MONAURAL SPEECH SEPARATION USING SOURCE-ADAPTED MODELS

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

LabROSA, Dept. of Electrical Engineering
Columbia University
{ronw,dpwe}@ee.columbia.edu

ABSTRACT

We propose a model-based source separation system for use on single channel speech mixtures where the precise source characteristics are not known *a priori*. We do this by representing the space of source 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 [1] and compared with separation using source-dependent models.

1. INTRODUCTION

Mixed signals containing multiple sources pose a significant problem for automatic signal analysis such as melody transcription or speech recognition, as well as for human listeners. Separating a mixture into its constituent sources is especially difficult when only a single channel input is available, making it impossible to use spatial constraints to separate the signals. In this paper we focus on the model-based approach to source separation which disambiguates the mixture based on statistical models for each source present in the mixture. Most previous work in this area, such as [2], uses source-specific models for separation (e.g. trained on the particular speaker to be separated). In [3] Ozerov et al. 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 source far better than using unadapted models, but requires a substantial amount of adaptation data. We consider adaptation when the data available is much less, 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. Finally, sections 5 and 6 contain experimental results and conclusions.

2. SOURCE MODELS

As shown in [2], incorporating temporal dynamics in source models can significantly improve separation performance, 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 [4], which models temporal dynamics only within each phone.

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 emits a GMM with 8 mixture components. The transitions from each phone to all others have equal probability. This allows us to incorporate some knowledge of speech structure without modeling the grammar.

The models were trained on the Speech Separation Challenge training data [1], downsampled to 16kHz and pre-emphasized. Spectral features were derived from a short-time Fourier transform with 40 ms window and 10 ms hop. 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; only the GMM means were updated during the SD training process.

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. 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 first orthogonalizing the SD models using principal component analysis (PCA), allowing each point in the space spanned by the different speakers to be represented using only a few "eigenvoice" weights [5].

Only the model means are adapted. The mean vectors of each state in the SD model for speaker j are concatenated into a mean supervector $\boldsymbol{\mu}_j$. Performing PCA on the set of 34 supervectors yields orthonormal basis vectors for the eigenvoice space. The mean for state s of a speaker-adapted model can then be written as a linear combination of these bases:

$$\boldsymbol{\mu}_s = \sum_{j=1}^{N} w_j \hat{\boldsymbol{\mu}}_{j,s} + \bar{\boldsymbol{\mu}}_s \qquad (1)$$

where w_j is the weight applied to the jth eigenvoice dimension, $\hat{\boldsymbol{\mu}}_{j,s}$, and $\bar{\boldsymbol{\mu}}_s$ is the average across all SD models of the mean for state s. Estimation of the eigenvoice parameters w_j is described in section 4.4. Note that for simplicity all equations in this paper describe the case of HMMs with Gaussian emissions. The extensions to mixture model emissions is straightforward.

*This work was supported by the National Science Foundation (NSF) under Grants No. 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.

4. SPEECH SEPARATION

Without strong temporal constraints, separation performance is poor when the same model is used for both sources. Separation can be improved with models matched to each source, but the adaptation procedure described in [5] requires clean source signals. We solve the problem of estimating the eigenvoice parameters for the two sources from the mixture using the following iterative algorithm:

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

4.1. Initialization

As with many iterative algorithms, this method can be slow to converge and is vulnerable to local optima. Good initialization is crucial to finding good solutions quickly. We start 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, so further steps are taken to differentiate the two speakers.

We use the speaker identification component of the Iroquois speech separation system [2] 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 will 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 each of them to be treated independently. So instead of learning the settings for each of 34 speakers, we quantize each dimension separately (e.g. w_1 can be quantized to -550, -20, or 600) to approximate the training cohort with just a few values, and then use the speaker identification algorithm described above to find the most likely settings of that dimension for the two sources. 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 dimensions with the most variance are 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 [6]. Each frame of the mixed signal $\mathbf{o}(t)$ is modeled by the combination of one state from each source model. The joint likelihood of each state combination is derived using the max approximation [7] which is based on the assumption that each time-frequency cell will be dominated by a single source.

If the emission for state s_i under model i has a Gaussian distribution with mean $\boldsymbol{\mu}_{i,s_i}$ and diagonal covariance Σ_{i,s_i}, this can be computed as follows:

$$P(\mathbf{o}(t)|s_1,s_2) = \mathcal{N}(\mathbf{o}(t); \max(\boldsymbol{\mu}_{1,s_1}, \boldsymbol{\mu}_{2,s_2}), \Sigma) \tag{2}$$

where $\max$ is the element-wise maximum and $\Sigma = \Sigma_{1,s_1}$ for dimensions where $\boldsymbol{\mu}_{1,s_1} > \boldsymbol{\mu}_{2,s_2}$ (i.e. where source 1 dominates the mixture) and $\Sigma = \Sigma_{2,s_2}$ otherwise.

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 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: $\hat{\mathbf{x}}_1(t) = E[\mathbf{x}_1(t)|s_1(t), s_2(t), \mathbf{o}(t)]$ where $s_1(t)$ and $s_2(t)$ correspond to the active state combination at time t in the Viterbi path. Each dimension d of the conditional mean is found using the max approximation:

$$E[x_1^d(t)|s_1(t), s_2(t), \mathbf{o}(t)] = \begin{cases} o^d(t), & \text{if } \mu^d_{1,s_1(t)} > \mu^d_{2,s_2(t)} \\ \mu^d_{1,s_1(t)}, & \text{otherwise} \end{cases} \tag{3}$$

The estimate for $\hat{\mathbf{x}}_2(t)$ follows the same derivation.

4.4. Eigenvoice parameter inference

Finally, the speaker models are updated to better match the source estimates. This is done using an extension of the maximum likelihood eigen-decomposition EM algorithm described in [5] that explicitly models the gain g applied to each source as well as the eigenvoice parameters w_j.

First the posterior probability of the source occupying state s at time t, $\gamma_s(t)$, is computed for all s and t. For increased efficiency, we do not use the dynamics of the HMMs for this computation (i.e. the models are reduced to GMMs). Given the posteriors, the eigenvoice weights and gain for source i can be found by solving the following set of simultaneous equations for w_j and g:

$$\begin{bmatrix} \mathbf{x} \\ a \end{bmatrix} = \begin{bmatrix} Y & \mathbf{z} \\ \mathbf{z}^T & b \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ g \end{bmatrix} \tag{4}$$

where

$$x_j = \sum_t \sum_s \gamma_s(t) \hat{\boldsymbol{\mu}}^T_{j,s} \Sigma^{-1}_{i,s} (\hat{\mathbf{x}}_i(t) - \bar{\boldsymbol{\mu}}_s) \tag{5}$$

$$Y_{j,k} = \sum_t \sum_s \gamma_s(t) \hat{\boldsymbol{\mu}}^T_{j,s} \Sigma^{-1}_{i,s} \hat{\boldsymbol{\mu}}_{k,s} \tag{6}$$

$$z_j = \sum_t \sum_s \gamma_s(t) \hat{\boldsymbol{\mu}}^T_{j,s} \Sigma^{-1}_{i,s} \mathbf{1} \tag{7}$$

$$a = \sum_t \sum_s \gamma_s(t) \mathbf{1}^T \Sigma^{-1}_{i,s} (\hat{\mathbf{x}}_i(t) - \bar{\boldsymbol{\mu}}_s) \tag{8}$$

$$b = \sum_t \sum_s \gamma_s(t) \mathbf{1}^T \Sigma^{-1}_{i,s} \mathbf{1} \tag{9}$$

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

The process is iterated for each source estimate $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ until it converges.

Figure 1 gives an example of the source separation and adaptation process. 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 reconstrnuction based on SD models (bottom pane) by the fifth iteration.

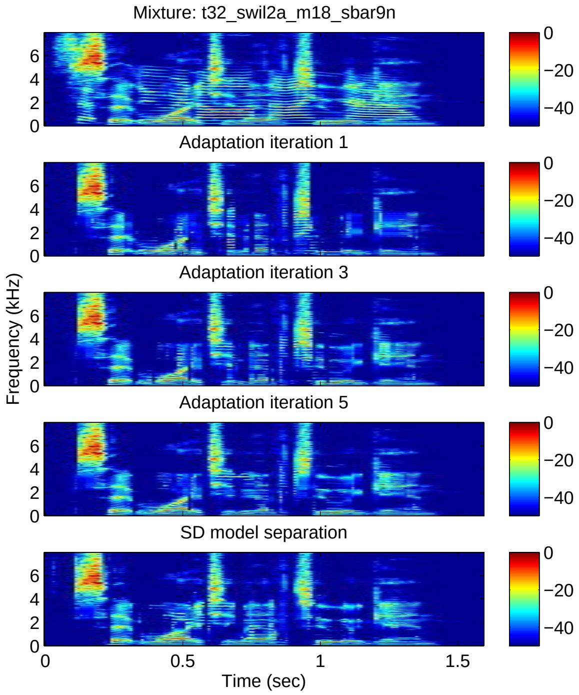

Figure 1: Separation using source-adapted models. The top plot shows the spectrogram of a mixture of female and male speakers. The middle three show the reconstructed target signal (“set white in l 2 again”) from the adapted models after iterations 1, 3, and 5. The bottom plot shows the result of separation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

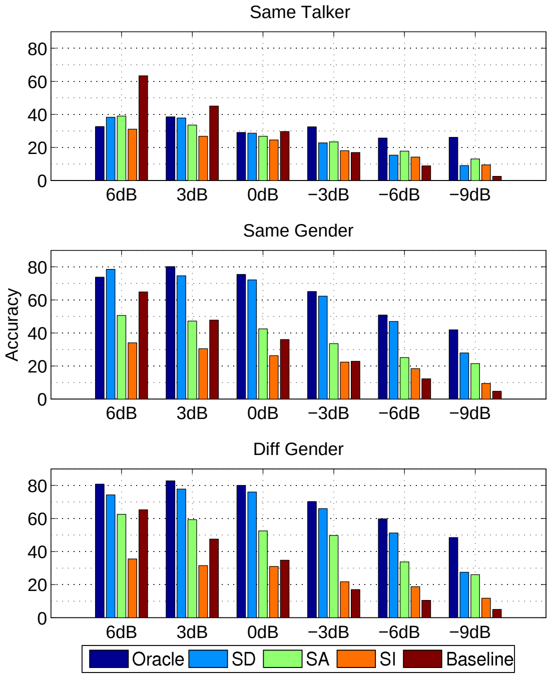

Figure 2: Separation performance using speaker-dependent (SD), speaker-adapted (SA), and speaker-independent (SI) models. The average accuracy over all conditions is 48% for the SD system, 36% for the SA system, and 23% for the SI system.

5. EXPERIMENTS

The system was evaluated on the test data from the 2006 Speech Separation Challenge [1]. 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 five iterations using eigenvoice speech models trained on all 34 speakers in the data set. The time-domain sources were reconstructed from the STFT magnitude estimates $\hat{\mathbf{x}}_i$ and the phase of the mixed signal. The two reconstructed signals are then passed to a speech recognizer; assuming one transcription contains “white”, it is taken as the target source. We used the default HTK speech recognizer provided by the challenge organizers, retrained on 16kHz data. Performance is measured using word accuracy of the letter and digit spoken by the target speaker.[1]

Figure 2 compares the performance of the source 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 [2]. 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 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, so such sources tend to have less overlap. On the other hand, the performance on the same talker task is quite poor. This is because the source models enforce limited dynamic constraints and the models used for each source are identical, except for the gain term. The lack of strong dynamic constraints allows for ambiguity in the Viterbi path through a factorial HMM composed of identical models [8]. 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.

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

[1] Sound examples of reconstructed sources are available at `http://www.ee.columbia.edu/~ronw/SSC.html`

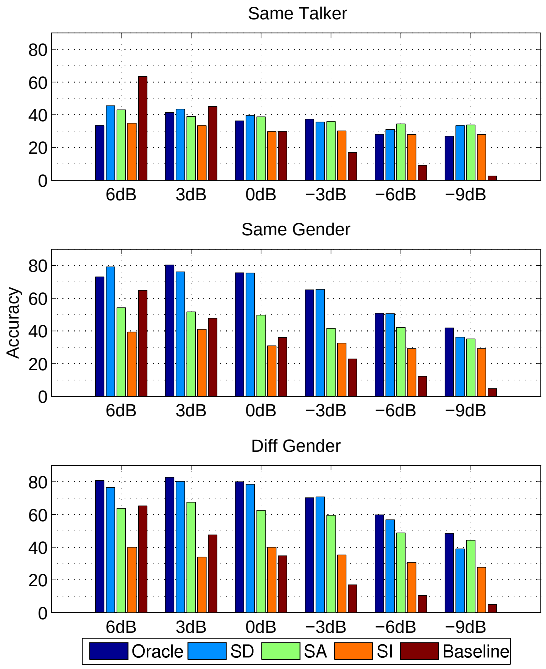

Figure 3: Separation performance when the target is chosen by picking the source that is closest to the target transcript.

in figure 1. 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. This is also why it performs significantly better on the different gender condition.

Figure 3 shows the performance of the same systems when the signal that most closely matches the target transcript is chosen as the target (i.e. a "cheating" condition). This metric is less sensitive to source permutations because it can correctly detect the estimated target source even when the color "white" is attributed to the wrong speaker. The performance of the SA and SI systems are improved under this metric, but the SA system naturally still falls short of the SD system.

Finally, figure 4 compares the performance of the SD system and SA system when data from only 10 speakers is used for training. These experiments were performed on a random subset of 50 mixtures that do not contain any of the subset of 10 speakers used to train the SD10 and SA10 systems. Performance of both systems suffers on held-out speakers, but the difference in performance between SD34 and SD10 is significantly larger than that between SA34 and SA10. In fact, SA10 tends to outperform SD10 at lower SNRs despite its problems with permutations. From this we can conclude that the performance of separation using eigenvoice speech models degrades more gracefully than SD model-based separation when presented with unseen data.

6. CONCLUSIONS

We propose a novel monaural source separation system based on adaptation of a generic source model to match the sources in the mixed signal. We use "eigenvoice" models to compactly define the space of speaker variation and use an iterative algorithm to infer the parameters for each source in a mixed signal. 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, but it 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 this system generalizes better to held-out speakers. Future work will address these issues by investigating better methods for inferring adaptation parameters.

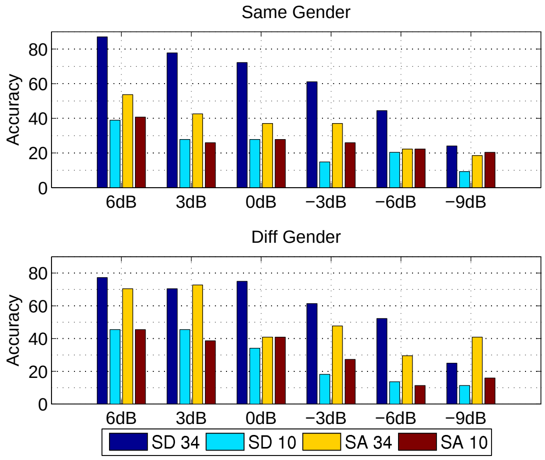

Figure 4: Performance on 50 mixtures from left-out speakers when a subset of 10 speakers are used for training compared to models trained on all 34 speakers.

7. REFERENCES

[1] M. Cooke and T. W. Lee, "The speech separation challenge," 2006. [Online]. Available: http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm

[2] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," in *Proceedings of Interspeech*, 2006.

[3] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005.

[4] J. Picone and G. R. Doddington, "A phonetic vocoder," in *Proceedings of ICASSP*, 1989, pp. 580 – 583.

[5] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transations on Speech and Audio Processing*, vol. 8, no. 6, pp. 695 – 707, November 2000.

[6] P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proceedings of ICASSP*, 1990.

[7] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of Eurospeech*, 2003.

[8] D. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. Brown, Eds. Wiley/IEEE Press, 2006, ch. 4, pp. 115–146.