CLOSE—A Data-Driven Approach to Speech Separation

Ji Ming, *Member, IEEE*, Ramji Srinivasan, *Member, IEEE*, Danny Crookes, *Senior Member, IEEE*, and Ayeh Jafari

***Abstract*—This paper studies single-channel speech separation, assuming unknown, arbitrary temporal dynamics for the speech signals to be separated. A data-driven approach is described, which matches each mixed speech segment against a composite training segment to separate the underlying clean speech segments. To advance the separation accuracy, the new approach seeks and separates the *longest* mixed speech segments with matching composite training segments. Lengthening the mixed speech segments to match reduces the uncertainty of the constituent training segments, and hence the error of separation. For convenience, we call the new approach *C*omposition of *Lo*ngest *Se*gments, or CLOSE. The CLOSE method includes a data-driven approach to model long-range temporal dynamics of speech signals, and a statistical approach to identify the longest mixed speech segments with matching composite training segments. Experiments are conducted on the Wall Street Journal database, for separating mixtures of two simultaneous large-vocabulary speech utterances spoken by two different speakers. The results are evaluated using various objective and subjective measures, including the challenge of large-vocabulary continuous speech recognition. It is shown that the new separation approach leads to significant improvement in all these measures.**

***Index Terms*—Co-channel speech, longest matching segment, speaker identification, speech recognition, speech separation, temporal dynamics.**

I. Introduction

We consider the problem of speech separation as falling into two categories: constrained and unconstrained. By constrained speech separation we mean that there is *a priori* knowledge about the vocabulary and grammar (or language model) of the speech utterances to be separated. For constrained speech separation, researchers have recently demonstrated a case of reaching near human performance, in the PASCAL Speech Separation Challenge (see, for example, [1], [8], [33]). The challenge was about the separation of two simultaneous speech utterances given single-channel mixed speech data; each utterance was formed from a small vocabulary obeying a command-sentence grammar, with both the vocabulary and grammar being known. This knowledge of vocabulary and grammar has been used to impose up to utterance-long constraints on the underlying speech signals, to restrict their allowable temporal-spectral structures and hence reduce their uncertainties. This has helped to correctly separate the underlying clean speech utterances. In this paper, we remove the requirement for prior information about the vocabulary, grammar or language model of the underlying speech utterances. Specifically, we deal with separation of two simultaneous utterances from two different speakers based on single-channel data, assuming unknown, arbitrary acoustic, lexical and language dynamics for both utterances. We call this problem unconstrained speech separation. We describe a system aiming to achieve the performance of constrained speech separation but for unconstrained speech.

In the past, model-based approaches have been heavily used to impose temporal constraints on speech signals for speech separation. The work in [7], for example, considered the phone-level dynamics by modeling phones using hidden Markov models (HMMs). The work in [28], for example, concatenated phone HMMs following a pronunciation dictionary, thereby extending the dynamics modeling to the lexicon level. Some PASCAL challenge methods considered word-level dynamics by using whole-word HMMs (e.g., [5], [12], [33]). Finally, many of the challenge methods went further to model utterance-level dynamics based on the known grammar [1]. For HMM-based modeling methods, the factorial HMM approach is typically used to model co-channel speech signals and perform separation (e.g., [2]–[8]). Two-dimensional Viterbi algorithms and approximations (e.g., iterative Viterbi or loopy belief) have been used to perform the inference [9]. As demonstrated in the PASCAL challenge task, imposing long-range temporal constraints helps separate speech from co-channel mixtures. However, modeling subword, word and sentence level dynamics requires transcribed training data and knowledge of the task. Without these, how to model long-range temporal dynamics of speech for speech separation remains an open research question. For separation of unconstrained speech (i.e., the speech with unconstrained acoustic, lexical and language dynamics), most current model-based systems use a Gaussian mixture model (GMM) or vector quantization (VQ) approach, which assumes independence between successive speech frames (see, for example, [10]–[16], [29]).

Other popular approaches suitable for unconstrained speech separation include computational auditory scene analysis (CASA) and basis-function based decomposition. CASA-based algorithms, for example [17]–[24], work in the time-frequency plane by segmenting the psychoacoustic cues such as pitch, onset/offset, temporal continuity, harmonic structures

Manuscript received October 10, 2011; revised January 09, 2012; accepted February 12, 2013. Date of publication March 07, 2013; date of current version March 22, 2013. This work was supported by the U.K. EPSRC under Grant EP/G001960/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chang D. Yoo.

The authors are with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K. (e-mail: j.ming@qub.ac.uk; r.srinivasan@qub.ac.uk; d.crookes@qub.ac.uk; ajafari01@qub.ac.uk).

Digital Object Identifier 10.1109/TASL.2013.2250959

1558-7916/$31.00 © 2013 IEEE

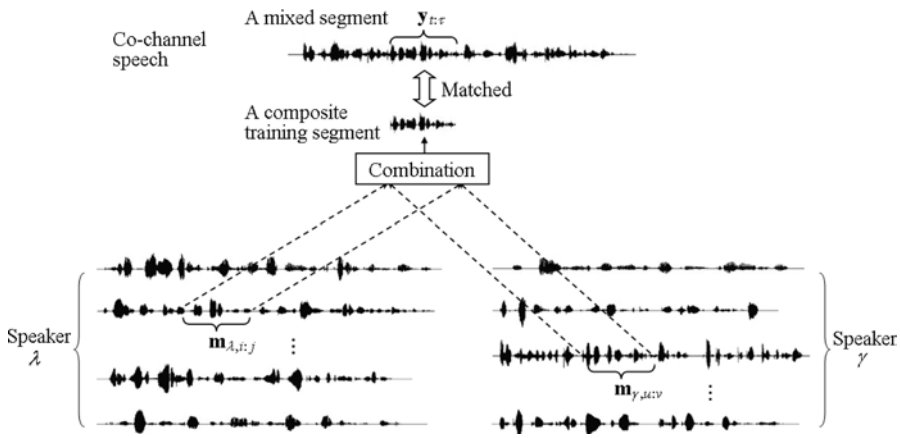


Fig. 1. Illustration of the proposed approach. Shown are a co-channel utterance containing mixture of two speech utterances, and the clean training utterances of the two constituent speakers. Also shown is the separation of a mixed speech segment by finding a composite training segment that matches the mixed segment. The composite training segment is formed by combining two clean training segments from the two speakers. We aim to identify the longest mixed speech segments with matching composite training segments for the separation. This will reduce the uncertainty, and hence the error, of the constituent training segments.

and modulation correlation into different sources, and performing separation by masking the interfering sources. Recent CASA-based algorithms also incorporate statistical models such as the HMM, GMM, and VQ into the segmentation process [12], [20], [21], [24]. In basis-function based decomposition, for example [25]–[29], a set of bases (or dictionary) is used to represent the short-time speech spectra of each constituent speaker; separation is performed by finding linear combinations of the constituent basis sets that match the given speech mixtures. Different methods have been used to derive the spectral basis functions, including nonnegative matrix factorization (NMF), VQ, GMM, and independent component analysis (ICA). The CASA and basis-function based approaches can be used for speech separation without requiring *a priori* knowledge such as vocabulary; the separation is usually performed on a frame-by-frame basis, or by capturing short-term dynamics (e.g., pitch contiguity) of speech [9].

In this paper, we study a new approach to unconstrained speech separation. We aim to improve the separation accuracy by imposing long-range temporal constraints on unconstrained speech signals. We achieve this by separating segments of consecutive frames as whole units, in a data-driven framework. Our approach is illustrated in Fig. 1. It shows a test utterance which is a co-channel mixture of two speech utterances with arbitrary temporal dynamics, and the clean training utterances of the two constituent speakers.[1]Fig. 1 also shows an example of separating a test segment of the co-channel speech by finding a composite training segment that best matches the test segment. The composite training segment is formed by combining two clean training segments from the two speakers. Knowing the make-up of the matching composite training segment we can separate the test segment into two clean speech segments, using the two constituent training segments. To enhance the separation accuracy, we aim to identify the *longest* test segments which can be accurately matched by composite training segments. The longer the test segments to match, the more specific the constituent training segments. Therefore separation based on the longest matching segments reduces the error of separation. The new approach represents a data-driven way to imposing long-range temporal constraints on the underlying speech signals, without requiring knowledge about the speech signals (e.g., vocabulary, language model, etc.), and transcripts of the training data. This work is an extension of our previous work [35] for noisy speech enhancement. For convenience, we call our new approach CLOSE (*C*omposition of *Lo*ngest *Se*gments).

The paper is organized as follows. In Section II, we describe the new CLOSE method. Two algorithms are described: an "exact" algorithm and an approximation; the latter bears a substantially reduced computational load and hence is the main algorithm used in our experiments. It is shown that the conventional GMM-based separation algorithm is a special case of the new algorithm. Section III presents more details of implementing the CLOSE method, including the identification of the constituent speakers given a speech mixture, and the reconstruction of the clean speech utterances based on the longest matching constituent training segments found. Experimental studies for separating unconstrained, large-vocabulary co-channel speech are presented in Section IV. Finally, conclusions are drawn in Section V.

II. The CLOSE Approach to Speech Separation

The new approach consists of two main parts. The first part is a data-driven approach for modeling the training speech utterances of the constituent speakers; the model facilitates the comparison of long-range temporal dynamics between speech utterances with unconstrained temporal dynamics. The second part is a method for identifying the longest segments of co-channel speech with matching composite training segments, for separating the underlying clean speech. The following provides the details.

A. Modeling Training Utterances

For each test utterance, we use clean training utterances as examples of the underlying clean speech. We aim to identify long matching segments (i.e., long matching temporal dynamics) for the separation. For underlying speech with unknown, arbitrary temporal dynamics, we use a data-driven approach to perform

[1]Identifying the two constituent speakers given the speech mixture is part of the separation problem and will be discussed in the paper.

the identification. First, we model the *complete* temporal dynamics in each training utterance. As such, any segment of any length in a training utterance, up to the complete training utterance, can be used as a whole unit to identify a corresponding underlying speech segment, for separating the segment. This modeling approach is similar to that described in [34], [35].

Let $\mathbf{x}_{\lambda,1:T_\lambda} = \{x_{\lambda,t} : t = 1, 2, \ldots, T_\lambda\}$ represent a training utterance for speaker λ, where T_λ is the number of frames in this utterance (which can be variable from utterance to utterance) and $x_{\lambda,t}$ is the feature vector of the frame at time t. We take two steps to build a model for each training utterance $\mathbf{x}_{\lambda,1:T_\lambda}$. First, we train a GMM for the feature vectors of each speaker by using all the training utterances from the speaker. Denote by G_λ the GMM for speaker λ, of M_λ Gaussian components, trained using all the training utterances $\mathbf{x}_{\lambda,1:T_\lambda}$. This can be expressed as

$$G_\lambda = \{g_\lambda(x|m), w_\lambda(m) : m = 1, 2, \ldots, M_\lambda\} \tag{1}$$

where $g_\lambda(x|m)$ is the m'th Gaussian component and $w_\lambda(m)$ is the corresponding weight, for speaker λ. Second, based on G_λ, we build a model for each training utterance $\mathbf{x}_{\lambda,1:T_\lambda}$ by taking each frame from $\mathbf{x}_{\lambda,1:T_\lambda}$ and finding the Gaussian component in G_λ that produces the maximum likelihood for the frame. As such, $\mathbf{x}_{\lambda,1:T_\lambda}$ can be alternatively represented by a corresponding time sequence of Gaussian components $\{g_\lambda(x|m_{\lambda,t}) : t = 1, 2, \ldots, T_\lambda\}$, where $g_\lambda(x|m_{\lambda,t})$ is a Gaussian component taken from G_λ with index $m_{\lambda,t}$, which produces the maximum likelihood for the training frame $x_{\lambda,t}$. This time sequence of Gaussian components can be fully characterized by the corresponding time sequence of Gaussian indexes, which we write as $\mathbf{m}_{\lambda,1:T_\lambda}$, where

$$\mathbf{m}_{\lambda,1:T_\lambda} = \{m_{\lambda,t} : t = 1, 2, \ldots, T_\lambda\}. \tag{2}$$

We call (2) an utterance model, for the training utterance $\mathbf{x}_{\lambda,1:T_\lambda}$. In the training stage, we create a model $\mathbf{m}_{\lambda,1:T_\lambda}$ for each training utterance $\mathbf{x}_{\lambda,1:T_\lambda}$ of each speaker λ. All the training utterance models for a speaker together form a speech model of the speaker, to be used in the CLOSE system for separation.

As can be noticed, the above utterance model $\mathbf{m}_{\lambda,1:T_\lambda}$ shares characteristics with a template, in the sense that both capture the full temporal dynamics, from acoustic to lexical and to language, that join together the appropriate short-time frames to form a specific training utterance. However, the model (2) provides a smoother, and hence more robust, representation than templates by representing each speech frame $x_{\lambda,t}$, which is subject to random variation, using a Gaussian component $g_\lambda(x|m_{\lambda,t})$. Other advantages of the model over templates are the reduced memory space for storing the training data, and the reduced computation for frame matching. By mapping all the training utterances to a GMM, the complexity of finding a match of a test frame among all the training frames is scaled down to the calculation of the M_λ Gaussians of the test frame. The model was first introduced in [34] for speech segmentation and recognition, and has lately been further explored for speech enhancement [35], speaker recognition [36] and speech recognition [38].

B. Separation Based on Composition of Longest Segments—CLOSE

Let $\mathbf{y}_{1:T} = \{y_t : t = 1, 2, \ldots, T\}$ be a test utterance (a co-channel speech mixture) with T frames, containing two speech utterances spoken by speaker λ and speaker γ (later we will discuss an algorithm for identifying these two speakers given the test utterance). In our system, the problem of speech separation can be stated as: for each test frame y_t, identifying one training frame $m_{\lambda,i}$ from speaker λ and another training frame $m_{\gamma,u}$ from speaker γ, such that their combination matches y_t. This separates the two clean speech frames forming the test frame; the two clean frames can be reconstructed by using the corresponding clean training frames, modeled by the Gaussian components $g_\lambda(x|m_{\lambda,i})$ and $g_\gamma(x|m_{\gamma,u})$.

Because of the short duration of a frame, given a test frame of co-channel speech, there could be many different choices of the two constituent training frames in terms of producing a similar composite frame matching the test frame. Therefore uncertainty remains over the correct constituent training frames. We solve this problem by matching test segments and composite training segments, both consisting of consecutive frames (see Fig. 1). The longer the test segment to match, the more specific the constituent training segments, because of the increasingly distinct temporal dynamics. Therefore separation based on matching long test segments reduces the uncertainty of the correct constituent training frames for each test frame, and hence the error of separation. The following describes the CLOSE algorithm, which aims to match the *longest* test segments for separation given the training and test data. The training data are modeled by using the training utterance model (1) and (2) described in the last section.

Let $\mathbf{y}_{t:\tau} = \{y_\epsilon : \epsilon = t, t+1, \ldots, \tau\}$ represent a test segment of the co-channel speech taken from the test utterance $\mathbf{y}_{1:T}$ and consisting of consecutive frames from time t to τ. In a similar notation, let $\mathbf{m}_{\lambda,i:j} = \{m_{\lambda,t} : t = i, i+1, \ldots, j\}$ represent a training segment taken from the model $\mathbf{m}_{\lambda,1:T_\lambda}$ and modeling consecutive frames from i to j in the training utterance $\mathbf{x}_{\lambda,1:T_\lambda}$ of speaker λ. Given $\mathbf{y}_{t:\tau}$, we identify the two matching constituent training segments $\mathbf{m}_{\lambda,i:j}$ and $\mathbf{m}_{\gamma,u:v}$ (see Fig. 1) by using the posterior probability $P(\mathbf{m}_{\lambda,i:j}, \mathbf{m}_{\gamma,u:v}|\mathbf{y}_{t:\tau})$. Assume an equal prior probability P for all possible constituent segments $(\mathbf{s}_\lambda, \mathbf{s}_\gamma)$ from the two speakers. This posterior probability can be expressed as shown in equation (3) at the bottom of the next page, where $p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j}, \mathbf{m}_{\gamma,u:v})$ is the likelihood that the given test segment $\mathbf{y}_{t:\tau}$ is matched by the two constituent training segments $\mathbf{m}_{\lambda,i:j}$ and $\mathbf{m}_{\gamma,u:v}$. Assuming that the frames within a segment are conditionally independent (conditioned on the segment), this segmental likelihood function can be written as

$$p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j}, \mathbf{m}_{\gamma,u:v}) = \prod_{\epsilon=t}^{\tau} p(y_\epsilon|m_{\lambda,\zeta(\epsilon)}, m_{\gamma,\eta(\epsilon)}) \tag{4}$$

where $p(y_\epsilon|m_{\lambda,\zeta(\epsilon)}, m_{\gamma,\eta(\epsilon)})$ is the likelihood that the given test frame y_ϵ is matched by the two constituent training frames $m_{\lambda,\zeta(\epsilon)}$ and $m_{\gamma,\eta(\epsilon)}$. In our experiments reported in this paper, we use a log-max model to calculate $p(y_\epsilon|m_{\lambda,\zeta(\epsilon)}, m_{\gamma,\eta(\epsilon)})$, discussed in Section III-A. In (4), $\zeta(\epsilon)$ and $\eta(\epsilon)$ represent the time

warping functions between the test segment $\mathbf{y}_{t:\tau}$ and the two constituent training segments $\mathbf{m}_{\lambda,i:j}$ and $\mathbf{m}_{\gamma,u:v}$, in forming the match. We assume a fixed-endpoint condition: $\zeta(t)=i$, $\zeta(\tau)=j$, and $\eta(t)=u$, $\eta(\tau)=v$. Furthermore, to speed up the algorithm, in our experiments we only compare equal-length segments with linear time warping. In other words, we only search temporally identical segments for the matching/separation.

In (3), the denominator is expressed as the sum of two terms. The first term is the average likelihood that the given test segment $\mathbf{y}_{t:\tau}$ is matched by a composite segment with both of its constituent segments found in the training data; this likelihood is calculated over all possible training segments of the two speakers. The second term, denoted by $p(\mathbf{y}_{t:\tau}|\phi_{t:\tau})$, represents the average likelihood that the test segment $\mathbf{y}_{t:\tau}$ is matched by a composite segment with either or both of its constituent segments not found in the training data. This likelihood, associated with unseen constituent segments, can be expressed by using a mixture model, allowing for temporally independent combinations of the training frames to simulate arbitrary unseen speech segments (similar to the use of a temporally independent GMM to model text-independent speech). Combining the two speakers' GMMs [i.e., (1)], we use the expression

$$p(\mathbf{y}_{t:\tau}|\phi_{t:\tau}) \simeq \prod_{\epsilon=t}^{\tau}\left[\sum_{m_\lambda=1}^{M_\lambda}\sum_{m_\gamma=1}^{M_\gamma} w_\lambda(m_\lambda)w_\gamma(m_\gamma)p(y_\epsilon|m_\lambda,m_\gamma)\right]. \tag{5}$$

The sums inside the brackets provide a mixture-based likelihood for the test frame y_ϵ, assuming that it will match one of the composite frames taking into consideration all possible combinations of frames between the two speakers. Equation (5) further assumes statistical independence between consecutive frames, so that it can simulate test segments with arbitrary temporal dynamics. In other words, if we view the segmental temporal dynamics as "text" dependence, then (4) gives a "text-dependent" likelihood of the test segment, dependent on the temporal dynamics of both constituent training segments, while (5) gives a "text-independent" likelihood of the test segment. Test segments with mismatched constituent training segments will result in low "text-dependent" likelihoods [i.e.,(4)] but not necessarily low "text-independent" likelihoods [i.e.,(5)], and hence low posterior probabilities of match [i.e.,(3)]. For test segment $\mathbf{y}_{t:\tau}$ with matched constituent training segments $\mathbf{m}_{\lambda,i:j}$ and $\mathbf{m}_{\gamma,u:v}$, we can assume that the "text-dependent" likelihood is greater than the "text-independent" likelihood, i.e., $p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v}) \geq p(\mathbf{y}_{t:\tau}|\phi_{t:\tau})$. This is because

$$\begin{aligned} p(\mathbf{y}_{t:\tau}|\phi_{t:\tau}) &\simeq \prod_{\epsilon=t}^{\tau}\max_{m_\lambda}\max_{m_\gamma} w_\lambda(m_\lambda)w_\gamma(m_\gamma)p(y_\epsilon|m_\lambda,m_\gamma) \\ &\simeq \prod_{\epsilon=t}^{\tau} w_\lambda(m_{\lambda,\zeta(\epsilon)})w_\gamma(m_{\gamma,\eta(\epsilon)})p(y_\epsilon|m_{\lambda,\zeta(\epsilon)},m_{\gamma,\eta(\epsilon)}) \\ &\leq \prod_{\epsilon=t}^{\tau} p(y_\epsilon|m_{\lambda,\zeta(\epsilon)},m_{\gamma,\eta(\epsilon)}). \end{aligned} \tag{6}$$

The second approximation is based on the assumption that matching and hence highly likely constituent training frames dominate the mixture-based likelihood. Therefore, with (3) and (5), we can obtain a larger posterior probability for matching constituent training segments, and a smaller posterior probability for mismatching constituent training segments, for the given test segment.

The posterior probability formulation (3) has another important characteristic: it favors the continuity of match and produces larger probabilities for the constituent training segments matching longer test segments. Assume that the test segment $\mathbf{y}_{t:\tau}$ and the two constituent training segments $\mathbf{m}_{\lambda,i:j}$ and $\mathbf{m}_{\gamma,u:v}$ are matching, in the sense that the segmental likelihoods $p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v}) \geq p(\mathbf{y}_{t:\tau}|\mathbf{m}'_{\lambda,i':j'},\mathbf{m}'_{\gamma,u':v'})$ for any $(\mathbf{m}'_{\lambda,i':j'},\mathbf{m}'_{\gamma,u':v'}) \neq (\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v})$, and $p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v}) \geq p(\mathbf{y}_{t:\tau}|\phi_{t:\tau})$. Then we can have the following inequality concerning the posterior probabilities of the matching constituent and test segments with different lengths

$$P(\mathbf{m}_{\lambda,i:\zeta(\epsilon)},\mathbf{m}_{\gamma,u:\eta(\epsilon)}|\mathbf{y}_{t:\epsilon}) \leq P(\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v}|\mathbf{y}_{t:\tau}) \tag{7}$$

where $\mathbf{y}_{t:\epsilon}$, with $\epsilon \leq \tau$, is a test segment starting at the same time as $\mathbf{y}_{t:\tau}$ but not lasting as long and $\mathbf{m}_{\lambda,i:\zeta(\epsilon)}$ and $\mathbf{m}_{\gamma,u:\eta(\epsilon)}$ are the corresponding constituent training subsegments matching the shorter test segment $\mathbf{y}_{t:\epsilon}$. The inequality indicates that larger posterior probabilities are obtained when longer test segments are matched. A similar inequality, concerning the posterior probability of match between a test segment and a single training segment for speech enhancement from noise, is proven in [35]. In this paper, we have extended the proof to

$$\begin{aligned} P(\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v}|\mathbf{y}_{t:\tau}) &= \frac{p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v})P}{p(\mathbf{y}_{t:\tau})} \\ &= \frac{p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v})P}{\sum_{\mathbf{s}_\lambda,\mathbf{s}_\gamma\in\text{Training}} p(\mathbf{y}_{t:\tau}|\mathbf{s}_\lambda,\mathbf{s}_\gamma)P + \sum_{\mathbf{s}_\lambda,\mathbf{s}_\gamma\notin\text{Training}} p(\mathbf{y}_{t:\tau}|\mathbf{s}_\lambda,\mathbf{s}_\gamma)P} \\ &= \frac{p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v})}{\sum_{\mathbf{m}'_{\lambda,1:T_\lambda}}\sum_{i':j'}\sum_{\mathbf{m}'_{\gamma,1:T_\gamma}}\sum_{u':v'} p(\mathbf{y}_{t:\tau}|\mathbf{m}'_{\lambda,i':j'},\mathbf{m}'_{\gamma,u':v'}) + p(\mathbf{y}_{t:\tau}|\phi_{t:\tau})} \end{aligned} \tag{3}$$

the match between a test segment and two constituent training segments, as indicated above. For clarity of presentation, the proof is included in Appendix A.

Based on (7), therefore, we can use the maximum values of the posterior probability to locate the longest test segments with matching constituent training segments, to be used to separate the test utterance into clean utterances. Consider a test utterance $\mathbf{y}_{1:T} = \{y_t : t = 1, 2, \ldots, T\}$. At each frame time t, we can find a longest test segment, denoted by $\mathbf{y}_{t:\tau_{\max}}$, and the corresponding matching constituent training segments, denoted by $\mathbf{m}^t_{\lambda,i:j}$ and $\mathbf{m}^t_{\gamma,u:v}$, by maximizing the posterior probability, i.e.,

$$\mathbf{y}_{t:\tau_{\max}}, (\mathbf{m}^t_{\lambda,i:j}, \mathbf{m}^t_{\gamma,u:v}) = \arg\max_{\tau} \max_{\mathbf{m}'_{\lambda,i':j'}, \mathbf{m}'_{\gamma,u':v'}} P(\mathbf{m}'_{\lambda,i':j'}, \mathbf{m}'_{\gamma,u':v'} | \mathbf{y}_{t:\tau}). \tag{8}$$

That is, the longest $\mathbf{y}_{t:\tau_{\max}}$ and the matching $(\mathbf{m}^t_{\lambda,i:j}, \mathbf{m}^t_{\gamma,u:v})$ are found by first finding for each fixed-length test segment $\mathbf{y}_{t:\tau}$ the most-likely constituent training segments, and then finding the test segment with maximum length (i.e., $\tau_{\max}$) that results in the maximum posterior probability. Before discussing more details of implementing this algorithm for speech separation, we consider two special cases.

C. Special Cases

In (8), by forcing each test segment to contain only a single frame, i.e., $\tau = t$ for all t, we obtain a system which finds two constituent training frames for each test frame independently of the other test/training frames in the sequence. Noting that in our system each constituent training frame corresponds to a Gaussian component, this frame-by-frame matching system is effectively identical to the conventional GMM-based separation system (e.g., [13]–[16]), which performs unconstrained speech separation assuming temporal independence between speech frames. We will include this GMM-based separation system in our experimental comparison, to demonstrate the effect of segment matching, based on the CLOSE algorithm, on unconstrained speech separation.

Equation (8) corresponds to a "two-sided" constrained separation system, in which both the constituent training frames for each test frame are constrained temporally by their respective longest matching constituent training segments. To find the two matching constituent training segments for a test segment, this system needs to search $N_\lambda \times N_\gamma$ possible combinations, where N_λ and N_γ represent the number of training segments from speaker λ and speaker γ, respectively. Between the GMM-based system in which there is no temporal constraint on successive constituent frames, and the two-sided constrained system in which both constituent frames are constrained temporally by longest matching constituent segments, there is a third system in which one constituent frame is constrained by longest matching constituent segment and the other constituent frame is left unconstrained temporally. We call the third system a "one-sided" constrained system, with the formulation presented below. As will be demonstrated experimentally in the paper, the one-sided constrained system has the potential to offer a good balance between the accuracy of separation and the computational load of separation, in comparison to the two-sided constrained system.

Reconsider the segmental likelihood function (4) of a test segment $\mathbf{y}_{t:\tau}$, now associated with a temporally constrained constituent training segment $\mathbf{m}_{\lambda,i:j}$ from speaker λ, and a temporally unconstrained constituent training segment from speaker γ. Denote the unconstrained constituent training segment as $*_{\gamma,t:\tau}$. This likelihood function can be expressed as

$$p(\mathbf{y}_{t:\tau} | \mathbf{m}_{\lambda,i:j}, *_{\gamma,t:\tau}) = \prod_{\epsilon=t}^{\tau} \max_{1 \le m_\gamma \le M_\gamma} p(y_\epsilon | m_{\lambda,\zeta(\epsilon)}, m_\gamma). \tag{9}$$

The unconstrained constituent training segment is formed by choosing the frames freely from the training data of speaker γ (mapped to GMM G_γ), to maximize the likelihood with the constrained constituent training segment. Thus, $*_{\gamma,t:\tau} = (\hat{m}_{\gamma,t}, \hat{m}_{\gamma,t+1}, \ldots, \hat{m}_{\gamma,\tau})$ where each $\hat{m}_{\gamma,\epsilon} = \arg\max_{1 \le m_\gamma \le M_\gamma} p(y_\epsilon | m_{\lambda,\zeta(\epsilon)}, m_\gamma)$. Equation (9) gives a "text-dependent" likelihood of the test segment, dependent on the temporal dynamics of the constituent training segment $\mathbf{m}_{\lambda,i:j}$. Substituting $p(\mathbf{y}_{t:\tau} | \mathbf{m}_{\lambda,i:j}, *_{\gamma,t:\tau})$ into (3) in place of the two-sided constrained likelihood function, we can obtain the one-sided constrained posterior probability. We use the expression

$$P(\mathbf{m}_{\lambda,i:j}, *_{\gamma,t:\tau} | \mathbf{y}_{t:\tau}) = \frac{p(\mathbf{y}_{t:\tau} | \mathbf{m}_{\lambda,i:j}, *_{\gamma,t:\tau})}{\sum_{\mathbf{m}'_{\lambda,1:T_\lambda}} \sum_{i',j'} p(\mathbf{y}_{t:\tau} | \mathbf{m}'_{\lambda,i':j'}, *'_{\gamma,t:\tau}) + p(\mathbf{y}_{t:\tau} | \phi_{t:\tau})}. \tag{10}$$

Equation (10) is only a function of the temporally constrained training segment $\mathbf{m}_{\lambda,i:j}$. Similar to (8), we can locate the longest test segments with matching temporally constrained training segments, by maximizing the posterior probabilities. At each frame time t, we obtain the longest test segment $\mathbf{y}_{t:\tau_{\max}}$ and the corresponding matching temporally constrained training segment $\mathbf{m}^t_{\lambda,i:j}$ by first finding for each fixed-length test segment $\mathbf{y}_{t:\tau}$ the most-likely temporally constrained training segment, and then finding the test segment with maximum length (i.e., $\tau_{\max}$) that results in the maximum posterior probability, i.e.,

$$\mathbf{y}_{t:\tau_{\max}}, \mathbf{m}^t_{\lambda,i:j} = \arg\max_{\tau} \max_{\mathbf{m}'_{\lambda,i':j'}} P(\mathbf{m}'_{\lambda,i':j'}, *_{\gamma,t:\tau} | \mathbf{y}_{t:\tau}). \tag{11}$$

Equation (11) shows the estimation of the temporally constrained training segments for speaker λ. By switching the temporal constraint from speaker λ to speaker γ, the same system can be used to identify the temporally constrained training segments for speaker γ.

Therefore, the two-sided constrained problem (8) can be reduced to two one-sided constrained problems (11), each dealing with the estimation of a clean speech utterance from the test utterance. To find the two temporally constrained constituent training segments for a test segment, the one-sided constrained system has a search complexity of about $N_\lambda + N_\gamma$ possible combinations, which can be significantly less than $N_\lambda \times N_\gamma$ required for the two-sided constrained system, for large numbers of training segments N_λ and N_γ. See Fig. 2 for a pseudo-program description of the one-sided constrained CLOSE algorithm. Further details of this algorithm, including its computational complexity, will be revealed in Sections III–V.

	Given a co-channel speech mixture $\mathbf{y}_{1:T} = \{y_t : t = 1,2,...,T\}$
A	For $t = 1, 2, ..., T$ Calculate mixed frame likelihoods $\max_{1 \le m_\gamma \le M_\gamma} p(y_t \mid m_\lambda, m_\gamma)$, for $1 \le m_\lambda \le M_\lambda$, using (14)
	For $t = 1, 2, ..., T$ For $\tau = t, t+1, ..., T$ Calculate $p(\mathbf{y}_{t:\tau} \mid \phi_{t:\tau})$ using (5) For each training utterance $\mathbf{m}_{\lambda,1:T_\lambda}$ from speaker λ
B	For each frame i in $\mathbf{m}_{\lambda,1:T_\lambda}$ and segment $\mathbf{m}_{\lambda,i:j}$ from i with a length equal to $\mathbf{y}_{t:\tau}$ Calculate segmental likelihood $p(\mathbf{y}_{t:\tau} \mid \mathbf{m}_{\lambda,i:j}, *_{\gamma,t:\tau})$ using (9)
	For each training segment $\mathbf{m}_{\lambda,i:j}$
C	Prune $\mathbf{m}_{\lambda,i:j}$ with small $p(\mathbf{y}_{t:\tau} \mid \mathbf{m}_{\lambda,i:j}, *_{\gamma,t:\tau})$ and all further $\mathbf{m}_{\lambda,i:j}$ from i
	Calculate posterior probability $P(\mathbf{m}_{\lambda,i:j}, *_{\gamma,t:\tau} \mid \mathbf{y}_{t:\tau})$ using (10) Obtain the longest matching segments $(\mathbf{y}_{t:\tau_{\max}}, \mathbf{m}^t_{\lambda,i:j})$ at t using (11) Reconstruct the clean utterance from speaker λ using (18)

Fig. 2. Outline of the one-sided constrained CLOSE algorithm for separating one clean speech utterance from a co-channel speech mixture. Part A, B and C include the major strategies for accelerating the algorithm, detailed in Section IV-D.

III. More Implementation Details

A. Likelihood of Mixed Frame and Gain Modeling

In the above algorithms, (4), (5), (9), we need to calculate the likelihood of a test frame associated with two constituent training frames $p(y_t|m_\lambda, m_\gamma)$, where m_λ and m_γ each correspond to a Gaussian component in the appropriate speaker's GMM, i.e., $g_\lambda(x|m_\lambda)$ and $g_\gamma(x|m_\gamma)$, which model the probability distributions of the two constituent frames. Given the probability distribution of each constituent frame, and given the assumption that the test frame y_t is an additive mixture of the two constituent frames, there can be several methods, for example, log-max, Algonquin, lifted max or parallel model combination [2], [9], [30]–[32], that can be used to derive the likelihood of the test frame. In this paper, we use a simple method, the log-max model, to obtain this likelihood.

For each frame, we calculate its log power spectrum as the feature. Assume that the log power spectrum of y_t can be expressed in F distinct frequency channels, i.e., $y_t = \{y_{t,f} : f = 1, 2, \ldots, F\}$, where $y_{t,f}$ is the log power of the fth channel. Then $p(y_t|m_\lambda, m_\gamma)$ can be expressed as

$$p(y_t|m_\lambda, m_\gamma) = \prod_{f=1}^{F} p(y_{t,f}|m_\lambda, m_\gamma) \tag{12}$$

where $p(y_{t,f}|m_\lambda, m_\gamma)$ is the likelihood of the log power of the fth channel. For simplicity, in (12) we assume independence between the frequency channels. Let $x_{\lambda,f}$ and $x_{\gamma,f}$ represent the log powers of the same channel of the two constituent frames, subject to probability distributions $g_\lambda(x|m_\lambda)$ and $g_\gamma(x|m_\gamma)$. We can have $y_{t,f} \simeq \max(x_{\lambda,f}, x_{\gamma,f})$ [2], [30]. Thus, $p(y_{t,f}|m_\lambda, m_\gamma)$ can be written as

$$\begin{aligned} &p(y_{t,f}|m_\lambda, m_\gamma) \\ &\quad = g_\lambda(y_{t,f}|m_\lambda)P_\gamma(y_{t,f}|m_\gamma) + g_\gamma(y_{t,f}|m_\gamma)P_\lambda(y_{t,f}|m_\lambda) \end{aligned} \tag{13}$$

where $P_\gamma(y_{t,f}|m_\gamma) = \int_{-\infty}^{y_{t,f}} g_\gamma(x_f|m_\gamma)dx_f$, and likewise for $P_\lambda(y_{t,f}|m_\lambda)$.

In the separation, we need to model constituent speakers/frames with gains different from the training data. Rewrite the constituent-frame Gaussians as $g_\lambda(x|m_\lambda, a_\lambda)$ and $g_\gamma(x|m_\gamma, a_\gamma)$, where a_λ and a_γ are the gain updates (in dB) for speaker λ and speaker γ, respectively, and $g_\lambda(x|m_\lambda, a_\lambda) = \mathcal{N}(x; \mu_{m_\lambda} + a_\lambda, \Sigma_{m_\lambda})$, where μ_{m_λ} and Σ_{m_λ} are the training-data based mean vector and covariance matrix of the appropriate Gaussian. For any given test utterance, we calculate the gain updates a_λ and a_γ at the frame level on a frame-by-frame basis, by maximizing the test frame likelihood $p(y_t|m_\lambda, m_\gamma)$ against a set of predefined update values for each constituent speaker. The gain-optimized test frame likelihood can be expressed as

$$p(y_t|m_\lambda, m_\gamma) = \max_{a_\lambda \in \mathcal{G}_\lambda, a_\gamma \in \mathcal{G}_\gamma} \prod_{f=1}^{F} p(y_{t,f}|m_\lambda, m_\gamma, a_\lambda, a_\gamma) \tag{14}$$

where $\mathcal{G}_\lambda$ and $\mathcal{G}_\gamma$ are the predefined gain-update value sets for speaker λ and γ, and $p(y_{t,f}|m_\lambda, m_\gamma, a_\lambda, a_\gamma)$ is the local channel likelihood (13) with each component Gaussian including a corresponding gain update.

B. Speaker-Pair Identification

In the above discussions, we have assumed that for each test utterance the identities of the two constituent speakers are known. Actually, in our experiments, we assume no prior knowledge about the speakers' identities. The following details the algorithm which we use to automatically identify the constituent speakers for each given test utterance.

Assume that a test utterance contains frames/segments which are dominated by the individual speakers. Therefore, the problem can be viewed as one to identify the two constituent speakers using a noisy utterance, with *partial temporal* corruption (corresponding to those heavily mixed features not matching any single speaker's feature). We describe a new approach for extracting the single-speaker dominated features for the identification. The new approach is an extension of our previous approach [37] for speech recognition using signals with partial temporal corruption.

Given a test utterance $\mathbf{y}_{1:T} = \{y_t : t = 1, 2, \ldots, T\}$, we use the following GMM-based expression to calculate its frame

likelihood associated with speaker λ and speaker γ with respective gains a_λ and a_γ:

$$\begin{aligned} &q(y_t|\lambda,\gamma,a_\lambda,a_\gamma) \\ &\quad=\frac{1}{2}\sum_{m=1}^{M_\lambda} w_\lambda(m)g_\lambda(y_t|m,a_\lambda) \\ &\qquad+\frac{1}{2}\sum_{m=1}^{M_\gamma} w_\gamma(m)g_\gamma(y_t|m,a_\gamma). \end{aligned} \tag{15}$$

When the test frame y_t is dominated by a single speaker, λ or γ, and has the correct gain, $q(y_t|\lambda,\gamma,a_\lambda,a_\gamma)$ should be large. Therefore, we can identify the speaker pair by using the frames producing large likelihoods, assuming that they are likely to correspond to the single-speaker dominated frames of the two speakers. Denote by $q(y_t^*|\lambda,\gamma,a_\lambda,a_\gamma)$ the test-frame likelihoods sorted in descending order, with y_t^*, $t=1,2,\ldots,T$, corresponding to the test frames from the highest likelihood to the lowest likelihood associated with speaker pair λ, γ and gains a_λ, a_γ. To select the optimal frames for identification, we formulate a posterior probability for each speaker pair using the corresponding $q(y_t^*|\lambda,\gamma,a_\lambda,a_\gamma)$ for each pair, as a function of the number of test frames with the highest likelihoods. This posterior probability can be expressed as

$$\begin{aligned} &Q(\lambda,\gamma|\mathbf{y}_{1:T},\mathcal{T}) \\ &\quad=\frac{\displaystyle\sum_{a_\lambda\in\mathcal{G}_\lambda,a_\gamma\in\mathcal{G}_\gamma}\prod_{t=1}^{\mathcal{T}} q(y_t^*|\lambda,\gamma,a_\lambda,a_\gamma)}{\displaystyle\sum_{\lambda',\gamma'}\sum_{a_{\lambda'}\in\mathcal{G}_{\lambda'},a_{\gamma'}\in\mathcal{G}_{\gamma'}}\prod_{t=1}^{\mathcal{T}} q(y_t^{*'}|\lambda',\gamma',a_{\lambda'},a_{\gamma'})} \\ &\qquad\qquad \mathcal{T}=1,2,\ldots,T \end{aligned} \tag{16}$$

where we assume an equal prior probability for all the speaker pairs and gains, and $\mathcal{T}$ is the number of the highest-likelihood frames used in forming the posterior probability. Thus, the most-likely speaker pair can be obtained by jointly maximizing $Q(\lambda,\gamma|\mathbf{y}_{1:T},\mathcal{T})$ over all speaker pairs and all possible numbers of the highest-likelihood frames $\mathcal{T}$, i.e.,

$$\hat{\lambda},\hat{\gamma}=\arg\max_{\lambda,\gamma,\mathcal{T}} Q(\lambda,\gamma|\mathbf{y}_{1:T},\mathcal{T}). \tag{17}$$

We have found that it is helpful to impose a constraint on the minimum value of the optimal frame number $\mathcal{T}$. The constraint reflects a balance between retaining sufficient features for identification and ignoring noisy features for robustness. In our experiments, we forced $\mathcal{T}\geq T/2$, i.e., at least half of the frames from a test utterance are used to identify the constituent speakers.

C. Clean Utterance Reconstruction

Given test utterance $\mathbf{y}_{1:T}=\{y_t : t=1,2,\ldots,T\}$, after finding the longest test segment $\mathbf{y}_{t:\tau_{\max}}$ and the matching constituent training segments $\mathbf{m}_{\lambda,i:j}^t$ and $\mathbf{m}_{\gamma,u:v}^t$ at each time t [i.e., (8) or (11)], we use $\mathbf{m}_{\lambda,i:j}^t$ and $\mathbf{m}_{\gamma,u:v}^t$ to estimate the two underlying clean speech utterances forming the test utterance. In the following, we describe the algorithm which uses $\mathbf{m}_{\lambda,i:j}^t$ to estimate the clean utterance from speaker λ. The same algorithm can be used to estimate the clean utterance from speaker γ, by replacing $\mathbf{m}_{\lambda,i:j}^t$ with $\mathbf{m}_{\gamma,u:v}^t$.

Let $s_{\lambda,\epsilon}$ represent the clean frame of speaker λ at time ϵ, $\epsilon=1,2,\ldots,T$, and $S_{\lambda,\epsilon}$ be the magnitude spectrum of the frame. We can obtain an estimate of $S_{\lambda,\epsilon}$ by taking all the longest matching training segments that contain $s_{\lambda,\epsilon}$ and averaging over the corresponding training frames. In the average, we use the posterior probability, obtained in (8) or (11), as a confidence score. We use the expression

$$\begin{aligned} &\hat{S}_{\lambda,\epsilon} \\ &\quad=\frac{\displaystyle\sum_t A(m_{\lambda,\zeta(\epsilon)}^t)\sqrt{\exp(a_{\lambda,\zeta(\epsilon)}^t)}P(\mathbf{m}_{\lambda,i:j}^t,\mathbf{m}_{\gamma,u:v}^t|\mathbf{y}_{t:\tau_{\max}})}{\bar{P}} \end{aligned} \tag{18}$$

where the sum is over all test segments $\mathbf{y}_{t:\tau_{\max}}$ that contain frame $s_{\lambda,\epsilon}$; $m_{\lambda,\zeta(\epsilon)}^t$ is the training frame and $a_{\lambda,\zeta(\epsilon)}^t$ is the corresponding gain [obtained using (14)] corresponding to $s_{\lambda,\epsilon}$, taken from the longest matching training segment $\mathbf{m}_{\lambda,i:j}^t$; $A(m_{\lambda,\zeta(\epsilon)}^t)$ represents a magnitude spectrum corresponding to training frame $m_{\lambda,\zeta(\epsilon)}^t$. As shown in (18), each clean frame is estimated through identification of a longest matching training segment, and each estimate is smoothed over successive longest matching training segments. This improves both accuracy for frame estimation and robustness to imperfect segment match. Frames within the same segment share a common confidence score which is the posterior probability of the segment. In (18), $\bar{P}$ is a normalization term. In our experiments, the following expression is found to be suitable:

$$\bar{P}=\begin{cases}\sum_t P(\mathbf{m}_{\lambda,i:j}^t,\mathbf{m}_{\gamma,u:v}^t|\mathbf{y}_{t:\tau_{\max}}) \\ \qquad\text{if } \sum_t P(\mathbf{m}_{\lambda,i:j}^t,\mathbf{m}_{\gamma,u:v}^t|\mathbf{y}_{t:\tau_{\max}})>1 \\ 1 \qquad\text{if } \sum_t P(\mathbf{m}_{\lambda,i:j}^t,\mathbf{m}_{\gamma,u:v}^t|\mathbf{y}_{t:\tau_{\max}})\leq 1.\end{cases} \tag{19}$$

The last condition prevents small posterior probabilities being scaled up to give a false emphasis. If we use the one-sided constrained system (11), the posterior probabilities in (18) and (19) should be replaced by $P(\mathbf{m}_{\lambda,i:j}^t,*_{\gamma,t:\tau_{\max}}|\mathbf{y}_{t:\tau_{\max}})$, as defined in (11).

In our system, we use the DFT (discrete Fourier transform) magnitudes of the training frames as the magnitude spectra $A(m_{\lambda,i})$ to form the estimate. Given the index of a training frame $m_{\lambda,i}$, we can have two different approaches to calculate $A(m_{\lambda,i})$. First, $A(m_{\lambda,i})$ can be calculated directly using the specific training speech frame $x_{\lambda,i}$ corresponding to $m_{\lambda,i}$ [see the definition of $m_{\lambda,i}$ in (2)]. Alternatively, $A(m_{\lambda,i})$ can be calculated as an average DFT magnitude over all the training speech frames used to form the Gaussian component $g_\lambda(x|m_{\lambda,i})$ in the speaker's GMM [see (1)]. In the latter case, $A(m_{\lambda,i})$ corresponds to the mean vector of $g_\lambda(x|m_{\lambda,i})$ in the DFT magnitude format. For convenience, we call the estimate (18) based entirely on the training data a codeword-based estimate, by viewing each training frame or Gaussian component as a codeword and the corresponding training data set or GMM as a codebook.

Alternatively, we can form an estimate for each clean utterance by directly suppressing the crosstalk noise in the test utterance. In this approach, we use the codeword-based estimate

(18) to form an optimal filter. In our system, we use a Wiener filter of the form:

$$H_{\lambda,\epsilon} = \frac{\hat{S}^2_{\lambda,\epsilon}}{\hat{S}^2_{\lambda,\epsilon} + \hat{N}^2_{\lambda,\epsilon}} \tag{20}$$

where $H_{\lambda,\epsilon}$ represents the filter function at time ϵ, and $\hat{N}^2_{\lambda,\epsilon}$ is an estimate of the crosstalk noise power spectral density, which can be obtained by using the test speech periodogram and the clean speech power spectral density estimate in a smoothed recursion:

$$\hat{N}^2_{\lambda,\epsilon} = \alpha \hat{N}^2_{\lambda,\epsilon-1} + (1-\alpha)\max\left[Y^2_{\epsilon} - \hat{S}^2_{\lambda,\epsilon}, 0\right] \tag{21}$$

where α is a smoothing constant ($\alpha = 0.95$ in our experiments) and Y^2_{ϵ} represents the test speech periodogram at time ϵ. For convenience, we call the estimate based on (20) a filter-based estimate. In our experiments, both estimates based on the codewords and on the filter produced similar separation quality. The filter-based estimates are used in the evaluation.

IV. Experimental Studies

A. Test Data, Systems and Performance Measures

The large-vocabulary continuous speech recognition Wall Street Journal Phase I (WSJ0) database [39] was used in the experiments. In the database, there are 101 speakers providing short-term data for speaker-independent training (SI-TR-S). From these, we selected 20 speakers (10 male, 10 female) to construct our experimental data set. Each speaker has about 140 utterances, with an average utterance duration of about 7 s. For each speaker, we chose two utterances to be used to form mixed speech, or co-channel speech, for separation test, and used the remaining (about 138) utterances for training; the training utterances and test utterances had no sentence texts in common (a simulation of the unconstrained speech scenario). All the test utterances of all the speakers were chosen to have a similar duration (about 9 s), with an average of 20.4 words per utterance (see Appendix B for more details).

The two test utterances of each speaker (target) were mixed with the two test utterances of each of the other 19 speakers (masker), first utterance to first utterance, and second utterance to second utterance, at five different target-to-masker ratios (TMRs): 10, 5, 0, -5, and -10 dB, measured on the utterance level (i.e., the target speakers vary from being dominant to background). Therefore, at each TMR level, there were $20 \times 19 \times 2 = 760$ co-channel utterances, each co-channel utterance containing two speech utterances, for separation test. In other words, at each TMR level, every speaker was used as target, with every of the other speakers being used as masker, for the mixture; two sets of such speech mixtures were generated for the test, each set containing a different utterance for each speaker.

To build the CLOSE system, first, we trained a GMM [i.e., (1)] for each speaker using the training utterances of the speaker. In our experiments, each speaker's GMM contained 512 Gaussian components with diagonal covariance matrices. Then, we took the GMM and the training utterances of each speaker and obtained an utterance model [i.e., (2)] for each training utterance, to be used in the CLOSE algorithm for segment matching. The speech signals, sampled at 16 kHz, were divided into frames of 20 ms with a frame period of 10 ms. In our experiments, for identifying matching segments, we represented each frame in the form of Mel-frequency log filterbank power spectrum. We have tested filterbanks of variable numbers of channels, from 26 as typically used in speech analysis for speech recognition, to some higher resolutions up to 128. In general, a higher-resolution power spectrum representation gave improved results, but also resulted in higher computational load. For the experiments in this paper, we used a 50-channel filterbank representation, which appeared to provide a good balance. As described in Section III-C, when matching training segments are found, the clean speech frames are reconstructed using the DFT magnitudes of the corresponding training frames (with phases taken from the test frames of co-channel speech). In the CLOSE system for the experiments, we used a gain-update set [-12, -9, -6, -3, 0, 3, 6, 9, 12] to account for the variable gain changes from the training data, without assuming specific knowledge of the TMR in each test utterance. This set corresponds to $\mathcal{G}_\lambda$ and $\mathcal{G}_\gamma$ in (14) and (16).

In the CLOSE system, for identifying the longest matching segments, we have implemented both the two-sided constrained algorithm (8) and the one-sided constrained algorithm (11). Our experiments were performed mainly using the one-sided algorithm, for its much lower computational complexity. We have compared the two-sided algorithm and one-sided algorithm using a smaller number of training utterances for each speaker, and found that they achieved similar separation performance. We chose the GMM-based separation system as a baseline system for comparison. As described in Section II-C, the GMM-based system is a special case of the CLOSE system, which assumes independence between consecutive speech frames to account for the lack of knowledge of temporal dynamics of unconstrained speech. The comparison between the GMM-based system and the CLOSE system demonstrates the feasibility and benefits of identifying maximum-length matching segments between the training data and test data as a form of temporal constraint for unconstrained speech separation.

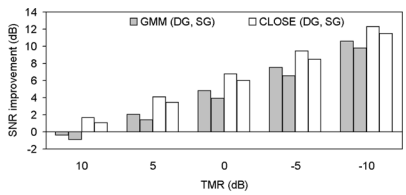
Both objective and subjective tests were conducted to evaluate the separation performance. The objective measures include sentence-level signal-to-noise ratio (SNR) improvement after separation, perceptual evaluation of speech quality (PESQ), and large-vocabulary continuous speech recognition (LVCSR) word accuracy. The subjective tests include the mean opinion score (MOS) for quality, for intelligibility, and subjective preference test.

B. Speaker Identification Evaluation

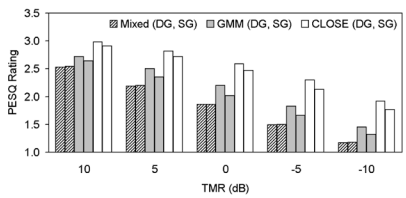
In the first set of experiments, we evaluated our algorithm for identifying the constituent speakers given co-channel speech. The algorithm uses optimal frame selection, as described in Section III-B. Table I presents the identification accuracy for the target speaker in each co-channel utterance, as a function of the utterance TMR. The results are averaged over the 760 co-channel utterances for test at each TMR condition, with all the 20 speakers appearing as targets as described above. To

TABLE I
TARGET SPEAKER IDENTIFICATION ACCURACY (%) GIVEN CO-CHANNEL SPEECH, AS A FUNCTION OF THE TARGET-TO-MASKER RATIO (TMR), BY THE PROPOSED SPEAKER-PAIR IDENTIFICATION ALGORITHM WITH OPTIMAL FRAME SELECTION, COMPARED TO NO FRAME SELECTION

TMR (dB)	Optimal frame selection	No frame selection
10	100	100
5	99.6	99.3
0	99.1	94.5
-5	94.1	82.8
-10	85.3	70.8



(a)



(b)

Fig. 3. Objective evaluation of the target speech given co-channel speech, as a function of the TMR, in different-gender (DG) and same-gender (SG) mixtures and separated utterances, by the GMM-based separation algorithm and the new CLOSE separation algorithm. (a) SNR improvement. (b) PESQ measure.

assess the effect of optimal frame selection on the identification, we also included the corresponding identification results based on all the frames (i.e., no frame selection). From Table I, it is evident that the optimal frame selection significantly improved the speaker identification accuracy, especially at the lower TMR conditions. All the experiments described below, including those for the GMM-based separation system, were based on the speaker identification results produced by the algorithm.

C. One-Sided Constrained CLOSE System Evaluation

Three types of objective evaluation were conducted. At each TMR condition, we divide the 760 co-channel utterances for test into two group: same-gender (SG) mixture, with 360 utterances, and different-gender (DG) mixture, with 400 utterances. The results presented below for each group are obtained by averaging over the utterances within the group. Fig. 3(a) shows the SNR improvement after separation by the CLOSE separation system and GMM-based separation system, as a function of the original co-channel utterance TMR. The CLOSE system improved over the GMM-based system in all the gender group and TMR conditions. The CLOSE system obtained positive SNR improvement even in the high TMR condition (10 dB) where the GMM-based system failed to show improvement.

Next, PESQ scores for the target speech in the co-channel speech and separated utterances were calculated using the code provided in [40] and the results are presented in Fig. 3(b). Again, both separation systems improved the PESQ scores

TABLE II
LARGE-VOCABULARY CONTINUOUS SPEECH RECOGNITION WORD ACCURACY (%) FOR THE TARGET SPEECH GIVEN CO-CHANNEL SPEECH, AS A FUNCTION OF THE TMR, IN DIFFERENT-GENDER (DG) AND SAME-GENDER (SG) MIXTURES AND SEPARATED UTTERANCES, BY THE GMM-BASED SEPARATION ALGORITHM AND THE NEW CLOSE ALGORITHM

TMR (dB)	Mixed			GMM			CLOSE		
	DG	SG	AVG	DG	SG	AVG	DG	SG	AVG
Clean			81.5						
10	31.7	30.9	31.4	37.8	30.8	34.5	61.3	57.8	59.6
5	12.3	10.7	11.5	31.2	23.4	27.5	57.8	51.5	54.8
0	-1.5	-1.5	-1.5	22.6	13.8	18.4	51.8	43.6	47.9
-5	-6.6	-8.1	-7.3	11.9	6.3	9.3	39.9	30.1	35.2
-10	-8.9	-10.3	-9.5	5.9	2.2	4.2	25.5	18.3	22.1

of the target utterances, and the CLOSE system obtained the highest scores in all the gender group and TMR conditions. It is also observed that as the TMR decreased, the CLOSE algorithm suffered a slower degradation in the PESQ score than suffered by the co-channel utterances and the GMM-based separation algorithm, for both the SG and DG groups.

A more challenging objective measure was the accuracy rate achieved by a large-vocabulary continuous speech recognition (LVCSR) system. The LVCSR system was built following the HTK WSJ Training Recipe [41], trained using the full set of WSJ0 and WSJ1 training data, with TIMIT-bootstrapped monophones. Slightly different from the recipe system, in our system we dropped the zero'th cepstral coefficient (C0) to account for the variable gain changes of the target speech. For validation, the system was tested on the November'92 ARPA WSJ 5k-vocabulary test set, with 330 test utterances from eight untrained speakers, and achieved $\sim 92\%$ word accuracy. In the speech separation evaluation, the separated target utterances were passed to the system for recognition without any compensation for the likely acoustic mismatch caused by the separation and reconstruction processes. Table II shows the word recognition accuracy for the target utterances when they were clean, mixed and separated, as a function of the TMR. First, the recognition system obtained only 81.5% word accuracy for the 40 clean WSJ0 utterances which we used to form the co-channel utterances for test. This may indicate that these 40 utterances are more difficult to recognize accurately than the average of the November'92 test utterances. Second, compared to the GMM-based separation system, the CLOSE system significantly improved the word accuracy, especial with lower TMR levels. Compared to the co-channel utterances, the GMM-based system also improved the word recognition accuracy in almost all test conditions, except for one test condition with same-gender mixture at $\mathrm{TMR} = 10$ dB, in which it failed to improve the word recognition accuracy. All the above three objective measures indicate that it is generally more difficult to correctly separate the utterances when the speakers are of the same gender. This is shown by the lower scores in the SNR, PESQ, and word recognition accuracy for the separated target speech from same-gender mixtures than from different-gender mixtures.

Three types of subjective listening tests were conducted. A group of eighteen volunteers (four female and fourteen male) participated in the tests. The test samples were also prepared in two groups: same-gender mixture and different-gender mixture. From the 20 constituent speakers, we selected 12 speakers

TABLE III
FORMATION OF THE DG AND SG SPEECH MIXTURES FOR SUBJECTIVE LISTENING TESTS. AFTER SEPARATION, THE ESTIMATED TARGET SPEECH IN EACH MIXTURE IS PRESENTED TO THE SUBJECTS FOR EVALUATION

Group	Constituent speaker (gender, number)	Mixing	
		Target, Masker	TMR (dB)
DG	F1	F1, M1	10
	M1	M1, F1	-10
	F2	F2, M2	5
	M2	M2, F2	-5
	F3	F3, M3	0
	M3	M3, F3	0
SG	F4	F4, F5	10
	F5	F5, F4	-10
	F6	F6, F7	5
	F7	F7, F6	-5
	M4	M4, M5	0
	M5	M5, M4	0

(7 female and 5 male), each speaker with one utterance, to form the co-channel data for test. Table III gives the details of the formation. As shown in the table, from the 12 speakers we created 6 groups of co-channel utterances; each group consisted of two co-channel utterances, differing in the target/masker specification, and formed by two speakers at one of the TMR levels 10/-10, 5/-5, and 0/0 dB. After separation, the separated target utterances of all the groups were used for evaluation. Therefore, for each separation system (GMM and CLOSE), there were a total of 12 separated target utterances for evaluation which covered all the 12 constituent speakers and the full range of TMR from 10 to -10 dB.

The *quality* MOS test for the target speech was conducted by closely adhering to the standard ITU-T P.800 [42]. The 24 separated target utterances produced by the two separation systems were presented to each listener in random order. The listeners assessed the quality of each utterance by rating it on one of the five scales: 1-poor, 2-bad, 3-fair, 4-good, and 5-excellent. Finally, the mean opinion score for each utterance was obtained by averaging its ratings over all the listeners. Fig. 4(a) shows the results, as a function of the co-channel speech TMR. The results reveal that the CLOSE system outperformed the GMM-based system for both the DG and SG groups at all the TMR conditions. Again, it can be found that the improvement by the CLOSE system was greater in the lower TMR conditions than in the higher TMR conditions. In the subjective tests, we have seen cases in which higher scores were given to the target speech separated from same-gender mixtures than from different-gender mixtures, in both separation systems.

The *intelligibility* MOS test for the target speech was conducted using an approach similar to that described in [43], [44]. In [43], a five-scale system was used to assess the intelligibility of a speech utterance, and in [44] a seven-scale system was used to assess the difficulty in understanding a speech utterance. In our test, we tried to combine these two rating systems. We used a seven-scale rating system to assess the intelligibility of each separated target utterance: 1-not intelligible, 2-slightly intelligible, 3-somewhat intelligible, 4-mostly intelligible only if I concentrate, 5-mostly intelligible, 6-completely intelligible only if I concentrate, and 7-completely intelligible. We found that the use of a higher-resolution rating system is helpful to reduce the ambiguity in assessing the intelligibility by the listeners. The 24 separated target utterances produced by the two separation systems were presented to each listener in random order and the final score for each utterance was obtained by averaging over all the listeners. Fig. 4(b) presents the results. The CLOSE system outperformed the GMM-based system in all the test conditions. At the very low $\mathrm{TMR} = -10$ dB, the target utterances estimated by the CLOSE system were rated an average score of about 4.5, between 'mostly intelligible only if I concentrate' and 'mostly intelligible,' while the target utterances estimated by the GMM-based system were rated an average score of about 2.6, between 'slightly intelligible' and 'somewhat intelligible.'

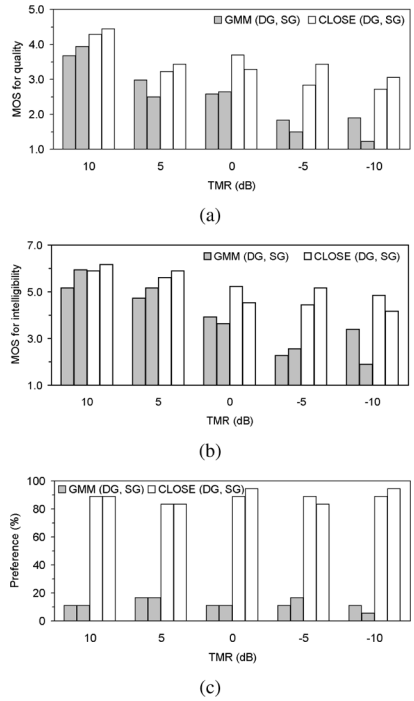


Fig. 4. Subjective evaluation of the target speech estimated by the GMM-based separation algorithm and the new CLOSE algorithm. (a) MOS for quality. (b) MOS for intelligibility. (c) Preference percentage.

A further informal subjective evaluation, in the form of a preference test, was conducted. The same target utterances, one estimated by the GMM-based system and the other by CLOSE, were paired. The 12 pairs were then presented in random order, with the two utterances in each pair also in random order, to each listener. The results are presented in Fig. 4(c), which shows the percentage of the listeners preferring the utterances/separation systems in all the utterance pairs.

D. CLOSE Algorithm Analysis

We conducted experiments to compare the two versions of the CLOSE algorithm: the two-sided constrained algorithm (8) and the one-sided constrained algorithm (11). The first is an "exact" algorithm while the second is an approximation. The second is more efficient in computation and was used to produce all the experimental results described above. From the above full test data set we chose a subset to conduct the comparison experiments. We randomly selected eight speakers (four male and four female) from the 20 speakers, each speaker taking one utterance,

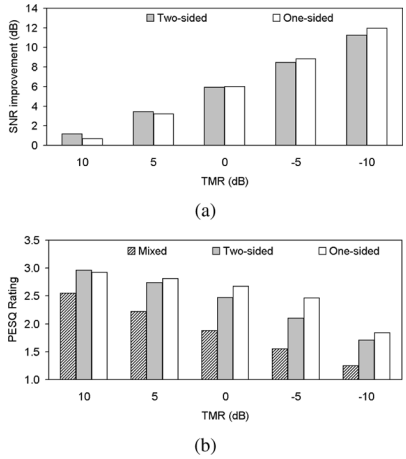


Fig. 5. Objective evaluation of the target speech given co-channel speech and separated utterances, by the two-sided constrained CLOSE algorithm and one-sided constrained CLOSE algorithm, for a smaller test data set. Ten training utterances from each speaker were used to provide segment examples for segment matching. (a) SNR improvement. (b) PESQ measure.

to form the co-channel utterances for test. For each TMR condition, the full combination between the eight speakers resulted in 56 co-channel utterances to be separated by each algorithm.

As discussed in Section II-C, for each test segment of the co-channel speech, the two-sided algorithm needs to compare all possible combinations between the training segments of the two constituent speakers. For the WSJ0 database used in our experiments, in which each constituent speaker has about 138 training utterances with an average utterance duration about 7 s (an average overall duration about 16 min), the two-sided algorithm requires an average of ~ 9.3 billion comparisons and ~ 8 GB memory for finding the matching constituent training segments for each test segment. This amount of computation and memory usage was found to be impractical in our experiments. Based on practicality reasons, we reduced the number of the training utterances used to provide segment examples from ~ 138 to 10 (about 1 min speech) for each constituent speaker. These 10 training utterances were selected randomly for each speaker from his/her training utterances. Both the two-sided and one-sided algorithms were compared on this new system with reduced numbers of segment examples for segment matching.

Fig. 5 and Table IV present the results of the comparison using three objective measures: SNR improvement, PESQ score and LVCSR word accuracy. We can see that the three measures are well correlated, all showing slightly better performance for the two-sided algorithm at the higher TMR conditions, and all showing slightly better performance for the one-sided algorithm at the lower TMR conditions. It was observed that the one-sided algorithm actually found a greater number of long matching segments between the co-channel speech and the 10 constituent training utterances, than the two-sided algorithm. This gave the one-sided algorithm better robustness to separate the lower-TMR utterances. Given limited choices of constituent training segments, the probability of finding simultaneously two long constituent training segments that match a randomly given long co-channel speech segment could be small. Therefore with limited training segments the two-sided algorithm performed separation mainly based on

TABLE IV
LARGE-VOCABULARY CONTINUOUS SPEECH RECOGNITION WORD ACCURACY (%) FOR THE TARGET SPEECH GIVEN CO-CHANNEL SPEECH AND SEPARATED UTTERANCES, BY THE TWO-SIDED CONSTRAINED CLOSE ALGORITHM AND ONE-SIDED CONSTRAINED CLOSE ALGORITHM. TEN TRAINING UTTERANCES FROM EACH SPEAKER WERE USED TO PROVIDE SEGMENT EXAMPLES FOR SEGMENT MATCHING

TMR (dB)	Mixed	Two-sided	One-sided
Clean	82.7		
10	30.1	58.8	58.6
5	10.2	52.6	51.8
0	-4.4	38.3	42.4
-5	-10.4	24.2	28.1
-10	-13.3	12.4	15.9

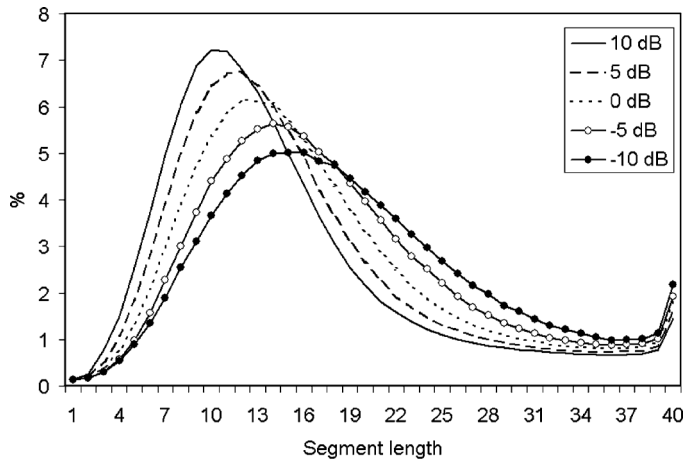


Fig. 6. Histogram of the length (in number of frames) of the longest matching segments found by the CLOSE algorithm as a function of the co-channel speech TMR.

matching short co-channel speech segments, which explains its poorer robustness to the lower-TMR utterances. The above experiments indicate that the one-sided algorithm is an effective alternative to the two-sided algorithm, for obtaining comparable separation accuracy with significantly reduced computational complexity. We also compared the results in Fig. 5 and Table IV for the one-sided algorithm to the results based on full set training shown in Fig. 3 and Table II, on the same test data set. We noticed only moderate performance reduction by reducing the number of training utterances for segment matching from ~ 138 to 10.

Fig. 6 shows the histogram of the length of the longest matching co-channel speech segments found by the one-sided constrained algorithm for correctly identified target speakers, as a function of the TMR. The histograms are calculated over the full test data set (with 760 co-channel utterances for each TMR condition). It is interesting to note that the histogram follows a consistent pattern as the TMR decreases. For the target speech with a high TMR, sharp matching training segments can be found in the corresponding clean training utterances; the number of these "true" matching segments which are long is limited given the limited training data. As the TMR decreases, ambiguity increases towards the identity of the true matching segments; this is indicated by the increased number of longer, but less sharp, matching segments. Across the TMR conditions, over 97% of the matching segments found are four or more frames long, with a mean length from about fifteen frames

to about nineteen frames as the TMR varies from 10 dB to -10 dB. The rising tails of the histograms are mostly due to the matches of the long beginning/ending silences between the co-channel utterances and the training utterances.

Finally, in our experiments several strategies were used to make the one-sided constrained CLOSE algorithm computationally efficient. These strategies are outlined in Fig. 2. The first step of the algorithm, calculating the test frame likelihoods (Part A in Fig. 2), has the same computational complexity as the conventional GMM-based separation algorithm. This step produces all the frame likelihoods required by the CLOSE algorithm for segment matching; the maximization over m_γ is required in (9), which is taken here to reduce the memory required for saving the likelihoods from two-dimensional $M_\lambda \times M_\gamma$ to one-dimensional M_λ. The subsequent search for the longest matching segments is made computationally efficient mainly in two steps. First, we only search equal-length or temporally identical training/test segments using linear time warping for matching (Part B). Second, pruning is used to remove those unlikely training segments after comparing their first few frames with the test segment (Part C); this can significantly reduce the computation without noticeable loss of performance. Combining these steps, the complexity of the algorithm scales linearly or less with the number of training segments used for segment matching. With the WSJ0 database used in our experiments, using the full training set for each speaker, our experiments indicate a comparison of 2.6:1 for the average time taken by the CLOSE algorithm compared to that taken by the GMM-based algorithm, for separating the two clean utterances of a co-channel utterance. This demonstrates the computational feasibility of the CLOSE method.

V. Concluding Remarks

In this paper, we presented a new approach, namely CLOSE, to single-channel speech separation. The CLOSE approach aimed to improve the separation accuracy by imposing long-range temporal constraints on the speech signals, without assuming knowledge about the vocabulary, grammar, or language model of the speech signals to be estimated. This was achieved by using a data-driven framework. Given co-channel speech and the training data of the individual speakers, we seek the longest co-channel speech segments with matching composite training segments to perform the separation. Our conjecture was that this would help reduce the uncertainty of the matching constituent training segments and hence the error of separation. A statistical method was presented for identifying the longest matching segments within the CLOSE system.

Experiments were conducted on the WSJ database, for separating mixtures of large-vocabulary speech utterances spoken by different speakers, without assuming knowledge about the task's vocabulary and language model, and transcripts of the training data. Various objective and subjective measures were used to evaluate the performance, including large-vocabulary continuous speech recognition. The results have demonstrated the significance of matching longest speech segments for speech separation, in terms of improving performance over conventional frame-by-frame separation algorithms for all the measures. We have also demonstrated the computational feasibility of the new method. Presently, we are studying the direct incorporation of the CLOSE algorithm into a speech recognition system, for further optimized recognition performance.

Appendix A
Proof of Inequality (7)

Express $\mathbf{y}_{t:\tau}$ as a union of two consecutive subsegments $\mathbf{y}_{t:\epsilon}$ and the complement $\mathbf{y}_{\epsilon+1:\tau}$, and express the two matching constituent segments $\mathbf{m}_{\lambda,i:j}$ and $\mathbf{m}_{\gamma,u:v}$ each as a union of the corresponding constituent subsegments, i.e., $\mathbf{m}_{\lambda,i:j} = \mathbf{m}_{\lambda,i:\zeta(\epsilon)} \cup \mathbf{m}_{\lambda,\zeta(\epsilon+1):j}$, and $\mathbf{m}_{\gamma,u:v} = \mathbf{m}_{\gamma,u:\eta(\epsilon)} \cup \mathbf{m}_{\gamma,\eta(\epsilon+1):v}$. We have the likelihood-ratio inequality (22) shown at the bottom of the page. The last inequality is obtained based on the assumption that the subsegment $\mathbf{y}_{\epsilon+1:\tau}$ and the corresponding constituents $(\mathbf{m}_{\lambda,\zeta(\epsilon+1):j},\ \mathbf{m}_{\gamma,\eta(\epsilon+1):v})$ are matching and hence $p(\mathbf{y}_{\epsilon+1:\tau}|\mathbf{m}_{\lambda,\zeta(\epsilon+1):j},\mathbf{m}_{\gamma,\eta(\epsilon+1):v}) \geq p(\mathbf{y}_{\epsilon+1:\tau}|\mathbf{m}'_{\lambda,\zeta'(\epsilon+1):j'},\mathbf{m}'_{\gamma,\eta'(\epsilon+1):v'})$ for any $(\mathbf{m}'_{\lambda,\zeta'(\epsilon+1):j'},\mathbf{m}'_{\gamma,\eta'(\epsilon+1):v'}) \neq (\mathbf{m}_{\lambda,\zeta(\epsilon+1):j},\mathbf{m}_{\gamma,\eta(\epsilon+1):v})$. Based on (6), we can have a similar inequality concerning the likelihood ratio associated with the unseen constituent segments:

$$\frac{p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v})}{p(\mathbf{y}_{t:\tau}|\phi_{t:\tau})} \geq \frac{p(\mathbf{y}_{t:\epsilon}|\mathbf{m}_{\lambda,i:\zeta(\epsilon)},\mathbf{m}_{\gamma,u:\eta(\epsilon)})}{p(\mathbf{y}_{t:\epsilon}|\phi_{t:\epsilon})}. \quad (23)$$

Dividing both the numerator and denominator of (3) by $p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v})$, and applying the above two likelihood-ratio inequalities to the expression, we can obtain the posterior probability inequality (7).

$$\begin{aligned}
&\frac{p(\mathbf{y}_{t:\tau}|\mathbf{m}_{\lambda,i:j},\mathbf{m}_{\gamma,u:v})}{p(\mathbf{y}_{t:\tau}|\mathbf{m}'_{\lambda,i':j'},\mathbf{m}'_{\gamma,u':v'})} \\
&\quad= \frac{p(\mathbf{y}_{t:\epsilon}|\mathbf{m}_{\lambda,i:\zeta(\epsilon)},\mathbf{m}_{\gamma,u:\eta(\epsilon)})p(\mathbf{y}_{\epsilon+1:\tau}|\mathbf{m}_{\lambda,\zeta(\epsilon+1):j},\mathbf{m}_{\gamma,\eta(\epsilon+1):v})}{p(\mathbf{y}_{t:\epsilon}|\mathbf{m}'_{\lambda,i':\zeta'(\epsilon)},\mathbf{m}'_{\gamma,u':\eta'(\epsilon)})p(\mathbf{y}_{\epsilon+1:\tau}|\mathbf{m}'_{\lambda,\zeta'(\epsilon+1):j'},\mathbf{m}'_{\gamma,\eta'(\epsilon+1):v'})} \\
&\quad\geq \frac{p(\mathbf{y}_{t:\epsilon}|\mathbf{m}_{\lambda,i:\zeta(\epsilon)},\mathbf{m}_{\gamma,u:\eta(\epsilon)})}{p(\mathbf{y}_{t:\epsilon}|\mathbf{m}'_{\lambda,i':\zeta'(\epsilon)},\mathbf{m}'_{\gamma,u':\eta'(\epsilon)})}. \quad (22)
\end{aligned}$$

APPENDIX B
THE WSJ0 SPEAKERS AND TEST UTTERANCES USED IN THE EXPERIMENTS

013c0202 013c021f 01gc020x 01gc0219 01kc021a 01kc021b 01lc020u 01lc021d 01oc0218 01oc021d 01sc020b 01sc020c 01uc0208 01uo030c 01vc020j 01vc021f 01xc0205 01xc020n 022c0203 022c0218 02bc0207 02bc0213 02dc020t 02dc021b 403c020f 403c021b 404c020h 404c020z 406c0206 406c020g 407c020s 407c021e 408c020g 408c021a 40fc020g 40fc0212 40gc0201 40gc0204 40jc020g 40jc0216

REFERENCES

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Comput. Speech Lang., Special Iss. Speech Separat. Recognit.*, vol. 24, pp. 1–15, 2010.

[2] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP'90*, Albuquerque, NM, USA, 1990, pp. 845–848.

[3] S. T. Roweis, "One microphone source separation," in *Proc. Neural Inf. Process. Syst.*, Denver, CO, USA, 2000, pp. 793–799.

[4] A. N. Deoras and M. Hasegawa-Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel," in *Proc. ICASSP'04*, Montreal, QC, Canada, 2004, pp. 861–864.

[5] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. Interspeech'06*, Pittsburgh, PA, USA, 2006, pp. 89–92.

[6] M. Reyes-Gomez and N. Jojic, "Speech separation by efficient combinatorial decoding of speech mixtures," in *Proc. ICMEP'09*, New York, NY, USA, 2009, pp. 498–505.

[7] R. J. Weis and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Comput. Speech Lang.*, vol. 24, pp. 16–29, 2010, Special Iss. Speech Separat. and Recognit..

[8] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang.*, vol. 24, pp. 45–66, 2010, Special Iss. Speech Separat. and Recognit..

[9] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Process. Mag.*, vol. 27, pt. 6, pp. 66–80, Nov. 2010.

[10] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech'03*, Geneva, Switzerland, 2003, pp. 1009–1012.

[11] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Monaural speech segregation based on fusion of source-driven with model-driven techniques," *Speech Commun.*, vol. 49, pp. 464–476, 2007.

[12] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monoaural speech separation based on MAXVQ and CASA for robust speech recognition," *Comput. Speech Lang.*, vol. 24, pp. 30–44, 2010, Special Iss. Speech Separat. and Recognit..

[13] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. ICASSP'04*, Montreal, QC, Canada, 2004, pp. 817–820.

[14] A. M. Reddy and B. Raj, "Soft mask methods for single channel speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1766–1776, Nov. 2007.

[15] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[16] P. Mowlaee, R. Saeidi, Z.-H. Tan, M. G. Christensen, P. Franti, and S. H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. ICASSP'10*, Dallas, TX, USA, 2010, pp. 4430–4433.

[17] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.

[18] M. Wu, D. L. Wang, and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," *IEEE Trans. Audio Speech Process.*, vol. 11, no. 3, pp. 229–241, May 2003.

[19] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[20] R. Han, P. Zhao, Q. Gao, Z. Zhang, H. Wu, and X. Wu, "CASA based speech separation for robust speech recognition," in *Proc. Interspeech'06*, Pittsburgh, PA, USA, 2006, pp. 78–81.

[21] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 289–298, Jan. 2006.

[22] S. M. Schimmel, L. E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Proc. ICASSP'07*, Honolulu, HI, USA, 2007, pp. 605–608.

[23] Y. P. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Commun.*, vol. 51, pp. 230–239, 2009.

[24] J. Barker, N. Ma, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Comput. Speech Lang.*, vol. 24, pp. 94–111, 2010, Special Iss. Speech Separat. and Recognit..

[25] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *J. Mach. Learn. Res.*, vol. 7, pp. 1963–2001, 2006.

[26] G. J. Jang and T. W. Lee, "A maximum likelihood approach to single-channel source separation," *J. Mach. Learn. Res.*, vol. 4, pp. 1365–1392, 2003.

[27] B. Raj, R. Singh, and P. Smaragdis, "Recognizing speech from simultaneous speakers," in *Proc. Interspeech'05*, Lisbon, Portugal, 2005, pp. 3317–3320.

[28] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. Interspeech'06*, Pittsburgh, PA, USA, 2006, pp. 2614–2617.

[29] B. Raj and P. Smaragdis, "Latent variable de-composition of spectrograms for single channel speaker separation," in *Proc. ICASSP'05*, Philadelphia, PA, USA, 2005, pp. 17–20.

[30] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," in *Proc. ICASSP'88*, New York, NY, USA, 1988, pp. 517–520.

[31] B. J. Frey, L. Deng, A. Acero, and T. T. Kristjansson, "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. Eurospeech'01*, Aalborg, Denmark, 2001, pp. 901–904.

[32] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[33] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modelling for speech separation," *Comput. Speech Lang.*, vol. 24, pp. 67–76, 2010, Special Iss. Speech Separat. and Recognit..

[34] J. Ming, "Maximizing the continuity in segmentation—A new approach to model, segment and recognize speech," in *Proc. ICASSP'09*, Taipei, Taiwan, 2009, pp. 3849–3852.

[35] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 822–836, May 2011.

[36] A. Jafari, R. Srinivasan, D. Crookes, and J. Ming, "A longest matching segment approach for text-independent speaker recognition," in *Proc. Interspeech'10*, Makuhari, Japan, 2010, pp. 1469–1472.

[37] J. Ming and F. J. Smith, "Speech recognition with unknown partial feature corruption—A review of the union model," *Comput. Speech Lang.*, vol. 17, pp. 287–305, 2003.

[38] X. Sun and Y. Zhao, "Intergrate template matching and statistical modeling for speech recognition," in *Proc. Interspeech'10*, Makuhari, Japan, 2010, pp. 74–77.

[39] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. 5th DARPA Speech and Natural Lang. Workshop*, 1992, pp. 357–362.

[40] P. Loizou, *Speech Enhancement: Theory and Practice*. New York, NY, USA: CRC, Taylor & Francis, 2007.

[41] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Cavendish Laboratory, Tech. Rep., 2006.

[42] "Methods for subjective determination of transmission quality," Geneva, Switzerland, ITU-T Rec. P.800, Aug. 1996.

[43] C. Bowen, *Children's Speech Sound Disorders*. Oxford, U.K.: Wiley-Blackwell, 2009.

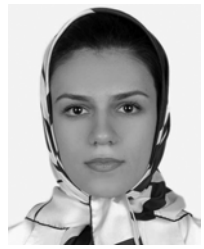
[44] B. N. Gover and J. S. Bradley, "Comparison of subjective and objective ratings of intelligibility of speech recordings," in *Proc. Canadian Acoust. Assoc. Conf.*, Montreal, QC, Canada, 2007, pp. 1–2.

Ji Ming (M'97) received the B.Sc. degree from Sichuan University, Chengdu, China, in 1982, the M.Phil. degree from Changsha Institute of Technology, Changsha, China, in 1985, and the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 1988, all in electronic engineering.

He was Associate Professor with the Department of Electronic Engineering, Changsha Institute of Technology, from 1990 to 1993. Since 1993, he has been with the Queen's University Belfast, Belfast, U.K., where he is currently a Professor in the School of Electronics, Electrical Engineering and Computer Science. From 2005 to 2006, he was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge. His research interests include speech and language processing, image processing, signal processing and pattern recognition.

Ramji Srinivasan (M'08) received the B.E. degree in electrical and electronics engineering from Madurai Kamaraj University, India, the M.Tech. degree in process control and instrumentation from Regional Engineering College, Trichy, India, and the Ph.D. degree in electrical engineering from Anna University, Chennai, India, in 2000, 2002, and 2008, respectively. He started his research career with the Fluid Control Research Institute, India, as a Post Graduate Research Fellow in 2002 and later joined the National Institute of Ocean Technology, Chennai, as a Scientist and worked there from 2003 to 2008. He was a Research Fellow in the Institute of Electronics, Communications, and Information Technology, Queen's University Belfast, Belfast, U.K., from 2009 to 2012. He is currently working as a Senior Research Engineer in Advanced Audio Research group at Cambridge Silicon Radio limited. His research interests include speech, audio and acoustic signal processing; instrumentation system design and integration.

Danny Crookes (SM'12) was appointed to the Chair of Computer Engineering in 1993 at Queen's University Belfast, Belfast, U.K., and was Head of Computer Science from 1993–2002. He is currently Director of Research for Speech, Image and Vision Systems at the Institute of Electronics, Communications and Information Technology, Queen's University Belfast. His current research interests include the use of novel architectures (especially GPUs) for high performance speech and image processing. Professor Crookes is currently involved in projects in automatic shoeprint recognition, speech separation and enhancement, and processing of 4D confocal microscopy imagery. Professor Crookes has over 200 scientific papers in journals and international conferences.



Ayeh Jafari received the B.Sc. degree in Electrical-Electronics Engineering from Azad University, Iran, in 2003, and the M.Sc. degree (distinction) in Telecommunication and Information Systems from the University of Essex, U.K., in 2008, and the Ph.D. degree in Computer Science from Queen's University Belfast, Belfast, U.K., in 2011. In 2008, she received the Telecom Prize for her performance in the M.Sc. project. She is now working in Andor Technology on a scientific software package for managing microscopy experiments and performing image analysis. Her research interests are in the fields of networks and signal processing. Her Ph.D. work was focused on the topics concerning robust speaker recognition, speaker clustering and speech enhancement and separation.