A COUPLED HMM FOR SOLVING THE PERMUTATION PROBLEM IN FREQUENCY DOMAIN BSS

Saeid Sanei, Wenwu Wang, and Jonathon A. Chambers

Centre for DSP Research, King's College London, UK, [saeid.sanei, wenwu.wang, Jonathon.chambers]@kcl.ac.uk

ABSTRACT

Permutation of the outputs at different frequency bins remains as a major problem in the convolutive blind source separation (BSS). In this work a coupled Hidden Markov model (CHMM) effectively exploits the psychoacoustic characteristics of signals to mitigate such permutation. A joint diagonalization algorithm for convolutive BSS, which incorporates a non-unitary penalty term within the cross-power spectrum-based cost function in the frequency domain, has been used. The proposed CHMM system couples a number of conventional HMMs, equivalent to the number of outputs, by making state transitions in each model dependent not only on its own previous state, but also on some aspects of the state of the other models. Using this method the permutation effect has been substantially reduced, and demonstrated using a number of simulation studies.

1. INTRODUCTION

Convolutive BSS of nonstationary signals has been introduced recently [1] [2]. In practical situations such as in radio telecommunications, telemetry, radar, sonar, and especially in the speech context the sources are often nonstationary. A number of methods have been presented to solve BSS for convolutive mixtures: (1) performing blind separation in the time domain by extending the existing instantaneous algorithms. There are, however, two major problems with this method; first, it cannot cope with the nonstationary signals efficiently, and second the unmixing matrix may not be causal [3]. The later problem prevents an online separation of the sources. (2) Decomposing the problem rather than to learn the possibly huge filter all at once, i.e. the decomposition approach [4]; (3) exploiting the statistical special structure contained within the source signals to formulate various separation criteria [1]; (4) Transferring the mixtures into the frequency domain and apply BSS in each frequency bin, as an easy, effective and straightforward way to separate the nonstationary convolutive mixtures [5] [6] [2]. Assuming short-term stationarity of the data, a short term Fourier transform (STFT) is utilized to transform the signal segments into the frequency domain. In this case the convolutive BSS problem is totally or partially transformed into multiple short-term instantaneous problems. The instantaneous mixtures are then separated in every frequency bin. As for the other BSS methods, there are ambiguities due to the change in sign, scale, spectral shape, and permutation, but all except permutation can essentially be ignored. The permutation problem has been addressed in the literature and some solutions have been given [7]. In this paper a new method based on CHMM is developed. CHMMs have been introduced to better model multiple interacting time series processes [8]. The proposed CHMM system readjusts the permuted outputs by coupling a number of conventional HMMs, equivalent to the number of outputs, by making state transitions in each model dependent not only on its own previous state, but also on some aspects of the state of the other models.

2. CONVOLUTIVE BSS IN FREQUENCY DOMAIN

Consider N source signals are received by M sensors, where $M \geq N$. The output of the jth sensor is modelled as a weighted sum of convolutions of the source signals corrupted by additive noise, that is

$$x_j(n) = \sum_{i=1}^{N} \sum_{p=0}^{P-1} h_{jip} s_i(n-p) + v_j(n) \qquad (1)$$

where h_{jip} is the P-point impulse response from source i to sensor j ($j = 1, \ldots, M$), s_i is the ith source signal, x_j is the received mixture by the jth sensor, v_j is the additive noise, and n is the discrete time index. x_j are converted into frequency-domain time-series, $X_j(\omega,t)$, using the Discrete Fourier Transform. Assuming the mixing and the unmixing systems are time invariant [1], a linear convolution can be approximated by circular convolution if $P \ll T$;

$$\mathrm{X}(\omega,t) = \mathrm{H}(\omega)\mathrm{S}(\omega,t) + \mathrm{V}(\omega,t) \qquad (3)$$

where $\mathrm{S}(\omega,t) = [S_1(\omega,t), \cdots, S_N(\omega,t)]^T$ and $\mathrm{X}(\omega,t) = [X_1(\omega,t), \cdots, X_M(\omega,t)]^T$ are the time-frequency representations of the source signals and the observed signals respectively. An unmixing matrix is then developed in order to reconstruct the source signals as

$$\mathrm{Y}(\omega,t) = \mathrm{W}(\omega)\mathrm{X}(\omega,t) \quad (4)$$

Here $\mathrm{Y}(\omega,t) = [Y_1(\omega,t),\cdots,Y_N(\omega,t)]^T$ is the time-frequency representation of the output signals. The parameters of $\mathrm{W}(\omega)$ are determined so that the outputs are mutually independent.

Based on the separation in the frequency domain the multiple covariance matrices estimated at different time lags are simultaneously approximately diagonalized for the transformed convolutive mixtures. The separation criterion, or the cost function, is a minimisation of the squared error between the covariance matrix of $\mathrm{Y}(\omega,t)$ and the diagonal covariance matrix of the source signals $\mathrm{S}(\omega,t)$, which is approximated by the diagonal covariance matrix of the output signals $\mathrm{Y}(\omega,t)$ i.e.

$$J(\mathrm{W}) = \arg\min_{\mathrm{W}} \sum_{\omega=1}^{T} \sum_{k=1}^{K} J_M(\mathrm{W})(\omega,k) \quad (5)$$

where $J_M(\mathrm{W})(\omega,k)$ is defined as

$$J_M(\mathbf{W})(\omega,k) = \left\| \mathbf{R}_Y(\omega,k) - diag[\mathbf{R}_Y(\omega,k)] \right\|_F^2 \quad (6)$$

where $\left\| \cdot \right\|_F^2$ is the squared Frobenius norm, $\mathbf{R}_Y(\omega,k)$ is the output covariance matrix, and diag(.) is an operator which zeros the off-diagonal elements of the matrix. Since $\mathbf{W}(\omega) = 0$ leads to a trivial solution, the cost function is modified by effectively incorporating a penalty term using a constraint on $\mathrm{W}(\omega)$ to prevent this degenerate solution at each iteration. Using a non-unitary matrix constraint with the form

$$J_c(\mathrm{W})(\omega,k) = diag[\mathrm{W}(\omega) - \mathrm{I}][\eta\mathrm{I} - (1-\eta)\mathrm{W}(\omega)] \quad (7)$$

where $\mathbf{I}$ is an $M \times M$ unitary matrix and η is a Lagrange multiplier. Then we have

$$J(\mathrm{W}) = \arg\min_{\mathrm{W}} \sum_{\omega=1}^{T} \sum_{k=1}^{K} \left\{ J_M(\mathrm{W})(\omega,k) + \lambda J_c(\mathrm{W})(\omega,k) \right\}$$

where λ is a weighting factor. The parameter η provides a compromise between the separation performance and the convergence speed [2]. Regarding the least squares (LS) solution to minimise the above cost function the following update equation is achieved.

$$\mathrm{W}_{l+1}(\omega) = \mathrm{W}_l(\omega) - \mu(\omega).\frac{\partial J(\mathrm{W})}{\partial \mathrm{W}_l^*(\omega)} \quad (9)$$

Some criteria have also been introduced for adaptation of the iteration step size $\mu(\omega)$ [2].

Although the algorithm effectively separates the independent components there is still indeterminacy in separating the actual sources due to the inherent permutation problem. In above method, when we try to combine the results from the individual frequency bins in the time domain, the permutation problem occurs because of the inherent permutation ambiguity in the rows of $\mathbf{W}(\omega)$. The existing methods try to solve the problem in the following ways: (1) Constraints on the filter models in the frequency domain [7] [1]; (2) exploiting the continuity of the spectra of the recovered signals [9]; (3) co-modulation of different frequency bins [10]; (4) using a time-frequency source model [7] and finally (5) using a beamforming view to align solutions [11]. Short-term stationarity of the signals is efficiently exploited here in construction of a CHMM model by coupling the sequential frames of the output signals.

3. SOLUTION TO PERMUTATION PROBLEM USING CHMM

The frequency-domain BSS (FD-BSS) algorithms are assumed to be invariant to scaling and permutation of the separated frequency bin signals. The scaling can cause the scaling of every frequency band to be different resulting in spectral deformation of the original sources. As suggested in [7] the scaling problem can be remedied by forcing the determinant of the unmixing matrices to unity. This prevents alteration of the spectral envelope, while preserving the separation. On the other hand permutation indeterminacy is still an open problem. In places where there is no severe spectral deformation and the number of sources is low, the uniformity of the spectrum may be exploited in readjusting the weights of the unmixing matrix to alleviate the problem. However, a systematic approach to the problem is required where the number of sources is high.

To develop an effective solution to the permutation problem an effective way is to take the psychoacoustic model of the speech signals into account. As a simple manifestation of such a model is that the pitch frequency of the speakers are almost fixed and different from each other's. Also, the third formant for each speaker does not vary dramatically, or it is slow varying. However, the position of the other formants can be predicted using a simple autoregressive model. The overall spectrum is then approximated. Here, a number of HMMs equivalent to the number of the sources, coupled to each other, can be used to effectively track the direction of separation and ultimately prevent permutation. The number of states in each layer is identical to the number of frequency bins. The proposed CHMM system is learned and classifies based on the peak value at each frequency bin. Figure 1 shows the model for a system of two sources.

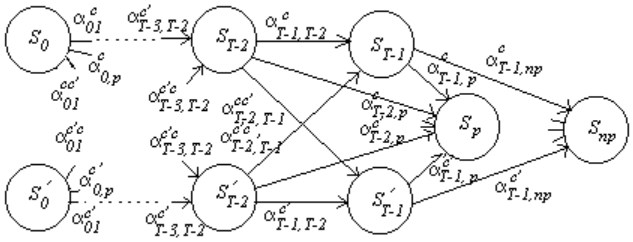

Fig. 1. The proposed CHMM model for solving the permutation problem for two sources (C=2). S_p denotes the permutation state and S_{np} refers to the state where there is no permutation.

The CHMM is trained based on the previous frames and the estimated spectrum of the current frame. T refers to the number of frequency bins in this case equivalent to the number of states in each layer. S_{np} is the state, which confirms that there is no permutation. Similarly, S_p is the state, which confirms that there is a permutation.

3.1. CHMM Formulation

The transition probabilities, a_{ij}, are determined as the result of a learning algorithm. In this model $P(S_t|S_{t-1})$, probability of being in state S_t at time t subject to being in state S_{t-1} at time t-1, for a standard HMM, is replaced by $P(S_t^{(c)}|S_{t-1}^{(1)},S_{t-1}^{(2)},\ldots,S_{t-1}^{(C)})$. The major problem here is to estimate this joint probability density function (pdf). The best way to simplify the problem is to replace the joint pdf by a linear combination of marginal probabilities as $\sum_{c'=1}^{C}\theta_{c'c}P(S_t^{(c)}\mid S_{t-1}^{(c')})$. $\theta_{c'c}$ s are the coupling parameters representing the coupling strengths between the two objects c' and c, $1\le c,c'\le C$, where C is our number of layers equivalent to the number of the speakers. In the case of having two sources, $\theta_{c'c}=\alpha_{k-1,k}^{c'c}$, $0\le k\le T-1$, and C=2. Thus the proposed CHMM is characterized by a quadruplet $\lambda=(\pi,A,B,\theta)$, where π is the initial condition, $A=\{\alpha_{ij}\}$ is the matrix of transition probabilities, $B=\{b_j\}$ is the symbol probability vector and $\theta=\{\theta_{c'c}\}$ is the new interaction parameter in the CHMM formulation. For C HMMs coupled together, the extended forward and backward variables should be defined jointly across C HMMs as

$$\alpha_t(j_1,\ldots,j_C)=P(o_0,\ldots,o_t,S_{t,j_1},\ldots,S_{t,j_C}|\lambda) \qquad (10)$$

and

$$\beta_t(j_1,\ldots,j_C)=P\left(o_{t+1},\ldots,o_{T-1}\middle|S_{t,j_1},\ldots,S_{t,j_C},\lambda\right). \qquad (11)$$

Since the conventional modified variables require high computational complexity the following modified iterative method is used to calculate the forward variables inductively [12].

1. Initialisation:

$$\alpha_1^{(c)}(j)=\pi_j^{(c)}.b_j^{(c)}(o_1^{(c)}) \qquad (12)$$

2. Induction:

$$\alpha_t^{(c)}(j)=b_j^{(c)}(o_t)\sum_{c'}\theta_{c'c}\sum_i\left(\alpha_{t-1}^{(c')}(i).a_{ij}^{(c',c)}\right),\ t>1 \qquad (13)$$

3. Termination:

$$P(O|\lambda)=\prod_c\left(\sum_j\alpha_T^{(c)}(j)\right) \qquad (14)$$

where $b_j^{(c)}(o_t)$ is the probability of observing o_t in state j.

3.2. Training the CHMM

Instead of using an EM algorithm [12], to avoid the computational complexity, an approach described by Baum [13] based on self-mapping transformation, for learning the CHMM is followed. The convergence of the algorithm has been guaranteed [13]. The transformation is motivated by the optimality condition of standard Lagrange multiplier method and leads to an iterative reestimation procedure. Based on the iterative optimisation procedure for learning the parameters [13] it can be verified that $P=P(O|\lambda)$ can be locally maximized when $\alpha_{ij}^{(c',c)}$ is transformed to

$$\alpha_{ij}^{(c',c)}\rightarrow\frac{\alpha_{ij}^{(c',c)}\partial P/\partial\alpha_{ij}^{(c',c)}}{\sum_k\alpha_{ik}^{(c',c)}\partial P/\partial\alpha_{ik}^{(c',c)}} \qquad (15)$$

By changing "$\rightarrow$" to the "=" sign the values for $a_{ij}^{(c',c)}$ are obtained. Similar procedures can be followed to find π, B, and θ parameters. The algorithm takes only a few iterations (on the order of 0.5 seconds on a P4 PC) to learn and a negligible time to classify.

4. EXPERIMENTAL RESULTS

Similar to [2], for artificially convolved mixtures, the source signals are downloaded from the website http://medi.uni-oldenburg.de. Both signals are sampled at 12kHz. The samples are 16-bit 2's complement in little endian format. The sources are mixed using $H_{11}(z)$ = 1+1.9 z^{-1} – 0.75 z^{-2}, $H_{21}(z)$ = - 0.7 z^{-5} – 0.3 z^{-6} + 0.2 z^{-7}, $H_{12}(z)$ = 0.5 z^{-5} + 0.3 z^{-6} 0.2 z^{-7}, $H_{22}(z)$ = 0.8 – 0.1 z^{-1}. For a frame length of 6000 samples, the weights are initialised at $W_0(\omega)$, a fixed μ =1, η =0.1 and λ = 0.01 (for the best result), we compared the results by comparing the error

as $\varepsilon^2 = E[\| y - s \|^2]$ with the results of the same method when the permutation is not considered, and also the results of Parra's algorithm ($\lambda = 0$) in the following table.

Table 1. The comparison between the three BSS systems, in terms of the estimation error:

	Parra's method ($\lambda = 0$)	Without CHMM ($\lambda = 0.1$)	With CHMM ($\lambda = 0.1$)
ε^2	-25 dB	-38 dB	-40 dB

A comparison between the spectrum of the separated signals without and with compensation of the permutation is given in Figure 2. From the figure it is clear that the permutation has been compensated for a number of bins; observe for example, the improved continuity in the spectrum of 2.(c) over the interval 1000-2000 Hz.

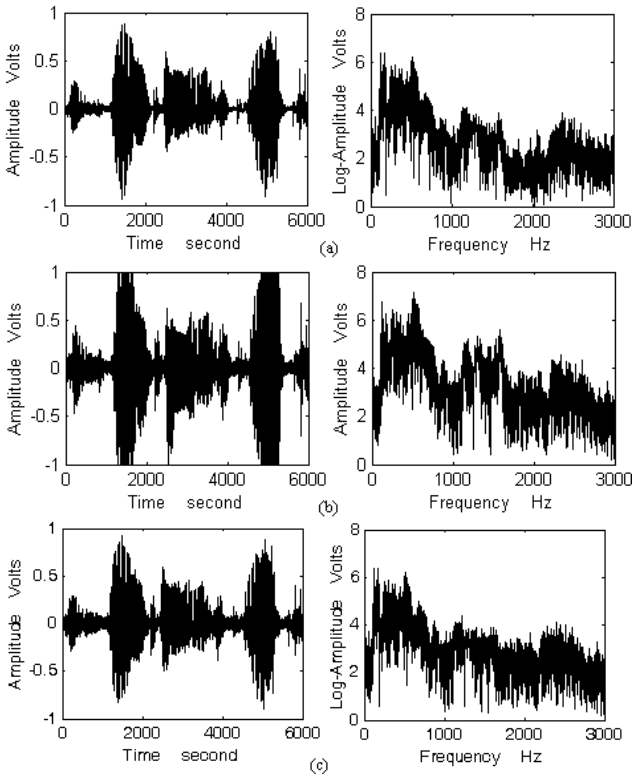

Fig. 2. A comparison between the signals (only one of the signals), (a) the original signal and its spectrums, (b) the reconstructed signal without CHMM, and (c) the separated signal after using CHMM.

For the real room recording the microphone sounds are downloaded from http://www.esp.ele.tue.nl/. The room size was a $3.4 \times 3.8 \times 5.2$ m^3, and the microphones spaced 58 cm apart. The sampling frequency and the bitrate were 12 kHz and 16 bits/sample respectively. The subjective comparison verifies the improvement achieved as a result of application of the proposed CHMM to avoid the permutation problem.

5. CONCLUSIONS AND FUTURE WORK

A new method based on a CHMM has been presented here for solving the permutation problem of the convolutive BSS of nonstationary sources in the frequency domain. The objective (for when the source signals are available) and subjective results show a remarkable improvement in the system performance. The proposed CHMM can be modified to take all the psychoacoustic parameters of the speech signals into account. This will result in a more accurate system at the price of an increase in complexity and the computation time. The efficacy of this method is likely to vary with the nature of the speech interval.

REFERENCES

[1] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. on SAP, pp. 320-327, 2000.*

[2] W. Wang, J. A. Chambers, and S. Sanei, "A joint diagonalization method for convolutive blind separation of nonstationary sources in the frequency domain," *Proc. of ICA2003, Nara, Japan, pp. 939-944, 2003.*

[3] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Novel on-line algorithms for blind deconvolution using natural gradient approach," *Proc. SYSID-97, Japan, pp. 1057-1062, 1997.*

[4] Mansour, A., Jutton C. and Loubaton P., "Adaptive subspace algorithm for blind separation of independent sources in convolutive mixtures," *IEEE Trans. On SP, vol. 48, pp. 583-586, Feb. 2000.*

[5] Smaragdis P., "Information theoretic approaches to source separation," *Master's thesis, MIT Media Lab, June 1997.*

[6] Schobben W.E. and Sommen C.W., "A frequency domain blind source separation method based on decorrelation," *IEEE Trans. on SP, vol. 50, pp. 1855-1865, Aug. 2002.*

[7] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing, vol. 22, pp. 21-34, 1998.*

[8] Razek I. and Roberts S.J. "Estimation of coupled hidden Markov models with application to biosignal interaction modelling," *Proc. IEEE Int. Conf. on Neural Network for Signal Processing, vol. 2, pp. 804-813, 2000.*

[9] V. Capdevielle, C. Serviere, and J. L. Lacoume, "Blind separation of wide-band sources in the frequency domain," *Proc. ICASSP95, pp. 2080-2083, 1995.*

[10] J. Anemuller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," *Proc. ICA2000, Helsinki, Finland, pp. 215-220, June 2000.*

[11] L. C. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. on SAP, vol. 10, no. 6, pp. 352-362, Sept. 2002.*

[12] L.K. Saul, and M. I. Jordan "Mixed memory Markov models: Decomposing complex processes as mixtures of simpler ones," *Machine Learning, vol. 37, pp. 75-87, 1999.*

[13] L.E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process," *Inequalities, vol. 3, pp. 1-8, 1969.*