Model-Based Sequential Organization in Cochannel Speech

Yang Shao, *Student Member, IEEE,* and DeLiang Wang, *Fellow, IEEE*

***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 energy of the same speaker across time into a single stream. In this paper, we focus on sequential organization in cochannel speech, or mixtures of two voices. We extract minimally corrupted segments, or usable speech, in cochannel speech using a robust multipitch tracking algorithm. The extracted usable speech is shown to capture speaker characteristics and improves speaker identification (SID) performance across various target-to-interferer ratios. To utilize speaker characteristics for sequential organization, we extend the traditional SID framework to cochannel speech and derive a joint objective for sequential grouping and SID, leading to a problem of search for the optimum hypothesis. Subsequently we propose a hypothesis pruning algorithm based on speaker models in order to make the search computationally efficient. Evaluation results show that the proposed system approaches the ceiling SID performance obtained with prior pitch information and yields significant improvement over alternative approaches to sequential organization.**

***Index Terms*—Auditory scene analysis, cochannel speech, model-based approach, sequential organization, speaker identification (SID), usable speech.**

I. Introduction

COCHANNEL speech is a combination of speech utterances from two talkers, usually produced when two speech signals are transmitted over a single communication channel. Unlike conversations, talkers from different channels are not aware of each other in cochannel speech. Consequently, speech from both channels has large overlap, which presents a considerable challenge to automatic speaker and speech recognition. On the other hand, for a cochannel recording that has comparable energies from both talkers [e.g., target-to-interferer ratio (TIR) is zero], human listeners can readily select and follow one speaker's voice [6]. 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 is not exceedingly low [4], [8], [12]. Bregman [4] describes this process of auditory perception as auditory scene analysis, which is composed of simultaneous organization and sequential organization. The former integrates concurrent sound components and the latter integrates components across time into the same perceptual stream. Most of the existing computational auditory scene analysis systems, e.g., [5] and [13], address only simultaneous organization. It is well known that human listeners use speaker characteristics, such as pitch and vocal tract information to identify a speaker's voice [23] and such characteristics have been incorporated in models of automatic speaker recognition [1], [10], [17], [20], [21].

In this paper, we study how to use speaker characteristics, particularly speaker models, for sequential organization of time–frequency energy of the same speaker into a single stream in cochannel speech. As a result of successful sequential organization, speaker recognition from cochannel mixtures should improve. Hence, we also study the potential benefits of sequential organization for cochannel speaker identification (SID).

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 [18], [19]. Zissman and Seward [32] examined pitch continuity in cochannel speech and assigned 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.* [18] estimated the dominant pitch and then reconstructed 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. Afterward, 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 the divergence measure proposed by Carlson and Clement [7], which is the symmetrized Kullback–Leibler divergence [15]. Because of the short-term processing, the spectral comparison is biased toward the comparison of phonetic information contained in a frame instead of speaker characteristics. Therefore, to capture speaker characteristics, it is desirable to base comparison on speaker homogeneous segments, which consists of a number of time frames dominated by one speaker.

In automatic speaker recognition, as pointed out in [16], the intelligibility and quality of extracted speech are not important. What the system needs are portions of the speech that

Manuscript received May 7, 2004; revised December 3, 2004. This research was supported in part by AFOSR under Grant FA9550-04-1-0117, in part by the National Science Foundation under Grant IIS-0081058, and in part by the AFRL under Grant FA8750-04-1-0093. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy J. Hazen.

The authors are with the Department of Computer Science and Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: shaoy@cse.ohio-state.edu; dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TSA.2005.854106

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 [16].

Previous studies [14], [16] find that voiced segments contain most of the information for SID and have developed criteria such as frame-level TIR and spectral autocorrelation ratio to extract usable speech in cochannel mixtures. Results show that 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 produces frames in which energy from one speaker is much stronger than that of the other. Spectral autocorrelation ratio estimates the ratio between dominant peak and valley in the autocorrelation of the spectrum in order to decide whether a frame is well structured (single-speaker speech) or unstructured (corrupted speech). Finally, the extracted usable segments are grouped using frame-level TIRs. It is a simple and effective method and shows a substantial improvement in SID performance. However, frame-level TIRs are hard to estimate from mixture speech. A further study in [27] explored a maximum-likelihood decision in an attempt to determine the speakers that generate usable speech segments.

Studies have been conducted on speaker detection and tracking in multispeaker environments such as conversational speech and broadcast news (see, e.g., [9] and [31]). Various methods, supervised or unsupervised, have been explored. A typical method [9] is to use log-likelihood ratio scores, calculated from trained Gaussian mixture models (GMM) for speakers and a universal background model, to partition a recording into homogeneous segments and then cluster the segments. However, such methods cannot be applied to cochannel speech because, as mentioned earlier, cochannel talkers strongly overlap, resulting in very short speaker-homogenous segments. In the case of 0-dB TIR, such segments typically last 30–300 ms, far shorter than the optimal segment length of around 2.5 s and the typical minimum length of 1 s for speaker clustering [9]. As pointed out in [16], a speaker recognizer's ability to identify talkers based on pooled frame-level scores is sharply reduced if available observation frames are limited in number, especially when the overall length is less than 500 ms. To verify this, we have explored segment clustering for sequential grouping ourselves; specifically, segments are iteratively clustered based on distance measures in the feature space, such as cepstral coefficients. The result is barely above the chance level of 50%, which is obtained by randomly putting each segment into one of the two clusters.

In this paper, we propose to sequentially organize automatically extracted usable speech, i.e., speaker-homogeneous segments, into streams. Our method employs a robust multipitch tracking algorithm proposed recently [28] for extraction. We develop a computational objective for joint cochannel SID and sequential grouping, or speaker assignment, of usable speech. Our formulation leads to a search problem to find an optimal hypothesis in the joint speaker and grouping space. Exhaustive search finds the optimal hypothesis though it is computationally extensive. We propose a hypothesis pruning method, which iteratively removes hypotheses with low probabilities and, thus, reduces the search space and computation time greatly. We show that the pruning method achieves a performance level close to that of exhaustive search.

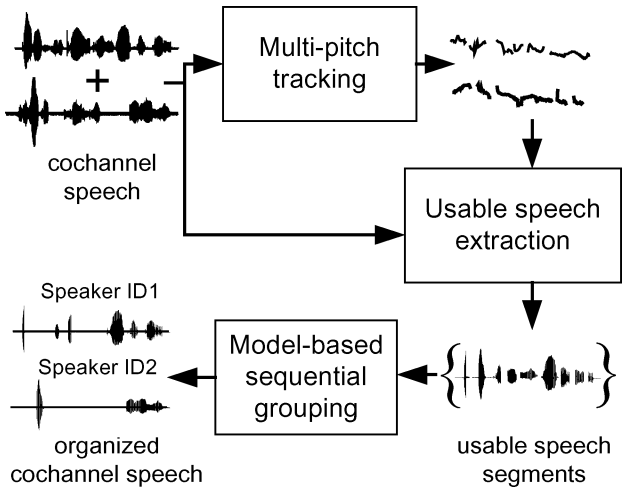


Fig. 1. Schematic diagram of the proposed system. First, cochannel speech is passed through a multipitch tracking algorithm and pitch contours are obtained. Then, usable speech segments are extracted based on the pitch information. Finally, a model-based sequential grouping algorithm organizes segments into two streams and corresponding speaker identities are also produced.

Our system is introduced in Section II. We describe how to extract usable speech using multipitch tracking in Section III. In Section IV, we develop the computational goal by extending the probabilistic framework of traditional SID to cochannel speech and detail our method to achieve the objective. Evaluation results and comparisons are given in Section V. Section VI concludes the paper.

II. System Overview

In this section, we give an overview of the processing stages of our system. As shown in Fig. 1, the proposed system consists of three stages. First, the multipitch tracking algorithm [28] is adapted and applied to cochannel speech and pitch contours for both speakers are produced. The algorithm filters the mixture signal into multiple frequency channels through an auditory filterbank; it then selects "clean" channels and peaks within each clean channel as pitch candidates at each time frame. Multiple pitch hypotheses are formed; the hypotheses are further integrated across the frequency channels. Afterwards, pitch contours are decoded as a sequence of most likely pitch hypotheses using a hidden Markov model (HMM) framework.

The second stage is used to extract usable speech from a cochannel mixture based on the pitch information [24]. Due to the nature of human voice, a speech utterance contains voiced portions, unvoiced portions and silence. Therefore, there are some portions (segments) of cochannel speech that contain only one speaker's voiced part or one speaker's voiced part plus another speaker's unvoiced part, the latter usually having much lower energy. The voiced spectra of these frames are minimally corrupted, and can be used to derive speaker features for SID. So, they form usable speech and are retained, while the portions

with overlapping pitch contours as well as silent portions are removed, resulting in a set of usable speech segments.

For any two segments in the usable speech set, whether they are from the same speaker is unknown. In the third stage, our model-based sequential grouping algorithm groups the segments into two speaker streams by searching for the optimal hypothesis in the joint speaker and grouping space. Our formulation is extended from the traditional SID probabilistic framework. Exhaustive search in the space is computationally extensive. Thus, we propose a hypothesis pruning algorithm to remove hypotheses of low likelihoods, which reduces computation time while resulting in comparable performance with exhaustive search. As a byproduct, speaker identities are also determined.

III. Usable Speech Extraction via Multipitch Tracking

We employ and adapt a recent multipitch tracking algorithm proposed by Wu *et al.* [28] for usable speech extraction. We chose this algorithm because it is designed to track two overlapping pitch contours, which fits our needs, and produces very good results.

First, an input mixture is passed through a bank of 128 gammatone filters in order to obtain a cochlear spectrogram, or cochleogram, representation. The envelopes in high-frequency channels (center frequency greater than 800 Hz) are calculated and normalized correlograms (autocorrelations) are computed for each frequency channel. The peaks of the correlogram in a frequency channel indicate the periodicity of the signal, but some peaks are inconsistent with the pitch because of pitch dynamics and the fact that harmonics are unresolved in high frequency channels. Also, in noisy conditions, the peaks in corrupted channels do not agree with the pitch. In order to minimize the effects introduced by these false peaks, corrupted channels are removed and the peaks are further selected in the retained clean channels.

A statistical model of pitch contours given the observed peaks is constructed as follows. A mixture of a Laplacian and a uniform distribution is employed to model the distribution of time-lag difference between the true pitch period and the closest peak in a selected channel. The distribution parameters are estimated from clean speech by maximum likelihood. Thus, the probability of a frequency channel supporting a pitch hypothesis is formulated. 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. A continuous HMM is used to model dynamic pitch contours. HMM states represent possible pitch states in every time frame and the transitions represent the probabilistic pitch dynamics, which models the pitch change in time and the jumps between zero-pitch, one-pitch, and two-pitch spaces. The observation probability is the observed conditional probability described above.

Fig. 2 shows an example of multipitch tracking. The cochannel speech is created by mixing two female utterances. The prior pitch points are obtained using Snack [26] (an open source version of ESPS/waves+) from premixing utterances.

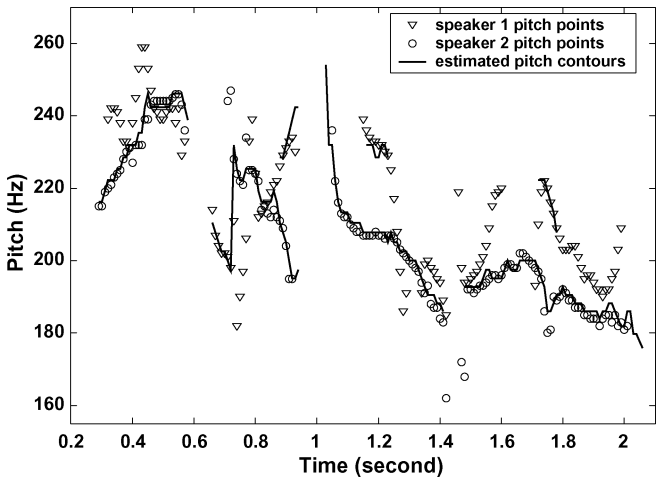


Fig. 2. Estimated pitch contours from multipitch tracking compared with single-speaker pitch points. The solid lines represent the pitch contours obtained from a female–female cochannel mixture using the multipitch tracking algorithm. The triangles and circles represent the pitch points obtained from the premixing utterances using Snack.

The algorithm produces the pitch contours that fit well the true pitch points, even though these two utterances have the same pitch range. 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. This kind of mistake, however, is rather benign as far as usable speech extraction is concerned.

Usable speech extraction means to determine what segments, i.e., sequences of frames containing only a single speaker's information, are usable for SID. Pitch contours overlap from time to time due to the nature of cochannel speech. Pitch-overlapping segments are not usable for SID because the energies of both talkers are strong, leading to the corruption of single-speaker features used in SID. In such a frame, more precisely, the harmonics and formants from both talkers are added together in the power spectrum domain and ruin the second frequency analysis process (discrete cosine transform) in the derivation of commonly used cepstral features. Speech enhancement methods such as spectral subtraction [3] are not effective here because human speech is highly nonstationary. Thus, we remove pitch-overlapping segments from cochannel speech.

For the segments with only one speaker's voiced speech, the other speaker is either silent or producing unvoiced speech. In the former case, the power spectrum is intact; in the latter case, usually the energy of unvoiced speech is much lower than voiced speech and the voiced power spectrum is contaminated much less than in the voiced–voiced situation. Thus, we consider these segments with single-pitch contours as usable speech. The remaining segments are considered unusable and removed. To ensure the homogeneity of a usable speech segment, if estimated pitch values of neighboring frames change abruptly, we consider that a speaker change occurs. Specifically, if this change is above 10 Hz, the segment is split into two shorter segments.

IV. Model-Based Sequential Organization

Maximum-likelihood classification is well established for SID [20]. However, in order to recognize talkers in cochannel speech, the traditional probability framework for a single speaker needs to be extended to multiple speakers.

A. Speaker Identification

Given a set of reference speaker models $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$, the goal of SID is to find the speaker model that maximizes the posterior probability for an observation sequence $O = \{o_1, o_2, \ldots, o_M\}$. Cepstral features, such as mel-frequency cepstral coefficients (MFCCs), are widely used as observations for speech signals. The SID decision rule is

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} P(\lambda|O). \tag{1}$$

Applying the Bayesian rule, we have

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} \frac{P(O|\lambda)P(\lambda)}{P(O)}. \tag{2}$$

Typically, prior probabilities of speakers are assumed equal, and the maximization over λ is not affected by $P(O)$. Hence, $P(\lambda)$ and $P(O)$ can be dropped. Using pretrained speaker models and assuming independence between observations at different times, (2) can be rewritten as

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} \sum_{m=1}^{M} \log p(o_m|\lambda) \tag{3}$$

after taking the log operation. Here, m indexes observations. $p(o|\lambda)$ is the standard Gaussian mixture model estimated from training speech of specific talkers using the EM algorithm [20]. In the following 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 [30]. Note that no background model is used.

B. Extension to Cochannel Speech

Cochannel SID aims to find two speaker models that maximize the posterior probability for the observations. For a cochannel mixture, our usable speech extraction method extracts N speech segments, $X = \{S_1, S_2, \ldots, S_i, \ldots, S_N\}$, each of which is a segment of consecutive speech frames, $S_i = \{x\}$, with a single-pitch contour. Given X, (1) can be modified as follows:

$$\hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}} = \arg\max_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda} P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}|X) \tag{4}$$

which is to find 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 $\{S_1^0, S_2^1, \ldots, S_i^1, \ldots, S_N^0\}$, 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, $\hat{y}$, that jointly maximize the posterior probability

$$\hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}}, \hat{y} = \arg\max_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, y \in Y} P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, y|X) \tag{5}$$

where Y is the assignment space, which includes all possible assignments (labelings) of the segments.

C. Derivation

The posterior probability in (5) can be rewritten as

$$\begin{aligned} P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, y|X) &= \frac{P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, y, X)}{P(X)} \\ &= P(X|y, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) P(y|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \frac{P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})}{P(X)}. \end{aligned} \tag{6}$$

Since the assignment is independent of specific models, $P(y|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$ becomes $P(y)$, 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 SID that prior probabilities of speaker models are the same, we insert (6) into (5) 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:

$$\hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}}, \hat{y} = \arg\max_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, y \in Y} P(X|y, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}). \tag{7}$$

Note the conditional probability is essentially the joint SID score of assigned segments. Given y, the labeling, we denote X^0 as the subset of usable speech segments labeled 0, and X^1 the subset labeled 1. Since X^0 and X^1 are complementary, the probability term in (7) can be written as follows:

$$P(X|y, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) = P(X^0, X^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}). \tag{8}$$

The y 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 (8) can be written as

$$\begin{aligned} P(X^0, X^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) &= P(X^0|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) P(X^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \\ &= \prod_{S_i \in X^0} P(S_i|\lambda_{\mathrm{I}}) \prod_{S_j \in X^1} P(S_j|\lambda_{\mathrm{II}}). \end{aligned} \tag{9}$$

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

$$P(S|\lambda) = \prod_{x \in S} p(x|\lambda). \tag{10}$$

D. Computational Method

The computational objective in (7) is to find two speakers and one assignment that yield the maximal probability using (8)–(10). Given the extracted usable speech segments and individual speaker models trained from clean speech, the maximization amounts to a search for the globally optimal hypothesis in the joint speaker and assignment space Λ and Y.

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, $P(X|y, \lambda_{\rm I}, \lambda_{\rm II})$, for every possible pair out of K speakers in Λ and every assignment in Y. Let the calculation of $P(X|y, \lambda_{\rm I}, \lambda_{\rm II})$ take a unit time, then total computation time is on the order of $O(K^2 \cdot 2^N)$. However, according to (7)–(10), 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 leads 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 complexity of search is reduced to $O(K^2 \cdot N)$.

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 in (7) 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. Appendix I gives the details of the algorithm. The 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 $P(X|y, \lambda_{\rm I}, \lambda_{\rm II})$ 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.

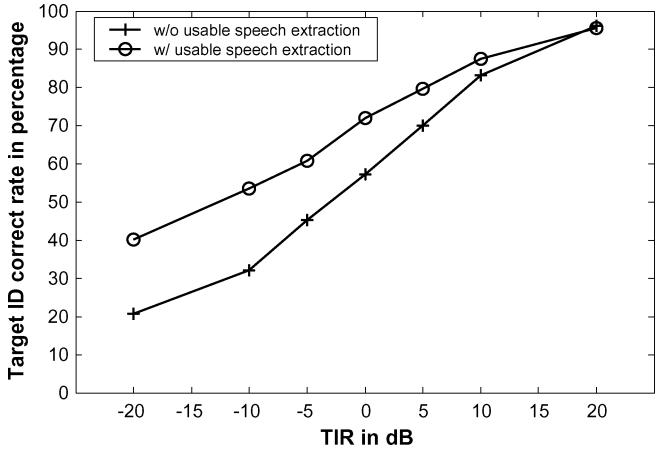


Fig. 3. Target SID correct rate before and after usable speech extraction. SID is considered correct when the target speaker is identified from cochannel speech. Sequential grouping is done using *a priori* pitch information.

V. Evaluation and Comparison

A. Data Preparation

As in Lovekin *et al.* [16], we employ the evaluation data from the TIMIT speech corpus. The speaker set consists of 38 speakers from the "DR1" dialect region, 14 of which are female and the rest are male. Each speaker has ten utterance files, ranging from about 1.5 to 6.2 s in length. For each speaker, five out of ten files are used for training and the remaining five files are used to create cochannel mixtures for testing. For each speaker deemed as the target speaker, one out of five 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 overall TIR of the speech mixture is calculated as the ratio of the target speech energy over the interfering speech energy

$$\mathrm{TIR} = 10 \log_{10} \left(\frac{\sum_n \left(s_T^2[n] \right)}{\sum_n \left(s_I^2[n] \right)} \right) \tag{11}$$

in which s_T and s_I are the speech samples of target and interfering speakers in the time domain. The interfering utterance is either cropped or concatenated with itself to match the length of the target utterance. Speech is scaled to create the mixtures at different TIRs: -20, -10, -5, 0, 5, 10, and 20 dB. 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.

B. Usable Speech Evaluation

Our first experiment evaluates how the new method works for usable speech extraction. To show the effectiveness of our method, SID on usable speech is performed after the extracted segments are ideally assigned. In many situations, one is interested in the target speaker, and the speech signal from the other speaker is considered noise. Hence, we choose the target speaker SID as our evaluation criterion (see [25] for SID results on recognizing either speaker). Fig. 3 gives the target speaker recognition rate under various TIRs. As a baseline, a conventional SID system is applied to the cochannel speech to recognize the

TABLE I
CORRECT ASSIGNMENT RATE FOR SEQUENTIAL ORGANIZATION AND COCHANNEL SPEAKER IDENTIFICATION CORRECT RATE

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.

target speaker. The baseline performance documents the top two identified speakers. The correct rate degrades sharply when TIR decreases because the target speech is increasingly corrupted. Yantorno *et al.* [29] obtained comparable results in a similar study to understand how cochannel speech impacts SID performance. As a comparison, usable speech segments are extracted from cochannel mixture as described previously. Here, we assume that pitch information of individual speakers is known *a priori* and segments are ideally grouped into speaker streams based on *a priori* pitch. Specifically, a segment takes the label that is taken by the majority of the frames in it, which is determined by comparing the detected pitch with the *a priori* pitch. The first observation from Fig. 3 is that, under cochannel situations, usable speech extraction improves SID performances; the average improvement is about 12% in terms of absolute correct rate. Second, the improvements are consistent across all TIR levels. Improvement decreases at higher TIRs because the designated target speaker dominates the mixture. However, the target speaker is dominated by the interferer at lower TIRs, resulting in better performance after usable speech extraction.

C. Sequential Grouping Evaluation

Here, we evaluate 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 I, including the results from the alternative methods we will describe in the following sections.

The second column in Table I 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 third and fourth columns in Table I 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 correct rate. The second criterion records the percentage of mixtures where both speakers are correctly identified; this is the more stringent criterion (see [25] for another criterion based on recognizing either speaker).

In Table I, the baseline rate of correct grouping corresponding to random labeling of each usable frame is 50.0%. The second 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.

In terms of SID accuracy, the baseline performance is taken to be identification accuracy by recognizing individual speakers directly (see the next subsection for a method of recognizing speaker pairs using combined GMM). 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 to 0.037 s/file on a Pentium III workstation (the computation time for the exponential version of exhaustive search is on average 7 min/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 I). 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 in Section V-E should be explored instead.

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 [11]. 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.

D. Alternative Methods

We have explored a number of variations of our hypothesis pruning algorithm. Because the algorithm prunes certain paths, it resembles beam search [22]. In the simplest case where the beam width is 1, the algorithm keeps only one hypothesis at each iteration. In this case, beam search obtains the correct assignment rate of 66.0% and the SID results are presented in Table I. The results are significantly worse than the proposed algorithm. In the case where the beam width is 2, it is very similar to the proposed algorithm except that the latter already keeps two possible labels at each iteration. The results are given in Table I, and they are indeed very close to those obtained by the hypothesis pruning algorithm.

If the main objective is cochannel SID, 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 y in (5) 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: $p(O|\lambda_{\rm I},\lambda_{\rm II}) = 0.5(p(O|\lambda_{\rm I}) + p(O|\lambda_{\rm II}))$. 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 its SID performance is given in Table I. It achieves SID performance 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.

E. Comparison

We have shown the system's ability to extract usable speech and improve both cochannel SID and sequential organization performance. In this section, we compare with alternative sequential grouping methods, namely one that employs pitch dynamics and one based on spectral divergence.

One reasonable alternative is to utilize pitch information, particularly since pitch contours have already been obtained. Previous studies have demonstrated the importance of pitch contours for speaker recognition, e.g., [1]. We collect pitch differences between the end-point of a segment and the start point of the following segment from the training data. Considering that the longer is the gap between two segments the less likely they belong to the same speaker, we multiply the difference by the time lag between them. The resulting product describes the pitch change dynamics between neighboring segments. A Gaussian-like peak can be observed centered on 0 in the histogram and maximum-likelihood estimation is employed to obtain the statistics of the distribution, which is modeled as a mixture of Gaussian and uniform distribution [28]. When grouping, for each segment from S_1 to S_N, the pitch dynamics product is obtained and a local decision is made regarding whether the current segment comes from the same speaker as the previous segment by comparing the likelihoods of the dynamics feature given the distribution. After the assignment is done, a search for the two most probable speakers is applied. So, it is obvious that the computational complexity is $O(K)$ for this method. From the results given in Table I, this method clearly performs worse than the pruning algorithm, but gives a significant improvement over the baseline case without usable speech processing.

We have also compared our algorithm with a spectrum-based method, specifically the speaker assignment algorithm of Morgan *et al.* [18] that also addresses sequential organization. Their system aims to enhance cochannel speech by separating two talkers and subsequently assigning separated speech components to two speaker streams. The assignment of intermittent

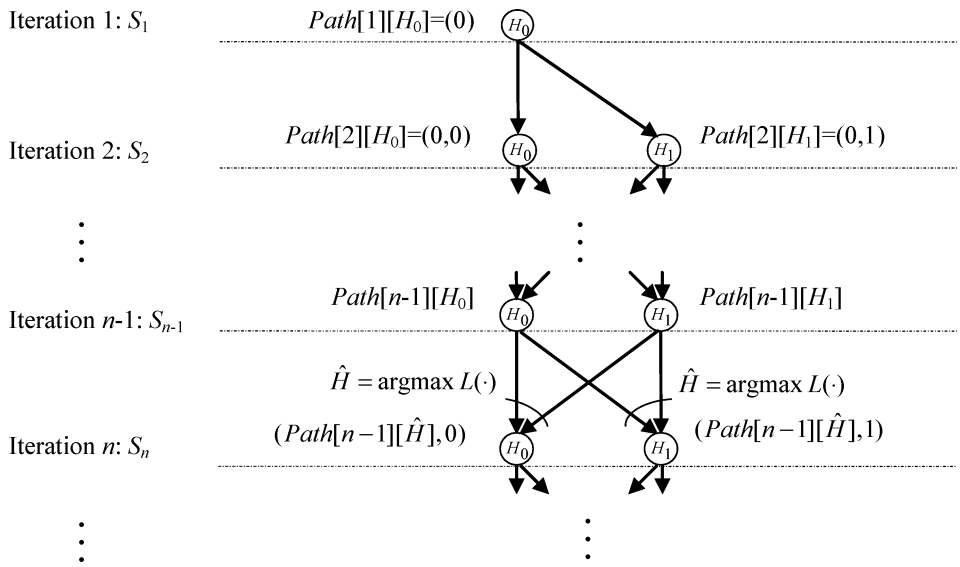


Fig. 4. Illustration of the hypothesis pruning algorithm. The algorithm is executed segment by segment. Every segment is hypothesized to be either H_0 or H_1 and labeled with 0 or 1, respectively, except that S_1 is identified with hypothesis H_0. $Path$ records the best label path. For either hypothesis of the segment to be considered, the better label path from the preceding iteration is chosen by comparing $L(.)$ defined in (12), and its label path is copied to the current path. The algorithm repeats until the last segment is processed.

voiced components, essentially the beginning frames of segments, is based on a frame-level spectral comparison with 50 recently assigned frames using the spectral divergence measure of Carlson and Clement [7]. Since our system considers a usable segment to belong to one speaker, we adapt and employ the algorithm to perform only speaker assignment; that is, segments are organized using their spectrum-based method. Specifically, the initial 50 frames of each speaker stream are *a priori* assigned, and then the subsequent segments are sequentially assigned according to their divergence measure. The assignment result is shown in Table I. With 66.2% correct rate, the spectral method is 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.

VI. Concluding Remarks

Sequential organization groups sound components of the same source across time into the same stream. In this paper, we have proposed a model-based approach for sequential organization, to assign the extracted usable speech segments into speaker streams. Our usable speech extraction method produces segments useful for cochannel SID across various TIR conditions. We have shown that the proposed hypothesis pruning algorithm achieves SID performance close to the ceiling performance with prior pitch information or exhaustive search, and it performs significantly better than alternative approaches to speaker assignment.

It is worth noting that our sequential grouping algorithm can handle the situation where only one speaker is present in a cochannel mixture. Since segments may all take the same label after assignment, our algorithm can produce only one speaker identity. Also, when evaluating the likelihood of assigned segments in (7), the same speaker model could be two top choices for both subsets, which signals that only one speaker is present. The probabilistic framework proposed in here can be extended to situations with more than two speakers in a mixture.

By extracting voiced speech as usable speech, the speaker information carried in unvoiced speech segments is removed. How much does usable speech extraction impact the performance of single-speaker recognition? For the trained GMMs and the test corpus described in Section V, the SID correct rate is about 99.5% without usable speech extraction. After the extraction of single-pitch segments, the SID correct rate by performing SID on extracted segments only degrades to 92.4%. Lovekin *et al.* [16] reported similar degradation when voiced speech is tested on normal-trained speaker models. We note that speaker models could be trained with extracted usable speech directly instead of entire speech files as suggested in [16], which could not only improve SID performance but also reduce the amount of training data. The study in [16] observed some SID improvement by doing so. This will be investigated in future work.

The speech decoding model of Barker *et al.* [2] also addresses sequential integration, and their formulation is extended from the statistical framework of automatic speech recognition. Their model searches for the most likely word sequence and additionally determines the set of signal fragments that compose the speech signal, leaving the rest as the noise background. Our model is analogous to theirs in the emphasis of recognition-based organization. However, the domain of cochannel speaker recognition where our model is derived differs from their speech recognition domain, and as a result the computational methods used in the two models are very different. It is not clear, for example, how their model can address the cochannel situation where the interfering noise is also speech.

Appendix I
Hypothesis Pruning Algorithm

We give the detailed algorithm below. See Section IV-D for notations.

Step 0) Order the segments in $X = \{S1, S2, \ldots, SN\}$ sequentially in time.

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

Step 2) For S_2 in X, 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 X^0, and H_1 assumes that the current segment belongs to X^1. The label paths are

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

$Path[n][.]$ 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:

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

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

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

$$L\left(Path[n-1][H], l\right) = \max_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda} P\left(X \middle| \left(Path[n-1][H], l\right), \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}\right) \quad (12)$$

$l = 0$ 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 (12) and the segment assignment for this hypothesis.

The L function in (12) is the same as (7) in the main text except that L only considers the partial segment assignment from S_1 to S_n. Fig. 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 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 (12). The better path is then chosen. The algorithm repeats until the last segment is processed.

ACKNOWLEDGMENT

The authors would like to thank M. Wu for his assistance in using the multipitch tracking algorithm, J. Barker for a suggestion regarding exhaustive search complexity, and three anonymous referees for extensive and helpful comments.

REFERENCES

[1] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1687–1697, 1972.

[2] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, no. 1, pp. 5–25, Jan. 2005.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, 1979, pp. 113–120.

[4] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[5] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.

[6] D. S. Brungart, "Information and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Amer.*, vol. 109, pp. 1101–1109, 2001.

[7] B. A. Carlson and M. A. Clements, "A computationally compact divergence measure for speech processing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 1, pp. 1–6, Jan. 1991.

[8] E. C. Cherry, "Some experiments on the recognition of speech with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–979, 1953.

[9] R. B. Dunn, D. A. Reynolds, and T. F. Quatieri, "Approaches to speaker detection and tracking in conversational speech," *Digit. Signal Process.*, vol. 10, pp. 93–112, 2000.

[10] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*. New York: Marcel-Dekker, 2001.

[11] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.

[12] H. Helmholtz, *On the Sensation of Tone*. New York: Dover, 1863.

[13] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, pp. 1135–1150, 2004.

[14] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under cochannel conditions," presented at the *Int. Symp. Intelligent Signal Processing and Communication Systems*, 2000.

[15] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.

[16] J. M. Lovekin, R. E. Yantorno, K. R. Krishnamachari, D. S. Benincasa, and S. J. Wenndt, "Developing usable speech criteria for speaker identification," in *Proc. ICASSP*, 2001, pp. 421–424.

[17] T. Matsui and S. Furui, "Text-independent speaker recognition using vocal tract and pitch information," in *Proc. ICASSP*, 1990, pp. 137–140.

[18] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 407–424, May 1997.

[19] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, no. 1, pp. 56–69, Jan. 1990.

[20] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, 1995.

[21] D. A. Reynolds *et al.*, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP*, 2003, pp. 784–787.

[22] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2003.

[23] A. Schmidt-Nielsen and T. H. Crystal, "Human v.s. machine speaker identification with telephone speech," presented at the ICSLP, 1998.

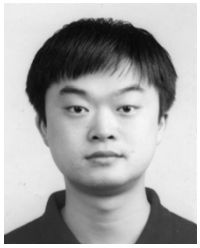
[24] Y. Shao and D. L. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. ICASSP*, vol. 2, 2003, pp. 205–208.

[25] ——, "Model-based sequential organization for cochannel speaker identification," in *Proc. ICSLP*, 2004, pp. 2593–2596.

[26] K. Sjolander and J. Beskow, "Wavesurfer—An open source speech tool," in *Proc. ICSLP*, vol. IV, 2000, pp. 464–467.

[27] B. Y. Smolenski, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Co-channel speaker segment separation," in *Proc. ICASSP*, 2002, pp. 125–129.

[28] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 299–241, May 2003.
[29] R. E. Yantorno, D. Benincasa, and S. Wenndt, "Effects of co-channel interference on speaker identification," in *Proc. SPIE Int. Symp. Technologies for Law Enforcement*, vol. 4232, 2001, pp. 258–261.
[30] S. Young, D. Kershaw, J. Odell, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0)*. Redmond, WA: Microsoft Corp., 2000.
[31] G. Yu and H. Gish, "Identification of speakers engaged in dialog," in *Proc. ICASSP*, 1993, pp. 383–386.
[32] M. A. Zissman and D. C. Seward, "Two-talker pitch tracking for co-channel talker interference suppression," Tech. Rep., Lincoln Lab., Mass. Inst. Technol., Cambridge, 1992.



Yang Shao (S'03) received the B.S. degree in computer science from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, and the M.S. degree in computer science from Fudan University, Shanghai, China. He is currently pursuing the Ph.D. degree in computer science and engineering at The Ohio State University, Columbus.

His research interests include computational auditory scene analysis, speech processing, and automatic speech and speaker recognition.

DeLiang Wang (M'90–SM'01–F'04) received the B.S. and M.S. degrees in computer science in 1983 and 1986, respectively, from Peking (Beijing) University, Beijing, China, and the Ph.D. degree in computer science from the University of Southern California, Los Angeles, in 1991.

From July 1986 to December 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer Science and Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, where he is currently a Professor. From October 1998 to September 1999, he was a Visiting Scholar in the Department of Psychology, Harvard University, Cambridge, MA. His research interests include machine perception and neurodynamics.

Dr. Wang currently chairs the IEEE Computational Intelligence Society Neural Networks Technical Committee and is a member of the Governing Board of the International Neural Network Society and the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee. He is a recipient of the 1996 U.S. Office of Naval Research Young Investigator Award.