A VARIATIONAL EM ALGORITHM FOR LEARNING EIGENVOICE PARAMETERS IN MIXED SIGNALS

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

LabROSA, Dept. of Electrical Engineering
Columbia University
New York NY 10027 USA
{ronw,dpwe}@ee.columbia.edu

ABSTRACT

We derive an efficient learning algorithm 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 factor-analyzed hidden Markov models (HMM) where source specific characteristics are captured by an "eigenvoice" speaker subspace model. The proposed algorithm is able to learn adaptation parameters for two speech sources when only a mixture of signals is observed. We evaluate the algorithm on the 2006 Speech Separation Challenge data set and show that it is significantly faster than our earlier system at a small cost in terms of performance.

Index Terms— Eigenvoices, model-based source separation, variational EM

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 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 largely 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 [1]. 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 [2, 3].

Good performance of such model-based source separate systems requires source models with high frequency resolution to capture speaker-dependent aspects of the signal [4]. It is precisely the speaker-dependent characteristics, mainly specific fundamental and formant frequencies, that enable such approaches to identify time-frequency regions dominated by a particular source. In [3], Kristjansson et al. describe a model-based separation system that assumes that the observed sources come from a closed set of talkers for which prior training data is available. However, in most applications it is reasonable to assume that the separation system will not have prior knowledge of which specific sources are present in a particular mixture. Weiss and Ellis describe a similar model-based approach in [4] that relaxes this assumption by constructing a parametric speech model based on eigenvoice modeling [5] that is able to adapt to the sources present in a particular mixture. In this paper, we derive a more principled algorithm to estimate the adaptation parameters based on variational expectation maximization (EM) learning in a factorial hidden Markov model [6]. The new approach is about four times faster than [4].

The remainder of the paper is organized as follows. In sections 2 and 3 we describe our speaker-adaptation model and mixed signal model respectively. The variational EM adaptation algorithm is described in section 4. Experimental results on a subset of the 2006 Speech Separation Challenge dataset [7] are given in section 5. Finally, we conclude in section 6.

2. SPEAKER SUBSPACE MODEL

We model the log power spectrum of the speech signal produced by speaker i, $\mathbf{x}_i(t)$ using a hidden Markov model (HMM) trained over clean speech data from that speaker. The joint likelihood of the observations $\mathbf{x}(1..T)$ and all possible state sequences $s(1..T)$ can be written as follows:

$$P(s_i(1..T) \,|\, \mathbf{x}_i(1..T)) \propto \prod_t P(s(t) \,|\, s(t{-}1)) \, P(\mathbf{x}_i(t) \,|\, s(t)) \quad (1)$$

$$P(\mathbf{x}_i(t) \,|\, s) = \mathcal{N}(\mathbf{x}_i(t);\, \boldsymbol{\mu}_{i,s},\, \Sigma_{i,s}) \quad (2)$$

where $\boldsymbol{\mu}_{i,s}$ and $\Sigma_{i,s}$ refers to the Gaussian mean and covariance matrix corresponding to state s in the model for speaker i.

Each of the 35 phones used in the Speech Separation Challenge task grammar are modeled using a standard 3-state forward HMM topology. Each state's emissions are modeled by a Gaussian mixture model (GMM) with 8 mixture components, but to simplify the notation we assume that this has been converted to a model with Gaussian emissions (i.e. each GMM component is treated as a separate state). The transitions from each phone to all others have equal probability, which was found to work as well as more phonotactically-informed transitions. This structure allows us to incorporate some knowledge of speech dynamics without being specific to any grammar.

We used the HTK toolkit [8] to train the models on the Speech Separation Challenge training data [7], downsampled to 16 kHz and pre-emphasized as in the Iroquois system. The training data for all

[*] This work was supported by the NSF under Grants No. IIS-0238301 and IIS-0535168, and by EU project AMIDA. 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 Sponsors.

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 state ordering across all speaker models is needed for the speaker adaptation process we describe now.

We use this set of SD models to construct an eigenvoice speaker subspace model which can be adapted to correspond to a particular speaker in the training set. This is very similar to the factor analysis parameterization of speaker models commonly used for speaker verification [9]. Detailed discussions of this approach can be found in [5] and [4]. The only difference in this work is that we adapt the covariance parameters as well as the mean parameters because it was shown in [4] that adapting the covariance parameters could potentially lead to significant performance gains.

If we concatenate the SD parameters – consisting of the Gaussian means, U_i and the log-covariances, $\log S_i$, for all states for speaker i – into a parameter supervector $P_i = [U_i;\ \log S_i]$, we can consider any speaker model to be a point in this very high dimensional space. The space spanned by all K training speakers can then be described by the matrix $P = [P_1, P_2, \ldots P_K]$. Performing principal component analysis (PCA) on this matrix yields a set of orthonormal basis vectors for the speaker subspace which allows any particular speaker model to be described as a linear combination of these bases:

$$\boldsymbol{\mu}_{i,s} = \boldsymbol{\mu}_s(\mathbf{w_i}) = U_s\mathbf{w}_i + \bar{\boldsymbol{\mu}}_s \tag{3}$$

$$\log\Sigma_{i,s} = \log\Sigma_s(\mathbf{w_i}) = \log(S_s)\mathbf{w}_i + \log\bar{\Sigma}_s \tag{4}$$

where the (diagonal) covariance parameters are modeled in the log domain to guarantee positivity regardless of $\mathbf{w}_i$.

Essentially, the very high dimensional parameters for speaker i are represented as a function of a low dimensional vector $\mathbf{w}_i$. Because the number of parameters needed to describe a particular speaker is so small, this technique has the advantage of requiring very little adaptation data, make it suitable for our application of adapting models to a single utterance. Finally, because the speaker subspace parameters are continuous, this approach allows for smooth interpolation across the entire space, enabling it to capture a wider variety of SD models than were used in training.

3. MIXED SIGNAL MODEL

The graphical model for our mixed signal model is shown in figure 1. Each source signal $\mathbf{x}_i(t)$ is generated by the factor-analyzed HMM described in the previous section. The speaker-dependent characteristics of source i are compactly described by the parameters $\mathbf{w}_i$ which are used to generate the Gaussian means and covariances comprising the HMM emission distributions. Finally, the observed mixture $\mathbf{y}(t)$ is explained by the combination of the two source signals. Therefore, the overall observation is generated by a sequence of state combinations corresponding to the state sequences of the underlying clean source models.

We use the common "max" approximation [2] to describe the way two natural speech signals mix in the short-time Fourier transform (STFT) domain:

$$y(t) = \sum_{i=1}^{I} x_i(t) \tag{5}$$

$$\mathbf{y}(t) \approx \max_i \mathbf{x}_i(t) \tag{6}$$

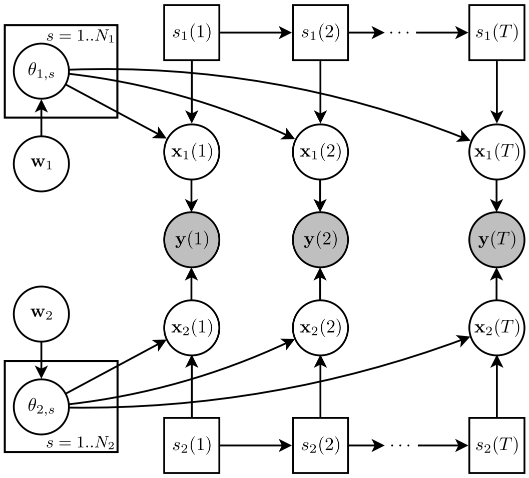

Fig. 1. Proposed mixed signal model. The mixture observations $\mathbf{y}(t)$ are explained as the combination of two hidden source signals $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$. Each source signal is modelled by a separate speaker-adapted hidden Markov model, that is derived from the speaker subspace model described in section 2.

where $\mathbf{y}(t)$ is the log power spectrum of the waveform $y(t)$.

As described above, each clean source signal is modeled using a hidden Markov model. The mixed signal can therefore be modeled by combining the separate speech models into a factorial HMM:

$$P(s_1(1..T), s_2(1..T) \,|\, \mathbf{y}(1..T)) \propto \prod_t P(\mathbf{y}(t) \,|\, s_1(t), s_2(t)) \\ P(s_1(t) \,|\, s_1(t{-}1)) P(s_2(t) \,|\, s_2(t{-}1)) \tag{7}$$

Using the max approximation, the likelihood of the mixed signal under state combination s_1, s_2 can be written as follows when using diagonal covariances:

$$P(\mathbf{y}(t) \,|\, s_1, s_2) \\ = \mathcal{N}(\mathbf{y}(t);\ M_1\boldsymbol{\mu}_{1,s_1} + M_2\boldsymbol{\mu}_{2,s_2},\ M_1\Sigma_{1,s_1} + M_2\Sigma_{2,s_2}) \tag{8}$$

where M_i behaves as a binary mask that selects frequency bands dominated by source i. It is a diagonal matrix containing ones for dimensions where model i is bigger than the other model ($\mu_{i,s_i} > \mu_{2,s_2(t)}$) and zeros elsewhere. Similarly, $M_2 = I - M_1$.

Given this model for the mixed signal, we separate a speech mixture in two stages. First, the subspace parameters are derived for each source in the mixture, producing a set of speaker-adapted models capturing the speaker-dependent statistics of the constituent talkers. Then, given the adapted models, the clean source signals are reconstructed by finding the minimum mean square error reconstruction of the signals given the model. This is done by finding the Viterbi path through the factorial HMM as described in [4].

The adaptation process involves using the mixed signal to learn the parameters $\mathbf{w}_i$ that define the speaker-adapted parameters. It is possible to derive an EM algorithm for this, similar to that in [6], but the exact computation of the posterior probabilities in the E-step is intractable due to the combinatorial nature of the state space. I.e. if speaker HMM i contains N_i states, the statistics needed by the full EM algorithm must take into account all possible state combinations from all speakers leading to an equivalent state space containing $\prod_i N_i$ states. Instead, we derive an approximate E-step with a complexity of $\sum_i N_i$ states based on the variational approximation presented in [6]. This is described in detail in the following section.

4. VARIATIONAL LEARNING

We approximate the full distribution over the hidden variables with an approximate distribution in which the HMM chains for each speaker are decoupled from the observations. This is equivalent to removing the arrows pointing from $\mathbf{x}_i(t)$ to $\mathbf{y}(t)$ in the graphical model shown in figure 1. We introduce a variational parameter $h_{i,s}(t)$ for each arrow removed from the graphical model. These parameters act as pseudo-observation likelihoods which serve to couple the two chains. Given the variational parameters, the approximate distribution can be written as follows:

$$Q(s_1(1..T), s_2(1..T) \,|\, \mathbf{y}(1..T)) \propto \prod_{i,t} h_{i,s_i}(t)\, P(s_i(t)|s_i(t{-}1)) \quad (9)$$

Q has the same form as the single chain HMM in (1), with observation probabilities given by the variational parameters.

An outline of the overall variational learning algorithm is described below. Details are given in the following sections.

- E-step: Iteratively learn the posterior distribution over state combinations of both speaker models.
    1. Run the HMM forward-backward algorithm separately for each model, using the variational pseudo-observations to compute $\gamma_{i,s}(t)$, forward-backward lattice posteriors for each model,
    2. Compute variational parameters $h_{s_i}(t)$ based on $\gamma_{i,s}(t)$ and iterate.
- M-step: Update the model parameters $\mathbf{w}_1, \mathbf{w}_2$ using the posteriors computed in the E-step.

4.1. E-step

The optimal variational parameters can be computed by minimizing the Kullback-Leibler divergence between the approximation distribution Q and the full distribution P:

$$KL(Q||P) = \sum_{t,s_1,s_2} \gamma_{1,s_1}(t)\, \gamma_{2,s_2}(t)\, (\log P(\mathbf{y}(t) \,|\, s1, s2) + \log h_{1,s_1}(t) + \log h_{2,s_2}(t)) + c \quad (10)$$

This implies the following updates for the variational parameters:

$$\log h_{1,s_1}(t) = \sum_{s_2} \gamma_{2,s_2}(t)\; \log P(\mathbf{y}(t) \,|\, s_1, s_2) \quad (11)$$

$$\log h_{2,s_2}(t) = \sum_{s_1} \gamma_{1,s_1}(t)\; \log P(\mathbf{y}(t) \,|\, s_1, s_2) \quad (12)$$

Because $\gamma_{i,s_i}(t)$ is generally quite sparse (i.e. very few states per frame have significant probability mass), the expectations in (11) and (12) are fast to compute. The overall complexity is reduced from $O(N_1 N_2)$ for computing the full $P(\mathbf{y}(t) \mid s_1, s_2)$ to $O(N_1 + N_2)$. The process effectively holds one chain constant while updating the other. The final joint posterior of being jointly in s_1 and s_2 is obtained by simply combining the two marginal distributions:

$$\gamma_{s_1,s_2}(t) = \gamma_{1,s_1}(t)\gamma_{2,s_2}(t) \quad (13)$$

4.2. M-step

Given the posterior distribution over the hidden state sequences, the speaker model parameters $\theta = \{\mathbf{w}_i\}$ can be updated by maximizing the expected log likelihood of the model:

$$\mathcal{L}(\theta) = \sum_{s_1,s_2} \gamma_{s_1,s_2}(t)\; \log P(\mathbf{y}(t) \,|\, s_1, s_2) + k \quad (14)$$

As shown in [10], this objective function is not convex when both the Gaussian means and covariances depend on the subspace parameters being optimized. Instead, as suggested in [5], we derive an update based only on the mean statistics and rely on the correlation between the mean and covariance parameters implicit in the learned subspace to adapt the model covariances. The simplified objective can be written as follows:

$$\mathcal{L}(\theta) = -\frac{1}{2} \sum_{t,s_1,s_2} \gamma_{s_1,s_2}(t)\; \mathcal{M}(\mathbf{y}(t) - \boldsymbol{\mu}_{s_1 s_2}(\theta),\; \Sigma_{s_1 s_2}) \quad (15)$$

where

$$\begin{aligned} \mathcal{M}(\mathbf{a}, B) &= \mathbf{a}^T B^{-1} \mathbf{a} && (16) \\ \boldsymbol{\mu}_{s_1 s_2}(\theta) &= M_1 \boldsymbol{\mu}_{s_1}(\mathbf{w}_1) + (I - M_1)\boldsymbol{\mu}_{s_2}(\mathbf{w}_2) && (17) \\ \Sigma_{s_1 s_2} &= M_1 \bar{\Sigma}_{s_1} + (I - M_1)\bar{\Sigma}_{s_2} && (18) \end{aligned}$$

A further complication results from the fact that the step function (i.e. the binary mask) inherent in the max approximation implied in equation (8) makes the objective function non-differentiable. This makes it difficult to maximize exactly. Instead we hold the masks M_1 constant in the optimization. Because of this approximation, the log likelihood is not always guaranteed to increase, but in practice it works quite well.

The resulting weights can be found solving the following set of simultaneous equations for $\mathbf{w}_1$ and $\mathbf{w}_2$:

$$\sum_{t,s_1,s_2} \gamma_{s_1 s_2}(t)\; U_{s_1}^T\, M_1\, \Sigma_{s_1 s_2}^{-1}(\mathbf{y}(t) - U_{s_1}\mathbf{w}_1 - \bar{\boldsymbol{\mu}}_{s_1}) = 0 \quad (19)$$

$$\sum_{t,s_1,s_2} \gamma_{s_1 s_2}(t)\; U_{s_2}^T\, M_2\, \Sigma_{s_1 s_2}^{-1}(\mathbf{y}(t) - U_{s_2}\mathbf{w}_2 - \bar{\boldsymbol{\mu}}_{s_2}) = 0 \quad (20)$$

These updates are quite similar to the clean signal eigenvoice updates derived in [5], except for the binary masks M_i which partition the observations into regions dominated by a single talker, causing the algorithm to ignore interference-dominated time-frequency regions when updating the parameters for a particular talker.

5. EXPERIMENTS

We evaluate the proposed algorithm on the 0 dB SNR subset of the 2006 Speech Separation Challenge [7] data set. This consists of 200 single-channel mixtures of two talkers of different gender, and 179 mixtures of two talkers of the same gender, mixed at 0 dB SNR. 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”.

We compare a number of separation algorithm using a common framework. Given a mixed signal, each system is used to generate an STFT representation of each source. The time-domain sources are reconstructed from the STFT magnitude estimates and the phase of the mixed signal. The two reconstructed signals are then passed to a speech recognizer; assuming one transcription contains “white”,

	Mean Only		Mean + Covar	
Algorithm	Same Gender	Diff Gender	Same Gender	Diff Gender
Variational EM	47.49%	61.75%	58.10%	69.75%
Iterative separation/adaptation [4]	56.15%	66.75%	60.06%	78.75%
Speaker-dependent model selection [3]	72.07%	76.00%	83.52%	80.00%
Baseline	36.03%	34.75%	36.03%	34.75%

Table 1. Digit-letter recognition accuracy on the 0dB SNR two-talker subset of the 2006 Speech Separation Challenge data set.

it is taken as the target source. We used the default HTK speech recognizer provided by [7], retrained on 16 kHz data.

The proposed variational EM algorithm is compared to our previous method based on iterative separation and adaptation in [4], to our implementation of the Iroquois system [3] based on model selection from a closed set of speaker-dependent models, and to the baseline recognition results obtained by running the speech recognizer over the mixture. All systems were evaluated using models where only the means were speaker-dependent (Mean Only) as in [4] as well as using models where both the means and covariances were speaker-dependent (Mean + Covar).

The results are summarized in table 1. All of the evaluated separation systems show very large improvements over the baseline recognizer run on the mixtures without any other processing. The proposed system performs almost as well as the iterative separation/adaptation algorithm from [4], particularly on same gender mixtures when covariance is adapted. Qualitatively, the main difference between the two algorithms is that the EM approach considers all possible paths through the joint state space of the speech models whereas the algorithm in [4] chooses the most likely path. This might result in differing convergence behavior of the two algorithms. Both were only run for 15 iterations, which was shown to work well for the approach in [4]. The variational EM algorithm might simply take longer to converge because it evaluates more state combinations.

The advantage to the algorithm proposed in this paper is that the nature of the approximation allows it to run significantly faster than the old system which ran the Viterbi algorithm over the factorial HMM state space for every iteration. Our Matlab implementation of the new algorithm runs about 3-5 times faster than our previous optimized, pruned, C-coded Viterbi search.

The system based on selection of speaker-dependent models performs best, significantly outperforming the adaptation based systems on same gender mixtures. The advantage on different gender mixtures is not as pronounced. This is because same-gender sources have more overlap, which makes it more difficult to segregate them, which in turn makes it difficult for the adaptation algorithm to isolate regions unique to a single source. Instead, the adaptation based systems sometimes converge on solutions which are partial matches for both speakers, leading to separations which contain phone permutations across sources as described in [4]. We suspect that this is a result of the fact that only a short utterance is available for adaptation. If more adaptation data was available, it is likely that the algorithm would be able to find more clean glimpses of each speaker, leading to more robust adaptation.

Finally, as predicted in [4], the addition of speaker-adapted covariance parameters gives a significant performance improvement of between 5% and 10% absolute to all systems under all conditions. The improvements tend to be larger for different gender mixtures for the same reasons described earlier. Because same gender mixtures tend to overlap more in our STFT representation, the algorithm initialization tends not to be as robust.

6. CONCLUSIONS

We have described a model for speaker adaptation and separation of a mixed signal based on a compact speaker subspace model. We derive a fast an efficient learning algorithm based on a variational approximation to the factorial hidden Markov model. Although performance is not quite as good as that obtained using our previous approach, the proposed algorithm is significantly faster. We also show that a very simple extension to the subspace model to allow it to adapt the model covariances as well as the model means yields very significant performance improvements for all evaluated systems.

7. REFERENCES

[1] S. Srinivasan, Y. Shao, Z. Jin, and D. Wang, “A computational auditory scene analysis system for robust speech recognition,” in *Proceedings of Interspeech*, September 2006, pp. 73–76.

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

[3] 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, pp. 97–100.

[4] R. J. Weiss and D. P. W. Ellis, “Speech separation using speaker-adapted eigenvoice speech models,” *Computer Speech and Language*, 2008 (in press).

[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] Z. Ghahramani and M. Jordan, “Factorial hidden markov models,” *Machine Learning*, vol. 29, no. 2-3, pp. 245–273, November 1997.

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

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

[9] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of interspeaker variability in speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, July 2008.

[10] C.-H. Huang, J.-T. Chien, and H.-M. Wang, “A new eigenvoice approach to speaker adaptation,” in *Proceedings of ISCSLP*, 2004.