MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Multi-Channel Source Separation by Beamforming Trained with Factorial HMMS

Manuel J. Reyes-Gomez[*] Bhiksha Raj Daniel P. W. Ellis[†]

TR-2003-088 July 2004

Abstract

Speaker separation has conventionally been treated as a problem of Blind Source Separtion (BSS). This approach does not utilize any knowledge of teh statistical characteristics of the signals to be separated, relying mainly on the independence between the various signals to separate them. Maximum-likelihood techniques, on the other hand, utilize knowledge of the a priori probability distributions of the signals from the speakers, in order to effect separation. In [5] we presented a Mizimum-likelihood speaker separation technique that utilizes detailed statistical information about the signals to be separated, represented in the form of hidden Markov models (HMMs), to estimate the parameters of a filter-and-sum process for signal separation. In this paper we show that the filters that are estimated for any utterance by a speaker generalize well to other utterances by the same speaker, provided the locations of the various speakers remains constant. Thus, filters that have been estimated using a ”training” utterance of known transcript can be used to separate all future signals by the speaker from mixtures of speech signals in an unsupervised manner. On the other hand, the filters are ineffective for other speakers, indicating that they capture the spatio-frequency characteristics of the speaker.

[*] Columbia University
[†] Columbia University

Publication History:

1. First printing, TR-2003-088, July 2004

MULTI-CHANNEL SOURCE SEPARATION BY BEAMFORMING TRAINED WITH FACTORIAL HMMS

Manuel J. Reyes-Gomez[1]*, Bhiksha Raj*[2]*, Daniel P. W. Ellis*[1]

[1]Dept. of Electrical Engineering, Columbia University
[2]Mitsubishi Electric Research Laboratories

ABSTRACT

Speaker separation has conventionally been treated as a problem of Blind Source Separation (BSS). This approach does not utilize any knowledge of the statistical characteristics of the signals to be separated, relying mainly on the independence between the various signals to separate them. Maximum-likelihood techniques, on the other hand, utilize knowledge of the *a priori* probability distributions of the signals from the speakers, in order to effect separation. In [5] we presented a Maximum-likelihood speaker separation technique that utilizes detailed statistical information about the signals to be separated, represented in the form of hidden Markov models (HMMs), to estimate the parameters of a filter-and-sum processor for signal separation. In this paper we show that the filters that are estimated for any utterance by a speaker generalize well to other utterances by the same speaker, provided the locations of the various speakers remains constant. Thus, filters that have been estimated using a "training" utterance of known transcript can be used to separate all future signals by the speaker from mixtures of speech signals in an unsupervised manner. On the other hand, the filters are ineffective for other speakers, indicating that they capture the spatio-frequency characteristics of the speaker.

1. INTRODUCTION

There are several situations where two or more speakers speak simultaneously, and it is necessary to be able to separate the speech from the individual speakers from recordings of the simultaneous speech. Conventionally, this is referred to as the *speaker-separation* or *source-separation* problem. One approach to this problem is through the use of a time-varying filter on single-channel recordings of speech spoken simultaneously by two or more speakers [1, 2]. This approach uses extensive and speaker specific prior information about the statistical nature of speech from the different speakers, usually represented by dynamic models like the hidden Markov model (HMM), to compute the time-varying filters. The utility of time-varying filter approach is limited by the fact that the amount of information present in a single recording is usually insufficient to do effective speaker separation.

A second, more popular approach to speaker separation is through the use of signals recorded using multiple microphones. The algorithms involved typically require at least as many microphones as the number of signal sources. The problem of speaker separation is then treated as one of *Blind* Source Separation (BSS), which is performed using standard techniques like Independent Component Analysis (ICA). In this approach, no *a priori* knowledge of the signals is assumed. Instead, the component signals are estimated as a weighted combination of current and past samples from the multiple recordings of the mixed signals. The weights are estimated to optimize an objective function that measures the independence of the estimated component signals [3]. The blind multiple-microphone based approach ignores the known *a priori* probability distribution of the speakers, a potentially important source of information.

Maximum-likelihood source separation techniques (e.g. [4]), on the other hand, utilize the *a priori* probability distributions of individual signals to separate the signals from multiple-microphone recordings. Typically, the *a priori* distribution of the time-domain signals are considered. For the purpose of modelling these distributions, the samples of the time-domain signal are usually considered to be independent and identically distributed.

In [5] we report a Maximum-likelihood speaker separation technique where we model the *a priori* distribution of *frequency-domain* representations, *i.e.* log-spectra of the signals from the various speakers using HMMs, which capture the temporal characteristics of the signal. The actual signal separation is performed in the time domain, using the filter-and-sum method [6] described in Section 2. The algorithm can hence be viewed as beamforming that is performed using statistical information about the expected signals. The parameters of the filters estimated using an EM algorithm that maximizes the likelihood of log-spectral features computed from the separated signals. The iterations of the EM algorithm require the distribution of mixed signal, which is composed from the marginal distributions of the signals using locally linear transformations of the state output densities of the individual signals.

In [5] we report signal separation results when the HMMs for the signals are composed with full knowledge of the word sequences uttered by the various speakers. The delay-and-sum filters estimated using these HMMs are shown to be highly effective at separating the signals for which they were trained. In this paper we extend the results reported in [5], and evaluate the generalizability of the filters estimated for a given utterance. We establish experimentally that the estimated filters actually capture the spatio-frequency characteristics of the individual speakers, and do not merely over fit to the specific utterances for which they were trained. Thus, filters estimated for a given utterance from a speaker at any location are also effective at separating other utterances by the same speaker from that location. The filters can hence be *trained* using an utterance for which the transcriptions are known, and then used for subsequent utterances by the same speaker. On the other hand, they are ineffective for other speakers at the same location.

The rest of this paper is arranged as follows: In Section 2 we outline the filter-and-sum array processing methodology. In Sections 3 and 4 the EM algorithm for estimating the filters in the filter-and-sum array is briefly outlined. In Section 5 we describe our experiments and present our experimental results. Finally in Section 5 we present our conclusions.

2. FILTER-AND-SUM MICROPHONE PROCESSING

In this section we will describe the filter-and-sum array processing to be used for developing the current algorithm for speaker separation. The only assumption we make in this context is that the number of speakers is known. For each of the speakers, a separate filter-and-sum array is designed. The signal from each microphone is filtered by a microphone-specific filter. The various filtered signals are summed to obtain the final processed signal. Thus, the output signal for speaker i, $y_i[n]$, is obtained as:

$$y_i[n] = \sum_{j=1}^{L} h_{ij}[n] * x_j[n] \qquad (1)$$

where L is the number of microphones in the array, $x_j[n]$ is the signal at the j^{th} microphone and $h_{iv}[n]$ is the filter applied to the j^{th} filter for speaker i. The filter impulse responses $h_{ij}[n]$ must be optimized such that the resultant output $y_i[n]$ is the separated signal from the i^{th} speaker.

3. TRAINING THE FILTERS FOR A SPEAKER

In the training phase, the filters for any speaker are optimized using the available information about their speech. The information used is based on the assumption that the correct transcription of the speech from the speaker whose signal is to be extracted is known for a short training signal. It is assumed that this training signal has the same characteristics in terms of speakers and their relative positions with respect to the microphones as the combined signals that the filters are intended to separate. We further assume that we have access to a speaker-independent hidden Markov model (HMM) based speech recognition system that has been trained on a 40-dimensional Mel-spectral representation of the speech signal. The recognition system includes HMMs for the various sound units that the language comprises. From these, known transcription for the speaker's training utterance, we first construct an HMM for the utterance. Following this, the filters for the speaker are estimated to maximize the likelihood of the sequence of 40-dimensional Mel-spectral vectors computed from the output of the filter-and-sum processed signal, on the utterance HMM.

For the purpose of optimization, we must express the Mel-spectral vectors as a function of the filter parameters as follows: We concatenate the filter parameters for the i^{th} speaker, for all channels, into a single vector $\mathbf{h}_i$. Let Z_i represent the sequence of Mel-spectral vectors computed from the output of the array for the i^{th} speaker. Let z_{it} be the t^{th} spectral vector in Z_i. z_{it} is related to $\mathbf{h}_i$ by the following equation:

$$z_{it} = log(\mathbf{M}|DFT(\mathbf{y}_{it})|^2) = log(\mathbf{M}(diag(\mathbf{F}\mathbf{X}_t\mathbf{h}_i\mathbf{h}_i^T\mathbf{X}_t^T\mathbf{F}^H))) \qquad (2)$$

where $\mathbf{y}_{it}$ is a vector representing the sequence of samples from $y_i[n]$ that are used to compute z_it, $\mathbf{M}$ is the matrix of the weighting coefficients for the Mel filters, $\mathbf{F}$ is the Fourier transform matrix and $\mathbf{X}_t$ is a supermatrix formed by the channel inputs and their shifted versions.

Let Λ_i represent the set of parameters for the HMM for the utterance from the i^{th} speaker. In order to optimize the filters for the i^{th} speaker, we maximize $L_i(Z_i) = log(P(Z_i|\Lambda_i))$, the log-likelihood of Z_i on the HMM for that speaker. $L_i(Z_i)$ must be computed over all possible state sequences through the utterance HMM. However, in order to simplify the optimization, we assume that the overall likelihood of Z_i is largely represented by the likelihood of the most likely state sequence through the HMM, *i.e.*, $P(Z_i|\Lambda_i) \approx P(Z_i, \mathbf{S}_i|\Lambda_i)$, where $\mathbf{S}_i$ represents the most likely state sequence through the HMM. Under this assumption, we get

$$L_i(Z_i) = \sum_{t=1}^{T} log(P(z_{it} \mid \mathbf{s}_{it})) + log(P(\mathbf{s}_{i1}, \mathbf{s}_{i2}, .., \mathbf{s}_{iT})) \qquad (3)$$

where T represents the total number of vectors in Z_i, and $\mathbf{s}_{it}$ represents the state at time t in the most likely state sequence for the i^{th} speaker.
$log(P(\mathbf{s}_{i1}, \mathbf{s}_{i2}, .., \mathbf{s}_{iT}))$ does not depend on z_{it} or the filter parameters, and therefore does not affect the optimization, hence maximizing equation 3 is the same as maximizing $\sum log(P(z_{it} \mid \mathbf{s}_{it}))$. We make the simplifying assumption that this is equivalent to minimizing the distance between Z_i and the most likely sequence of vectors for the state sequence $\mathbf{S}_i$. When state output distributions in the HMM are modeled by a single Gaussian, the most likely sequence of vectors is simply the sequence of means for the states in the most likely state sequence. In the rest of this paper we will refer to this sequence of means as the *target* sequence for the speaker. We can now define the objective function to be optimized for the filter parameters as:

$$Q_i = \sum_{t=1}^{T}((z_{it} - m^i_{\mathbf{s}_{it}})^T(z_{it} - m^i_{\mathbf{s}_{it}})) \qquad (4)$$

where the t^{th} vector in the target sequence, $m^i_{\mathbf{s}_{it}}$ is the mean of $\mathbf{s}_{it}$, the t^{th} state, in the most likely state sequence $\mathbf{S}_i$.

It is clear from equations 2 and 4 that Q_i is a function of $\mathbf{h}_i$. Direct optimization of Q_i with respect to $\mathbf{h}_i$ is, however, not possible due to the highly non-linear relationship between the two. We therefore optimize Q using the method of conjugate gradient descent.
The filter optimization algorithm proceeds iteratively by alternately estimating the best target, and optimizing the filters. Further details of the filter optimization algorithm can be found in [5].
Since the algorithm aims to minimize the distance between the output of the array and the target, the choice of a good target becomes critical to its performance. The next section deals with the determination of the target sequences for the various speakers.

4. TARGET ESTIMATION

The ideal target would be a sequence of Mel-spectral vectors obtained from clean uncorrupted recordings of the speaker. All other targets must be considered approximations to the ideal target. In this work we attempt to derive the target from the HMM for that speaker's utterance. This is done by determining the best state sequence through the HMM from the current estimate of that speaker's signal. A direct approach to obtaining the state sequence would be to directly find the most likely state sequence for the sequence of Mel-spectral vectors for the signal. Unfortunately, in the early iterations of the algorithm, when the filters have not yet been fully optimized, the output of the filter-and-sum array for any speaker contains a significant fraction of the signal from other speakers as well. As a result, naive alignment of the output to the HMM results in poor estimates of the target.

Instead, we also take into consideration the fact that, at any iteration, the array output is a mixture of signals from all the speakers. The HMM that represents this signal is a *factorial* HMM (FHMM) that is the cross-product of the individual HMMs for the various speakers. In an FHMM each state is a composition of one state from the HMMs for each of the speakers, reflecting the fact that the individual speakers may have been in any of their respective states, and the final output is a combination of the output from these states. Figure 1 illustrates the dynamics of an FHMM for two speakers.

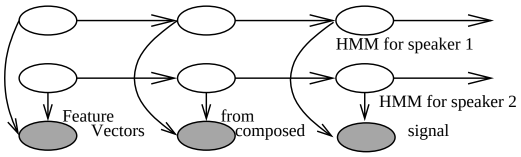

Fig. 1. Factorial HMM for two speakers (two chains).

For simplicity, we focus on the two-speaker case. Extension to more speakers is straightforward. Let S_i^k represent the i^{th} state of the HMM for the k^{th} speaker (where $k \in \{1,2\}$). Let S_{ij}^{kl} represent the factorial state obtained when the HMM for the k^{th} speaker is in state *i* and that for the l^{th} speaker is in state *j*. The output density of S_{ij}^{kl} is a function of the output densities of its component states:

$$P(X|S_{ij}^{kl}) = f(P(X|S_i^k), P(X|S_j^l)) \tag{5}$$

The precise nature of the function $f()$ depends on the proportions to which the signals from the speakers are mixed in the current estimate of the desired speaker's signal. This in turn depends on several factors including the original signal levels of the various speakers, and the degree of separation of the desired speaker effected by the current set of filters. Since these are difficult to determine in an unsupervised manner, $f()$ cannot be precisely determined.

We do not attempt to estimate $f()$. Instead, the HMMs for the individual speakers are constructed to have simple Gaussian state output densities. We assume that the state output density for any state of the FHMM is also a Gaussian whose mean is a linear combination of the means of the state output densities of the component states. We define m_{ij}^{kl}, the mean of the Gaussian state output density of S_{ij}^{kl} as:

$$m_{ij}^{kl} = \mathbf{A}^k m_i^k + \mathbf{A}^l m_j^l \tag{6}$$

where m_i^k represents the D dimensional mean vector for S_i^k and $\mathbf{A}^k$ is a $D \times D$ weighting matrix. The covariance of the factorial state S_{ij}^{kl} is also similarly modelled.

The various $\mathbf{A}^k$ values and the covariance's parameters are unknown and must be estimated from the current estimate of the speaker's signal. The estimation is performed using the expectation maximization (EM) algorithm. In the expectation (E) step of the algorithm, the *a posteriori* probabilities of the various factorial states, and thereby the *a posteriori* probabilities of the states of the HMMs for the speakers, are found. The factorial HMM has as many states as the product of the number of states in its component HMMs and direct computation of the E step is prohibitive. We therefore take the variational approach proposed by Ghahramani *et. al.* [7] for the computation.

Update formulae for all the parameters are obtained in the maximization (M) step. We forego the presentation of the mathematical details of the algorithm here and refer the interested reader to [5].

Once the EM algorithm converges and the covariance terms are computed, the best state sequence for the desired speaker can also be obtained from the FHMM, also using the variational approximation.

The overall system to determine the target for a speaker now works as follows: Using the feature vectors from the unprocessed signal and the HMMs found using the transcriptions, the different parameters are iteratively updated until the total log-likelihood converges.

Thereafter, the most likely state sequence through the desired speaker's HMM is found. Once the target is obtained, the filters are optimized, and the output of the filter-and-sum array is used to reestimate the target. The system is said to have converged when the target does not change on successive iterations.

A schematic of the overall system is shown in figure 2.

5. EXPERIMENTAL EVALUATION

In [5] we report experiments where we show that the proposed algorithm is highly effective at separating speech signals with known transcript. In those experiments, the signal transcripts were used to compose an utterance-specific HMM from the components of a large vocabulary speech recognizer. The HMMs were then used to estimate filters for the signals. The filter-and-sum processor incorporating the estimated filters were then applied to the signals obtain the separated signal. It was shown that the procedure was able to extract the signals from a background speaker that were 20dB below those from a foreground speaker.

The goal of the experiments reported in this section is to evaluate the generalizability of the estimated filters to other signals. For these experiments, simulated mixed-speaker recordings were generated using utterances from the test set of the Wall Street Journal(WSJ0) corpus. Room simulation impulse response filters were designed for a room 4m × 5m × 3m with a reverberation time of 200msec. The microphone array configuration consisted of 8 microphones placed around an imaginary 0.5m × 0.3m flat panel display on one of the walls. To obtain mixed recordings, two speech sources were placed in different locations in the room. A room impulse response filter was created for each source/microphone pair. The clean speech signals for both sources were passed through

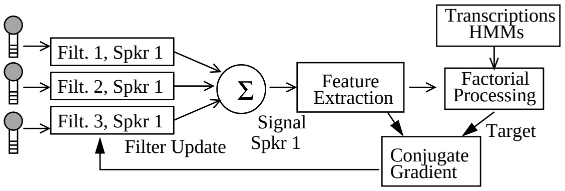

Fig. 2. Complete signal separation system for speaker 1.

each of the 8 speech source room impulse response filters and then added together.

Mixed training recordings were generated using two utterances (of known transcript), each from a different speaker. A different position in the room was assigned to each speaker. Filters were estimated for each of the speakers in the training mixture using the algorithm described in this paper. For the test data, mixed recordings were generated using other utterances, both with utterances from the same speakers as in the training recording, and with recordings from new speakers. The locations of the speakers in the test recordings were also varied, with recordings being generated both from the same locations as the training speakers, and from other locations.

Table 1 shows the separation results for four typical mixed recordings, obtained with filters estimated from a single training recording. The separation results for the training utterances are also shown. The table gives the ratio of the energy of the signal from the desired speaker to that from the competing speaker, measured in decibels, in the separated signals. We refer to this measurement as the “speaker-to-speaker ratio”, or the SSR. The higher this value, the higher the degree of separation obtained for the desired speaker.

The first row in 1 (labeled Delay & Sum) shows separation results obtained with a default comparator. Since the basic approach is that of beamforming, we designated simple delay-and-sum processing [6] as the comparator. Here the signals are simply aligned to cancel out the delays from the desired speaker to the microphone (computed here with full prior knowledge of speaker and microphone positions) and added. The second row in 1 shows the results for the filter-and-sum processing using the filter for the desired speaker. The columns in the table are arranged pair-wise. Each column reports the separation performance obtained for one of the two speakers. In all cases, the SSR for the desired speaker is reported.

In the experiment, we measure the similarity between training and test signals by two factors: relative distance to the speakers positions on the training signal and whether the test and training utterances are generated by the same speakers or not. These values are given in the table just above the “sp*/sp*” labels. The first set of test signals has 0.0 relative distance from the locations of the speakers in the training utterances, and is generated by the same speakers as in the training utterances. The separation results are as good as with the training signal. This shows that the filters are well able to generalize to other utterances by the same speakers in the same location. Test signal set 2 is also composed from utterances by the same speakers. However their relative position with respect to the speaker positions in the training signals has a difference of 1.48m. The separation results are still good, showing that the filters are relatively robust to minor fluctuations in speaker position. Test signal set 3 is also composed from utterances by the same speakers as in the training signal but the speaker positions were swapped. This had drastic influence on separation performance: here the filter sets for both speakers retrieved the signal from the foreground speaker. The last set of test signals corresponds to two different speakers placed in the same positions as in the training sequences. In this test both filter sets retrieved the signal from the background speaker.

The results suggest that the filters learn both speaker specific frequency characteristics, as well as the spatial characteristics of the speakers. Also, for a given set of speakers, the estimated filters are relatively robust to small variations in speaker location.

		Training		Unseen1		Unseen2		Unseen3		Unseen4	
Relative Distance	Same Speakers	0.00m, yes		0.00m, yes		1.48m, yes		2.54m, yes		0.00m, no	
		Sp1/Sp2	Sp2/Sp1	Sp1/Sp2	Sp2/Sp1	Sp1/Sp2	Sp2/Sp1	Sp1/Sp2	Sp2/Sp1	Sp1/Sp2	Sp2/Sp1
Delay&Sum		-11dB	+12dB	-11dB	+12dB	-12dB	+13dB	-12dB	+14dB	+2dB	+1dB
Filter&Sum		+36dB	+24dB	+35dB	+23dB	+34dB	+18dB	-40dB	+29dB	+46dB	-8dB

Table 1. SSRs obtained for different signals. For the training signal, sp1 represents for the background speaker, and sp2 for the foreground speaker.

6. CONCLUSIONS AND FUTURE WORK

The proposed algorithm is observed to result in filter-and-sum array processors that are specific to the speaker and the speaker locations represented by the training utterances, but not to the actual contents of the utterance themselves. Further they are observed to be robust to small variations in speaker location. While this immediately presents the possibility of using such a technique in situations such as meeting transcription, where speakers are relatively stationary and can be expected to be willing to record a calibration utterance, the greater implication is the feasibility of online algorithms that are based on the same principle. We note that the specific implementation described in this paper and [5] does not lend itself simply to online implementations. However, online implementations become feasible with relatively minor modification of the statistical models and the objective functions used in filter estimation, provided a good initial value is available for the filters. Hence, the observed speaker specificity and the robustness to minor variations in position, suggest that the algorithm can be extended to continually track a specific speaker provided the speaker location changes relatively slowly. Future work will explore this possibility.

7. REFERENCES

[1] S. Roweis, “One Microphone Source Separation.,” *Neural Information Processing Systems* 2000.

[2] J. Hershey and M. Casey “Audio Visual Sound Separation Via Hidden Markov Models,” *Neural Information Processing Systems* 2001.

[3] A. Hyvärinen, “Survey on Independent Component Analysis,” *Neural Computing Surveys* 1999.

[4] A. Belouchrani and J.-F. Cardoso “Maximum Likelihood Source Separation by the Expectation-Maximization Technique: Deterministic and Stochastic Implementation,” *Proc. 1995 International Symposium on Non-Linear Theory and Applications*, Las Vegas, NV, pp. 49-53, 1995.

[5] M.J. Reyes-Gomez, B.Raj and D. Ellis, “Multi-channel source separation by beamfroming trained with factorial HMMs” *ICASSP 2003*, Hong Kong 2003.

[6] D.H. Johnson and D.E. Dudgeon “Array signal Processing,” *Signal Processing Series, Prentice Hall* 1992.

[7] Z. Ghahramani and M.I. Jordan, “Factorial Hidden Markov Models,” *Machine Learning, Kluwer Academic Publishers*, Boston 1997.