Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 24 (2010) 16–29
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Speech separation using speaker-adapted eigenvoice speech models

Ron J. Weiss, Daniel P.W. Ellis *

LabROSA, Department of Electrical Engineering, Columbia University, 500 West 120th Street, Room 1300, Mailcode 4712, New York, NY 10027, United States

Received 30 June 2007; received in revised form 18 February 2008; accepted 3 March 2008
Available online 15 March 2008

Abstract

We present a system for model-based source separation for use on single channel speech mixtures where the precise source characteristics are not known *a priori*. The sources are modeled using hidden Markov models (HMM) and separated using factorial HMM methods. Without prior speaker models for the sources in the mixture it is difficult to exactly resolve the individual sources because there is no way to determine which state corresponds to which source at any point in time. This is solved to a small extent by the temporal constraints provided by the Markov models, but permutations between sources remains a significant problem. We overcome this by adapting the models to match the sources in the mixture. We do this by representing the space of speaker variation with a parametric signal model-based on the eigenvoice technique for rapid speaker adaptation. We present an algorithm to infer the characteristics of the sources present in a mixture, allowing for significantly improved separation performance over that obtained using unadapted source models. The algorithm is evaluated on the task defined in the 2006 Speech Separation Challenge [Cooke, M.P., Lee, T.-W., 2008. The 2006 Speech Separation Challenge. Computer Speech and Language] and compared with separation using source-dependent models. Although performance is not as good as with speaker-dependent models, we show that the system based on model adaptation is able to generalize better to held out speakers.

Keywords: Source separation; Model adaptation; Eigenvoice

1. Introduction

Recognition of signals containing contributions from multiple sources continues to pose a significant problem for automatic speech recognition as well as for human listeners. One solution to this problem is to separate the mixed signal into its constituent sources and then recognize each one separately. This approach is especially difficult when only a single channel input is available, making it impossible to utilize spatial constraints to separate the signals. Instead, most approaches to monaural source separation rely on prior knowledge

* Corresponding author. Tel.: +1 212 854 8928; fax: +1 212 932 9421.
E-mail addresses: ronw@ee.columbia.edu (R.J. Weiss), dpwe@ee.columbia.edu (D.P.W. Ellis).

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.03.003

about the nature of the sources present in the mixture to constrain the possible source reconstructions. Because natural audio sources tend to be sparsely distributed in time-frequency, a monaural mixture can be segregated simply by segmenting its spectrogram into regions dominated by each source. This can be done using perceptual cues as in systems based on computational auditory scene analysis (CASA) such as Srinivasan et al. (2006). Alternatively, given statistical models of the source characteristics for each source in the mixture, the signals can be reconstructed by performing a factorial search through all possible model combinations (Varga and Moore, 1990; Roweis, 2000; Kristjansson et al., 2004).

The 2006 Speech Separation Challenge (Cooke and Lee, 2008) was an organized effort to evaluate different approaches to monaural speech separation. The task was to recognize what was said by a target speaker in instantaneous two-talker mixtures composed of utterances from 34 different speakers. The entries included some based on computational auditory scene analysis (e.g. Srinivasan et al., 2006) as well as some based on model-based separation (e.g. Kristjansson et al., 2006). In general, model-based systems outperformed those based on CASA type heuristics. Like most previous work in this area, such as Kristjansson et al. (2004), these systems modeled each speaker using hidden Markov models (HMMs) for separation. The best performing systems incorporated a lot of task-specific knowledge into these models. The Iroquois system (Kristjansson et al., 2006) incorporated knowledge of the task grammar to constrain the reconstructed sources, and was able to outperform human listeners under some conditions. Instead of reconstructing each source prior to recognition, other model-based systems (Virtanen, 2006; Barker et al., 2006) attempted to recognize the two simultaneous utterances directly from the mixture.

We focus on the model-based approach to source separation when the precise source characteristics are not known *a priori* (Weiss and Ellis, 2007). Ozerov et al. (2005) propose the idea of beginning with a source-independent model and adapting it to the target source for monaural singing voice separation. This approach can separate previously unseen sources far better than using unadapted models, but requires a substantial amount of adaptation data. In this work we consider adaptation when there is much less data available, requiring a more constrained model space. The remainder of this paper is organized as follows: Section 2 reviews the source models used in our system. The technique for model adaptation is described in Section 3. Section 4 describes the detailed separation algorithm. Experimental results are reported in Section 5. Finally, Sections 6 and 7 conclude the paper with a discussion of the shortcomings of the algorithm and directions for future research.

2. Source models

An important observation for efficient model-based source separation is that audio sources tend to be sparsely distributed in time and frequency. Given the short-time Fourier transform magnitude of a mixture of two speech signals, it is empirically observed that over 80% of the time-frequency cells lie within 3 dB of the larger of the corresponding cells of the two constituent clean signals (Ellis, 2006). Selecting a representation that exploits this property allows for efficient inference because the full range of combinations of source model states need not be considered when their overlap is minimal (Roweis, 2003). We therefore follow the example of the Iroquois system and model the log-spectrum of each source derived from a short-time Fourier transform with 40 ms window and 10 ms hop.

As shown in Kristjansson et al. (2006), incorporating temporal dynamics in source models can significantly improve separation performance. This is especially true when all sources in a speech mixture use the same model, in which case separation depends on knowledge of the task grammar. We, however, are interested in creating a more generic speech model that is not specific to a given grammar, so we follow the "phonetic vocoder" approach (Picone and Doddington, 1989), which models temporal dynamics only within each phone. This is similar to the approach taken in Schmidt and Olsson (2006) which utilizes phone models trained using non-negative matrix factorization.

The log power spectrum of each source is modeled using a hidden Markov model (HMM) with Gaussian mixture model (GMM) emissions. Each of the 35 phones used in the task grammar are modeled using a standard 3-state forward HMM topology. Each state's emissions are modeled by a GMM with eight mixture components. The transitions from each phone to all others have equal probability, which was found to work as well as more phonotactically-informed values. This structure allows us to incorporate some knowledge of speech dynamics without being specific to any grammar.

We used the HTK toolkit (Young et al., 2006) to train the models on the Speech Separation Challenge training data (Cooke and Lee, 2008), downsampled to 16 kHz and pre-emphasized as in the Iroquois system. The training data for all 34 speakers was used to train a speaker-independent (SI) model. We also constructed speaker-dependent (SD) models for each speaker by bootstrapping from the SI model to ensure that each mixture component of the SD models corresponded directly to the same component in the SI model. The consistent ordering across all speaker models is needed for the speaker adaptation process described in the next section.

Only the GMM means were updated during the SD training process, so the likelihood of a given frame of signal $\mathbf{x}(t)$ under component c of state s of the model for speaker i can be written as

$$P(\mathbf{x}(t)|s,c,\boldsymbol{\mu}_i) = \mathcal{N}(\mathbf{x}(t);\boldsymbol{\mu}_{i,s,c},\Sigma_{s,c}) \tag{1}$$

where $\boldsymbol{\mu}_{i,s,c}$ denotes the mean for component c of state s in the model for speaker i, and $\Sigma_{s,c}$ denotes the corresponding covariance matrix from the speaker-independent model. Note that all models described in this paper use diagonal covariances, so for convenience we use the notation $\boldsymbol{\sigma}_{s,c}$ to denote the diagonal components of $\Sigma_{s,c}$.

3. Model adaptation

Because only a single utterance is available for model adaptation, there is insufficient data to use standard adaptation methods such as MLLR (Leggetter and Woodland, 1995). We solve this problem by using the SD models described above as *priors* on the space of speaker variation. Adapting to the observed source involves projecting the source onto the space spanned by these priors. This is done by orthogonalizing the SD models using principal component analysis (PCA), which allows each point in the space spanned by the different speakers to be represented as a point in a low dimensional "eigenvoice" space (Kuhn et al., 2000).

Only the model means are adapted. The mean vectors of each component of each state in the SD model for speaker i are concatenated into a mean supervector $\boldsymbol{\mu}_i$. These supervectors are constructed for all M speaker models and used to construct a matrix $U = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_M]$ that spans the space of speaker variation. Performing PCA on U yields orthonormal basis vectors for the eigenvoice space, $\hat{\boldsymbol{\mu}}_j$.

Although the ordering of states and mixture components in the supervectors is arbitrary, care must be taken to ensure that the ordering is consistent across all speakers. Because we are using GMM emissions, further complications are possible if there is no correspondence of mixture components across speaker models. This is possible if the speaker models are trained independently using different initializations or a process such as mixture splitting. The training procedure described in Section 2 is used to enforce a one-to-one mapping between mixture components in all speaker models and avoid such problems.

In addition to eigenvoice adaptation, we extend the model of Kuhn et al. (2000) to include a gain parameter tied across all states to account for any mismatch between signal level in the training and testing data. This compensation is especially important because the speech signals in the SSC dataset are mixed over a wide range of signal-to-noise ratios.

Using the new basis, a speaker-adapted model can be parametrized simply by a set of N eigenvoice weights, $\mathbf{w}$, and the gain parameter g. The mean for component c of state s of a speaker-adapted model can be written as a linear combination of these bases:

$$\boldsymbol{\mu}_{s,c}(\mathbf{w},g) = \sum_{j=1}^{N} w_j \hat{\boldsymbol{\mu}}_{j,s,c} + \bar{\boldsymbol{\mu}}_{s,c} + g \tag{2}$$

where w_j is the weight applied to the jth eigenvoice, $\hat{\boldsymbol{\mu}}_{j,s,c}$, and $\bar{\boldsymbol{\mu}}_{s,c}$ is the average across all SD models of the mean for component c of state s (i.e. the mean voice). We describe the process for inferring the adaptation parameters $\mathbf{w}$ and g for a particular speaker in Section 4.4.

In the experimental results reported in this paper we do not discard any low variance eigenvoice dimensions (i.e. $N = M - 1$) so it is possible to exactly reconstruct the original model for speaker i given the corresponding weights $\mathbf{w}_i$, i.e. $\boldsymbol{\mu}_i = \boldsymbol{\mu}(\mathbf{w}_i, 0)$. However, in practice the adapted models never match the original speaker models perfectly because the adaptation is based only on the relatively small set of model states found in a single utterance.

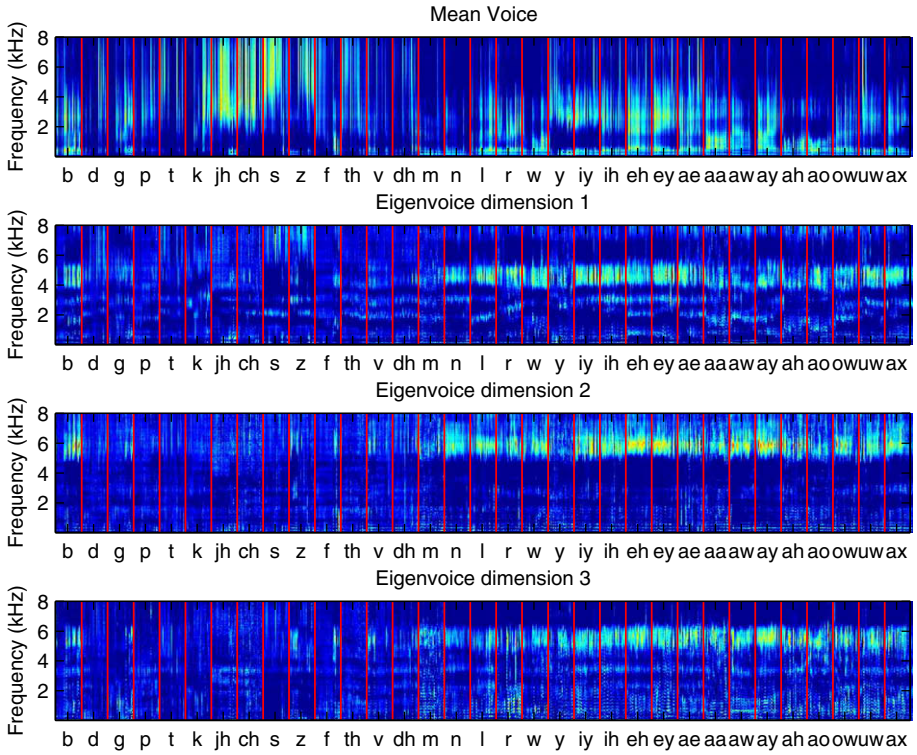

Fig. 1. Eigenvoice speech model. The top panel shows the mean voice $\bar{\boldsymbol{\mu}}$ which closely resembles the speaker-independent model. The remaining panels show the three eigenvoices with the largest variance, $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$, and $\hat{\boldsymbol{\mu}}_3$ respectively.

Fig. 1 shows the mean voice as well as the three eigenvoices with the highest variance learned from the training data. The mean voice is very similar to the speaker-independent model and very coarsely models the overall spectral shape characteristic of different phones. Successive eigenvoices are used to add additional high resolution detail to this model. Eigenvoice 1, $\hat{\boldsymbol{\mu}}_1$, emphasizes formant resonances that are characteristic of female speakers. In fact, as shown in Fig. 3, the corresponding eigenvoice weight is perfectly correlated with gender; female speakers have positive w_1 and male speakers have negative w_1. Eigenvoice 2 emphasizes different formants in consonant states and introduces some fundamental frequency information and high frequency resonances into the vowel states. Finally, $\hat{\boldsymbol{\mu}}_3$ incorporates additional pitch trajectory detail into voiced phones.

4. Separation algorithm

As in Kristjansson et al. (2006), we perform separation by finding the Viterbi path through a factorial HMM composed of models for each source. This is complicated by the fact that we do not have prior knowledge of the speaker models. Instead we must use the same speaker-independent model for both sources. However, separation performance is quite poor because the speech model does not enforce strong temporal constraints. This is due to ambiguity in the Viterbi path through a factorial HMM composed of identical models (Ellis, 2006). The state sequences can permute between sources whenever the Viterbi path passes through the same state in both models at the same time. Since our models only include basic phonetic constraints, the resulting separated signals can permute between sources whenever the two sources have (nearly) synchronous phone transitions. Fig. 2 shows an example of such permutations where each source is modeled using the same speaker-independent model.

This permutation problem can be solved using models matched to each source, but the adaptation procedure described in Section 4.4 requires clean source signals. Instead we use the following iterative algorithm to solve the problem of estimating the eigenvoice parameters for the two sources directly from the mixture:

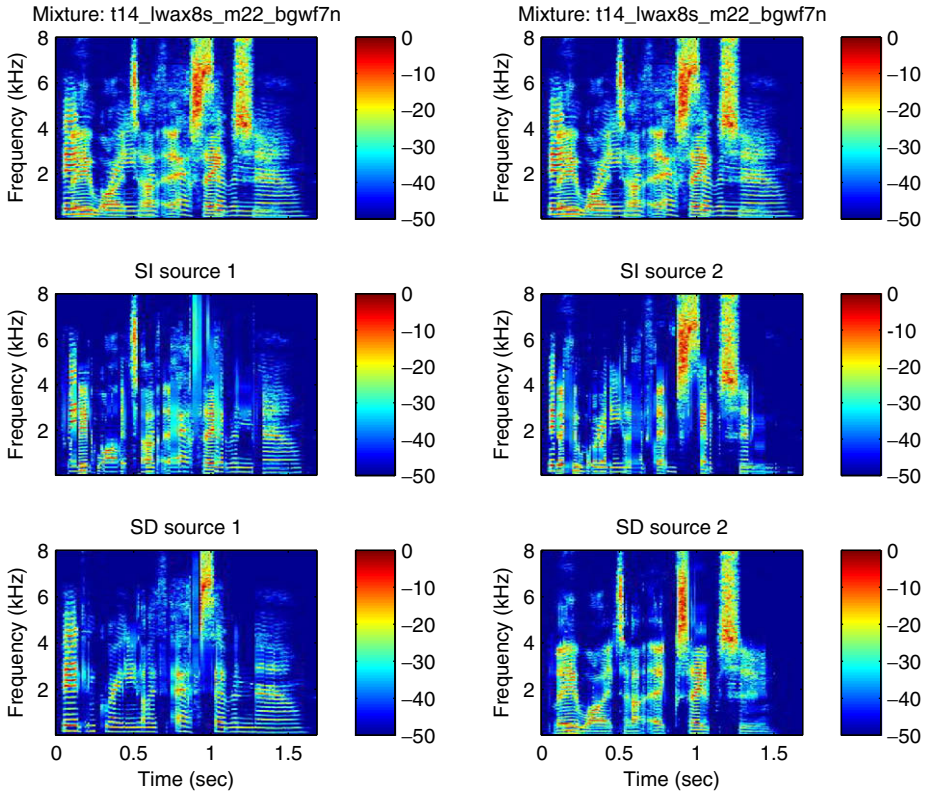

Fig. 2. Phone permutations found in the Viterbi path through a factorial HMM that models the mixture of two sources. Each source is modeled with the same speaker-independent (SI) models. Neither of the resulting source reconstructions (middle row) is a good match for either of the sources found by separation using speaker-dependent models (bottom row). Each SI reconstruction contains short regions from both of the SD sources. For example, the final phone in SI source 1 is quite close to that of the corresponding SD source, but the first half second of the signal is closer to SD source 2.

1. Obtain initial model estimates for each source.
2. Separate signals using factorial HMM decoding.
3. Reconstruct each source.
4. Update model parameters.
5. Repeat 2–4 until convergence.

The intuition behind the iterative approach is that each of the reconstructed source estimates will resemble one source more closely than the other (i.e. more than half of it will match one of the sources) even if the initial separation is quite poor. As a result, the model parameters inferred from these estimates will also be a better match to one source than to the other. This in turn should improve the separation in the next iteration.

Initially the dynamic constraints in the model partially make up for the lack of source-dependent feature constraints. But the reconstructions are still quite prone to permutations between sources. The permutations tend to get corrected as the algorithm iterates because the adaptation allows the models to better approximate the speaker characteristics. Unfortunately the algorithm is slow to converge, so this process takes many iterations.

4.1. Initialization

As with many iterative algorithms, this method is vulnerable to becoming stuck in local optima. Good initialization is crucial to finding good solutions quickly. We begin by projecting the mixed signal onto the

eigenvoice bases to set the parameters for both sources (see Section 4.4). Obviously these parameters will not be a good match to either isolated source and, as described earlier, using the same model for both sources will lead to poor performance. So further steps are taken to differentiate the two speakers.

We use the speaker identification component of the Iroquois speech separation system (Kristjansson et al., 2006) which chooses the most likely speaker model based on frames of the mixture that are dominated by a single source. This could be used directly to search through a set of adaptation parameter vectors corresponding to the speakers in the training set, in which case our system reduces to a variant of Iroquois. However this may not work well on sources that are not in the training set. Instead we note that by design the eigenvoice dimensions are decorrelated which allows us to treat each of them independently. The idea is to build an approximation of $\mathbf{w}$ for each source from the bottom up, beginning from w_1 and adding consecutive weights.

During training we learn prototype settings for each weight w_j by coarsely quantizing the corresponding weights of the training speakers to three quantization levels using the Lloyd–Max algorithm (Lloyd, 1982). This allows w_j for any given speaker to be approximated by one of the quantized values $\{\hat{w}_j^1, \hat{w}_j^2, \hat{w}_j^3\}$. The first two panels of Fig. 3 show example quantization levels for w_1 and w_2.

Given the mixed signal, we can approximate the eigenvoice adaptation parameters for each speaker, $\mathbf{w}_1$ and $\mathbf{w}_2$, using the following bottom up construction:

Initialize $\mathbf{w}_1$ and $\mathbf{w}_2$ to zero, i.e. set $\boldsymbol{\mu}(\mathbf{w}_1)$ and $\boldsymbol{\mu}(\mathbf{w}_2)$ to $\bar{\boldsymbol{\mu}}$.

For each speaker i and eigenvoice dimension j:

> Use the quantized weights to construct prototype models $\{\boldsymbol{\mu}^k(\mathbf{w}_i)\}_{1\leqslant k\leqslant 3}$ where $\boldsymbol{\mu}^k(\mathbf{w}_i) = \boldsymbol{\mu}(\mathbf{w}_i) + \hat{w}_j^k \hat{\boldsymbol{\mu}}_j$.
> Use the Iroquois speaker identification algorithm to select the most likely prototype model given the mixed signal and update $\mathbf{w}_i$ and $\boldsymbol{\mu}(\mathbf{w}_i)$ accordingly.

An example of this process is shown in Fig. 3. The first and second panels show the quantization levels of eigenvoice dimensions 1 and 2 respectively. The shaded regions show the prototypes chosen for speaker 1 (dark gray) and speaker 2 (light gray). Finally, the rightmost panel shows the joint selection of w_1 and w_2 for both speakers.

This is only done for the 3 eigenvoice dimensions with the highest variance. The remaining parameters are the same for both sources, set to match the mixture. This technique is not very accurate, but in most cases it suffices to differentiate the two sources. It works best at differentiating between male and female speakers because the eigenvoice dimension with the most variance is highly correlated with speaker gender.

4.2. Factorial HMM decoding

The mixed signal is modeled by a factorial HMM constructed from the two source models as in Varga and Moore (1990) and Roweis (2000). Each frame of the mixed signal $\mathbf{y}(t)$ is modeled by the combination of one state from each source model.

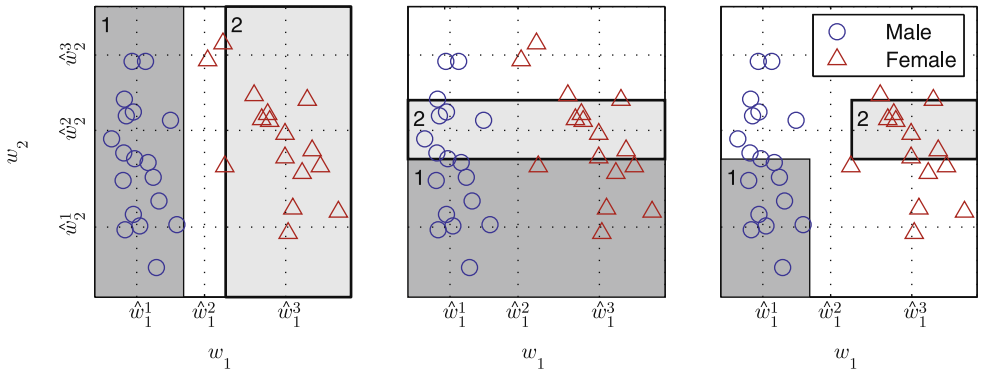

Fig. 3. Quantization of eigenvoice coefficients w_1 and w_2 across all training speakers.

The joint likelihood of each state combination is derived using the "max" approximation (Roweis, 2003) which relies on the sparsity of audio sources in the short-time Fourier transform representation. Assuming that each time-frequency cell of an audio mixture will be dominated by a single source, and that the GMM emissions use Gaussians with diagonal covariance, the joint likelihood of the mixed signal at time t can be computed as follows (Varga and Moore, 1990):

$$P(\mathbf{y}(t)|s_1,c_1,s_2,c_2) = \prod_d P(y^d(t)|s_1,c_1,s_2,c_2) \tag{3}$$

$$\begin{aligned} P(y^d(t)|s_1,c_1,s_2,c_2) = {} & \mathcal{N}(y^d(t);\mu^d_{1,s_1,c_1},\sigma^d_{s_1,c_1})\mathcal{C}(y^d(t);\mu^d_{2,s_2,c_2},\sigma^d_{s_2,c_2}) \\ & + \mathcal{N}(y^d(t);\mu^d_{2,s_2,c_2},\sigma^d_{s_2,c_2})\mathcal{C}(y^d(t);\mu^d_{1,s_1,c_1},\sigma^d_{s_1,c_1}) \end{aligned} \tag{4}$$

where $\mathcal{C}$ denotes the Gaussian cumulative distribution function.

Because the sources rarely overlap, it is possible to approximate the likelihood in a form that is significantly cheaper computationally:

$$P(\mathbf{y}(t)|s_1,c_1,s_2,c_2) \approx \mathcal{N}(\mathbf{y}(t);\max(\boldsymbol{\mu}_{1,s_1,c_1},\boldsymbol{\mu}_{2,s_2,c_2}),\boldsymbol{\sigma}) \tag{5}$$

where $\boldsymbol{\sigma} = \boldsymbol{\sigma}_{s_1,c_1}$ for dimensions where $\boldsymbol{\mu}_{1,s_1,c_1} > \boldsymbol{\mu}_{2,s_2,c_2}$ (i.e. where source 1 dominates the mixture) and $\boldsymbol{\sigma} = \boldsymbol{\sigma}_{s_2,c_2}$ otherwise. Eq. (5) is not accurate when μ^d_{1,s_1,c_1} is close to μ^d_{2,s_2,c_2} or when $\sigma^d_{s_1,c_1}$ and $\sigma^d_{s_2,c_2}$ are very different, but in practice we have found it to work well.

The sources are separated by finding the maximum likelihood path through this factorial HMM using the Viterbi algorithm. This process is quite slow since it involves searching through every possible state combination at each frame of the signal. To speed it up we prune the number of active state and component combinations at each frame to the 200 most likely.

4.3. MMSE source reconstruction

Model updates are performed on estimates of the spectral frames of each speaker. These are found using the minimum square error estimate: $\hat{\mathbf{x}}_1(t) = E[\mathbf{x}_1(t)|s_1,c_1,s_2,c_2,\mathbf{y}(t)]$ where (s_1,c_1) and (s_2,c_2) correspond to the active state and component combinations at time t in the Viterbi path. Each dimension d of the conditional mean is found using the same approximation as Eq. (5):

$$E[x^d_1(t)|s_1,c_1,s_2,c_2,y^d(t)] \approx \begin{cases} \mu^d_{1,s_1,c_1}, & \text{if } \mu^d_{1,s_1,c_1} < \mu^d_{2,s_2,c_2} \\ y^d(t), & \text{otherwise} \end{cases} \tag{6}$$

The estimate for $\hat{\mathbf{x}}_2(t)$ follows the same derivation. Because the factorial model of the mixture assumes that there is little overlap between the source signals, Eq. (6) simply assigns the observed frequency bin to the dominant source and uses the model mean wherever the source is masked.

4.4. Eigenvoice parameter inference

Finally, the speaker models are updated to better match the source estimates $\hat{\mathbf{x}}_1(t)$ and $\hat{\mathbf{x}}_2(t)$. The model parameters w_j and g can be estimated iteratively using an extension of the maximum likelihood eigen-decomposition (MLED) expectation maximization algorithm described in Kuhn et al. (2000). The derivation follows that of the MLLR transform described in Leggetter and Woodland (1995).

The E-step of the algorithm involves computing $\gamma_{s,c}(t)$, the posterior probability of the source occupying component c of state s at time t given the observations $\mathbf{x}_i(1\cdots t)$ and the model using the HMM forward-backward algorithm (Rabiner, 1989).

The M-step maximizes the likelihood of the observed sequence $\hat{\mathbf{x}}_i$ under the model. As in Kuhn et al. (2000) and Leggetter and Woodland (1995), this is done by maximizing the auxiliary function $\mathcal{L}(\mathbf{w},g)$:

$$\mathcal{L}(\mathbf{w},g) = -\sum_t \sum_s \sum_c \gamma_{s,c}(t)(\hat{\mathbf{x}}_i(t) - \boldsymbol{\mu}_{s,c}(\mathbf{w},g))^{\mathrm{T}} \Sigma^{-1}_{s,c}(\hat{\mathbf{x}}_i(t) - \boldsymbol{\mu}_{s,c}(\mathbf{w},g)) \tag{7}$$

The solution that maximizes (7) is found by solving the following set of simultaneous equations for w_j and g:

$$\sum_{t,s,c} \gamma_{s,c}(t)\hat{\boldsymbol{\mu}}_{j,s,c}^{\mathrm{T}}\Sigma_{s,c}^{-1}(\hat{\mathbf{x}}_i(t) - \bar{\mu}_{s,c} - g) = \sum_{t,s,c} \gamma_{s,c}(t) w_j \hat{\boldsymbol{\mu}}_{j,s,c}^{\mathrm{T}}\Sigma_{s,c}^{-1} \sum_k w_k \hat{\boldsymbol{\mu}}_{k,s,c} \tag{8}$$

$$\sum_{t,s,c} \gamma_{s,c}(t)\mathbf{1}^{\mathrm{T}}\Sigma_{s,c}^{-1}(\hat{\mathbf{x}}_i(t) - \bar{\mu}_{s,c} - g) = \sum_{t,s,c} \gamma_{s,c}(t)\mathbf{1}^{\mathrm{T}}\Sigma_{s,c}^{-1} \sum_j w_j \hat{\boldsymbol{\mu}}_{j,s,c} \tag{9}$$

where $\mathbf{1}$ is a vector of ones.

This solution can be written as a matrix inversion as follows:

$$\begin{bmatrix} \mathbf{w} \\ g \end{bmatrix} = \begin{bmatrix} A & \mathbf{b} \\ \mathbf{b}^{\mathrm{T}} & c \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{d} \\ e \end{bmatrix} \tag{10}$$

where

$$A_{j,k} = \sum_{t,s,c} \gamma_{s,c}(t)\hat{\boldsymbol{\mu}}_{j,s,c}^{\mathrm{T}}\Sigma_{s,c}^{-1}\hat{\boldsymbol{\mu}}_{k,s,c} \tag{11}$$

$$b_j = \sum_{t,s,c} \gamma_{s,c}(t)\hat{\boldsymbol{\mu}}_{j,s,c}^{\mathrm{T}}\Sigma_{s,c}^{-1}\mathbf{1} \tag{12}$$

$$c = \sum_{t,s,c} \gamma_{s,c}(t)\mathbf{1}^{\mathrm{T}}\Sigma_{s,c}^{-1}\mathbf{1} \tag{13}$$

$$d_j = \sum_{t,s,c} \gamma_{s,c}(t)\hat{\boldsymbol{\mu}}_{j,s,c}^{\mathrm{T}}\Sigma_{s,c}^{-1}(\hat{\mathbf{x}}_i(t) - \bar{\boldsymbol{\mu}}_{s,c}) \tag{14}$$

$$e = \sum_{t,s,c} \gamma_{s,c}(t)\mathbf{1}^{\mathrm{T}}\Sigma_{s,c}^{-1}(\hat{\mathbf{x}}_i(t) - \bar{\boldsymbol{\mu}}_{s,c}) \tag{15}$$

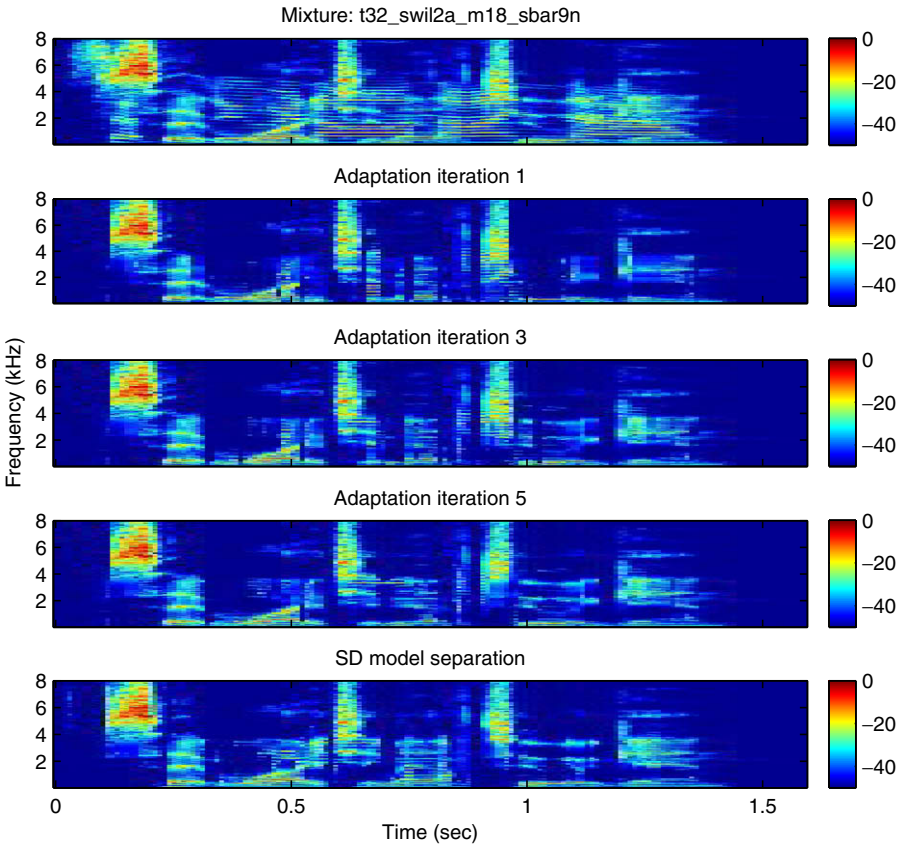

Fig. 4. Separation using speaker-adapted models. The top plot shows the spectrogram of a mixture of female and male speakers. The middle three show the reconstructed target signal ("set white in l 2 again") from the adapted models after iterations 1, 3, and 5. The bottom plot shows the result of separation using the speaker-dependent model for target speaker.

The EM algorithm is applied to each source estimate $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ to infer $\mathbf{w}$ and g for each source. For increased efficiency, we do not use the dynamics of the HMMs for the E-step computation (i.e. the models are reduced to GMMs). The updated source models are then used for the next iteration.

Fig. 4 gives an example of the source separation process using model adaptation. The initial separation does a reasonable job at isolating the target, but it make some errors. For example, the phone at $t = 1$ s is initially mostly attributed to the masking source. The reconstruction improves with subsequent iterations, getting quite close to the reconstruction based on SD models (bottom panel) by the fifth iteration.

5. Experiments

The system was evaluated on the test data from the 2006 Speech Separation Challenge (Cooke and Lee, 2008). This data set is composed of 600 artificial speech mixtures composed of utterances from 34 different speakers, each mixed at signal to interference ratios varying from −9 dB to 6 dB. Each utterance follows the pattern *command color preposition letter digit adverb*. The task is to determine the letter and digit spoken by the source whose color is "white".

The separation algorithm described above was run for fifteen iterations using eigenvoice speech models trained on all 34 speakers in the data set. All 33 eigenvoice dimensions were used for adaptation. The time-domain sources were reconstructed from the STFT magnitude estimates $\hat{\mathbf{x}}_i$ and the phase of the mixed signal. Sound examples of reconstructed sources are available at http://www.ee.columbia.edu/ ronw/SSC.html. The two reconstructed signals were then passed to a speech recognizer; assuming one transcription contains "white", it was taken as the target source. We used the default HTK speech recognizer provided by the challenge organizers (Cooke and Lee, 2008), retrained on 16 kHz data. The acoustic model consisted of whole-word HMMs based on MFCC, delta, and acceleration features. Performance is measured using word accuracy of the letter and digit spoken by the target speaker.

5.1. Results

Fig. 5 compares the performance of the speaker adaptation (SA) system to two comparison systems based on SD and SI models respectively. The SD system identifies the most likely pair of speakers present in the mixture by searching the set of SD models using the Iroquois speaker identification and gain adaptation technique (Kristjansson et al., 2006). The sources are separated by finding the maximum likelihood path through the factorial HMM composed of those two source models. We also compare this to performance when using oracle knowledge of the speaker identities and gains. Finally, we include baseline performance of the recognizer generating a single transcript of the original mixed signal. The overall performance of the SA and SD systems are also listed in Tables 1 and 2 respectively.

The performance of the SI system is not sensitive to the different speaker conditions because the same model is used for both sources. The other separation systems work best on mixtures of different genders because of the prominent differences between male and female vocal characteristics, which mean that such sources tend to have less overlap. Conversely, the performance on the same talker task is quite poor. This is because the models used for each source are identical (or close to it in the SA case) except for the gain term, and the models enforce only limited dynamic constraints. The performance of the SI system is quite poor in all conditions for the same reason. The marked difference between the SA and SI systems demonstrates that adapting the source models to match the source characteristics can do a lot to make up for the limited modeling of temporal dynamics.

Looking at general trends, we see that the SD models perform similarly whether using oracle or Iroquois-style speaker information. Both of these are significantly better than the SA system, itself better than the SI system and baseline. The reduced performance of the SA system in this task is mainly due to its vulnerability to permutations between sources, which reflects the sensitivity of the initial separation to initialization. The adaptation process is able to compensate for limited permutations, as in the final second in Fig. 4. However when the initialization does not sufficiently separate the sources, the system can get stuck in poor local optima where each of the estimated sources is only a partial match to the ground truth. In contrast, it performs significantly better on the different gender condition because the initial separation tends to be better. The errors

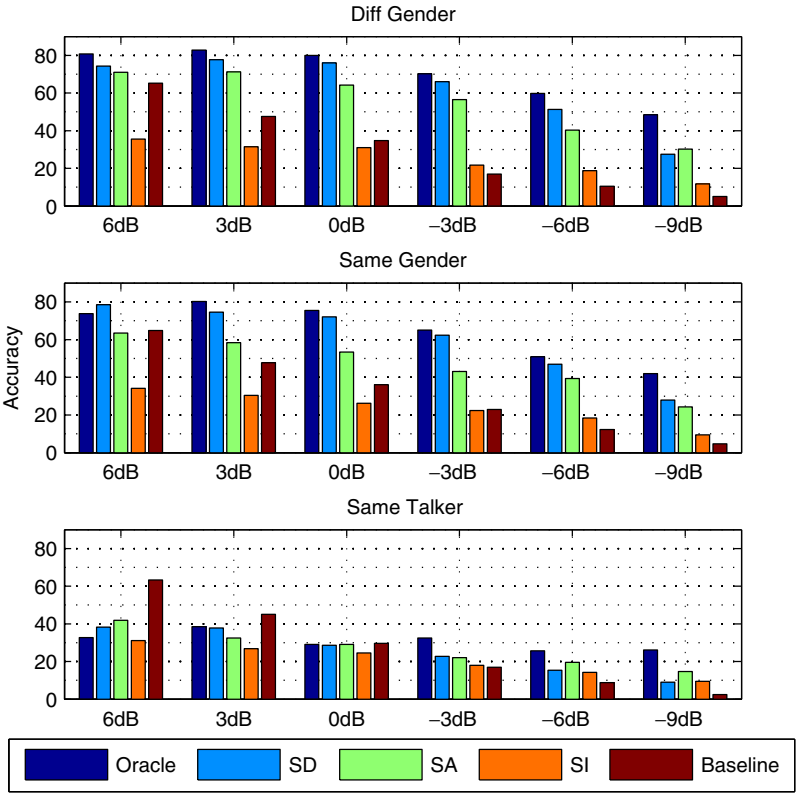

Fig. 5. Separation performance using speaker-dependent (SD), speaker-adapted (SA), and speaker-independent (SI) models. The target source is identified by choosing the source that contains the word "white". Also shown is performance of separation using oracle knowledge of speaker identities and relative gains (Oracle) and baseline performance of the recognizer on the mixed signal (Baseline).

Table 1
Recognition accuracy on the 2006 Speech Separation Challenge data test set using our source-adapted separation system

SNR (dB)	Same talker (%)	Same gender (%)	Different gender (%)	Average (%)
6	41.89	63.41	71.00	57.99
3	32.43	58.38	71.25	53.08
0	29.05	53.35	64.25	48.00
−3	22.07	43.02	56.50	39.77
−6	19.59	39.39	40.25	32.36
−9	14.64	24.30	30.25	22.71

Table 2
Recognition accuracy on the 2006 Speech Separation Challenge data test set using speaker-dependent models and Iroquois speaker identification on the 2006 Speech Separation Challenge test set

SNR (dB)	Same talker (%)	Same gender (%)	Different gender (%)	Average (%)
6	38.29	78.49	74.25	62.23
3	37.84	74.58	77.75	62.06
0	28.60	72.07	76.00	57.32
−3	22.75	62.29	66.00	48.92
−6	15.32	46.93	51.25	36.69
−9	9.01	27.93	27.50	20.80

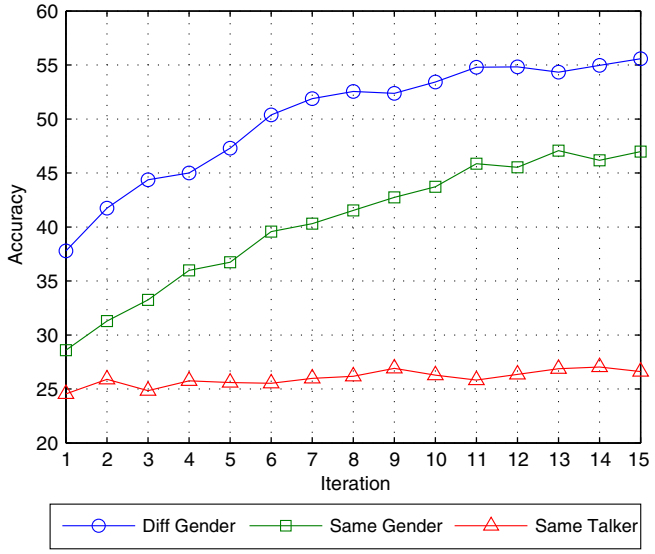

Fig. 6. Separation performance improvement averaged over all SNRs as the source adaptation/separation algorithm iterates.

get worse as SNR decreases because the stage of initialization that adapts to the mixed signal favors the louder source. Fortunately the algorithm is generally able to compensate for poor initialization as it iterates, however as shown in Fig. 6 this process is slow.

Fig. 6 shows how the word accuracy improves after each iteration averaged across all SNRs. It is quite clear from this figure that the iterative algorithm helps significantly, increasing the average accuracy by about 18% in both the same gender and different gender conditions after 15 iterations. The performance improvement for the same talker condition is more modest.

5.2. Held out speakers

Source separation systems based on speaker-dependent models are potentially at a disadvantage when presented with mixtures containing sources that are not represented in the SD model set. We expect that the SA system should be better suited to handling such cases. To evaluate this hypothesis we separated the SSC training data into random subsets of 10, 20, and 30 speakers and trained new eigenvoice models from each subset. All eigenvoice dimensions were retained for these models, e.g. the model trained from 10 speaker subset used 9 eigenvoice dimensions for adaptation, etc. A new test set was generated from utterances from the four speakers held out of all training subsets. The held out speakers were evenly split by gender. The test set consists of 400 mixtures at 0 dB SNR, broken up into 200 same gender mixtures and 200 different gender mixtures.

Fig. 7 compares the performance of the SD system and SA system on this data set. The SD models used were limited to the same speakers as were used to train the new eigenvoice models. Performance using smaller subsets of training speakers is compared to a baseline containing all 34 speakers from the training set. It is important to note that the mixtures in the test set were generated from portions of the clean data used to train the baseline 34 speaker SD and SA models. The performance would likely have been worse had there been enough clean data to properly generate a separate test set, so the accuracy shown for the 34 speaker set should be viewed as an upper bound.

Performance of both the SD and SA systems suffers on held out speakers, but the performance decrease relative to the use of models trained on all 34 speakers shown in the bottom row of the figure is much greater for the SD models. In fact, the SA system slightly outperforms the SD system in absolute accuracy in most of the held out speaker cases. It is clear that separation using eigenvoice speech models generalizes better to unseen data than separation based on model selection from a set of SD models.

Despite this, the performance drop on held out speakers for both systems is quite significant. We expect that this is because a relatively small set of speakers were used to train the systems. As the number of

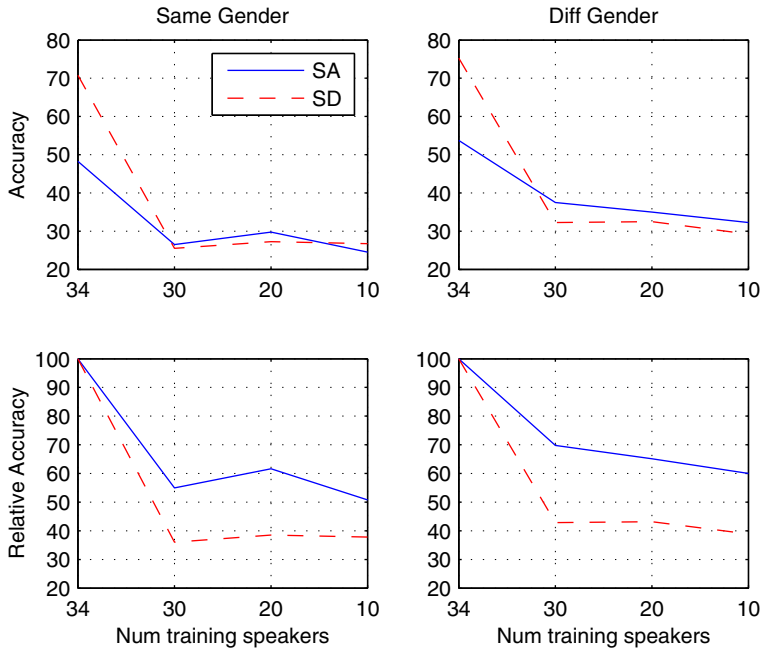

Fig. 7. Performance on mixtures of utterances from held out speakers using only subsets of 10, 20, and 30 speakers for training compared to models trained on all 34 speakers.

eigenvoice training examples increases we expect the model to better capture the characteristics of the general speaker space and thus be able to generalize better to unseen speakers. At the same time, as the number of training speakers grows it becomes increasingly likely that one of the models in the set will be a good match for a previously unseen speaker. Still, we expect that the performance of the SD system will not improve as quickly as the SA system as the size of the training set grows. This can be seen to some extent in the fact that the SD system have a flatter slope than the SA system as the number of models decreases in Fig. 7, especially in the different gender case.

In light of this, we note that these results are still preliminary. It is unclear how a system based on eigenvoice adaptation would compare to a system based on model selection from a large set of speaker-dependent models. In a future publication we plan to evaluate the ability of these systems to generalize to unseen speakers using a significantly larger data set containing utterances from hundreds of speakers.

6. Discussion

We have approached the task laid out in the 2006 Speech Separation Challenge using minimal prior knowledge about the signal content. Although the best performance requires the use of models that capture speaker specific characteristics, we have shown that good results are still possible using an approach based on speaker adaptation. The resulting system also has advantages in that it is better able to generalize to unseen speakers and it does not depend on knowledge of the task grammar.

The greatest weakness of our system is its tendency to permute between sources due to limited modeling of temporal dynamics. This is alleviated through the iterative eigenvoice re-estimation, but this process is slow. The permutation problem is caused by the ambiguity in the hard decisions made during the initial Viterbi decoding. Again, this arises because the initial source models computed as described in Section 4.1 are often very similar. Reducing the dependence of the initial adaptation on these hard assignments should decrease the severity of the permutation problem and potentially allow for faster convergence. This could be accomplished by extending the MLED algorithm to operate directly on the mixed signal, updating parameters based only on portions of the mixture dominated by a single source.

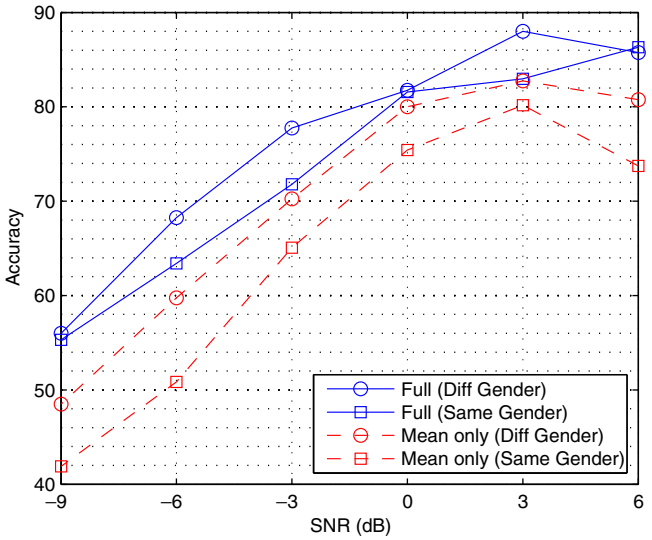

Fig. 8. Speaker-dependent separation performance where the full model is adapted (full) and where only the mean is updated (mean only). Performance on same gender mixtures is compared to performance on different gender mixtures.

We also note that only a subset of model parameters are being adapted to match the source statistics. Adapting the GMM covariances and HMM transition probabilities as well would make it easier to distinguish the sources. Fig. 8 compares separation performance using mean-adapted source models with fully adapted source models, including means, covariances, GMM priors, and transition probabilities. In both cases oracle knowledge of speaker identity and gains is used so this figure represents an upper bound on performance. It is clear that using fully adapted models can further improve performance over the mean-adapted models used in our system. This is especially true for the same gender conditions where the additional parameters make it easier to distinguish between the two speakers even if the corresponding model means are similar. As shown in the difference between the "full" and "mean only" same gender curves in the figure, this effect is most prominent at lower SNRs where the additional model constraints make up for the noisy observations. Potential extensions to the MLED inference algorithm to allow for adaptation of the remaining parameters are discussed in Kuhn et al. (2000).

7. Conclusion

In summary, we have described a novel monaural source separation system based on adaptation of a generic speech model to match each of the sources in a speech mixture. We use eigenvoice models to compactly define the space of speaker variation and propose an iterative algorithm to infer the parameters for each source in the mixture. The source-adapted models are used to separate the signal into its constituent sources. Source adaptation helps compensate for the limited temporal dynamics used in the speech model. However, it still does not perform as well as a system that uses speaker-dependent models, largely because it is prone to permutations between sources. Despite these shortcomings, we show that source adaptation based system shows promise in its ability to generalize better to held out speakers.

Acknowledgements

We wish to thank the reviewers for their helpful comments on this paper. A preliminary version of this work was presented at the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. This work was supported by the National Science Foundation (NSF) under Grants Nos. IIS-0238301 and IIS-0535168. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Barker, J., Coy, A., Ma, N., Cooke, M., 2006. Recent advances in speech fragment decoding techniques. In: Proceedings of Interspeech, pp. 85–88.

Cooke, M.P., Lee, T.-W., 2008. The 2006 Speech Separation Challenge. Computer Speech and Language.

Ellis, D., 2006. Model-based scene analysis. In: Wang, D., Brown, G. (Eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley/IEEE Press, pp. 115–146 (Chapter 4).

Kristjansson, T., Attias, H., Hershey, J., 2004. Single microphone source separation using high resolution signal reconstruction. In: Proceedings of ICASSP, pp. II–817–820.

Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., Gopinath, R., 2006. Super-human multi-talker speech recognition: the IBM 2006 Speech Separation Challenge system. In: Proceedings of Interspeech, pp. 97–100.

Kuhn, R., Junqua, J., Nguyen, P., Niedzielski, N., 2000. Rapid speaker adaptation in eigenvoice space. IEEE Transactions on Speech and Audio Processing 8 (6), 695–707.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, 171–185.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory 28 (2), 129–137.

Ozerov, A., Philippe, P., Gribonval, R., Bimbot, F., 2005. One microphone singing voice separation using source-adapted models. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 90–93.

Picone, J., Doddington, G.R., 1989. A phonetic vocoder. In: Proceedings of ICASSP, pp. 580–583.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (2), 257–286.

Roweis, S.T., 2000. One microphone source separation. In: Advances in Neural Information Processing Systems, pp. 793–799.

Roweis, S.T., 2003. Factorial models and refiltering for speech separation and denoising. In: Proceedings of Eurospeech, pp. 1009–1012.

Schmidt, M.N., Olsson, R.K., 2006. Single-channel speech separation using sparse non-negative matrix factorization. In: Proceedings of Interspeech, pp. 2614–2617.

Srinivasan, S., Shao, Y., Jin, Z., Wang, D., 2006. A computational auditory scene analysis system for robust speech recognition. In: Proceedings of Interspeech, pp. 73–76.

Varga, P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proceedings of ICASSP, pp. 845–848.

Virtanen, T., 2006. Speech recognition using factorial hidden Markov models for separation in the feature space. In: Proceedings of Interspeech, pp. 89–92.

Weiss, R.J., Ellis, D.P.W., 2007. Monaural speech separation using source-adapted models. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 114–117.

Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C., 2006. The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK.