Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2008) 2087–2097
NEUROCOMPUTING
www.elsevier.com/locate/neucom

An adaptive stereo basis method for convolutive blind audio source separation ☆

Maria G. Jafari[a], Emmanuel Vincent[b], Samer A. Abdallah[a], Mark D. Plumbley[a,*], Mike E. Davies[c]

[a]Centre for Digital Music, Department of Electronic Engineering, Queen Mary University of London, London E1 4NS, UK
[b]METISS Project, IRISA-INRIA, Campus de Beaulieu, 35042 Rennes cedex, France
[c]IDCOM and Joint Research Institute for Signal and Image Processing, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JL, UK

Available online 13 February 2008

Abstract

We consider the problem of convolutive blind source separation of stereo mixtures, where a pair of microphones records mixtures of sound sources that are convolved with the impulse response between each source and sensor. We propose an adaptive stereo basis (ASB) source separation method for such convolutive mixtures, using an adaptive transform basis which is learned from the stereo mixture pair. The stereo basis vector pairs of the transform are grouped according to the estimated relative delay between the left and right channels for each basis, and the sources are then extracted by projecting the transformed signal onto the subspace corresponding to each group of basis vector pairs. The performance of the proposed algorithm is compared with FD-ICA and DUET under different reverberation and noise conditions, using both objective distortion measures and formal listening tests. The results indicate that the proposed stereo coding method is competitive with both these algorithms at short and intermediate reverberation times, and offers significantly improved performance at low noise and short reverberation times.

Keywords: Blind source separation; Audio source separation; Independent component analysis; DUET algorithm; Adaptive basis; Sparse coding

1. Introduction

Convolutive blind audio source separation is a problem that arises when an array of microphones records mixtures of sound sources that are convolved with the impulse response between each source and sensor.

Several methods have been proposed to tackle this problem, either in the time domain or in the frequency domain. Time-domain methods mostly entail the extension of existing instantaneous blind source separation (BSS) algorithms to the convolutive case [5,14,35]. However, these techniques typically assume that the source signal samples are temporally independent, which can lead to over-whitening of the inputs.

Most work in audio BSS has concentrated on the frequency domain independent component analysis (FD-ICA) method [12,16,24,27,30,34]. This approach uses the short-time Fourier transform (STFT) to transform the convolved signal into the time–frequency domain, with instantaneous independent component analysis (ICA) performed separately in each frequency bin. This approach is typically simpler and computationally less complex than the time-domain approach, although it may require long STFT frames to successfully separate convolutively mixed signals. The use of separate ICA processes in each bin also introduces the well-known *permutation problem*, whereby the different frequency components of the signals become 'swapped' and require permutation to realign them.

Another approach that has been found to be successful in practical applications on stereo (two-microphone) anechoic mixtures is the degenerate unmixing estimation technique (DUET) [20,39]. Here the STFT is again used to transform the signal into the time–frequency domain.

☆This work was funded by EPSRC Grants GR/S85900/01, GR/R54620/01, and GR/S82213/01.
*Corresponding author.
E-mail address: mark.plumbley@elec.qmul.ac.uk (M.D. Plumbley).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.08.029

The relative amplitude and phase is used to estimate the dominant source in each time–frequency bin, and time–frequency masking is then used to extract the source components. While the DUET algorithm is not specifically designed for convolutive mixtures, some success has been observed if echoes are relatively minor. However, performance has been observed to degrade with increasingly echoic mixtures, and large microphone spacing can also cause problems in estimating the relative delay used by the algorithm.

In this article, we propose an adaptive stereo basis (ASB) source separation method for convolutive mixtures. Instead of using a fixed time–frequency transform such as the STFT, applied separately to each observation (microphone) channel, we learn an adaptive transform based on the observed stereo data that is applied to both channels together [2]. Many basis pairs of the resulting transform exhibit properties suggesting that they represent the components of individual sources, together with the filtering process from the sources to the microphone pair. In place of the permutation problem, in the ASB method we have a basis selection task to perform. We tackle this using the relative time delays between left and right channels of the stereo basis pairs, which correspond to different directions of arrival (DOAs) of the sources. We then have an association of each source with a subset of the stereo basis pairs, allowing us to estimate the separated sources.

We will show that this ASB method can give significantly better performance than FD-ICA and DUET for short reverberation times (RTs), and comparable performance to FD-ICA and DUET algorithm at intermediate RTs, even though it uses a smaller frame size than the FD-ICA and DUET algorithms.

The structure of this paper is as follows: the convolutive BSS problem and the FD-ICA algorithm are reviewed in Section 2, and our proposed ASB method is introduced in Section 3. The performance of the algorithm is evaluated in Section 4, followed by discussion and conclusions.

2. Convolutive BSS

2.1. Problem statement

Consider the problem of linear convolutive mixing, for example microphones recording mixed sound sources in a room with delays and echoes. Here each microphone records a linear combination of the source signals s_p, at several times and levels, as well as multipath copies (echoes) of the sources. This scenario can be modelled as a finite impulse response (FIR) convolutive mixture, given by [24]

$$x_q(n) = \sum_{p=1}^{P} \sum_{l=0}^{L_m-1} a_{qp}(l) s_p(n-l), \quad q = 1, \ldots, Q, \tag{1}$$

where $x_q(n)$ is the signal recorded at the qth microphone at time sample n, $s_p(n)$ is the pth source signal, $a_{qp}(l)$ denotes the impulse response of the mixing filter from source p to sensor q, and L_m is the maximum length of all impulse responses [32]. The source signals s_p are typically assumed to be independent. The aim of convolutive BSS is then to estimate the original source signals $s_p(n)$ and the mixing process $a_{qp}(n)$ given only the mixtures $x_q(n)$.

This problem can be approached by estimating a matrix of unmixing filters $w_{pq}(k)$ to produce an output

$$y_p(n) = \sum_{q=1}^{Q} \sum_{k=0}^{M-1} w_{pq}(k) x_q(n-k), \tag{2}$$

where $y_p(n)$ is an estimate of the original sources and M is the length of the unmixing filters, which are assumed to be sufficiently long to approximately deconvolve (1).

However, there is an inherent *filtering ambiguity* in this problem. Filtering operations in the pth source channel can typically either be considered to be part of the source s_p or in the mixing filters a_{qp} [27]. To avoid this ambiguity we instead consider the problem of estimating the *image* x_{qp} of the source s_p at the qth microphone, given by

$$x_{qp}(n) = \sum_{l=0}^{L_m-1} a_{qp}(l) s_p(n-l) \tag{3}$$

which is the contribution to

$$x_q(n) = \sum_{p} x_{qp}(n)$$

due to the pth source. While this source image approach does require the images at all Q microphones to be estimated for each of the P sources, it has the advantage that it is uniquely defined [27].

2.2. Frequency-domain ICA

Rather than attempting to construct the unmixing filters (2) directly in the time domain, a popular approach is to work in a time–frequency domain instead, leading to the approach known as FD-ICA. In FD-ICA, we divide the input sequence into frames, and approximate the mixing model (1) in the time–frequency domain by

$$\tilde{\mathbf{x}}(f,t) = \tilde{\mathbf{A}}(f)\tilde{\mathbf{s}}(f,t), \tag{4}$$

where $\tilde{\mathbf{s}}(f,t)$ and $\tilde{\mathbf{x}}(f,t)$ are the STFTs of the original sources and the observations, respectively, and $\tilde{\mathbf{A}}(f)$ is the matrix of mixing filters.

The unmixing model (2) is then approximated by

$$\tilde{\mathbf{y}}(f,t) = \tilde{\mathbf{W}}(f)\tilde{\mathbf{x}}(f,t), \tag{5}$$

where $\tilde{\mathbf{y}}(f,t)$ are the recovered source estimates in the frequency domain, and $\tilde{\mathbf{W}}(f)$ are the separating filters to be estimated. The convolutive BSS problem is thus transformed into multiple complex valued ICA problems in the time–frequency domain, with a suitable ICA algorithm (e.g. [4,11,15]) used to estimate $\tilde{\mathbf{W}}(f)$ separately in each frequency bin. Once we have the separated source estimates,