A GEOMETRICALLY CONSTRAINED MULTIMODAL APPROACH FOR CONVOLUTIVE BLIND SOURCE SEPARATION

S. Sanei, S. M. Naqvi, J. A. Chambers and Y. Hicks

Centre of Digital Signal Processing, Cardiff University, Cardiff, CF24 3AA, U. K.
Email: {saneis, naqvisr, chambersj, hicksya}@cardiff.ac.uk

ABSTRACT

A novel constrained multimodal approach for convolutive blind source separation is presented which incorporates video information related to geometrical position of both the speakers and the microphones, and the directionality of the speakers into the separation algorithm. The separation is performed in the frequency domain and the constraints are incorporated through a penalty function-based formulation. The separation results show a considerable improvement over traditional frequency domain convolutive BSS systems such as that developed by Parra and Spence. Importantly, the inherent permutation problem in the frequency domain BSS is potentially solved.

Index Terms— Frequency domain BSS, geometrical constraints and multimodal separation.

1. INTRODUCTION

Convolutive blind source separation (CBSS) has been a subject of considerable research recently since it attempts to address the inherent characteristics of (a real echoic) mixing environment. Generally, the main objective of BSS is to decompose the measurement signals into their constituent independent components as an estimation of the true sources which are assumed a priori to be independent.

CBSS has been conventionally developed in either the time [1] or frequency [2] [3] [4] domains. Frequency domain convolutive blind source separation (FDCBSS) however, has been more popular as the convolutive mixing is converted into a number of instantaneous mixing operations. The permutation problem inherent to FDCBSS is more severe and destructive than for time domain schemes [5]. In such systems there are no priori assumptions on the source statistics or the mixing system. On the other hand, in a multimodal approach the video system can capture the positions of the speakers and the directions they face [6]. The video information can thereby help to estimate the mixing matrix more accurately and ultimately increase the separation performance. Following this idea, the objective of this paper is to efficiently use such information in the enhancement of the separation results. The CBSS system can be described as follows: assume m statistically independent sources as $\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]^T$ where $[.]^T$ denotes transpose operation. The sources are convolved with a linear model of the physical medium (mixing matrix) which can be represented in the form of a multichannel FIR filter $\mathbf{H}$ to produce n sensor signals $\mathbf{x}(t) = [x_1(t), \ldots, x_n(t)]^T$ as

$$\mathbf{x}(t) = \sum_{\tau=0}^{P} \mathbf{H}(\tau)\mathbf{s}(t-\tau) + \mathbf{v}(t) \quad (1)$$

Work supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK.

where $\mathbf{v}(t) = [v_1(t), \ldots, v_n(t)]^T$ is the noise vector at discrete time sample t and $\mathbf{H} = [\mathbf{H}(0), \mathbf{H}(1), \ldots, \mathbf{H}(P)]$. Using time domain CBSS, the sources are estimated using a set of unmixing filters $\mathbf{W}(\tau), \tau = 0, .., Q$, such that

$$\mathbf{y}(t) = \sum_{\tau=0}^{Q} \mathbf{W}(\tau)\mathbf{x}(t-\tau) \quad (2)$$

where $\mathbf{y}(t) = [y_1(t), \ldots, y_m(t)]^T$ are the estimated sources. P and Q are respectively the lengths of the mixing and unmixing filters. The length of the signals is T. In FDBSS the problem is transferred into the frequency domain using the STFT. (1) and (2) then change respectively to:

$$\mathbf{X}(\omega, t) \approx \mathbf{H}(\omega)\mathbf{S}(\omega, t) + \mathbf{V}(\omega, t) \quad (3)$$

$$\mathbf{Y}(\omega, t) \approx \mathbf{W}(\omega)\mathbf{X}(\omega, t) \quad (4)$$

where ω denotes discrete normalized frequency. An inverse STFT is then used to find the estimated sources $\hat{\mathbf{s}}(t) = \mathbf{y}(t)$; however, this will be certainly affected by the permutation effects due to the variation of $\mathbf{W}(\omega_i)$ with ω_i. Parra's algorithm jointly diagonalizes the unmixing matrix for all the frequency bins by minimising the squared error (as the sum of off diagonal elements of the covariance matrix of the estimated sources) using the constrained gradient descent algorithm [7]. Considering the FDCBSS system developed by Parra and Spence the main cost function J_m is expressed in the form of

$$J_m = \sum_{\omega=0}^{T} \sum_{k=1}^{K} ||E(\omega, k)||_F^2 \quad (5)$$

where

$$E(\omega, k) = \mathbf{W}(\omega)[\mathbf{R}_x(\omega, k) - \mathbf{\Lambda}_v(\omega, k)]\mathbf{W}^H(\omega) - \mathbf{\Lambda}_s(\omega, k) \quad (6)$$

and $\mathbf{R}_x$, $\mathbf{\Lambda}_v$ and $\mathbf{\Lambda}_s$ are respectively the covariance matrices of the signals, noise and source signals spectra, and $\|.\|_F^2$ denotes the Frobenius norm. Ignoring the noise for simplicity, the main update equation for estimation of $\mathbf{W}(\omega_i)$ for the i-th FFT frequency is given as [2]

$$\mathbf{W}_{j+1}(\omega_i) = \mathbf{W}_j(\omega_i) - \mu \sum_{k=1}^{K} E(\omega_i, k)\mathbf{W}_j(\omega_i)\mathbf{R}_x(\omega_i, k) \quad (7)$$

where j, K and μ are the iteration index, the number of FFT points and learning rate respectively. $\mathbf{W}$ is updated for all the frequency bins ω_i and each time is initialized to the identity matrix. In the

following section we use the spacial information indicating the positions and directions of the sources using the "data" acquired instantaneously by a number of video cameras. The separation process is then constrained by such information. The comparison between the original Parra and Spence, and the proposed multimodal constrained FDCBSS, algorithms will be presented at the end.

2. THE CONSTRAINED PROBLEM

Given the position of the speakers and the microphones, the distances between the ith microphone and the jth speaker d_{ij}, and also their propagation times τ_{ij}, can be calculated (See Figure 1 for a simple two-speaker two-microphone case). Accordingly, in a homogenous medium such as air, the attenuation is related to the distances via

$$\alpha_{ij} = \frac{\kappa}{d_{ij}^2} \tag{8}$$

where κ is a constant representing the attenuation per unit length in a homogenous medium. Similarly, τ_{ij} in terms of the number of samples, is proportional to the sampling frequency f_s, sound velocity C, and the distance d_{ij} as:

$$\tau_{ij} = \frac{f_s}{C} d_{ij} \tag{9}$$

which is independent of the directionality. Both f_s and C are considered constant within each observation block for a block-based BSS system, or slowly varying in a real-time BSS process. However, in practical situations the speakers directions introduce another variable into the attenuation measurement. In the case of electronic loudspeakers (not humans) the directionality pattern depends on the type of loadspeaker. Here, we approximate this pattern as $\cos(\theta_{ij}/r)$ where $r > 2$, and has a smaller value for highly directional speakers and vice versa (an accurate profile can be easily measured using a SPL meter). Therefore, the attenuation parameters become

$$\alpha_{ij} = \frac{\kappa}{d_{ij}^2} \cos(\theta_{ij}/r) \tag{10}$$

If, for simplicity, only the direct path is considered the mixing filter is expected to have a form as:

$$\hat{H}(t) = \begin{bmatrix} \alpha_{11}\delta(t-\tau_{11}) & \alpha_{12}\delta(t-\tau_{12}) \\ \alpha_{21}\delta(t-\tau_{21}) & \alpha_{22}\delta(t-\tau_{22}) \end{bmatrix} \tag{11}$$

for which in the frequency domain the above filter has the form

$$\begin{aligned} \hat{\mathbf{H}} &= \begin{bmatrix} \alpha_{11}e^{-j\omega\tau_{11}} & \alpha_{12}e^{-j\omega\tau_{12}} \\ \alpha_{21}e^{-j\omega\tau_{21}} & \alpha_{22}e^{-j\omega\tau_{22}} \end{bmatrix} \\ &= \begin{bmatrix} \alpha_{11}z^{-\tau_{11}} & \alpha_{12}z^{-\tau_{12}} \\ \alpha_{21}z^{-\tau_{21}} & \alpha_{22}z^{-\tau_{22}} \end{bmatrix} \end{aligned} \tag{12}$$

Although the actual mixing matrix includes the reverberation terms related to the reflection of sounds by the obstacles and walls, in such a room environment it will always contain the direct path components as in the above equations. Therefore, we can consider $\hat{\mathbf{H}}$ as a biased estimate of the mixing filter and set the following constraint, which minimizes the Frobenius norm distance between the unmixing filter $\mathbf{W}$ and the permuted mixing filter $\hat{\mathbf{H}}$, i.e.

$$J_c = \|\mathbf{W} - \mathbf{P}\hat{\mathbf{H}}^{-1}\|_F^2 = \|vec(\mathbf{W} - \mathbf{P}\hat{\mathbf{H}}^{-1})\|_2^2 \tag{13}$$

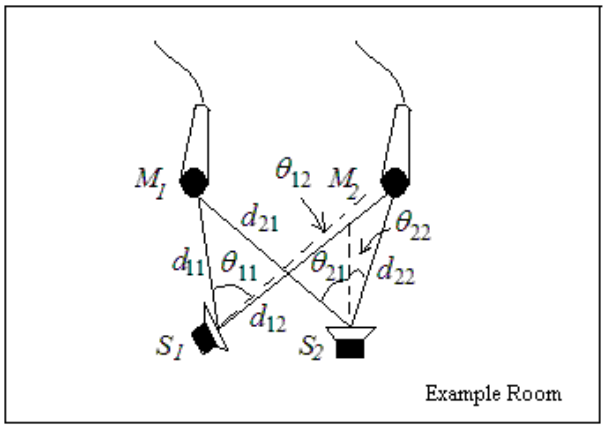

Fig. 1. A two-speaker two-microphone setup for recording within a reverberating (room) environment; only distances and angles between sources and microphones are shown.

where $\|.\|_2^2$ represent respectively, the Euclidean norm, $vec(.)$ converts a matrix argument column-wise into a column vector, and $\mathbf{P}$ is the permutation matrix. Ultimately, the cost function J_c has to be minimized with respect to both $\mathbf{W}$ and $\mathbf{P}$.

3. THE OVERALL CONSTRAINED BSS

In order to achieve the above goal, we need to minimize jointly J_m and J_c with respect to $\mathbf{W}$, and also minimise J_c with respect to the permutation matrix $\mathbf{P}$. The constrained optimisation problem can be changed to an unconstrained one using a Lagrangian approach or by means of a penalty function as in [8]. In this case

$$J(\mathbf{W}(\omega)) = J_m(\mathbf{W}(\omega)) + \lambda J_c(\mathbf{W}(\omega)) \tag{14}$$

where λ is the Lagrange multiplier. $\mathbf{W}$ and $\mathbf{P}$ are then found by minimizing the gradients of J and J_c respectively with respect to $\mathbf{W}$ and $\mathbf{P}$, i.e.

$$\mathbf{W}_{opt}(\omega) = \arg\min_{\mathbf{W}}\{J_m(\mathbf{W}(\omega)) + \lambda J_c(\mathbf{W}(\omega))\} \tag{15}$$

and

$$\mathbf{P}_{opt}(\omega) = \arg\min_{\mathbf{P}}\{J_c(\mathbf{W}(\omega))\} \tag{16}$$

Therefore, at each frequency bin ω_i the estimated sources will be aligned with the input source signals; as one of the major advantages of this algorithm there will not generally remain any permutation problem. Consequently, the update equations are obtained as:

$$\mathbf{W_{j+1}}(\omega) = \mathbf{W_j}(\omega) - \mu\nabla_{\mathbf{W}}(J(\mathbf{W_j}(\omega))) \tag{17}$$

$$\mathbf{P_{j+1}}(\omega) = \mathbf{P_j}(\omega) - \eta\nabla_{\mathbf{P}}(J_c(\mathbf{W_j}(\omega))) \tag{18}$$

where j is the iteration index, μ and η are the learning rates, and

$$\begin{aligned} \nabla_{\mathbf{W}^*}(J(\mathbf{W})) &= \nabla_{\mathbf{W}^*}(J_m(\mathbf{W})) + \lambda\nabla_{\mathbf{W}^*}(J_c(\mathbf{W})) \\ &= 2\sum_{k=1}^{K} E(\omega,k)\mathbf{W}(\omega)R_x(\omega,k) \\ &\quad +2\lambda[\mathbf{W}(\omega) - \mathbf{P}(\omega)\tilde{\mathbf{H}}^{-1}(\omega)] \end{aligned} \tag{19}$$

and

$$\nabla_{\mathbf{P}}(J_c(\mathbf{W})) = -2\tilde{\mathbf{H}}^{-1}(\omega)[\mathbf{W}(\omega) - \mathbf{P}(\omega)\tilde{\mathbf{H}}^{-1}(\omega)] \quad (20)$$

Before starting the update process $\tilde{\mathbf{H}}^{-1}(\omega)$ is normalised once using $\tilde{\mathbf{H}}^{-1}(\omega) \leftarrow \tilde{\mathbf{H}}^{-1}(\omega)/\|\tilde{\mathbf{H}}^{-1}(\omega)\|_F$ where $\|.\|_F$ denotes the Frobenius norm and after each iteration $\mathbf{W}(\omega_i)$ is also normalised. In the case of fractional filters where the distances between the speakers and the microphones are not integer multiples of the sampling interval, a previously developed algorithm to firstly estimate the fractional delay and then perform the **BSS** process [9] [10] can be used.

4. EXPERIMENTAL RESULTS

Two experiments were carried out; in the first experiment two single tones were used. The mixing matrix $\mathbf{H}$ was carefully chosen to model the room environment and $\tilde{\mathbf{H}}$ was selected to include only the direct path and the angle of departures θ (r considered to be 4). Both the Parra and Spence algorithm, and the proposed constrained FDCBSS were employed and the signal-to-interference ratio (SIR) was calculated as [2]

$$SIR = \frac{\Sigma_i \Sigma_\omega |H_{ii}(\omega)|^2 \langle |S_i(\omega)|^2 \rangle}{\Sigma_i \Sigma_{i \neq j} \Sigma_\omega |H_{ij}(\omega)|^2 \langle |S_j(\omega)|^2 \rangle} \quad (21)$$

Using the Parra and Spence algorithm the SIR was 6.1dB and using the CBSS the SIR achieved was 9.25dB. The 3dB superior performance is not only because of application of the geometrical constraints but also as a result of solving the permutation problem. Two major drawbacks of the system are the slight increase in the complexity and potential slower rate of convergence. In the second experiment, the Parra and Spence algorithm and the proposed CBSS were tested for a real room recording. The variables were selected as: $d_{11} = 24$ cm, $d_{12} = 50$ cm, $d_{21} = 40$ cm, $d_{22} = 18$ cm, $r = 4$, $\theta_{11} = 60^o$, $\theta_{12} = 5^o$, $\theta_{21} = 45^o$, and $\theta_{22} = 45^o$. λ is empirically chosen (here $\lambda = 0.15$) and the learning rates μ and η gradually decreased with respect to the iteration index j

$$\mu_j = \eta_j = \gamma \frac{0.02}{1 - (0.98)^j} \quad (22)$$

where γ is a constant equal to $\gamma = 0.01$. Figure 2 shows the original signals using a couple of microphones very close to the mouth of the speakers, the mixed signals, the separated signals using the Parra and Spence algorithm, and the estimated signals using our proposed CBSS method. SIRs for $P = Q = 1024$ have been calculated according to [2]. In this experiment we achieved $\mathbf{SIR}_{\mathbf{Parra's}} = 6.8$dB and $\mathbf{SIR}_{\mathbf{CBSS}} = 9.4$dB, which shows a marked improvement. In addition the filter length (DFT points) may be changed according to the room geometry to obtain even better results. Table 1 shows the SIR values for both experiments. Figure 3 illustrates the convergence graph of the cost function within the last frequency bin for both Parra and Spence, and the proposed CBSS method. As expected, by using the constraint term the convergence is slightly slower and the complexity of the system is higher.

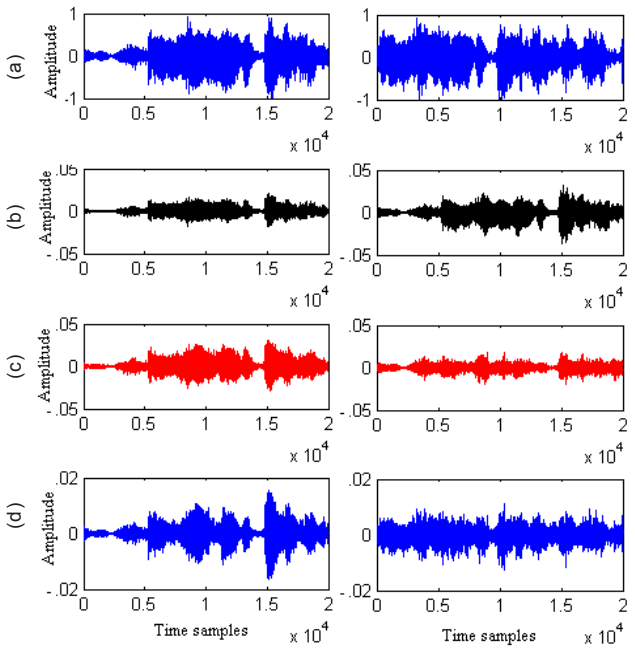

Fig. 2. (a) the original signals recorded by very close microphones, (b) the mixed signals (c) the separated sources using Parra and Spence algorithm, and (d) the estimated sources using the constrained FDCBSS.

Table 1. Comparison between Parra and Spence algorithm and the proposed method for different sets of mixtures.

SIR	Parra's Method/dB	Constrained FDCBSS/dB
Sinusoidal Signal	6.1	9.25
Speech Signal	6.8	9.4

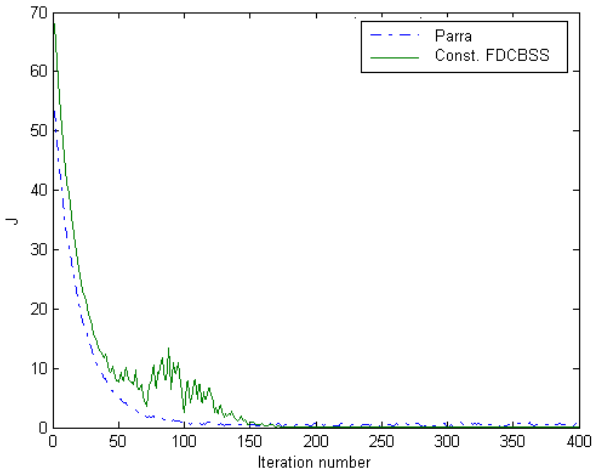

Fig. 3. The convergence graphs for both the Para and Spence, and the proposed constrained FDCBSS algorithms for only the last frequency bin.

5. SUMMARY AND CONCLUSIONS

In this paper the conventional FDCBSS algorithm has been modified by accommodating the geometrical information about the sources in a multi-modal BSS approach. The location and direction information have been obtained using a number of cameras equipped with a speaker tracking algorithm. The constrained problem has been partially changed to an unconstrained problem using Lagrange multipliers. The results show that the modified CBSS system enhances the performance of the traditional FDBSS system both objectively and subjectively. The outcome of this approach paves the way for establishing a multi-modal audio-video system for separation of speech and music signals.

6. REFERENCES

[1] A. S. Bregman, *Auditory scence analysis*, MIT Press, Cambridge, MA, 1990.

[2] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. On Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley, 2002.

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

[5] W. Wang, S. Sanei, and J. A. Chambers, "A joint diagonalization method for convolutive blind separation of nonstationary sources in the frequency domain," *Proc. ICA, Nara, Japan*, April 2003.

[6] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video assisted speech source separation," *Proc. IEEE ICASSP 2005, Pennsylvania, USA*, March 19-23.

[7] S. Haykin, *Adaptive filters*, John Wiley, 1994.

[8] W. Wang, S. Sanei, and J.A. Chambers, "Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 53, no. 5, pp. 1654–1669, 2005.

[9] C. Cheong Took, K. Nazarpour, S. Sanei, and J.A. Chambers, "Fractional differential delay estimation in local sparse component analysis of temporomandibular joint sounds," *Submitted to IEEE Transaction on Biomedical Engineering*, June 2006.

[10] C. Cheong Took, K. Nazarpour, S. Sanei, and J. Chambers, "Blind separation of temporomandibular joint sounds by incorporating fractional delay estimation," *Proc. of IEE IMA 2006, Cirincester, UK*.