporate the binary masks for

sequential grouping, we employ the same feature reconstruction and uncertainty decoding method as described in Section 3.4.2. Specifically, we reconstruct a GF frame using (3.13) and (3.14), and estimate its uncertainties using (3.15). The enhanced GF and uncertainty estimates are transformed into the GFCC domain for the calculation of the likelihood of $p(X \mid \lambda)$ using the uncertainty decoder in (3.17).

Besides the above modifications, the exhaustive search algorithm described in Section 4.2.2 is employed to group the simultaneous streams. The organized streams are then combined to produce binary T-F masks that represent segregated speaker streams.

4.3.2 Unvoiced segmentation and grouping

The simultaneous streams correspond to voiced speech. In natural speech, unvoiced speech constitutes a smaller portion of an overall utterance than voiced speech but it contains important phonetic information (Wang and Hu, 2006). Therefore, if automatic speech recognition is the intended application, it is also important to segment and group unvoiced speech.

Unvoiced speech lacks the harmonic structure, and as a result is more difficult to segment. Here we employ an onset/offset based segmentation system (Hu and Wang, 2007). This system has three processing stages: Smoothing, onset/offset detection, and multiscale integration. In the first stage, the system smoothes the cochleagram of an input using a Gaussian smoothing process. In the second stage, the system detects onsets and offsets in each filter channel and then merges simultaneous onsets and offsets from adjacent channels into onset and offset fronts, which are defined as vertical contours

connecting onset and offset candidates across frequency. Segments are generated by matching individual onset and offset fronts. The smoothing operation may blur event onsets and offsets of small T-F regions at a coarse scale, resulting in the loss of some true onsets and offset. On the other hand, the detection process may be sensitive to insignificant intensity fluctuations within events at a fine scale. Thus, the cochleagram may be under-segmented or over-segmented because of detection errors. In order to produce satisfactory segmentation, segments are produced at four different scales and integrated subsequently (for further details see Hu and Wang, 2007).

Since onsets and offsets correspond to sudden intensity increases and decreases which could be triggered by voiced speech or unvoiced speech, the obtained segments usually contain both speech types. Additionally, the mixing of sources leads to blurring and merging of onset/offset fronts. Thus, matching onset and offset fronts creates segments that may not be source homogeneous. Here, we extract the unvoiced segments from the onset/offset segments by removing those portions that overlap with the simultaneous streams.

Voiced speech likely plays a dominant role in sequential grouping and speaker recognition (see e.g. Shao and Wang, 2006a). Therefore, for a cochannel mixture, we first apply the model-based sequential grouping algorithm to organize the simultaneous streams, producing two binary masks (streams) and corresponding speaker identities. Secondly, unvoiced segments are grouped with the two streams using the above sequential grouping algorithm except that the system uses the detected speaker pair associated with the masks.

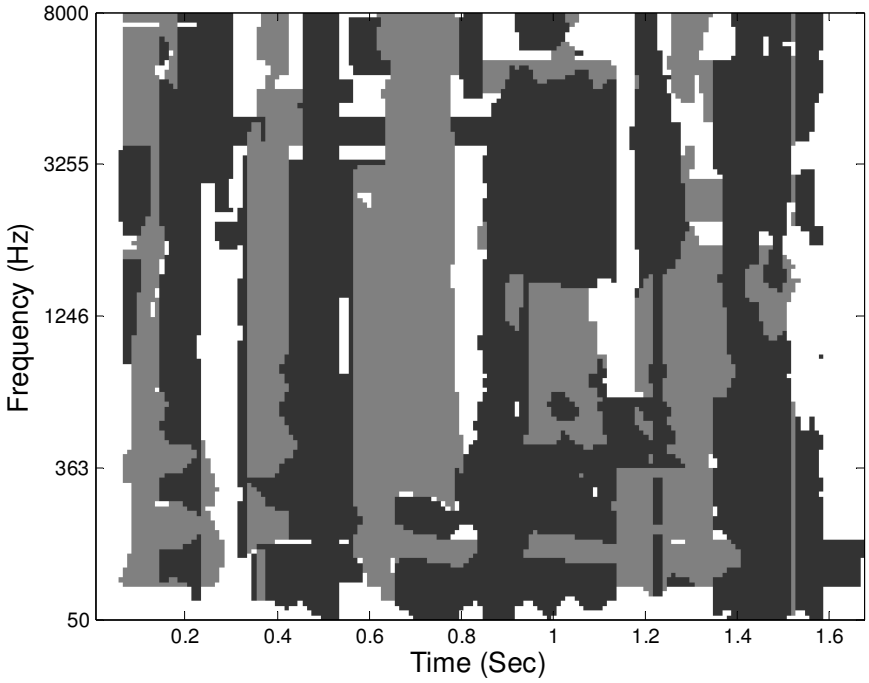

Figure 4.5: The estimated speaker streams after sequential grouping of simultaneous streams and unvoiced segments. White color shows the background. The two gray-colored regions represent two separated speaker streams.

We find that unvoiced segments are typically much smaller than simultaneous streams, thus resulting in poor likelihood estimation by GFCC reconstruction and uncertainty decoding. Therefore, likelihoods are calculated using the marginalization method which ignores the missing T-F units as specified in Equation (3.8). Figure 4.5 presents the separated speaker streams after grouping simultaneous streams and unvoiced segments. The two speaker streams are shown in two different gray colors.

We find that the onset/offset analysis does not capture all the speech segments. Therefore, to refine the binary masks, we apply a watershed algorithm (Vincent and Soille, 1991) to the cochleagram and extract fragments that comprise T-F units with similar energy values. A resulting segment is assigned to one of the aforementioned speaker streams if its mask largely (greater than two-thirds) overlaps with its binary mask. This step assumes that a small segment of connected T-F units with close energy values is produced by the same speaker. Subsequently, if a segment has not been merged, its overlapped portions with either of the two streams, if any, are removed from its mask. Finally, the remaining segments are grouped with the refined masks using the sequential grouping algorithm and the detected speaker pair.

4.3.3 Speech separation and recognition evaluation

We evaluate our system on the speech separation and recognition task (Cooke and Lee, 2006). One of the goals of this task is to recognize speech from a target talker in the presence of a competing speaker. This noisy condition is essentially cochannel speech. The signals are sampled at 25 kHz and every utterance follows a sentence grammar of

$command $color $preposition $letter $number $adverb.

There are 4 choices each for $command, $color, $preposition and $adverb, 25 choices for $letter (A-Z except W), and 10 choices for $number (1-9 and zero). For example, a valid utterance could be “Place blue at F 2 now”. The possible choices in each position are roughly uniformly distributed in the corpus. The two-talker task is to identify the letter and the number spoken by the talker who said the keyword color, “white”. The

speech recognition task is to identify the color, the letter and the number. The training data is drawn from a closed set of 34 talkers and consists of 17,000 utterances in total. The two-talker test data contains pairs of sentences mixed at 6 different target-to-masker ratios (TMRs): -9, -6, -3, 0, 3 and 6 dB. Note that TMR is the same as TIR. One third of this data consists of same talker (ST) mixtures, another third comprises of mixtures of different talkers of the same gender (SG), and the remaining third consists of different gender (DG) mixtures.

To build speaker models, we utilize the GFCC feature as described earlier. Each of the 34 speaker models comprises 64 mixtures of Gaussians. The speech prior model is trained on GF features and comprises 2048 Gaussian mixtures. This prior model and the binary masks are used in the cochleagram domain to reconstruct missing T-F units. The reconstructed GFs are then transformed into the GFCC domain using DCT. For recognition, we form the 60-dimensional feature vector of GFCC_D, including delta coefficients calculated using a sliding window of 5 frames. GF uncertainties are also transformed into the cepstral domain since DCT is a linear transformation. Whole-word HMM-based speaker-independent ASR models are then trained on clean speech; each word model comprises 8 states and 32 Gaussian mixtures with diagonal covariance in each state. The uncertainty decoder also uses diagonal covariance for uncertainties. During the recognition process, given the estimated uncertainties and the clean ASR, the uncertainty decoder calculates the likelihood of the reconstructed GFCC_D features and transcribes the speech.

Since the overall segregation system does not rely on the content information in an utterance, the system does not know which separated stream contains "white" in the two-talker task. In order to select the target, we employ a normalized scoring method. We let our uncertainty decoder recognize both segregated streams using two different grammars *W* and *NW*:

W: *$command white $preposition $letter $number $adverb*.

NW: *$command $non-white $preposition $letter $number $adverb*.

$non-white has 3 choices of colors except white. A normalized score is calculated for each stream by subtracting the recognition likelihood score of *NW* from the one using grammar *W*. The stream with a larger score is chosen as the target, i.e., stream 1 (s_1) is chosen as the target when

$$P_W(s_1) - P_{NW}(s_1) > P_W(s_2) - P_{NW}(s_2), \tag{4.10}$$

or stream 2 (s_2) if otherwise. This selection metric is actually the same as evaluating the joint likelihood score of one stream containing the keyword ``white'' while the other containing $non-white. (4.10) is the same as,

$$P_W(s_1) + P_{NW}(s_2) > P_{NW}(s_1) + P_W(s_2). \tag{4.11}$$

The evaluation results of our proposed speech segregation system on the two-talker task are summarized in Table 4.2. The performance is measured in terms of recognition accuracy of the relevant keywords at each TMR conditions. We report the results for the different gender (DG), the same gender (SG) and the same talker (ST) subcategories as well as the overall mean score (Avg.). For comparison, we also show the performance of a baseline system without segregation. The proposed system improves significantly over

TMR(dB)/Systems		DG	SG	ST	Avg.
6	Baseline	66.00	65.92	66.52	66.17
	Proposed	80.75	76.81	54.98	70.08
3	Baseline	51.25	49.44	51.58	50.83
	Proposed	78.50	72.63	39.14	62.25
0	Baseline	36.00	34.64	32.58	34.33
	Proposed	74.50	67.31	25.34	54.25
-3	Baseline	19.25	22.07	18.55	19.83
	Proposed	63.50	53.07	20.59	44.58
-6	Baseline	9.50	10.34	9.50	9.75
	Proposed	48.00	36.31	17.19	33.17
-9	Baseline	3.25	4.75	3.62	3.83
	Proposed	32.00	22.34	11.99	21.75

Table 4.2: Recognition accuracy (in %) of the baseline system and the proposed system on the two-talker task. DG, SG and ST refer to sub-conditions of "different gender", "same gender" and "same talker" respectively. Avg. is the mean accuracy.

the baseline system in terms of average accuracy across all TMR conditions. Larger improvements are observed in the DG and the SG conditions. However, the system does not perform nearly as well in the ST condition, which is not a realistic condition. This is primarily due to our use of speaker characteristics for sequential grouping. Note that for the ST condition, speaker characteristics are not distinctive for segregation. Figure 4.6 compares the system performance with (w/) and without (w/o) the ST condition. Note that baseline performance is nearly the same with ST and without ST. Our system achieves further absolute improvement of over 11% on average in the without-ST condition over the with-ST condition.

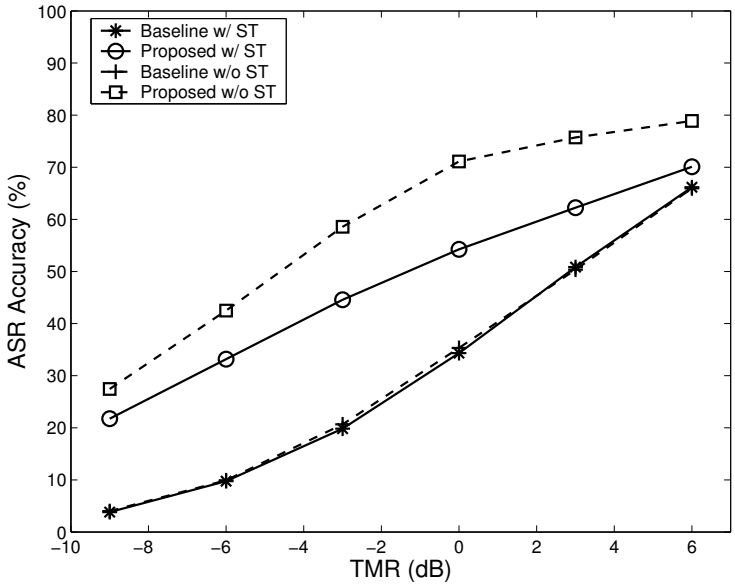

Figure 4.6: Recognition accuracy on the two-talker task. The solid star line represents our baseline recognition results. The dashed plus line shows the baseline performance without the same talker (ST) data. The results of the proposed system are given as the solid circle line. Its accuracy without the ST condition is presented as the dashed square line.

Since our sequential grouping algorithm also identifies the underlying speakers, we present the evaluation results of SID performance in Table 4.3. Note that for most of the TIR conditions, we achieve an accuracy of over 90% in recognizing the target speaker.

TMR(dB)	**-9**	**-6**	**-3**	**0**	**3**	**6**
Both SID	12.83	33.50	57.50	65.33	63.17	46.17
Target SID	57.17	89.50	98.17	99.50	99.83	99.33

Table 4.3: Speaker identification (SID) accuracies in the two-talker task. “Both SID” shows the accuracies when both speakers in a mixture are identified correctly. “Target SID” presents the accuracies when the target speaker is identified as either of the SID outputs.

CHAPTER 5

SEQUENTIAL GROUPING USING GENERIC MODELS

In the preceding chapter, we have shown model-based sequential grouping methods under cochannel speech conditions. In this chapter, we extend the sequential grouping system to acoustic conditions that include unknown interference sources. First, we generalize the grouping algorithm to deal with multi-talker scenes that are composed of more than two talkers. The algorithm is then extended to handle non-speech intrusion sources. We also show that the generalized system is able to function without interference models. Furthermore, good performance is achieved regardless of interference types and numbers. Finally we apply a speaker quantization method to a large speaker space, and use the obtained generic models for sequential organization of speech mixtures of unknown speakers. In other words, the system does not rely on any *a priori* knowledge of sources in an auditory scene. Our systematic evaluations show that the resulting performance is only moderately lower than the performance achieved with known speaker models.

Depending on task specifications, we have evaluated the sequential grouping methods using several performance metrics. For example, we have used the word accuracy of automatic speech recognition in a speech separation and recognition task. In this chapter, as an alternative, we evaluate grouping performance by comparing estimated binary masks to the ideal binary masks. Specifically, we adopt the SNR metric that compares the target signal $s(n)$ resynthesized from the ideal binary mask and the organized target signal $\hat{s}(n)$ resynthesized from an estimated binary mask in decibels (Hu and Wang, 2006). This measure directly compares signals in the time domain as

$$SNR = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n [s(n)-\hat{s}(n)]^2}, \tag{5.1}$$

where n indexes time.

5.1 General Modeling of Interferences

By definition, a cochannel mixture is composed of voices from two talkers. The voice of interest is designated as target and the other as interference. Under certain circumstances such as a meeting, there may be more than one interference speaker. To tackle such conditions, we extend the model-based sequential grouping method by replacing the speaker pair with a speaker triplet or a speaker quadruplet in Equation (4.3). We shall end up with a computational objective similar to (4.5) by applying the same derivation in (4.4). Specifically, given a set of K registered speaker models $\Lambda = \{\lambda_1, \lambda_2, \ldots,$

$\lambda_K\}$ and a set of simultaneous streams Y, an optimal segment assignment g to M speakers can be found as,

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}}, ... \hat{\lambda}_M = \underset{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, ... \lambda_M \in \Lambda, g \in G}{\arg\max} P(Y \mid g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, ... \lambda_M), \tag{5.2}$$

where $0 \le M \le K$. Naturally, the components of g take values from 1 to M. This objective leads to the same hypothesis search in Section 4.2.2 except that the speaker pair is replaced by M speakers.

The above extension makes an explicit assumption of speaker number in a mixture. However, this assumption may not always be satisfied, leading to open-set auditory scenes. To handle such conditions, we may further extend the formulation in (5.2) and include another search that evaluates different speaker numbers. In other words, the grouping algorithm evaluates the best hypotheses for one, two, three, … , and a sufficiently large number of speakers. The speaker number that yields the highest likelihood is chosen as the estimate. The grouping hypothesis associated with the speaker number estimate provides the optimal segment assignment.

Nevertheless, this extension is not scalable. For example, in the case of a cocktail party, there may be a large number of speakers in the background. Indeed, there are so many voices in the background that a listener perceives something more like babble noise. Hence, according to the complexity analysis in Section 4.2.2, searching through all the combinations of up to M speakers results in the complexity of $O(2^M)$. Furthermore, given a large M, it is unreasonable to assume knowing all the speakers *a priori* in a task. Therefore, this extension is not only unscalable from the algorithmic point of view but also unattainable from the practical point of view.

To deal with multi-talker conditions, we may gain some inspirations by studying how existing CASA systems treat interferences (Wang and Brown, 2006). Typically, target signal is segregated into a foreground (target) stream while the remaining of the input signal is organized into the background (interference) stream. This process holds regardless of actual interference source types or numbers. Hence, instead of modeling individual speakers, we propose to build a generic model that accounts for all interference sources. This generic model is constructed by training on a large sample pool of speakers. Conceptually, it is analogous to the universal background model (UBM) in Section 3.1.1. Thus, in the sequential grouping algorithm that searches speaker pairs, we replace one of them with the generic model and perform the search over the other speaker as follows.

$$\hat{g}, \hat{\lambda} = \arg\max_{\lambda \in \Lambda, g \in G} P(Y \mid g, \lambda, \lambda_{\mathrm{G}}) \tag{5.3}$$

5.1.1 Multi-talker intrusions

To simulate the multi-talker conditions, we create the test utterances based on the SSC corpus (Cooke and Lee, 2006). The SSC corpus provides 600 clean utterances in the test set. We use these utterances to generate mixtures of a target speaker and multiple interference speakers. The two-talker mixtures in the SSC corpus are not used here because SSC does not provide functions to create mixtures and our own method may differ in implementation details. For our tests, we construct two-talker, three-talker and four-talker mixture conditions. Specifically, for each utterance deemed as target, one, two or three utterances are randomly selected from other speakers in the clean set and mixed with the target. An interference utterance is either curtailed or appended with itself to

match the length of a target utterance. The multi-talker corpus comprises a wide range of TIRs, including -6 dB, 0 dB, 6 dB, 12 dB and 18 dB. Similar to Equation (3.6), TIR is calculated as follows,

$$TIR = 10\log\left(\sum_n s_T^2(n) \Big/ \sum_n \left(\sum_{c=1}^{C} s_{I,c}(n)\right)^2\right), \tag{5.4}$$

where n indexes time. C is the number of interference talkers and c is its index. s_T refers to target signal and $s_{I,c}$ refers to signal of the cth interference talker. This formula calculates the energy ratio of the target utterance and all the interferences combined. Note that the interference utterances have been scaled to have equally strong energy.

As in the study for the SSC task, we employ the voiced speech segregation system (Hu, 2006) for segmentation and simultaneous grouping (see Section 2.3.3 for details). This system estimates ideal binary T-F masks for the simultaneous streams and also produces corresponding pitch contours. In the sequential grouping algorithm, we employ the same GFCC reconstruction and uncertainty decoding method to calculate the likelihood score of a simultaneous stream given a model as described in Sections 3.4 and 4.3.

For performance evaluation, we construct the ideal binary masks based on Equation (2.1) and the mixture creation process as described at the beginning of this section. Based on the ideal mask, we generate an ideal sequential grouping (ISG) mask for each mixture by grouping simultaneous streams into the target stream according to its ideal binary mask. Specifically, a simultaneous stream is grouped as target if more than half of its

energy is retained by the ideal mask. This ISG mask presents the best mask that a sequential grouping algorithm can produce, thus reflecting an upper-bound performance.

We evaluate the effectiveness of our grouping algorithm using the SNR metric in Equation (5.1). The experimental results on the multi-talker corpus are presented in Tables 5.1 and 5.2. The first rows of the tables show the SNR results obtained by ideal sequential grouping. The second rows present baseline performance by randomly assigning a stream to either the target or the interference stream. An ISG mask ideally groups simultaneous streams, which are extracted from voiced regions of an input utterance (Hu, 2006). Hence, errors in simultaneous grouping, including the removal of unvoiced speech, are inherited in an ISG mask. Because of this, output SNRs of ISG masks are less than input SNRs under 12 dB and 18 dB conditions.

As feature-based methods, we perform sequential grouping using pitch information. We first evaluate their performance based on prior pitch information. Specifically, prior pitch contours are extracted from clean target utterances. A simultaneous stream is assigned to the target stream if the average difference of its pitch contour and a prior contour is within 5% range of the latter. The resulting performance is reported in the row denoted as 'Grouping Using Ideal Pitch'. This performance places an upper-bound for all the methods that utilize pitch. Since the pitch-based grouping algorithm in Section 4.1.1 cannot be directly applied under the multi-talker conditions, we employ a clustering method that is based on the mean pitch values of the segments. In addition, the number of clusters is set to the speaker number in a test mixture. SNR results are shown in the 'Pitch-based Grouping' row. The results are worse than the performance upper-bound

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	3.006	5.793	8.968	10.937	11.801
Random Grouping	-3.648	0.873	2.566	3.334	3.514
Grouping Using Ideal Pitch	2.298	3.931	5.552	6.860	8.039
Pitch-based Grouping	-0.622	3.121	5.233	6.158	6.443
Known Speaker No. w/ Aggregated Prior	2.112	5.157	7.792	9.510	10.122
Known Speaker No. w/ Individual Prior	2.021	5.005	7.797	9.802	10.562
Grouping Using Generic Model	1.686	4.533	7.396	9.639	10.577
Grouping Using Combined Generic Model	1.323	3.825	7.012	9.117	10.047
Generic Model w/o Interference Speakers	1.448	4.129	6.989	9.248	10.262
Open-set Generic Model	1.296	4.483	7.278	9.287	10.261

Table 5.1: Sequential grouping evaluation using the SNR metric. Numbers in the table show output SNR (dB) of segregated speech. The test utterances are mixtures of three talkers. In other words, there are one target and two interference speakers in a mixture.

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.529	5.492	8.965	11.004	11.835
Random Grouping	-3.648	0.873	2.566	3.334	3.514
Grouping Using Ideal Pitch	1.827	3.722	5.236	6.804	7.731
Pitch-based Grouping	-0.373	2.777	4.264	4.786	4.974
Known Speaker No. w/ Aggregated Prior	1.581	4.543	7.418	9.040	9.567
Known Speaker No. w/ Individual Prior	1.559	4.468	7.358	9.346	10.010
Grouping Using Generic Model	1.292	4.207	7.461	9.756	10.539
Grouping Using Combined Generic Model	1.718	4.145	6.895	9.120	10.181
Generic Model w/o Interference Speakers	1.448	4.129	6.989	9.248	10.262
Open-set Generic Model	0.636	4.169	7.355	9.314	10.148

Table 5.2: Sequential grouping evaluation using the SNR metric. Numbers in the table show output SNR (dB) of segregated speech. The test utterances are mixtures of four talkers. In other words, there are one target and three interference speakers in a mixture.

using ideal pitch because the clustering method is not able to differentiate speakers when their pitch contours are close to each other.

The following two rows in Table 5.1 and 5.2 present grouping performance based on speaker models. Here, the number of speakers in a mixture is assumed to be known. 'Aggregated Prior' refers to a speech prior that is estimated from the training data pooled from all the speakers, while 'Individual Prior' refers to a group of prior models that are estimated from individual speakers. The former approach requires less computation than the latter approach because the latter reconstructs missing T-F units using each of the 34 models instead of 1 for the former. For example, on a Dell PowerEdge 1850 server with 2 Xeon 3.4 GHz processors and 4 GB memory, the former method takes approximately 72 hours per TIR condition and the latter 4 hours. In terms of performance, both methods are comparable and the individual priors show some advantages over the aggregated prior at higher TIRs. This observation empirically suggests the use of an aggregated speech prior for GFCC reconstruction when computation time is an important factor for the task.

As we have proposed earlier, a generic model that incorporates all the interference speakers can be used to replace all the triplets or the quadruplets of speakers for sequential grouping. The SNR results are given in the row of 'Grouping Using Generic Model'. In our multi-talker corpus, a different generic model needs to be trained for a different target in a test mixture. It takes substantial time to construct all the possible generic models by training. For example, it takes one week on the aforementioned PowerEdge server to train such a single GMM prior with 2048 mixtures. After employing an iterative method that splits mixtures from 1 to 2048 in multiples of 2, it still takes

about 4 days per model. As an alternative, we construct a generic model by directly combining individual speaker models. Since a GMM is a summed group of Gaussian densities, we combine individual GMMs into the prior and reduce the Gaussian weights accordingly. Thus, for each test mixture, the algorithm can create a generic model that excludes the target in the runtime. With the same computing facilities, this method takes about 6 hours per TIR condition. The resulting performance is reported in the row of "Grouping Using Combined Generic Model'. Compared with the method that employs direct training, this approach achieves comparable performance at the high SNRs while performs moderately worse at low SNRs. This observation empirically justifies the use of the GMM-combination method for generic modeling.

The above methods assume the knowledge of interference speakers in the generic models. As we described earlier, this assumption may not always be satisfied. Here, to simulate the test conditions where a listener does not know any speakers other than the target, we remove the knowledge of interference speakers from sequential grouping. Specifically, for a multi-talker mixture, the models of interference speakers are not combined with other speakers to construct the generic model. The SNR results are reported in the row of 'Generic Model w/o Interference Speakers'. This method uses a single speech prior for GFCC reconstruction. Thus, the grouping algorithm still utilizes some knowledge of the interferences. To completely remove such knowledge, we employ the same GMM-combination method to construct priors, and remove the interference models. The SNR results are presented as 'Open-set Generic Model'. This method achieves comparable performance as the method that uses a single prior. It can be

observed from the tables that the model-based methods perform significantly better than the pitch-based methods under most of the TIR conditions. The only exception is -6 dB, where the T-F mask of a stream is too small for reliable GFCC reconstruction.

We also conduct similar experiments on the two-talker conditions and their results are shown in Table 5.3. The same observations made in the three-talker and the four-talker conditions hold in these two-talker conditions as well.

Methods \ Input SNR (dB)	**-6**	**0**	**6**	**12**	**18**
Ideal Sequential Grouping	3.604	6.483	8.287	8.865	9.084
Random Grouping	-2.459	0.487	2.474	2.699	3.051
Grouping Using Ideal Pitch	2.690	4.597	6.711	7.293	8.412
Pitch-based Grouping	0.598	4.167	6.013	6.527	7.102
Grouping Using Generic Model	2.475	5.097	6.923	7.929	8.271
Generic Model w/o Interference Speakers	2.091	4.719	6.660	7.632	7.979
Open-set Generic Model	2.545	5.065	6.708	7.623	8.004

Table 5.3: Sequential grouping evaluation using the SNR metric. Numbers in the table show output SNR (dB) of segregated speech. The test utterances are mixtures of two talkers. In other words, there are one target and one interference speaker in a mixture.

5.1.2 Non-speech intrusions

Unlike the preceding section, we deal with auditory scenes with non-speech intrusions in this section. We employ the noisy corpus that is derived from the SSC corpus and used for robust speaker recognition in Section 3.4. Specifically, the training

data is taken from SSC as given, and the test mixtures are generated by mixing clean utterances from SSC with four non-speech noise types: babble noise, destroyer (a navy ship) operation room noise, F-16 cockpit noise, and factory noise. The first two types contain a noisy background with many talkers speaking at the same time. They are classified as non-speech intrusions here because with so many voices together the signals do not exhibit clear speech patterns as our multi-talker mixtures do.

As in preceding experiments, we employ the same voiced speech segregation system for segmentation and simultaneous grouping (see Section 2.3.3 for details). This system produces simultaneous streams represented by binary T-F masks. Our grouping algorithm searches for the best segment assignment according to Equation (5.3), where the generic model is trained from pooled noise samples. Specifically, the noise samples include not only the four noise types in the test set but also fifteen other noise types (Hu, 2006) as follows,

- White noise
- Rock Music
- Siren
- Telephone
- Electric fan
- Clock alarm
- Traffic noise
- Bird chirp with water flowing
- Wind
- Rain
- Cocktail party noise
- Crowd noise at a playground
- Crowd noise with music
- Crowd noise with clap
- Babble noise (16 speakers).

This seeks to simulate a test condition where an actual noise source in a mixture originates from a large number of noise types. Given an unassigned segment, likelihood

scores are calculated for each registered speaker model and the generic model. Then, a joint likelihood is obtained by assigning all the segments to either a speaker or the generic model. The speaker that produces the maximum joint likelihood is selected as the target and its corresponding assignment of segments is returned as the grouping result. Here, the likelihood scores are calculated using the same method that is based on GFCC reconstruction and uncertainty decoding as in the multi-talker intrusion conditions.

We first evaluate contributions of GFCC reconstruction, uncertainty decoding and delta features to the overall grouping performance. Experimental results are shown in Table 5.4 for the babble, destroyer, F16 and factory noise types. The first row in the table presents SNR results obtained by ideal sequential grouping as defined in Section 5.1.1. 'Recon. only' and 'Recon. & UD' both calculate the likelihood scores using reconstructed GFCC features. The latter also uses the uncertainty decoder while the former does not. The last row, 'Recon., UD & Delta Feature', employs the delta features from Section 3.4 in addition to static GFCC features. All three methods assume known target identities. In other words, the algorithm does not search for the target speaker.

Compared with using enhanced GFCC alone, it is evident that UD improves SNR results. Delta features do not achieve performance improvement because one delta frame requires a neighboring window of five static frames as specified in Equation (3.19) and small segments lead to distorted delta features. Therefore, we will use UD for likelihood calculation in the following experiments and will not use delta features.

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.190	5.763	9.074	11.054	11.860
Recon. only	1.953	5.014	8.791	10.971	11.833
Recon. & UD	2.002	5.328	8.809	10.983	11.835
Recon., UD & Delta Feature	1.619	4.695	8.627	10.936	11.819

(a)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.670	6.486	9.693	11.262	11.883
Recon. only	2.075	5.415	9.285	11.063	11.835
Recon. & UD	2.209	5.599	9.298	11.070	11.830
Recon., UD & Delta Feature	1.864	5.178	9.130	11.064	11.822

(b)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	3.586	7.587	10.197	11.477	11.994
Recon. only	2.748	6.465	9.881	11.389	11.957
Recon. & UD	3.045	7.109	9.814	11.314	11.955
Recon., UD & Delta Feature	2.415	5.821	9.616	11.293	11.930

(c)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.958	7.063	9.855	11.420	11.866
Recon. only	2.477	5.982	9.410	11.308	11.819
Recon. & UD	2.693	6.712	9.441	11.296	11.825
Recon., UD & Delta Feature	2.147	5.386	9.190	11.264	11.798

(d)

Table 5.4: Sequential grouping evaluation using the SNR metric. Numbers in the table show output SNR (dB) of segregated speech. The test utterances contain babble noise in (a), destroyer noise in (b), F16 noise in (c) and factory noise in (d).

The above experiments evaluate different ways to calculate segment-level likelihoods. Here, we evaluate the grouping algorithm under conditions where target identities are unknown and under open conditions that exclude target models. The SNR results are given in Table 5.5. The first and third rows are taken from Table 5.4, showing results of ideal sequential grouping and grouping with known target identities. As before, the second row presents baseline performance by random assignment of segments.

The fourth row, 'Unknown Target', presents the condition with unknown targets. The grouping algorithm searches for the target and the best segment assignment as described earlier. This method achieves performance close to ideal sequential grouping and achieves comparable performance as the algorithm with known targets. Note that this condition tends to give higher likelihoods because of the maximum search even though the selected target may not be the correct one. This bias is reflected in the observation that the grouping algorithm recovers the segments that are missed (wrongly classified as non-speech) in the known target condition. On the other hand, the false-alarm errors do not increase much because of the significant differences between speech and noise in this experiment.

The last row in the table, 'Unregistered Target', shows the test configuration that removes target models from the registered speaker set, simulating a condition where a listener has not heard the voice of a target speaker before the test. In other words, the remaining speakers are regarded as generic models for the target speaker. Compared with the registered target condition as 'Unknown Target', the grouping performance here only degrades moderately. This observation implies that a set of 30~40 speakers likely

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.190	5.763	9.074	11.054	11.860
Random Grouping	0.349	2.159	2.752	3.212	3.566
Known Target	2.002	5.328	8.809	10.983	11.835
Unknown Target	1.065	5.302	8.849	11.002	11.831
Unregistered Target	0.802	4.238	7.617	10.294	11.401

(a)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.670	6.486	9.693	11.262	11.883
Random Grouping	-0.822	2.082	3.129	3.286	3.397
Known Target	2.209	5.599	9.298	11.070	11.830
Unknown Target	1.342	4.075	9.062	11.052	11.818
Unregistered Target	1.215	3.018	7.704	10.280	11.268

(b)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	3.586	7.587	10.197	11.477	11.994
Random Grouping	1.257	2.665	3.079	3.287	3.522
Known Target	3.045	7.109	9.814	11.314	11.955
Unknown Target	3.213	6.767	9.833	11.333	11.947
Unregistered Target	2.992	5.717	8.588	10.440	11.479

(c)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.958	7.063	9.855	11.420	11.866
Random Grouping	1.225	2.505	3.067	3.382	3.496
Known Target	2.693	6.712	9.441	11.296	11.825
Unknown Target	2.778	6.576	9.500	11.305	11.806
Unregistered Target	2.599	5.670	7.996	10.376	11.253

(d)

Table 5.5: Sequential grouping evaluation using the SNR metric. Numbers in the table show output SNR (dB) of segregated speech. The test utterances contain babble noise in (a), destroyer noise in (b), F16 noise in (c) and factory noise in (d).

contains a speaker that is acoustically close to the target. Thus, a small set of generic speakers might be sufficient for sequential grouping under general conditions. Furthermore, this observation inspires us to design an algorithm that systematically creates a set of generic speakers in the following section.

5.1.3 Unknown intrusion types

In this section, we deal with the intrusions in the preceding two sections together. Our sequential grouping algorithm is further extended to the conditions where noise sources are either speech or non-speech. By a direct extension, we construct a generic model that accounts for both speech and non-speech intrusions. More specifically, we employ the aforementioned GMM-combination method to combine the speech generic model from the multi-talker study in Section 5.1.1 and the non-speech generic model from Section 5.1.2. This combination method has achieved comparable results with standard training methods while saving substantial experimental time in the preceding studies. Sequential grouping uses the resulting generic model in Equation (5.3) to search for the best assignment of simultaneous streams.

The evaluation results are reported in Table 5.6 for the multi-talker conditions. Ideal sequential grouping results are given in the first row. The second row, 'Speech Generic', is directly taken from the last row in Table 5.1-5.3, which uses the multi-talker generic model. The last row presents the grouping performance by the combined generic model that includes non-speech noise types. It can be observed that the grouping performance is

Methods \ Input SNR (dB)	**-6**	**0**	**6**	**12**	**18**
Ideal Sequential Grouping	3.604	6.483	8.287	8.865	9.084
Speech Generic	2.545	5.065	6.708	7.623	8.004
Combined Generic Model	2.492	5.053	6.734	7.666	8.072

(a)

Methods \ Input SNR (dB)	**-6**	**0**	**6**	**12**	**18**
Ideal Sequential Grouping	3.006	5.793	8.968	10.937	11.801
Speech Generic	1.296	4.483	7.278	9.287	10.261
Combined Generic Model	0.994	4.495	7.372	9.342	10.309

(b)

Methods \ Input SNR (dB)	**-6**	**0**	**6**	**12**	**18**
Ideal Sequential Grouping	2.529	5.492	8.965	11.004	11.835
Speech Generic	0.637	4.169	7.355	9.314	10.148
Combined Generic Model	0.318	4.281	7.456	9.390	10.207

(c)

Table 5.6: Sequential grouping evaluation using the SNR metric. Numbers in the table show output SNR (dB) of segregated speech. The test utterances are mixtures of two talkers in (a), three talkers in (b) and four talkers in (c).

comparable before and after the noise model is combined. Such performance is expected because non-speech noise types are intrinsically different from speech.

In comparison, under the four non-speech noisy conditions as shown in Table 5.7, the resulting performance degrades moderately when the combined generic model is used. This is mainly because the speech generic model comprises features that are similar to the speaker models, and adding the speech generic model to the non-speech generic model leads to worse assignment of streams. For example, a target simultaneous stream that is correctly grouped to the target speaker stream under non-speech intrusion conditions may be wrongly assigned to the generic model stream because it is acoustically similar to one of the speakers in the generic model.

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.190	5.763	9.074	11.054	11.860
Non-speech Generic	1.065	5.302	8.849	11.002	11.831
Combined Generic Model	1.733	4.987	7.817	9.591	10.241

(a)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.670	6.486	9.693	11.262	11.883
Non-speech Generic	1.342	4.075	9.062	11.052	11.818
Combined Generic Model	2.021	4.823	7.822	9.323	9.926

(b)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	3.586	7.587	10.197	11.477	11.994
Non-speech Generic	3.213	6.767	9.833	11.333	11.947
Combined Generic Model	2.993	6.046	8.232	9.358	9.957

(c)

Methods \ Input SNR (dB)	-6	0	6	12	18
Ideal Sequential Grouping	2.958	7.063	9.855	11.420	11.866
Non-speech Generic	2.778	6.576	9.500	11.305	11.806
Combined Generic Model	2.594	5.706	8.133	9.390	9.933

(d)

Table 5.7: Sequential grouping evaluation using the SNR metric. Numbers in the table show output SNR (dB) of segregated speech. The test utterances contain babble noise in (a), destroyer noise in (b), F16 noise in (c) and factory noise in (d).

5.2 Generic Speaker Modeling for Sequential Grouping

Simultaneous streams obtained by segmentation and simultaneous grouping typically last for less than half a second. This data paucity restrains the utility of low-level features, such as pitch and spectral divergence, for sequential grouping. On the other hand, when interference speakers are not registered, the model-based grouping method still achieves good performance by using generic models. Since the grouping algorithm is based on maximization of likelihoods of segments given models, it essentially selects models that are acoustically close to unregistered speakers in input signal. Hence, under the extreme condition where none of the speakers in an auditory scene are registered, we propose employing a set of generic models that represent all the speakers in a domain for the grouping purpose.

Generic models have been proposed for unsupervised speaker indexing (Kwon and Narayanan, 2004; Kwon and Narayanan, 2005). Similar to the speaker detection and tracking studies described in Section 2.2, speaker indexing seeks to determine who is talking at a particular time in an audio stream. Such a task requires unsupervised methods when there is no prior information about the speakers in the input. Typical methods use generalized likelihood ratio test (Rice, 1995) to obtain speaker homogenous segments (Dunn *et al.*, 2000; Kwon and Narayanan, 2005). These segments are further clustered to index underlying speakers in the audio stream and construct models on-line. The optimal segment length of 2.5 *sec* and the typical minimum length of 1 *sec* (Dunn *et al.*, 2000) are found to be too short to obtain models that represent speakers well (Kwon and Narayanan, 2005), usually propagating clustering errors in the indexing process. A

number of methods have been proposed to create generic models from a large number of speakers and employ such models for unsupervised indexing (Kwon and Narayanan, 2005). These generic models include universal background model (UBM) and universal gender model (UGM). Kwon and Narayanan (2005) propose a different method that obtains generic models by quantizing a large speaker group. Specifically, it clusters speaker models based on the symmetrized Kullback-Leibler (K-L) divergence (Kullback, 1968). Each resulting cluster contributes a generic model that is randomly selected among the models in the cluster.

5.2.1 Speaker quantization

The basic idea of generic speaker modeling and speaker quantization is to identify and construct a small number of generic models that well represent a much larger speaker set. Generally speaking, quantization itself can be applied either in the feature space or in the model space. The former approach is popular for automatic speech recognition. However, without top-down constraints that model a speaker, a quantized model is more likely to reflect innate speech classes of the feature space instead of modeling speakers. Hence, we adopt the latter approach by performing quantization over speaker models. Specifically, we propose to use a speaker quantization method that is similar to the quantization method in speaker indexing (Kwon and Narayanan, 2005) to construct generic models for sequential grouping.

We first construct a large set of speaker models $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$. Pair-wise distances are obtained for each speaker pair in the set. Thus, the resulting distance matrix

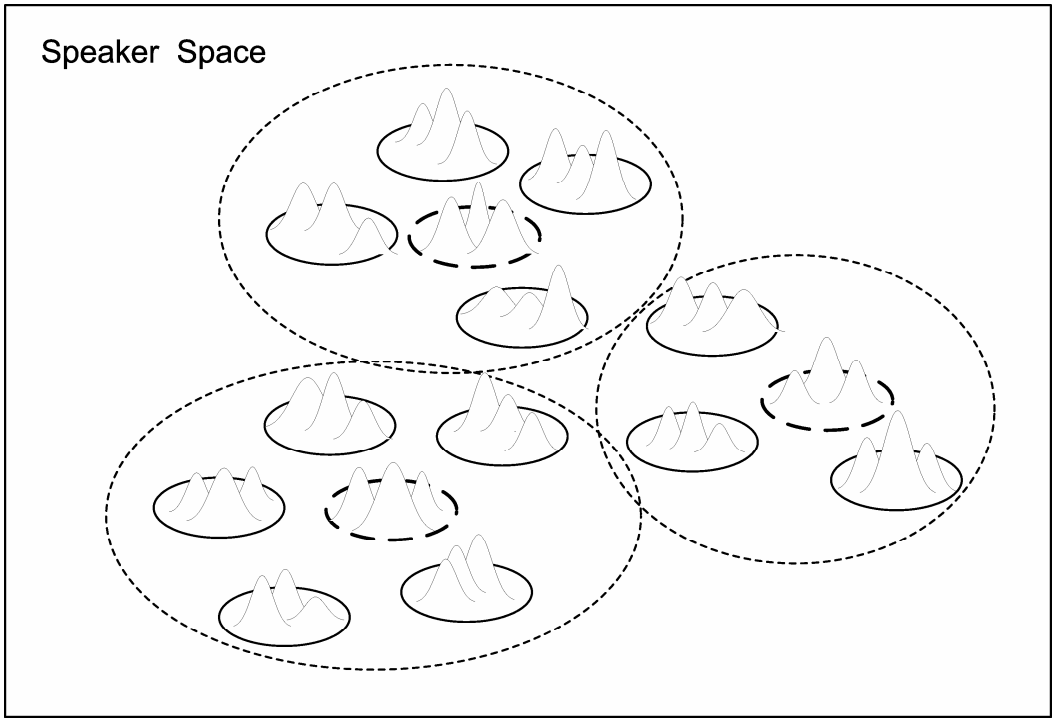

Figure 5.1: Illustration of speaker quantization. The solid circles represent individual speaker models. The dotted circles present clusters obtained by the speaker quantization method. The dashed circles denote the selected generic models from each cluster.

describes a distribution of all the models in this speaker space. Then, we apply a K-means clustering method (Duda *et al.*, 2001) to obtain a group of model clusters based on the distance matrix. Finally, within each cluster, the model that has the shortest average distance to the rest of the models in the cluster is selected as the generic model. Figure 5.1 illustrates quantized generic speakers.

As a speaker is usually modeled by a statistical distribution of its features, we employ the symmetrized K-L divergence (KLD) (Kullback, 1968) as the distance measure between two speaker models.

$$KL(f \| g) = \int f(x) \log \frac{f(x)}{g(x)} dx \tag{5.5}$$

defines the KLD, also known as the relative entropy, between two density functions, $f(x)$ and $g(x)$. The symmetric distance measure is,

$$D(f, g) = KL(f \| g) + KL(g \| f). \tag{5.6}$$

However, no closed-form solution exists for the KLD when $f(x)$ and $g(x)$ are GMMs (Li and King, 1999; Ben *et al.*, 2002; Vasconcelos, 2004; Goldberger and Aronowitz, 2005; Silva and Narayanan, 2006; Hershey and Olsen, 2007). Various methods have been proposed to approximate the KLD or estimate its upper-bound (Vasconcelos, 2004; Silva and Narayanan, 2006; Hershey and Olsen, 2007). The only method that asymptotically estimates the KLD is Monte Carlo simulation (Ben *et al.*, 2002; Vasconcelos, 2004; Hershey and Olsen, 2007). Here, we apply the Monte Carlo method to calculate the KLD between two GMMs. Specifically, we first draw N samples $\{x_i : i = 1...N\}$ from $f(x)$. KLD is estimated as,

$$KL(f \| g) \approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{f(x_i)}{g(x_i)}. \tag{5.7}$$

$KL(g \| f)$ is estimated in the same way with a set of samples drawn from $g(x)$. Thus, pair-wise symmetric K-L distances are calculated for all the speaker pairs and the resulting distance matrix defines the speaker space from which we perform quantization.

5.2.2 Evaluations

Generic models obtained by the speaker quantization method represent a large speaker space. For a mixture input of unknown speakers, our sequential grouping algorithm uses the generic models to organize simultaneous streams. These segments are generated by the same voiced speech segregation system as in the preceding experiments of this chapter. Segment likelihoods are obtained from reconstructed GFCCs by the uncertainty decoder. This section evaluates grouping performance of the algorithm on a simulated mixture corpus.

Our evaluation is based on the 2002 NIST Speaker Recognition Evaluation corpus (Przybocki and Martin, 2004). Unlike other experiments in the dissertation, this corpus is composed of telephone recordings. Specifically, we use the 1-speaker detection portion of the corpus. It contains 191 female and 139 male speakers, thus a total of 330 speakers. For each speaker, this corpus provides a 120 *sec* long recording of concatenated cell phone utterances. These utterances exhibit a slower speech rate than microphone recordings in TIMIT and SSC. To create noisy mixtures, the original recordings are sliced into short utterances of 4 *sec* each. Given the resulting 30 utterances for each speaker, four of them are randomly selected to simulate cochannel speech while the rest are retained for training. Cochannel mixtures are created at TIRs of -6 dB, 0 dB, 6 dB, 12 dB and 18 dB by mixing the selected four utterances with one randomly chosen utterance from every other speaker. Therefore, each TIR consists of 1320 test utterances.

The evaluation results are shown in Table 5.8. The first row presents SNR results by 'Ideal Sequential Grouping' that assigns input segments according to the ideal binary

Methods \ Input SNR (dB)	**-6**	**0**	**6**	**12**	**18**
Ideal Sequential Grouping	5.718	7.494	9.704	11.445	12.550
Known Speaker Identity	2.193	4.766	7.659	9.872	11.099
Random Grouping	-3.396	0.396	2.301	2.945	3.263
Exhaustive Search	1.515	4.397	7.270	9.442	10.384
Exhaustive Search (Subset 40)	1.808	4.637	7.443	9.590	10.488
Speaker Quantization (20)	-0.558	2.846	5.823	7.547	8.055
Speaker Quantization (40)	-0.314	2.931	5.868	7.618	8.261
Speaker Quantization (60)	-0.479	2.963	5.952	8.044	8.907
Speaker Quantization (80)	-0.493	2.948	5.996	8.058	8.972
Speaker Quantization (90)	-0.427	2.985	5.978	8.013	8.957
Speaker Quantization (100)	-0.534	2.941	6.035	8.139	9.124
Speaker Quantization (120)	-0.494	2.853	6.043	8.226	9.116
Speaker Quantization (140)	-0.319	3.093	6.117	8.215	8.934

Table 5.8: Sequential grouping evaluation using the SNR metric. Numbers in the table show output SNR (dB) of segregated speech. The test utterances are two-talker mixtures.

mask, giving an absolute upper-bound performance. 'Known Speaker Identity' denotes a condition where identities of speakers in a mixture are known *a priori*. In short, the algorithm degrades into a hypothesis test between the two speaker models. This actually places a performance upper-bound for all the model-based methods. Compared to ideal sequential grouping, grouping performance degrades faster with a decreasing TIR. This indicates that likelihood scores become less reliable when TIR decreases because there are more missing T-F units to be reconstructed from fewer reliable ones. The following

row gives baseline performance by randomly assigning the segments, thus setting a performance lower-bound.

Before evaluating quantized generic models, we also show how the grouping algorithm fares using the exhaustive search method as described in Section 4.2. Its SNR results are given in the 'Exhaustive Search' row. Basically, it sets M=2 and K=330 in (5.2), meaning that there are 2 streams to organize and $330 \times 290 / 2 = 37785$ speaker pairs to search. Apparently, this search requires substantial computation time. Hence, we also conduct an experiment that uses a reduced set of 40 speakers. Specifically, a reduced set is composed of the 2 underlying speakers in a mixture and 38 speakers that are randomly selected from the remaining 328 speakers. Its results are shown in 'Exhaustive Search (Subset 40)'. The exhaustive search over the complete set produces results almost as good as those obtained with known speaker identities, and on average the degradation is 0.5 dB. This observation empirically suggests the optimality of our model-based grouping algorithm. When the speaker number is reduced from 330 to 40, the performance improves slightly. This is because with a smaller number of speakers, models are less crowded in the speaker space and it is easier for the grouping algorithm to discriminate them.

The following rows in Table 5.8 present grouping results obtained by speaker quantization. For each cochannel input, we remove the two underlying speakers from the speaker set and perform speaker quantization on remaining 328 speakers. Thus, we create a different generic model set for each different speaker pair. This simulates the auditory scene where none of the speakers are registered. The number of generic models is a factor

that determines the trade-off between grouping performance and computation time. More generic models entail better matches between the models and unknown inputs while they require more computation time because of the increased search space. To observe how this factor affects grouping, we vary the number of quantized models in a range from 20 to 140. In Table 5.8, the number after 'Speaker Quantization' denotes this number.

SNR performance is significantly improved by increasing the number of generic models from 20 to 60. While it seems that the improvement stalls from 60 to 90, the performance is further improved above 90. On average, the performance with 140 generic models is about 1.9 dB worse than that of 'Known Speaker Identity', and about 1.4 dB worse than that of the exhaustive search within the complete speaker space. Since the core of the algorithm compares summarized likelihoods for every speaker pair with a complexity of $O(M^2)$, the computation time increases roughly 50 times by increasing the generic models from 20 to 140.

In our view, combining speaker quantization and generic modeling presents a promising approach for dealing with acoustic inputs of unknown speakers.

CHAPTER 6

SUMMARY

6.1 Contributions

CASA comprises simultaneous grouping and sequential grouping that together organize segments from different sources into corresponding streams. The former integrates concurrent segments and the latter integrates segments across time. This dissertation has presented a systematic study on sequential organization based on speaker characteristics. In addition, we have proposed CASA-based front-end processors for robust speaker recognition.

In Chapter 3, we have presented an extensive effort on robust speaker recognition. A novel usable speech extraction method has been proposed and shown to significantly improve speaker recognition accuracy in cochannel speech. Then, by combining missing-data recognition and the use of CASA-based segregation as a front-end processor, recognition performance is further improved under various noisy conditions. We have also proposed a general solution to robust speaker recognition in the presence of additive

noise. Novel speaker features are derived by auditory filtering and cepstral analysis. We employ an uncertainty decoder that accounts for front-end processing errors in conjunction with novel speaker features for robust speaker identification and verification. The proposed system has achieved substantial improvement over not only typical speaker features but also an advanced robust front-end processor for speech signals.

In Chapter 4, we have explored bottom-up grouping methods that employ features such as pitch and spectrum for speech organization. We have also presented sequential organization methods based on speaker models. A novel aspect of our study is the derivation of the computational objective for joint speaker recognition and sequential grouping. This formulation leads to the exhaustive search algorithm that finds the optimal assignment of simultaneous streams given the speaker models. In addition, we have proposed a hypothesis pruning method that reduces the search space and computation time while achieving a performance level close to that of exhaustive search. In the evaluations, the model-based methods yield significantly better accuracy in terms of segment assignment than alternative approaches. Furthermore, the grouping system is integrated with other CASA processes in a complete speech separation and recognition system and its evaluations show a significant improvement over the baseline performance in speech recognition for many noisy conditions.

In Chapter 5, we have incorporated the model-based sequential grouping algorithm with generic modeling methods to handle multiple interfering speakers and unknown noise types. By employing a generic model that takes different interference types into account, our methods achieve a level of performance close to that with registered

interference models although only the target speakers are registered. Subsequently, we have presented a speaker quantization method that constructs generic models by clustering a large set of speakers. These generic models are used for sequential grouping when none of the speakers in an auditory scene are registered. The systematic evaluations have shown that this approach gives only moderately worse performance than that obtained with registered speakers.

6.2 Insights Gained

During the course of this dissertation study, a number of insights have surfaced. A key insight is that speaker models encode potent information for sequential grouping. In Chapter 4, we have found that model-based methods perform significantly better than feature-based methods, which directly utilize features for speech organization. In other words, speaker characteristics are most effectively captured by statistical speaker models. In the CASA account, this means that schema-based grouping may play a more effective role than primitive grouping in sequential organization. This observation does not necessarily mean that CASA shall prefer the top-down process to bottom-up processes. Rather, given the feature distributions provided by models, it may be easier for a computational system to determine whether two separated segments of speech are sufficiently close in the space so that they should be grouped into the same stream.

Speech signals include versatile information sources such as linguistic content and speaker characteristics. From an information processing perspective, such information is captured and represented as features in a high-dimensional space. Modeling is a

supervised learning process that constructs schemas according to different sources, whether they are language-specific or speaker-specific. Essentially, the resulting schemas provide ways to observe the same data through different perspectives. In other words, the schemas are utilized to transform the same data into their corresponding information sources. In our view, compared with primitive grouping methods, schema-based methods yield superior performance because of such transformations.

When specific speaker schemas are not available, we have demonstrated how to perform sequential grouping using generic models in Chapter 5. A generic model is basically a broader and less specific schema or a class of schemas. The insight here is that replacing individual models with generic models incurs performance degradation because of loss of fine details but it still produces reasonable results. The model training process acquires consistency of schemas in the form of model structures and parameters, and such consistency facilitates approximations of individual models by generic ones. In the ASA account, our insight is that voices exhibit consistency and that a listener may have a way to grasp such consistency after years of exposure to daily auditory scenes. For example, spectral energy distributions relate to perceptual voice qualities and such distributions are determined by the vocal tract shape of a talker. Two speakers with similar shapes lead to similar voice qualities and we hypothesize that a listener employs such consistency for grouping either of the voices.

A key insight in Chapter 3 is that robust speaker recognition does not require a complete speech signal. What a recognition system needs are portions of the input signal that contain speaker characteristics. In other words, certain portions carry discriminative

information and are thus adequate for speaker recognition. This insight has been obtained from the robust recognition experiments that evaluate usable speech segments and binary T-F masks.

The usable speech for both speaker recognition and sequential grouping in Chapter 3-5 entails the use of missing-data methods for proper likelihood scoring. These methods include conditional density marginalization and missing-data reconstruction. One insight from the study is that the latter approach performs better when SNR is between 6 dB and 18 dB with non-speech intrusions. Under such conditions, there are usually more reliable components than unreliable ones within a time frame and unreliable components are reasonably estimated given the reliable ones. On the other hand, under low SNR conditions (-12 dB ~ 6 dB), the marginalization approach tends to yield superior performance. Because reconstruction imputes missing T-F units from reliable ones the paucity of reliable data under such low SNR conditions leads to unreliable likelihoods, hence inferior grouping performance.

6.3 Future Work

As described in Chapter 5, the speaker quantization approach selects the model that is closest to the remaining models in a cluster as the generic model. When the number of generic models is set small, the resulting models may be too sparse in the speaker space for reliable grouping. There are several problems in this situation. One is that sparse generic models do not represent the overall speaker space well. The other is that with few

generic models multiple distinct speakers in a mixture are more likely to be associated with a single cluster. In this case, our sequential grouping method will fail.

The basic idea to tackle the first problem is to incorporate more information within a cluster. One approach is to retrain the generic model by pooling speech samples from all the speakers in a cluster. This approach may cost substantial computational resources. An efficient alternative is to adapt the selected model to the samples of other speakers. Speaker recognition studies have used maximum likelihood linear regression or maximum *a posteriori* adaptation methods in the case of insufficient training data (Reynolds, 2002; Furui, 2005). Another approach to deal with the first problem is to combine the speaker models within a cluster instead of modeling from samples. In the case of GMMs, the Gaussian mixtures from different speakers may be summed together and mixture weights may be discounted in a way that is proportional to the K-L distance of a speaker GMM from the cluster center.

The second problem does not limit itself to generic modeling. In the SSC task, our grouping method is not able to handle same talker mixtures. This problem is intrinsic to model-based sequential grouping approach itself. Essentially, different speakers may produce very similar sounds and a speaker model may overlap with other models in the feature space. Given a short segment, its likelihood may indicate that it originates from one model while it truly belongs to another. Given the decision framework of sequential grouping in Chapter 4, a grouping algorithm is not able to make a correct decision without prior knowledge about the models or the dependence of the segment on other segments. Hence, one solution is to infer the prior probabilities of the models from the

mixture. The other solution is to model segment-level dependence. Such dependence may require linguistic information of the segments.

In the derivation of computational goal for sequential organization in Section 4.2.1, we assume that segment assignments are uniformly distributed, meaning that we do not have prior knowledge about segment labels. However, this uniform assumption is not satisfied under either high or low TIR conditions, where one speaker in a cochannel mixture dominates the other. One solution here is to model the prior probability of segment assignments based on input TIR. This solution cannot be easily applied when input TIR is unknown. On the other hand, TIR can be estimated from segregated speaker streams at output (see Hu, 2006). Thus, a complete solution would combine the above two processes in an iterative manner. More specifically, given an input TIR estimate, we construct a probability distribution for segment assignments and conduct sequential grouping accordingly. Then, we re-estimate the input TIR using output streams and repeat the previous step. This process iterates itself until the TIR estimate converges. The initial TIR value can be set to 0 dB, corresponding to a uniform distribution of segment assignments. Hu (2006) describes such an iterative process for pitch tracking and voiced speech segregation.

Throughout this dissertation, the adverse conditions mainly contain additive noise sources. Future work needs to address other distortions such as convolutive noise from telephone transmission and a combination of both noise types. One example is the distortion introduced by room reverberation. Here, what a system receives is an input signal comprises many delayed versions of the original clean signal that have been

bounced back. While it may seem straightforward to extend our model-based sequential grouping algorithms to reverberant conditions, there are several empirical questions that need to be addressed. The first question is what is target signal in a reverberant mixture? Is it clean target or reverberant target? Depending on the answer, speaker models in sequential grouping may need to be retrained. But it is not easy to conduct retraining since there are likely too many reverberant conditions to consider in the training phase. Another problem arises from the fact that simultaneous streams from different sources are more likely to overlap with each other under echoic conditions than anechoic conditions. Thus, in reverberant conditions, a simultaneous stream of target might contain strong energy from intrusions. One solution is to weight likelihoods from different T-F units according to a metric that measures how severe reverberation is.

BIBLIOGRAPHY

Assmann, P. F., and Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of Acoustical Society of America, 88*(2), 680-697.

Assmann, P. F., and Summerfield, Q. (1994). The contribution of waveform interactions to the perception of concurrent vowels. *Journal of Acoustical Society of America, 95*, 471-484.

Assmann, P. F., and Summerfield, Q. (2004). The perception of speech under adverse acoustic conditions. In *Speech processing in the auditory system*, S. Greenberg*, et al.*, ed.,

American Standards Association (1960). *Acoustical terminology SI, 1-1960*. New York: American Standards Association.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *Journal of Acoustical Society of America, 52*, 1687-1697.

Atienza, M., Cantero, J. L., Grau, C., Gomez, C., Dominguez-Marin, E., and Escera, C. (2003). Effects of temporal encoding on auditory object formation: A mismatch negativity study. *Cognitive Brain Research, 16*, 359-371.

Bach, F. R., and Jordan, M. I. (2004). Blind one-microphone speech separation: A spectral learning approach. In: *Proceedings of Neural Information Processing Systems Conference (NIPS)*.

Baker, B., and Sridharan, S. (2006). Speaker verification using hidden Markov models in a multilingual text-constrained framework. In: *Proceedings of IEEE Odyssey 2006, Speaker and Language Recognition Workshop*.

Barger, P., and Sridharan, S. (1997). Robust speaker identification using multi-microphone systems. In: *Proceedings of IEEE TENCON*, 261-264.

Barker, J., and Cooke, M. (1998). Is the sine-wave speech cocktail party worth attending? *Speech Communication, 27*, 159-174.

Barker, J., Cooke, M., and Ellis, D. (2005). Decoding speech in the presence of other sources. *Speech Communication, 45*, 5-25.

Ben, M., Blouet, R., and Bimbot, F. (2002). A monte carlo method for score normalization in automatic speakerverification using kullback-leibler distances. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *1*, 689-692.

Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 113-120.

Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., and Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing* (4), 430-451.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* (5:9/10), 341-345.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). A speech corpus for multitalker communications research. *Journal of Acoustical Society of America, 107*, 1065-1066.

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge MA: MIT Press.

Brown, G. J., and Cooke, M. (1994a). Computational auditory scene analysis. *Computer Speech and Language, 8*, 297-336.

Brown, G. J., and Cooke, M. (1994b). Perceptual grouping of musical sounds: A computational model. *Journal of New Music Research, 23*, 107-132.

Brungart, D. S. (2001). Information and energetic masking effects in the perception of two simultaneous talkers. *Journal of Acoustical Society of America, 109*, 1101-1109.

Brungart, D. S., Chang, P., Simpson, B. D., and Wang, D. L. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of Acoustical Society of America, 120*, 4007-4018.

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE, 85*, 1437-1462.

Campbell, W., Assaleh, K., and Broun, C. (2002). Speaker recognition with polynomial classifiers. *IEEE Transactions on Speech and Audio Processing, 10*, 205-212.

Carlson, B. A., and Clements, M. A. (1991). A computationally compact divergence measure for speech processing. *IEEE Transactions on Pattern Analysis Machine Intelligence, 13*, 1-6.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences, 8*(10), 465-471.

Chang, J. E., Bai, J. Y., and Zeng, F. G. (2006). Unintelligible low-frequency sound enhances simulated cochlear-implant speech recognition in noise. *IEEE Transactions on Biomedical Engineering, 53*(12), 2598-2601.

Cherry, E. C. (1953). Some experiments on the recognition of speech with one and with two ears. *Journal of Acoustical Society of America, 25*, 975-979.

Cooke, M. P. (2003). Glimpsing speech. *Journal of Phonetics, 31*, 579-584.

Cooke, M. P., and Ellis, D. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication, 35*, 141-177.

Cooke, M. P., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication, 34*, 267-385.

Cooke, M. P., and Lee, T. W. (2006). Speech separation and recognition competition. Available at http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm.

Culling, J., and Darwin, C. (1993). The role of timbre in the segregation of simultaneous voices with intersecting contours. *Perception and psychophysics, 54*, 303-309.

Dang, J., and Honda, K. (2002). Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. *Journal of Phonetics, 30*, 511-532.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *Journal of Acoustical Society of America, 114*(5), 2913-2923.

Darwin, C. J., and Hukin, R. (2000). Effectiveness of spatial cues, prosody and talker characteristics in selective attention. *Journal of Acoustical Society of America, 107*, 970-977.

Deng, L., Droppo, J., and Acero, A. (2005). Dynamic compensation of hmm variants using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE Transactions on Speech and Audio Processing, 13*, 412-421.

Divenyi, P. (2005). *Speech separation by humans and machines*. Norwell, Mass.: Kluwer Academic.

Drullman, R., and Bronkhorst, A. W. (2004). Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers. *Journal of Acoustical Society of America, 116*, 3090-3098.

Drygajlo, A., and El-Maliki, M. (1998). Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 121-124.

Drygajlo, A., and El-Maliki, M. (2001). Integration and imputation methods for unreliable feature compensation in gmm based speaker verification. In: *Proceedings of 2001: A Speaker Odyssey - The Speaker Recognition Workshop*, 107-112.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*, 2nd ed. New York: Wiley & Sons.

Dunn, R. B., Reynolds, D. A., and Quatieri, T. F. (2000). Approaches to speaker detection and tracking in conversational speech. *Digital Signal Processing, 10*, 93-112.

Ellis, D. P. W. (1996). Prediction-driven computational auditory scene analysis. Ph.D. Dissertation, MIT Department of Electrical Engineering and Computer Science.

Ellis, D. P. W. (2006). Model-based scene analysis. In *Computational auditory scene analysis: Principles, algorithms, and applications*, D. L. Wang, and G. J. Brown, ed., 115-146. Hoboken, NJ: Wiley-IEEE Press.

Foo, S. W., and Lim, E. G. (2002). Speaker recognition using adaptive boosted decision tree classifier. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *I*, 157-160.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 29*, 254-272.

Furui, S. (1989). *Digital speech processing, synthesis, and recognition*. New York: Marcel Dekker.

Furui, S. (1991). Speaker-dependent-feature extraction, recognition and processing techniques. *Speech Communication, 10*, 505-520.

Furui, S. (1994). An overview of speaker recognition technology. In: *Proceedings of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1-9.

Furui, S. (2001). *Digital speech processing, synthesis, and recognition*. New York: Marcel Dekker.

Furui, S. (2005). 50 years of progress in speech and speaker recognition. In: *Proceedings of International Conference on Speech and Computer (SPECOM)*, 1-9.

Gillick, L., and Cox, S. J. (1989). Some statistical issues in the comparison of speech recognition algorithms. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 532-535.

Godsmark, D., and Brown, G. J. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication, 27*, 351-366.

Goldberger, J., and Aronowitz, H. (2005). A distance measure between gmms based on the unscented transform and its application to speaker recognition. In: *Proceedings of Interspeech*, 1985-1988.

Gong, Y. (2002). Noise-robust open-set speaker recognition using noise-dependent gaussian mixture classifier. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 133-136.

Grose, J. H., and Hall, J. W. (1996). Perceptual organization of sequential stimuli in listeners with cochlear hearing loss. *Journal of Speech and Hearing Research, 39*(6), 1149-1159.

Hartmann, W. M. (1996). Pitch, periodicity, and auditory organization. *Journal of Acoustical Society of America, 100*(6), 3491-3502.

Helmholtz, H. (1863). *On the sensation of tone* (Translated by A. J. Ellis), Second English ed. New York: Dover Publishers.

Hermansky, H., and Morgan, N. (1994). Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing, 2*(4), 578-589.

Hershey, J. R., and Olsen, P. A. (2007). Approximating the kullback leibler divergence between gaussian mixture models. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *IV*, 317-320.

Hu, G. (2006). Monaural speech organization and segregation. Ph.D. dissertation, The Ohio State University.

Hu, G., and Wang, D. L. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks, 15*, 1135-1150.

Hu, G., and Wang, D. L. (2006). An auditory scene analysis approach to monaural speech separation. In *Topics in acoustic echo and noise control*, E. Hansler, and G. Schmidt, ed., 485-515. Heidelberg: Springer.

Hu, G., and Wang, D. L. (2007). Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech and Language Processing, 15*, 396-405.

Huang, X., Acero, A., and Hon, H. (2001). *Spoken language processing*. Upper Saddle River: Prentice Hall.

Krishnamachari, K. R., Yantorno, R. E., Benincasa, D. S., and Wenndt, S. J. (2000). Spectral autocorrelation ratio as a usability measure of speech segments under cochannel conditions. In: *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems*.

Kullback, S. (1968). *Information theory and statistics*. New York: Dover Publications.

Kwon, S., and Narayanan, S. (2004). Speaker model quantization for unsupervised speaker indexing. In: *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 1517-1520.

Kwon, S., and Narayanan, S. (2005). Unsupervised speaker indexing using generic models. *IEEE Transactions on Speech and Audio Processing, 13*(5), 1004-1013.

Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L. (1978). Generating vocal tract shapes from formant frequencies. *Journal of Acoustical Society of America, 64*, 1027-1035.

Li, X. Q., and King, I. (1999). Gaussian mixture distance for information retrieval. In: *Proceedings of International Conference on Neural Networks*, 2544-2549.

Liberman, A. M. (1982). On the finding that speech is special. *American Psychologist, 37*(2), 148-167.

Lovekin, J. M., Yantorno, R. E., Krishnamachari, K. R., Benincasa, D. S., and Wenndt, S. J. (2001). Developing usable speech criteria for speaker identification. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 421-424.

Mackersie, C. L. (2003). Talker separation and sequential stream segregation in listeners with hearing loss: Patterns associated with talker gender. *Journal of Speech Language and Hearing Research, 46*, 912-918.

Mackersie, C. L., Prida, T. L., and Stiles, D. (2001). The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by

listeners with sensorineural hearing loss. *Journal of Speech Language and Hearing Research, 44*, 19-28.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The det curve in assessment of detection task performance. In: *Proceedings of Eurospeech*, *4*, 1899-1903.

Martin, A. F., and Przybocki, M. A. (2001). The nist speaker recognition evaluations: 1996-2001. In: *Proceedings of A Speaker Odyssey, The Speaker Recognition Workshop*.

Matsui, T., and Furui, S. (1990). Text-independent speaker recognition using vocal tract and pitch information. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 137-140.

Matsui, T., and Furui, S. (1996). Speaker recognition using hmm composition in noisy environments. *Computer Speech and Language, 10*, 107-116.

Miller, G. A., and Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of Acoustical Society of America, 22*(2), 167-173.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing*, 5th ed. San Diego: Academic.

Moore, B. C. J., and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica united with Acustica, 88*, 320-332.

Morgan, D. P., George, E. B., Lee, L. T., and Kay, S. M. (1997). Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Transactions on Speech and Audio Processing, 5*, 407-424.

Naik, J. M. (1990). Speaker verification: A tutorial. *IEEE Communications Magazine*, 42-48.

Naik, J. M., Netsch, L. P., and Doddington, G. R. (1989). Speaker verification over long distance telephone lines. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 524-527.

Necioglu, B. F., Clements, M. A., and Barnwell III, T. P. (2000). Unsupervised estimation of the human vocal tract length over sentence level utterance. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1319-1322.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (1999). *Discrete-time signal processing*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall.

Ortega-Garcia, J., and Gonzalez-Rodriguez, J. (1997). Providing single and multi-channel acoustical robustness to speaker identification systems. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1107-1110.

Paige, A., and Zue, V. (1969). Calculation of vocal tract length. *IEEE Transactions on Audio and Electroacoustics, 18*, 268-270.

Parihar, N., and Picone, J. (2003). Analysis of the aurora large vocabulary evalutions. In: *Proceedings of Eurospeech*, 337-340.

Patterson, R. D., Holdsworth, J., and Allerhand, M. (1992). Auditory models as preprocessors for speech recognition. In *The auditory processing of speech: From sounds to words.*, M. E. H. Schouten, ed., 67-83. Berlin, Germany: Mouton de Gruyter.

Przybocki, M. A., and Martin, A. F. (2004). NIST speaker recognition evaluation chronicles. In: *Proceedings of Odyssey 2004*.

Przybocki, M. A., Martin, A. F., and Le, A. N. (2006). NIST speaker recognition evaluation chronicles-part 2. In: *Proceedings of IEEE Odyssey 2006, Speaker and Language Recognition Workshop*.

Quatieri, T. F., and Danisewicz, R. G. (1990). An approach to co-channel talker interference suppression using a sinusoidal model for speech. *IEEE Transactions on Acoustics, Speech and Signal Processing, 38*, 56-69.

Raj, B., Seltzer, M. L., and Stern, R. M. (2004). Reconstruction of missing features for robust speech recognition. *Speech Communication, 43*, 275-296.

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of Acoustical Society of America, 55*(3), 678-680.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101*(1), 129-156.

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing, 2*, 639-643.

Reynolds, D. A. (1995). Speaker identification and verification using gaussian mixture speaker models. *Speech Communication, 17*, 91-108.

Reynolds, D. A. (1997). Comparison of background normalization methods for text-independent speaker verification. In: *Proceedings of Eurospeech*, 936-967.

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *IV*, 4072-4075.

Reynolds, D. A.*, et al*. (2003). The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 784-787.

Rice, J. A. (1995). *Mathematical statistics and data analysis*. Belmont, CA: Duxbury Press.

Roman, N., Wang, D. L., and Brown, G. J. (2003). Speech segregation based on sound localization. *Journal of Acoustical Society of America, 114*, 2236-2252.

Rose, R. C., Hofstetter, E. M., and Reynolds, D. A. (1994). Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech and Audio Processing, 2*(2), 245-257.

Rosenberg, A., Lee, C., and Soong, F. (1990). Sub-word unit talker verification using hidden Markov models. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 269-272.

Russell, S., and Norvig, P. (2003). *Artificial intelligence: A modern approach*, 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Savic, M., and Gupta, S. (1990). Variable parameter speaker verification system based on hidden Markov modeling. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 281-284.

Schmidt-Nielsen, A., and Crystal, T. H. (1998). Human v.s. Machine speaker identification with telephone speech. In: *Proceedings of International Conference on Spoken Language Processing (ICSLP)*.

Schroeder, M. R. (1966). Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of Acoustical Society of America, 41*, 1002-1010.

Shao, Y., Srinivasan, S., and Wang, D. L. (2007). Incorporating auditory feature uncertainties in robust speaker identification. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *IV*, 277-280.

Shao, Y., and Wang, D. L. (2003). Co-channel speaker identification using usable speech extraction based on multi-pitch tracking. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *2*, 205-208.

Shao, Y., and Wang, D. L. (2006a). Model-based sequential organization in cochannel speech. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(1), 289-298.

Shao, Y., and Wang, D. L. (2006b). Robust speaker recognition using binary time-frequency masks. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *I*, 645-648.

Silva, J., and Narayanan, S. (2006). Average divergence distance as a statistical discrimination measure for hidden Markov models. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(3), 890-906.

Sivakumaran, P., and Ariyaeeinia, A. M. (2000). The use of sub-band cepstrum in speaker verification. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1073-1076.

Sjolander, K., and Beskow, J. (2000). Wavesurfer - an open source speech tool. In: *Proceedings of International Conference on Spoken Language Processing*.

Smolenski, B. Y., Yantorno, R. E., Benincasa, D. S., and Wenndt, S. J. (2002). Co-channel speaker segment separation. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 125-129.

Srinivasan, S., and Wang, D. L. (2007). Transforming binary uncertainties for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7), 2130-2140.

Stern, R. M., Brown, G. J., and Wang, D. L. (2006). Binaural sound localization. In *Computational auditory scene analysis: Principles, algorithms, and applications*, D. L. Wang, and G. J. Brown, ed., 147-186. Hoboken, NJ: Wiley-IEEE Press.

STQ-AURORA. (2005-11). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. In *ETSI ES 202 050 v1.1.4*, ed., European Telecommunications Standards Institute (ETSI).

Varga, A., and Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: Ii. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication, 12*(3), 247-251.

Vasconcelos, N. (2004). On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Transactions on Information Theory, 50*(7), 1482-1496.

Vincent, L., and Soille, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulation simulations. *IEEE Transactions on Pattern Analysis Machine Intelligence, 13*(6), 583-598.

Vliegen, J., and Moore, B. C. J. (1999). The role of spectral and periodicty cues in auditory stream segregation, measured using a temporal discrimination task. *Journal of Acoustical Society of America, 106*(2), 938-945.

Vliegen, J., and Oxenham, A. J. (1998). Sequential stream segregation in the absence of spectral cues. *Journal of Acoustical Society of America, 105*(1), 339-346.

Wakita, H. (1973). Direct estimation of the vocal-tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics, 21*, 417-427.

Wakita, H. (1977). Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 25*, 183-192.

Wan, V., and Renals, S. (2002). Evaluation of kernel methods for speaker verification and identification. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *I*, 669-672.

Wang, D. L. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines*, P. Divenyi, ed., 181-197. Norwell MA: Kluwer Academic.

Wang, D. L. (2006). Feature-based speech segregation. In *Computational auditory scene analysis: Principles, algorithms, and applications*, D. L. Wang, and G. J. Brown, ed., 81-114. Hoboken, NJ: Wiley-IEEE Press.

Wang, D. L., and Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks, 10*, 684-697.

Wang, D. L., and Brown, G. J. (ed., 2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. Hoboken, NJ: Wiley-IEEE Press.

Wang, D. L., and Hu, G. (2006). Unvoiced speech segregation. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *V*, 953-956.

Ward, D. B., Lehmann, E. A., and Williamson, R. C. (2003). Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing, 11*(6), 826-836.

Warren, R. M., Healy, E. W., and Chalikia, M. H. (1996). The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms. *Journal of Acoustical Society of America, 100*(4), 2452-2461.

Warren, R. M., Bashford, J. A., and Gardner, D. A. (1990). Tweaking the lexicon: Organization of vowel sequences into words. *Perception and Psychophysics, 47*(5), 423-432.

Weintraub, M. (1985). A theory and computational model of auditory monaural sound separation. Ph.D. Dissertation, Standford University.

Wong, L. P., and Russell, M. (2001). Text-dependent speaker verification under noisy conditions using parallel model combination. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 457-460.

Wu, M., Wang, D. L., and Brown, G. J. (2003). A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech Audio Process., 11*(3), 299-241.

Yabe, H., Sato, Y., Sutoh, T., Hiruma, T., Shinozaki, N., Nashida, T., Saito, F., and Kaneko, S. (1999). The duration of the integrating window in auditory sensory memory. *Electroencephalography and Clinical Neurophysiology, 49*, 166-169.

Yantorno, R. E. (1999). Final report for summer research faculty program. Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs.

Yantorno, R. E., Benincasa, D., and Wenndt, S. (2001). Effects of co-channel interference on speaker identification. In: *Proceedings of SPIE International Symposium on Technologies for Law Enforcement*, *4232*, 258-261.

Yegnanarayana, B., Sharat, K., and Kishore, S. P. (2001). Source and system features for speaker recognition using AANN models. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 409-412.

Yoma, N. B., and Villar, M. (2002). Speaker verification in noise using a stochastic version of the weighted viterbi algorithm. *IEEE Transactions on Speech and Audio Processing, 10*(3), 158-166.

Yoshida, K., Takagi, K., and Ozeki, K. (2001). A multi-snr subband model for speaker identification under noisy environments. In: *Proceedings of Eurospeech*, 2849-2852.

Young, S., Kershaw, D., Odell, J., Valtchev, V., and Woodland, P. (2000). *The HTK book (for HTK version 3.0)*. Microsoft Corporation.

Yu, G., and Gish, H. (1993). Identification of speakers engaged in dialog. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 383-386.

Zissman, M. A., and Seward, D. C. (1992). Two-talker pitch tracking for co-channel talker interference suppression. Technical Report, MIT Lincoln Laboratory.