Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 51 (2009) 657–667

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Sequential organization of speech in computational auditory scene analysis

Yang Shao [a,*], DeLiang Wang [a,b]

[a] Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Ave., Columbus, OH 43210, USA
[b] Center for Cognitive Science, The Ohio State University, Columbus, OH 43210, USA

Received 5 November 2007; received in revised form 12 October 2008; accepted 19 February 2009

Abstract

A human listener has the ability to follow a speaker's voice over time in the presence of other talkers and non-speech interference. This paper proposes a general system for sequential organization of speech based on speaker models. By training a general background model, the proposed system is shown to function well with both interfering talkers and non-speech intrusions. To deal with situations where prior information about specific speakers is not available, a speaker quantization method is employed to extract representative models from a large speaker space and obtained generic models are used to perform sequential grouping. Our systematic evaluations show that grouping performance using generic models is only moderately lower than the performance level achieved with known speaker models.

Keywords: Sequential organization; Computational auditory scene analysis; Speaker quantization; Binary time–frequency mask

1. Introduction

A daily auditory scene typically comprises multiple sound sources. Usually there is a target source that one is listening to, such as a radio host or a piece of music. Meanwhile, there are acoustic events from other sound sources that are of little interest to the listener, such as a ventilation fan in an office or a car passing by on the street. Cochannel speech, for example, is a combination of utterances from two speakers transmitted over a single communication channel (Quatieri and Danisewicz, 1990). Unlike conversations, speakers are usually not aware of each other under cochannel conditions, leading to large speech overlaps that present a considerable challenge to applications such as automatic speaker or speech recognition. On the other hand, for a cochannel mixture that has comparable energies from both talkers, human listeners can readily select and follow one speaker's voice (Brungart, 2001). Even in more adverse scenarios such as a cocktail party, listeners can segregate 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). According to Bregman (1990), the human ability to function well in complex acoustic environments is due to a perceptual process termed auditory scene analysis (ASA), which produces a perceptual representation of an individual source in an acoustic mixture.

Organization in ASA, according to Bregman (1990), takes place in two main processes: segmentation and grouping. Segmentation decomposes an auditory scene into groups of contiguous time–frequency (T–F) units or segments, each of which primarily originates from a single sound source (Wang and Brown, 2006). A T–F unit denotes the signal at a particular time frame and frequency. Grouping combines the segments that likely arise from the same source together into a single stream. Thus, each of the formed streams gives a perceptual representation of a sound source. Grouping itself is composed of simultaneous and sequential organization. Simultaneous organization groups segments that overlap in time, and sequential

[*] Corresponding author. Tel.: +1 614 292 7402; fax: +1 614 292 2911.
E-mail addresses: shao.19@osu.edu (Y. Shao), dwang@cse.ohio-state.edu (D. Wang).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.02.003

organization refers to grouping across time. A computational auditory scene analysis (CASA) system that segregates target speech is desirable for many applications. We address sequential grouping in this paper.

Previous systems utilize speech models from automatic speech recognition for speech organization (Barker et al., 2005; Ellis, 2006). However, by listening to a cochannel mixture, one seems to be able to follow the voice of either speaker even when they are speaking in a foreign language as concluded in an informal listening test reported by Wang (2006). In the test, subjects listened to cochannel mixtures of equally loud utterances in languages totally unknown to them, and the tested languages included French, German, Hindi, Japanese, Mandarin, and Spanish. We recently proposed a model-based system that employs voice characteristics and performs sequential organization in cochannel mixtures (Shao and Wang, 2006). This system maximizes the likelihood of grouped speech segments given a set of speaker models. It is reasonable to assume, under some conditions, that models of all the speakers that can appear in an input scene are available in advance. This is known as a close-set situation. Under some other conditions, a system can only assume familiarity with the voice that it is supposed to segregate. This presents an open-set situation where only the target speaker model is available. Furthermore, listening to foreign language mixtures indicates that a listener may not need any knowledge of the talkers to attend to a target voice (Wang, 2006). This is a completely open-set condition where none of the speaker models in the input are available. A major limitation of our previous model-based sequential grouping system is the requirement that input speakers come from a closed set of registered speakers. We seek to extend the grouping system to handle open-set conditions.

Our strategy is to model characteristics of unknown speakers in a systematic way when only target models are available. In speaker verification studies, a universal background model, typically constructed from a large number of speaker models to form a non-target model, has proven to be effective for facilitating the decision of whether the input signal is produced by a claimed speaker (Reynolds et al., 2000; Bimbot et al., 2004). Here, we employ a background model to incorporate a number of intrusion types so as to contrast with a target speaker for sequential grouping.

When target speaker models are not available, we can select speakers that are similar to those unregistered ones. In speaker indexing tasks, the latter processing is called generic modeling (Kwon and Narayanan, 2005). Similar to speaker detection and tracking studies (Dunn et al., 2000), speaker indexing determines who is talking at a particular time in an audio stream. Such a task uses unsupervised methods when there is no prior information about speakers in the input. Typical methods first use a generalized likelihood ratio test (Rice, 1995) to obtain speaker homogenous segments (Dunn et al., 2000; Kwon and Narayanan, 2005), which are then clustered to index underlying speakers in an audio stream and construct corresponding models. However, these systems usually require segments with a minimum length of one second (Dunn et al., 2000) and shorter segments do not represent a speaker well (Kwon and Narayanan, 2005). The data paucity problem caused by short segments usually propagates clustering errors in the indexing process. To tackle this problem, 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, 2004; Kwon and Narayanan, 2005). One way to obtain generic models is to quantize a speaker set (Kwon and Narayanan, 2005). This method clusters speaker models based on the symmetrized Kullback–Leibler (K–L) divergence (Kullback, 1968). Within each cluster, a speaker is randomly selected as a generic model that represents the cluster.

In this paper, we systematically study sequential organization of speech in mixtures that contain speech and nonspeech intrusions. As a special case, we first describe model-based sequential grouping that organizes simultaneous streams under cochannel conditions. Simultaneous streams are obtained by a voiced speech segregation system that performs segmentation and simultaneous grouping (Hu, 2006; Hu and Wang, 2006). Such streams primarily contain T–F energy from a single speaker within a short time period and they are separated in time. We then extend the model-based sequential organization framework from cochannel speech to mixtures that contain more than one interfering talker and non-speech interference. The extension incorporates background models that account for known and unknown interferences. We show that the system is able to function well when only target speaker models are available. Finally, we generalize the system to deal with unregistered target and interfering speakers. More specifically, we employ a speaker quantization method to derive generic models and use them for sequential organization under these open-set conditions. This quantization method performs clustering in a large speaker space based on the K–L divergence measure. Our grouping system then replaces individual speaker models with obtained generic models for sequential organization. Evaluations show that grouping performance using generic models is only moderately lower than that achieved with individual speaker models.

The rest of the paper is organized as follows. Section 2 describes the proposed organization system including extraction of simultaneous streams, model-based sequential grouping, background modeling, and generic modeling. Sequential organization evaluations are presented in Section 3. Section 4 concludes the paper.

2. Sequential organization

As described earlier, in the ASA account, the goal of sequential organization is to integrate separated speech from the same speaker across time. From the CASA perspective, the separated speech refers to simultaneous

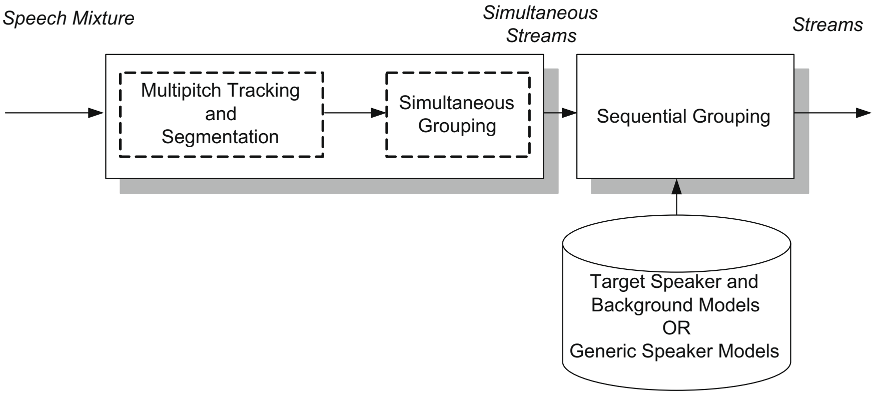
Speech Mixture
Simultaneous Streams
Streams
Multipitch Tracking and Segmentation
Simultaneous Grouping
Sequential Grouping
Target Speaker and Background Models OR Generic Speaker Models

Fig. 1. Diagram of the proposed CASA system for sequential organization of speech.

streams, each of which is composed of segments of contiguous T–F units that primarily originate from a single source. These streams are extracted from the mixture input by segmentation and simultaneous grouping. Thus, the goal of sequential organization in CASA is to organize simultaneous streams into their corresponding sources. In other words, it amounts to assigning simultaneous streams to their sources.

Fig. 1 presents a diagram of the proposed system. First, a voiced speech segregation system generates binary T–F masks that represent simultaneous streams. Then, sequential grouping searches for optimal assignment of simultaneous streams and organizes them into corresponding source streams. For the situations when only the target speaker models are available, the grouping system employs background modeling to handle multi-talker mixtures and mixtures with non-speech intrusions. In the case when none of the speakers in the input are registered, the system utilizes generic models derived by a speaker quantization method for grouping.

2.1. Segmentation and simultaneous grouping

To obtain simultaneous streams, we employ a pitch-based speech segregation system (Hu, 2006; Hu and Wang, 2006). We adopt this system because it makes relatively few assumptions about underlying noise and has been shown to significantly improve the SNR of segregated speech under various noisy conditions. The speech segregation system decomposes input signals into the T–F domain through a bank of gammatone filters (Patterson et al., 1988). This system performs segmentation by merging T–F units using cross-channel correlation and temporal continuity (Hu, 2006; Hu and Wang, 2006). Specifically, in the low frequency range, segments are formed by merging neighboring T–F units with sufficiently high cross-channel correlation in a correlogram, which consists of autocorrelations of filter responses (Wang and Brown, 2006). Since a gammatone filter responds to multiple harmonics in the high frequency range, segments are constructed on the basis of cross-channel correlation of response envelopes in high frequency.

For simultaneous grouping of obtained T–F segments, the speech segregation system estimates pitch contours based on the aforementioned correlogram since pitch is a primary cue for grouping (Bregman, 1990; Wang and Brown, 2006). A segment is grouped into a simultaneous stream that corresponds to a pitch contour if more than half of its T–F units exhibit periodicities that are consistent with the pitch contour. In the low-frequency range where harmonics are resolved, a T–F unit is labeled as consistent if it shows a large response at the estimated pitch period; otherwise it is labeled as inconsistent. For high-frequency channels, the consistency is determined by checking whether the envelope of a unit response shows a variation at a rate close to the estimated pitch period (Hu and Wang, 2006). Subsequently, a simultaneous stream is further expanded to include neighboring units that have the same label.

The speech segregation system outputs simultaneous streams as binary T–F masks, which are estimates of an ideal binary T–F mask (Wang, 2005). As a computational goal of CASA (Wang, 2005), the ideal binary mask is 1 if target energy is stronger than interference energy in the corresponding T–F unit and 0 otherwise. The ideal binary mask is motivated by the human auditory masking phenomenon (Moore, 2003), and under certain conditions provides the maximum SNR gain of all the binary masks (Hu and Wang, 2004). Since a simultaneous stream consists of contiguous T–F regions dominated by a speaker, a binary mask produced by the segregation system is an estimate of the ideal binary mask for the underlying speaker within the corresponding time interval of the stream. Thus, given a speech mixture, the segregation system generates a group of binary T–F masks, which represent simultaneous streams. Fig. 2b shows a collection of such streams obtained from a cochannel mixture in Fig. 2a by the speech segregation 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 yet to be grouped in time, which is the task of sequential grouping.

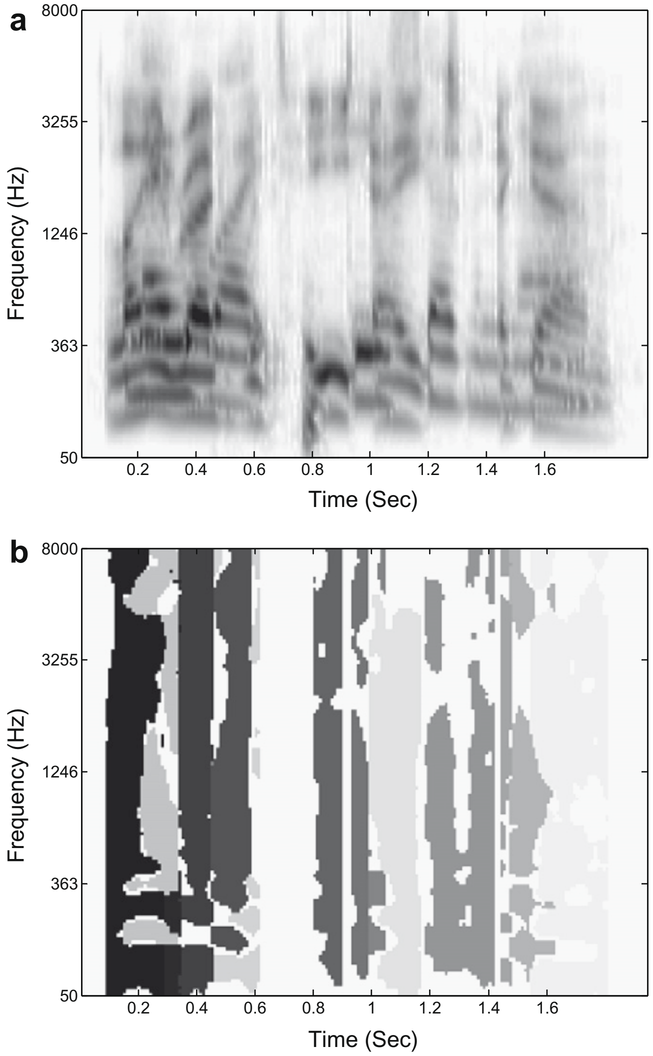

Fig. 2. Illustrations of noisy speech and estimated simultaneous streams. Plot (a) shows a cochleagram of a two-talker utterance mixed at 0 dB SNR. Darker color indicates stronger energy within the corresponding time–frequency unit. Plot (b) presents simultaneous streams derived from utterance in (a). White color shows the background. Different gray-colored regions indicate simultaneous streams that have been grouped across frequency but not across time.

2.2. Model-based sequential organization

In this section, we briefly describe our speaker-model based sequential grouping framework (Shao and Wang, 2006). We derive a computational objective in the context of cochannel speaker recognition. Assume that there are a set of K registered speaker models $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$, and they are constructed as Gaussian mixture models (GMM) using an EM algorithm (Reynolds, 1995). Given a cochannel input, CASA is used to generate N simultaneous streams, $Y = \{S_1, S_2, \ldots, S_i, \ldots, S_N\}$, each of which is deemed to primarily originate from a single speaker and represented by a binary T–F mask.

In cochannel speech, simultaneous streams must be organized into two speaker streams by sequential organization. For example, a possible stream assignment (grouping) may look like $\{S_1^0, S_2^1, \ldots, S_i^1, \ldots, S_N^0\}$, where superscripts, 0 and 1, do not represent the speaker identities but only denote that those simultaneous streams marked with the same label are from the same speaker. Therefore, we formulate the joint computational objective of sequential grouping and speaker identification (SID) as finding a pair of speaker models $\hat{\lambda}_{\mathrm{I}}$ and $\hat{\lambda}_{\mathrm{II}}$ together with a simultaneous stream assignment $\hat{g}$ that jointly maximize the posterior probability:

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}}, = \mathop{\arg\max}_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, g \in G} P(g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} | Y), \tag{1}$$

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

Assuming that the assignment is independent of specific models and that speaker models are independent of each other, it has been shown that (1) becomes (Shao and Wang, 2006)

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}} = \mathop{\arg\max}_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, g \in G} P(Y | g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}). \tag{2}$$

The computational objective in (2) is to find the optimal hypothesis of two speakers and one assignment that yield the maximal conditional probability. Given the simultaneous streams and individual speaker models trained from clean speech, the likelihood maximization amounts to searching for the globally optimal hypothesis in the joint speaker and assignment space, Λ and G. Note that the conditional probability is essentially the joint SID score of assigned simultaneous streams. Given an assignment g, we denote Y^0 as the subset of simultaneous streams labeled 0, and Y^1 the subset labeled 1. Y^0 and Y^1 are complementary.

In speaker recognition studies (Reynolds, 1995; Furui, 2001), feature vectors extracted from individual frames are usually assumed to be independent of one another for text-independent tasks. Since we are interested in sequential organization of speech independent of text information, we adopt a similar assumption that two simultaneous streams, S_i and S_j, are independent of each other given the speaker models. In addition, different labels of simultaneous streams (superscripts 0 and 1) correspond to different speakers. Hence, the conditional probability in (2) can be written below (Shao and Wang, 2006):

$$\begin{aligned} P(Y | g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) &= P(Y^0 | \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) P(Y^1 | \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \\ &= \prod_{S_i \in Y^0} P(S_i | \lambda_{\mathrm{I}}) \prod_{S_j \in Y^1} P(S_j | \lambda_{\mathrm{II}}). \end{aligned} \tag{3}$$

The likelihood in (3) calculates the probability of having a simultaneous stream, S_i or S_j, belong to a speaker model λ. This likelihood is not directly obtainable using conventional methods (Huang et al., 2001) that calculate the probability of a complete feature frame given a model because a binary T–F mask that represents a simultaneous stream includes both reliable and unreliable T–F units within a frame. To incorporate the binary masks for sequential grouping, we apply a feature reconstruction and uncer-

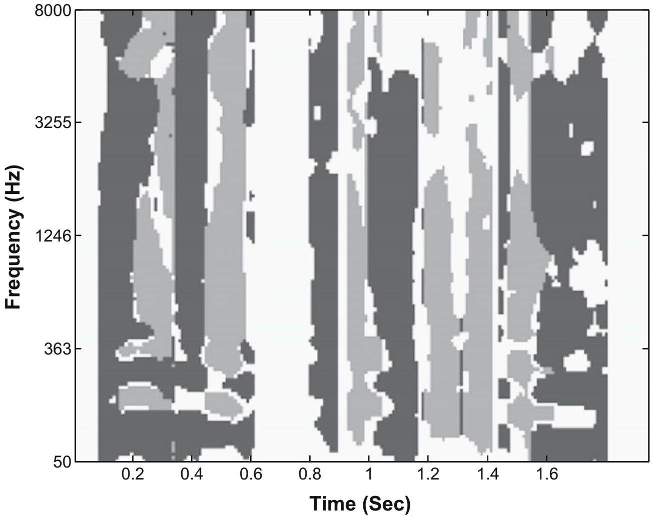

Fig. 3. Segregated speaker streams after sequential grouping of the simultaneous streams shown in Fig. 2b. White color shows the background. The two gray-colored sets of T–F regions represent two separated speaker streams.

tainty decoding method described in Shao et al. (2007). More specifically, we employ a novel auditory feature, gammatone frequency cepstral coefficients (GFCC), which are derived from gammatone filtering and cepstral analysis (Shao et al., 2007). The unreliable T–F components indicated by 0 in a binary mask are reconstructed using a speech prior (Raj et al., 2004) to enhance GFCCs. Moreover, reconstruction uncertainties are estimated to compensate for reconstruction errors (Shao et al., 2007; Srinivasan and Wang, 2007). The enhanced GFCCs are then used in conjunction with uncertainty estimates by an uncertainty decoder (Deng et al., 2005) to calculate the likelihood of a stream given a speaker in (3).

According to (2) and (3), given a pair of speakers, the best assignment of simultaneous streams by the two speaker models is determined by comparing the aforementioned stream likelihoods for all the simultaneous streams (Shao and Wang, 2006). This can be achieved efficiently since streams are assumed to be independent. Then, we iterate through all the possible pairs of speakers and find the optimal speaker pair and stream assignment (Shao and Wang, 2006). Fig. 3 illustrates two segregated speaker streams after sequential organization of the simultaneous streams in Fig. 2b. The two speaker streams are shown as two different gray colors.

2.3. Modeling of background

Under multi-talker conditions such as cochannel speech, the input to the system is a mixture composed of voices from multiple speakers. The voice of interest is designated as target and the others as interferences. In certain circumstances such as a meeting, there can be more than one interfering speaker. To tackle such conditions, we can extend the model-based sequential grouping framework by replacing the speaker pair with a speaker triplet, a speaker quadruplet, etc., in (1). This will end up with a computational objective similar to (2) by applying the same derivation in Section 2.2. An optimal stream assignment g to M speakers can be formulated as,

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}}, \ldots \hat{\lambda}_M = \operatorname*{arg\,max}_{\lambda_{\mathrm{I}},\lambda_{\mathrm{II}},\ldots\lambda_M \in \Lambda, g \in G} P(Y|g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, \ldots \lambda_M), \qquad (4)$$

where $M \leqslant K$. Naturally, the components of g take values from 1 to M. This extension makes an explicit assumption of the speaker number in a mixture. Without this assumption, one could further extend the formulation in (4) by including 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 would be chosen as the estimate, and the grouping hypothesis associated with the speaker number estimate would provide the optimal assignment of simultaneous streams. Nevertheless, this extension has the problem of scalability. For example, in the case of a cocktail party, there may be a large number of active speakers in the background. Indeed, there can be so many voices in the background that a listener perceives something more like babble noise. Searching through all the combinations of up to M speakers greatly increases computational time.

To deal with multi-talker conditions, consider how existing CASA systems treat interference (Wang and Brown, 2006). Typically, the target signal is segregated into a foreground (target) stream while the remaining parts of the input signal are organized into the background (interference) stream. This process applies regardless of actual interference types or numbers. Hence, instead of modeling individual speakers, we propose to build a background model that accounts for all interfering speakers as well as unregistered ones. This background model is constructed as a GMM by training on a large sample pool of speakers. Conceptually, it is analogous to a universal background model in speaker verification studies (Bimbot et al., 2004). Thus, we replace one speaker in (2) with a general background model λ_{B} and perform the search over the target speaker as follows:

$$\hat{g}, \hat{\lambda} = \operatorname*{arg\,max}_{\lambda \in \Lambda, g \in G} P(Y|g, \lambda, \lambda_{\mathrm{B}}). \qquad (5)$$

The formulation in (5) handles multi-talker conditions where the grouping system possesses the models of the target speakers but not others in an input mixture. This formulation is also able to deal with acoustic conditions with non-speech intrusion sources. To perform such a task, a background model is constructed using a large sample pool of non-speech intrusions; the existence of typical noise corpora such as the Noisex (Varga and Steeneken, 1993) and environmental sounds (Hu, 2006) facilitates such construction. The grouping system then organizes target speech into a foreground stream while assigning the remaining intrusions to the background. Moreover, in the case where interfering sources could be either speech or non-speech, our formulation of the background model can be extended to combine both multi-talker and non-speech interference conditions.

2.4. Generic modeling with speaker quantization

The sequential grouping approach described in the preceding subsection requires the prior knowledge of target speakers. Here, we extend the grouping algorithm to handle acoustic conditions where none of the speakers in an input mixture are registered. Specifically, we employ a speaker quantization method to derive generic models for this purpose. The basic idea of generic speaker modeling and speaker quantization is to identify and construct a small number of models that represent a much larger speaker set. Generally speaking, quantization can be applied either in the feature space or in the model space. The former approach is widely used in automatic speech recognition (Huang et al., 2001). However, without top-down constraints that model a speaker, a quantized model produced by this approach more likely reflects intrinsic speech classes of the feature space instead of speakers. Hence, we adopt the latter approach that performs quantization over speaker models. We propose to use a speaker quantization method that is similar to a quantization method used in speaker indexing (Kwon and Narayanan, 2005) to construct generic models for sequential grouping.

We first construct a large set of speaker models $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$, each of which is, once again, a GMM. Pair-wise distances between two models are calculated for each speaker pair within the set. Thus, the resulting distance matrix describes a distribution of all the models within the speaker space. Then, we apply a K-means clustering method (Duda et al., 2001) to obtain a number of clusters based on the distance matrix. Finally, within each cluster, the model that has the shortest average distance to the remaining models is selected as a generic model. Fig. 4 illustrates quantized generic models.

Since 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,

$$KL(f\|g) = \int f(x)\log\frac{f(x)}{g(x)}dx \tag{6}$$

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

$$D(f,g) = KL(f\|g) + KL(g\|f) \tag{7}$$

However, no closed-form solution exists for the KLD when $f(x)$ and $g(x)$ are GMMs (Vasconcelos, 2004; Silva and Narayanan, 2006). 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 known method that asymptotically estimates the KLD is Monte Carlo simulation (Ben et al., 2002; Vasconcelos, 2004; Hershey and Olsen, 2007). Here, we apply a Monte Carlo method to calculate the KLD between two GMMs. First, we draw N samples $\{x_i: i = 1, \ldots, N\}$ from $f(x)$. KLD is then approximated as,

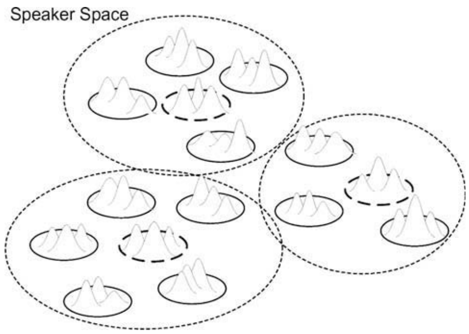

Fig. 4. Illustration of speaker quantization. The solid circles represent individual speakers, which are modeled as GMMs. The thin dashed circles represent clusters obtained by the speaker quantization method. The thick dashed circles denote selected generic models within each cluster.

$$KL(f\|g) \approx \frac{1}{N}\sum_{i=1}^{N}\log\frac{f(x_i)}{g(x_i)}. \tag{8}$$

$KL(g\|f)$ is estimated in the same way using 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 where we apply quantization.

3. Evaluation and comparison

This section systematically evaluates the performance of the sequential organization system. We adopt a performance metric that measures SNR of segregated speech after it is resynthesized in the time domain (Hu and Wang, 2004). This metric 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. This measure directly compares signals in the time domain as

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n (s(n)-\hat{s}(n))^2}, \tag{9}$$

where n indexes time.

3.1. Evaluation on speech background

To simulate multi-talker conditions, we create test utterances from the speech separation challenge (SSC) corpus (Cooke and Lee, 2006). This corpus provides speech materials from 34 speakers. Our training data is taken from the training part of SSC corpus and each of the 34 speakers is modeled as a 64-component GMM of 30-dimensional

GFCCs (Shao et al., 2007). The corpus also provides 600 clean utterances in the test set, so we use these utterances to generate two-talker, three-talker, and four-talker mixtures. 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. For each mixture, interfering utterances are either cut or appended with themselves to match the length of a target utterance. Interfering utterances are also scaled to have equally strong energy. Every multi-talker condition comprises an SNR range of −6 dB, 0 dB, 6 dB, and 12 dB that provides a wide range of noisy environments. Each of the SNR conditions contains 600 mixtures.

Evaluation results are presented in Tables 1–3 for two-talker, three-talker, and four-talker conditions, respectively. To establish a performance upper-bound, we first construct the ideal binary mask of each mixture (see Section 2.1). Then, we find an ideal sequential grouping (ISG) mask for the mixture by grouping simultaneous streams into the target stream according to its ideal binary mask. 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. The first rows of the tables show the SNR results of ISG. ISG significantly improves SNRs at −6 dB, 0 dB, and 6 dB. However, since an ISG mask is generated by grouping simultaneous streams which are produced by a voiced speech segregation system (Hu, 2006; Hu and Wang, 2006), 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 the 12 dB condition.

The second rows in the tables present the baseline performance by randomly assigning a stream to either the target or the background stream. When an input SNR is positive, the output SNR is expected to be lower because simultaneous streams are randomly assigned. On the other hand, with a negative input SNR, random grouping tends to produce a higher output SNR as in the case of −6 dB.

Similar to our previous study (Shao and Wang, 2006), we also conduct sequential grouping using pitch information as an alternative approach. We first evaluate grouping performance based on prior pitch, which is 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 third rows. This performance places an upper-bound for all sequential grouping methods that utilize pitch. To implement a pitch-based approach, we employ a clustering method that is based on the mean pitch values of simultaneous streams. The number of clusters is set to the speaker number in a test mixture. SNR results are shown in the 'Pitch-based Grouping' rows. The results are worse than the performance upper-bound using prior pitch because the clustering method is not able to differentiate speakers when their pitch contours are close to each other

Table 1
Evaluation of sequential grouping of two-talker mixtures. Numbers in the table show output SNR (dB) of segregated speech.

Methods	Input SNR (dB)			
	−6	0	6	12
Ideal sequential grouping	3.604	6.483	8.287	8.865
Random grouping	−2.459	0.487	2.474	2.699
Grouping using prior pitch	2.690	4.597	6.711	7.293
Pitch-based grouping	0.598	4.167	6.013	6.527
Background modeling	2.545	5.065	6.708	7.623

Table 2
Evaluation of sequential grouping of three-talker mixtures. Numbers in the table show output SNR (dB) of segregated speech.

Methods	Input SNR (dB)			
	−6	0	6	12
Ideal sequential grouping	3.006	5.793	8.968	10.937
Random grouping	−3.648	0.873	2.566	3.334
Grouping using prior pitch	2.298	3.931	5.552	6.860
Pitch-based grouping	−0.622	3.121	5.233	6.158
Background modeling	1.296	4.483	7.278	9.287

Table 3
Evaluation of sequential grouping of four-talker mixtures. Numbers in the table show output SNR (dB) of segregated speech.

Methods	Input SNR (dB)			
	−6	0	6	12
Ideal sequential grouping	2.529	5.492	8.965	11.004
Random grouping	−3.177	1.121	2.554	3.302
Grouping using prior pitch	1.827	3.722	5.236	6.804
Pitch-based grouping	−0.373	2.777	4.264	4.786
Background modeling	0.636	4.169	7.355	9.314

The last rows in the tables present sequential grouping results using general background models. Here, the grouping system only assumes information about target speakers. A general background model is trained on all the speakers other than the target and the interfering speakers. Specifically, for each input mixture, we randomly select a group of 10 speakers from the SSC corpus as the target speaker set. In other words, this simulates the acoustic condition where a system is only familiar with the voice of a target speaker. It can be observed from the tables that background modeling performs significantly better than the baseline and the pitch-based methods. Except at −6 dB, where the binary mask of a simultaneous stream is too sparse for reliable feature reconstruction and likelihood calculation, our model-based approach performs better even than grouping based on prior pitch. It is worth emphasizing that background modeling achieves a performance level that is only moderately lower than the upper-bound obtained by ISG.

Table 4
Sequential grouping of mixtures with non-speech interferences. Numbers in the table show output SNR (dB) of segregated speech. The test mixtures contain babble noise in (a), destroyer noise in (b), F16 noise in (c) and factory noise in (d).

Methods	Input SNR (dB)			
	−6	0	6	12
(a) Babble				
Ideal sequential grouping	2.190	5.763	9.074	11.054
Random grouping	0.349	2.159	2.752	3.212
Background modeling	1.065	5.302	8.849	11.002
Unregistered target	0.802	4.238	7.617	10.294
(b) Destroyer				
Ideal sequential grouping	2.670	6.486	9.693	11.262
Random grouping	−0.822	2.082	3.129	3.286
Background mdeling	1.342	4.075	9.062	11.052
Unregistered target	1.215	3.018	7.704	10.280
(c) F16				
Ideal sequential grouping	3.586	7.587	10.197	11.477
Random grouping	1.257	2.665	3.079	3.287
Background modeling	3.213	6.767	9.833	11.333
Unregistered target	2.992	5.717	8.588	10.440
(d) Factory				
Ideal sequential grouping	2.958	7.063	9.855	11.420
Random grouping	1.225	2.505	3.067	3.382
Background modeling	2.778	6.576	9.500	11.305
Unregistered target	2.599	5.670	7.996	10.376

3.2. Evaluation on non-speech background

In this section, we evaluate the background modeling method to deal with non-speech interferences. Similar to the preceding subsection, we create test mixtures by mixing clean test utterances of 34 speakers from SSC with four non-speech noise types: babble noise (100 speakers), destroyer (a navy ship) operation room noise, F16 cockpit noise, and factory noise. These four noise types are selected from the Noisex 92 corpus (Varga and Steeneken, 1993), which is widely used for robust speech recognition studies. The first two types contain a noisy background with many talkers speaking at the same time. They are considered as non-speech here because with so many voices the background does not exhibit clear speech patterns.

Evaluation results are shown in Table 4 for the babble, destroyer, F16 and factory noise types separately. The first two rows in each part of the table present the upper-bound performance and the baseline performance obtained by ISG and random grouping, respectively. The third rows show SNR results by employing background modeling. A background model is trained from pooled noise samples. These noise samples include the 4 noise types and 15 other non-speech noise types (Hu, 2006): 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, and babble noise (16 speakers). This simulates a test condition where an actual noise source in a mixture originates from a number of possible noise types. The resulting performance is close to that of ISG under most of the SNR conditions since the trained models of target speech and non-speech interferences are quite different.

The last rows in the table, ‘Unregistered Target’, show a test configuration that removes the target speaker 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 system performs grouping by automatically selecting the most likely speaker out of the remaining speakers as the target. Compared with the registered target condition in the third rows, the grouping performance here degrades only moderately. These results imply that a set of 30–40 speakers likely contains a speaker that is acoustically close to the target. Thus, a small set of speaker models might be sufficient for sequential grouping when input speakers are not registered. This observation led us to design an algorithm presented in Section 2.4 that creates a set of generic speaker models.

In the preceding evaluations, the grouping system assumes the knowledge of whether input mixtures contain multi-talker or non-speech intrusions. Our formulation of the background model can be extended to handle acoustic conditions that contain both speech and non-speech intrusions (Shao, 2007). As a direct extension, we can construct a general background model by incorporating both the multi-talker and the non-speech background models. A simpler approach is to combine these two GMM background models together and adjusting the weights of their Gaussian components accordingly. The grouping system then employs the combined model in (5) for sequential organization.

3.3. Evaluation on speaker quantization

Since speaker quantization requires a large number of speakers, we adopt the 2002 NIST Speaker Recognition Evaluation corpus (Przybocki and Martin, 2004) for evaluation. Unlike our previous evaluations and most of the evaluations in CASA studies (Wang and Brown, 2006), this corpus is composed of telephone recordings, which have a narrower bandwidth than typical microphone recordings. We use the ‘1-speaker detection task’ 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 s recording of concatenated cell phone utterances.

Considering computational complexity, our evaluations are focused on cochannel mixtures; regarding other kinds of mixtures inference could be made from our earlier evaluations. First, the original recordings are sliced into short utterances of 4 s each. Given the resulting 30 utterances for each speaker, 26 of them are randomly selected to construct speaker models while the remaining four are used for testing. Each of the 330 speakers is modeled as a 32-component GMM of 30-dimensional GFCCs. Then, cochannel mixtures are created at SNRs of −6 dB, 0 dB, 6 dB, and 12 dB. For each speaker designated as target, those four test utterances are mixed with randomly chosen test utter-

Table 5
Sequential grouping evaluation with generic models. Numbers in the table show output SNR (dB) of segregated speech. The test utterances are two-talker mixtures. Numbers in the parentheses refer to the number of generic models.

Methods	Input SNR (dB)			
	−6	0	6	12
Ideal sequential grouping	5.718	7.494	9.704	11.445
Known speaker identity	2.193	4.766	7.659	9.872
Random grouping	−3.396	0.396	2.301	2.945
Exhaustive search	1.515	4.397	7.270	9.442
Exhaustive search with subset of 40	1.808	4.637	7.443	9.590
Speaker quantization (20)	−0.558	2.846	5.823	7.547
Speaker quantization (140)	−0.319	3.093	6.117	8.215

ances from the remaining speakers. Therefore, each SNR consists of 1320 cochannel mixtures.

Evaluation results are shown in Table 5. The first row presents SNR results obtained by ISG. Like the results presented in earlier tables, errors in simultaneous grouping cause the output SNR to be lower than the input SNR under the 12 dB condition. 'Known speaker identity' denotes a condition where identities of speakers in a mixture are provided to the system beforehand. In short, the grouping algorithm reduces to a hypothesis test between the two speaker models. This places an actual performance upper-bound for all the model-based methods. Compared to ideal sequential grouping, grouping performance degrades faster with decreasing SNR. This indicates that likelihood scores become less reliable when SNR decreases because there are more missing T–F units to be reconstructed from fewer reliable units. The following row, 'Random Grouping', randomly assigns simultaneous streams, thus setting a performance lower-bound.

Before evaluating quantized generic models, we also show how the grouping system fares using individual speaker models. SNR results are given in the 'Exhaustive Search' row. This search requires a large amount of computation time with the complete set of 330 speakers (see Table 6). Since our previous study performs evaluation using a smaller set of 38 speakers (Shao and Wang, 2006), we also conduct an experiment that uses a reduced set of 40 speakers. The reduced set is composed of two underlying speakers in an input mixture and 38 speakers that are randomly selected from the remaining 328 speakers. Evaluation results are shown in the following row. 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 about 0.5 dB. This suggests the effectiveness of our model-based grouping method. 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 system to differentiate them.

The last two rows present grouping results obtained by performing speaker quantization with 20 and 140 speaker clusters respectively. Since the cochannel mixtures in the test set are created from utterances from the 330 speakers, we simulate the acoustic conditions that none of the test speakers are registered by employing a method similar to cross validation (Russell and Norvig, 2003). More specifically, for each cochannel input, we remove the two underlying speakers from the speaker set and perform speaker quantization on the remaining 328 speakers. Thus, we create a different generic model set for each test speaker pair. 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.

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 generic models and unregistered speakers in the input 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, and grouping results and average computation times per test file are presented in Table 6. The reported times were recorded from Matlab implementation on a Dell PowerEdge 1850 server with dual Xeon 3.4 GHz processor and 4 GB memory. In the table, a number after 'Speaker quantization' denotes the number of generic models. SNR performance is significantly improved by increasing the number of generic models from 20 to 60. While the improvement stalls from 60 to 90, the performance is further improved beyond 90. 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 21 times by increasing the number of generic models from 20 to 140.

4. Discussion and conclusion

Sequential organization groups sound components of the same source across time into the same stream. In this paper, we have extended a model-based sequential grouping framework (Shao and Wang, 2006) to include general background modeling in order to handle multiple interfering speakers and non-speech intrusions. By employing a general background model that takes different interference types into account, our system achieves a level of performance close to that with registered interference models. 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.

In general our system produces better output SNRs when interfering signals are non-speech. This is to be expected because even though each speaker's voice is unique it is still easier to discriminate speech from non-speech

Table 6
Sequential grouping evaluation for different numbers of generic models. Numbers in the table show output SNR (dB) of segregated speech as well as computing times. Numbers in parentheses refer to the number of generic models.

Methods	Input SNR (dB)				
	−6	0	6	12	Computation time (s)
Speaker quantization (20)	−0.558	2.846	5.823	7.547	74.8
Speaker quantization (40)	−0.314	2.931	5.868	7.618	159.8
Speaker quantization (60)	−0.479	2.963	5.952	8.044	298.3
Speaker quantization (80)	−0.493	2.948	5.996	8.058	472.5
Speaker quantization (90)	−0.427	2.985	5.978	8.013	636.7
Speaker quantization (100)	−0.534	2.941	6.035	8.139	802.5
Speaker quantization (120)	−0.494	2.853	6.043	8.226	1133.6
Speaker quantization (140)	−0.319	3.093	6.117	8.215	1614.1

than telling apart different voices. For non-speech interference, other methods can be applied. For example, the decoding model of Barker et al. (2005) uses a speech recognizer to organize segments into speech and non-speech ones. Recently, Hu and Wang (2008) introduced a classification method to decide whether segments during unvoiced intervals belong to speech or interference. In multi-talker situations, the performance of our system degrades when the number of interfering speakers increases. However, when this number becomes large, the combined signal of many interfering talkers approaches babble noise, which becomes easier for sequential organization.

Since our approach is model based, it requires a training process with training data from both target and interfering sources. Like other model-based methods, the requirement of prior training poses certain limitations on the potential application of the system. Indeed, it may be impractical to collect all the possible intrusion types in the world. However, as evidenced in the evaluations, our approach is able to deal with conditions with hundreds of speakers. Even though a training corpus does not account for all the possible non-speech noise types, a trained background model may generalize to noises not included in the corpus as long as the corpus is reasonably representative of the kinds of interference encountered in an application domain. In addition, voice characteristics might be quite unique compared to acoustic properties of non-speech signals, hence placing only modest demands on the accuracy of a background model.

Our sequential grouping system organizes simultaneous streams that are produced by segmentation and simultaneous grouping. Thus, errors in simultaneous grouping, particularly the omission of unvoiced speech, will propagate to the sequential organization process. Such errors have limited the performance of our model in high input SNR conditions. Unvoiced speech segregation has been addressed in a recent study (Hu and Wang, 2008). Unfortunately, this study only deals with non-speech interference. A system that is capable of segregating unvoiced speech from a general background containing both speech and non-speech interference is yet to be developed. Future research also needs to address how to integrate sequential grouping and simultaneous grouping and optimize CASA performance as a whole.

Acknowledgement

This research was supported in part by an AFOSR Grant (FA9550-04-1-0117), an NSF Grant (IIS-0081058), and an AFRL Grant (FA8750-04-1-0093).

References

Barker, J., Cooke, M., Ellis, D., 2005. Decoding speech in the presence of other sources. Speech Comm. 45 (1), 5–25.

Ben, M., Blouet, R., Bimbot, F., 2002. A Monte Carlo method for score normalization in automatic speaker verification using Kullback–Leibler distances. In: Proc. ICASSP, Vol. I, pp. 689–692.

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

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambridge, MA.

Brungart, D.S., 2001. Information and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Amer. 109, 1101–1109.

Cherry, E.C., 1953. Some experiments on the recognition of speech with one and with two ears. J. Acoust. Soc. Amer. 25, 975–979.

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

Deng, L., Droppo, J., Acero, A., 2005. Dynamic compensation of Hmm variants using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Trans. Speech Audio Process. 13, 412–421.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. Wiley, New York.

Dunn, R.B., Reynolds, D.A., Quatieri, T.F., 2000. Approaches to speaker detection and tracking in conversational speech. Digital Signal Process. 10, 93–112.

Ellis, D.P.W., 2006. Model-based scene analysis. In: Wang, D.L., Brown, G.J. (Eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken, NJ, pp. 115–146.

Furui, S., 2001. Digital Speech Processing, Synthesis, and Recognition. Marcel Dekker, New York.

Helmholtz, H., 1863. On the Sensation of Tone (A.J. Ellis, Trans.), Second English ed., Dover Publishers, New York.

Hershey, J.R., Olsen, P.A., 2007. Approximating the Kullback–Leibler divergence between Gaussian mixture models. In: Proc. ICASSP, Vol. IV, pp. 317–320.

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

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

Hu, G., Wang, D.L., 2006. An auditory scene analysis approach to monaural speech separation. In: Hansler, E., Schmidt, G. (Eds.), Topics in Acoustic Echo and Noise Control. Springer, Heidelberg, pp. 485–515.

Hu, G., Wang, D.L., 2008. Segregation of unvoiced speech from nonspeech interference. J. Acoust. Soc. Amer. 124, 1306–1319.

Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing. Prentice Hall, Upper Saddle River.

Kullback, S., 1968. Information Theory and Statistics. Dover, New York.

Kwon, S., Narayanan, S., 2004. Speaker model quantization for unsupervised speaker indexing. In: Proc. ICSLP, pp. 1517–1520.

Kwon, S., Narayanan, S., 2005. Unsupervised speaker indexing using generic models. IEEE Trans. Speech Audio Process. 13 (5), 1004–1013.

Moore, B.C.J., 2003. An Introduction to the Psychology of Hearing, fifth ed. Academic, San Diego.

Patterson, R.D., Nimmo-Smith I., Holdsworth J., Rice P., 1988. An efficient auditory filterbank based on the gammatone function. APU Report 2341, Cambridge, UK, MRC Applied Psychology Unit.

Przybocki, M.A., Martin, A.F., 2004. NIST Speaker Recognition Evaluation Chronicles. In: Proc. Odyssey 2004.

Quatieri, T.F., Danisewicz, R.G., 1990. An approach to co-channel talker interference suppression using a sinusoidal model for speech. IEEE Trans. Acoust. Speech Signal Process. 38, 56–69.

Raj, B., Seltzer, M.L., Stern, R.M., 2004. Reconstruction of missing features for robust speech recognition. Speech Comm. 43, 275–296.

Reynolds, D.A., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Comm. 17, 91–108.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10, 19–41.

Rice, J.A., 1995. Mathematical Statistics and Data Analysis. Duxbury Press, Belmont, CA.

Russell, S., Norvig, P., 2003. Artificial Intelligence: A Modern Approach, second ed. Prentice Hall, Upper Saddle River, NJ.

Shao, Y., 2007. Sequential organization in computational auditory scene analysis. Ph.D. Dissertation, The Ohio State University.

Shao, Y., Srinivasan, S., Wang, D.L., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In: Proc. ICASSP, Vol. IV, pp. 277–280.

Shao, Y., Wang, D.L., 2006. Model-based sequential organization in cochannel speech. IEEE Trans. Audio Speech Lang. Process. 14 (1), 289–298.

Silva, J., Narayanan, S., 2006. Average divergence distance as a statistical discrimination measure for hidden Markov models. IEEE Trans. Audio Speech Lang. Process. 14 (3), 890–906.

Srinivasan, S., Wang, D.L., 2007. Transforming binary uncertainties for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 15 (7), 2130–2140.

Varga, A., 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 Comm. 12 (3), 247–251.

Vasconcelos, N., 2004. On the efficient evaluation of probabilistic similarity functions for image retrieval. IEEE Trans. Inform. Theory 50 (7), 1482–1496.

Wang, D.L., 2005. On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (Ed.), Speech Separation by Humans and Machines. Kluwer Academic, Norwell, MA, pp. 181–197.

Wang, D.L., 2006. Feature-based speech segregation. In: Wang, D.L., Brown, G.J. (Eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken, NJ, pp. 81–114.

Wang, D.L., Brown, G.J. (Eds.), 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken, NJ.