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 & Joint Research Institute for Signal and Image Processing, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JL, UK*

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.

Preprint submitted to Elsevier Science *10 August 2007*

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.

Key words: 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 [1–3]. However, these techniques typically assume that the source signals samples are temporally independent, which can lead to over-whitening of the inputs.

Most work in audio blind source separation has concentrated on the frequency domain independent component analysis (FD-ICA) method [4–9]. This approach uses the short-time Fourier transform (STFT) to transform the convolved signal into the time-frequency domain, with instantaneous independent

⋆ This work was funded by EPSRC grants GR/S85900/01, GR/R54620/01, and GR/S82213/01.

* Corresponding author.
Email address: mark.plumbley@elec.qmul.ac.uk (Mark D. Plumbley).

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) [10,11]. Here the STFT is again used to transform the signal into the time-frequency domain. 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 [12]. 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, and comparable performance to FD-ICA and DUET algorithm at intermediate reverberation times, 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 Adaptive Stereo Basis method is introduced in Section 3. The performance of the algorithm is evaluated in Section 4, followed by discussion and conclusions.

2 Convolutive Blind Source Separation

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 [4]

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

where $x_q(n)$ is the signal recorded at the q-th microphone at time sample n, $s_p(n)$ is the p-th 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 [13]. The source signals s_p are typically assumed to be independent. The aim of convolutive blind source separation 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 p-th source channel can typically either be considered to be part of the source s_p or in the mixing filters a_{qp} [9]. To avoid this ambiguity we instead consider the problem of estimating the *image* x_{qp} of the source s_p at the q-th 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 p-th 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 [9].

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 *frequency-domain ICA* (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 short-time Fourier transforms (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.* [14–16]) used to estimate $\tilde{\mathbf{W}}(f)$ separately in each frequency bin. Once we have the separated source estimates, we can estimate the source images $\hat{\tilde{\mathbf{x}}}_{qp}(f,t)$ using the estimate $\hat{\tilde{\mathbf{A}}}(f) = \tilde{\mathbf{W}}^{-1}(f)$ for the mixing process [9].

The use of separate ICA algorithms for each frequency bin f in (5) leads

to the well-known *permutation problem*. Due to the inherent ambiguity in the identification of the sources, any ICA algorithm can only find a set of original sources relative to some unknown permutation. Since these are applied independently to each frequency bin, a further process is required to match the source estimates $\tilde{\mathbf{y}}(f,t)$ at a particular frequency bin f with those at other frequency bins.

A wide variety of methods have been proposed to address this permutation problem [5–8]. One interesting approach is to consider the spatial arrangement of the source and microphones: a *beamforming* approach [17–19]. If most of the signal observed at the microphones arrives from the direction of the direct path from the source, the time delay between the microphones will correspond to the *direction of arrival* (DOA) of the source. The source estimates can then be permuted so that their DOAs are aligned [17,18].

When using the beamforming approach to the permutation problem, we need to take care to avoid *spacial aliasing*. Due to the narrowband nature of the signals in each frequency bin, to ensure the estimated direction of arrival is unique, the inter-microphone spacing must satisfy $d < \lambda_{\min}/2 = c/(2f_{\max})$ where $f_{\max}$ is the maximum frequency to be aligned. If all frequency bins are to be aligned, $f_{\max}$ will be the Nyquist limit, *i.e.* half the sampling frequency. For example, with $f_{\max} = 8\,\text{kHz}$ and $c = 340\,\text{m/s}$ we get $d \leq (340/16000)\,\text{m} \approx 2.1\,\text{cm}$ [13]. If uniqueness is not satisfied, for example when the microphone spacing is too large (*e.g.* $d \approx 1\,\text{m}$), then several DOAs may correspond to a given delay, and we will have spatial aliasing. When $f > f_{\max}$ we can overcome the spacial aliasing problem either by performing DOA estimation using only the lower band of frequencies $f < f_{\max}$ [20], or by using a 'peakier' directivity pattern method based on the MuSIC algorithm [21], as proposed by

Mitianoudis and Davies [22]. We use the latter method in our comparative evaluation later in this article.

2.3 Towards an adaptive basis method

In FD-ICA the STFT was used to transform the mixture signal into the time-frequency domain to approximate the convolutive mixing process (1) by a set of parallel instantaneous narrowband mixing processes (4). A side-effect of the STFT is that many signals are *sparse* in the time-frequency domain: *i.e.* signals are zero or very small more often than it might be expected from their variances [23]. It has been noted that many ICA algorithms have improved performance when sources are sparse [24].

The method that we propose in this article is based on the search for a transform that will directly allow us to partition the transform components into subsets corresponding to each source. If we could achieve this with the single-channel STFT, this would be a simple filtering operation, assigning frequency bands (subsets of frequency bins) to each source. However, since the sources we are considering do not occupy disjoint frequency bands, we use an *adaptive* transform.

In fact, we can use ICA to learn such an adaptive transform, but instead of using it across mixtures to separate sources, we use it across time samples to search for interesting structure in the data. In an early application of this method, Bell and Sejnowski [25] found that ICA trained on time-frames of monophonic recordings of 'tooth taps' discovered features (basis vectors) exhibiting localized time and phase structure, while those learned by *e.g.* prin-

cipal components analysis (PCA) did not. Other studies on monophonic audio signals have reported that the basis vectors learned by ICA from speech signals are mostly well localised in time and frequency, yielding a representation that exhibits wavelet-like bases [26,27]. The resulting representation of the sounds transformed into this learned basis are sparse, *i.e.* with most coefficients close to zero, giving a representation reminiscent of that of auditory nerve fibres [27].

In a preliminary study [12], we investigated an extension of this technique to stereo signals, applying an ICA algorithm to sequences of stereo time frames. We found that many of the resulting basis vectors typically exhibited the wavelet-like localized time and frequency representation as for the monophonic case. However, while the frequency representation of a typical basis vector is *localized* around a particular centre frequency, it is not *narrowband* as is the case for STFT basis vectors, and a time-domain centre is normally observed. Furthermore, many bases also displayed relative amplitude differences and time delays between the two channels, suggesting that the basis vectors discovered by the algorithm represent the components of individual sources and the filtering process from the sources to each of the microphones. If this is the case, then by partitioning these bases into subsets corresponding to each of the sources, it should be possible to separate the original source signals from each other. This is the principle behind the proposed Adaptive Stereo Basis (ASB) method.

3 Adaptive Stereo Basis method

The essence of the Adaptive Stereo Basis (ASB) method is that we wish to find a basis transform of the stereo observation sequence, where the transform is such that the sources are disjointly represented in the transform space. Thus we can consider the method to be a *multidimensional ICA* (MICA) method [28], also known as independent subspace analysis (ISA). We are attempting to find a transform matrix (basis matrix) where each of the basis vectors (columns of the basis matrix) lies within an independent subspace occupied by one of the sources [29]. By grouping the transform basis vectors appropriately we can then extract the sources estimates. The method therefore uses the following sequence of steps:

(1) Reshape the observed vector sequence
(2) Learn the basis matrix
(3) Group the basis components
(4) Extract source image estimates

Each of these steps is detailed below.

3.1 Reshaping the observed vector sequence

The ASB method attempts to find a basis set that encodes both spatial and temporal correlations in the observed data. Therefore we need to reshape the sequence of stereo vectors $\mathbf{x}(n)$ into a matrix $\bar{\mathbf{X}}$, such that several stereo sample pairs $\mathbf{x}(n_1), \ldots, \mathbf{x}(n_2)$ are 'stacked' to form each vector $\bar{\mathbf{x}}(k)$ of $\bar{\mathbf{X}}$. Reshaping the input in this way allows both correlations between microphones

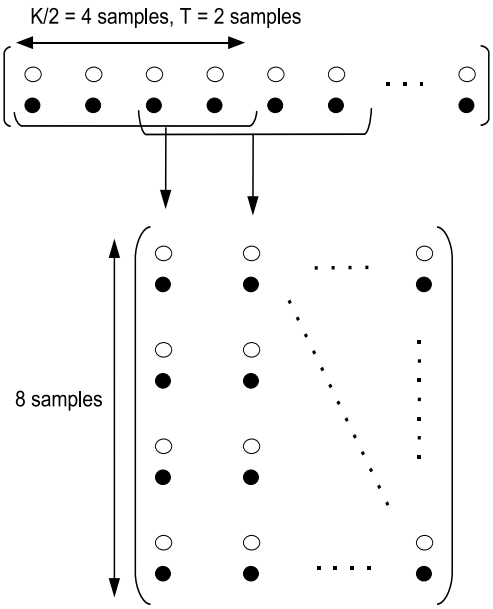

Fig. 1. Reshaping of the sensor vector prior to training with ICA. In this illustration, we have $K/2 = 4$ sample pairs per frame, with an overlap of $T = 2$ samples.

and correlations across time to be modelled.

To make this specific, the observed stereo vector sequence $\mathbf{x}(n)$ is reshaped into a $K \times k_{\max}$ matrix, where successive frames of $K/2$ stereo sample vectors are taken from each mixture, with an overlap of T samples (Figure 1). Thus, the (i,k)-th element of the new matrix, $\bar{\mathbf{X}}$, is given by

$$[\bar{\mathbf{X}}]_{i,k} = \begin{cases} x_1\left((k-1)Z + (i+1)/2\right) & : i \text{ odd} \\ x_2\left((k-1)Z + i/2\right) & : i \text{ even} \end{cases} \tag{6}$$

where $Z = K/2 - T$, and $i \in \{1, \ldots, K\}$, and $k \in \{1, \ldots, k_{\max}\}$.

3.2 Learning the basis matrix

We now wish to construct an unmixing matrix $\bar{\mathbf{W}} \in \mathbb{R}^{K \times K}$ so that each of the components of the vector sequence $\bar{\mathbf{y}}(k) = \bar{\mathbf{W}}\bar{\mathbf{x}}(k)$ will contain activity from only one of the underlying P sources. We would like the activity of each source to be represented by some subset of components of $\bar{\mathbf{y}}$, where these component subsets are mutually exclusive. Therefore this is an multidimensional ICA (independent subspace analysis) problem. To solve this multidimensional ICA problem, we use an ICA algorithm to find the unmixing matrix $\bar{\mathbf{W}}$, followed by a clustering algorithm to group the rows of $\bar{\mathbf{W}}$ into subsets corresponding to each source.

For the ICA algorithm we use the natural gradient maximum likelihood (ML) algorithm [12]:

$$\Delta\bar{\mathbf{W}} = \eta \left(\mathbf{I} - E\{\mathbf{f}(\bar{\mathbf{y}})\bar{\mathbf{y}}^T\}\right) \bar{\mathbf{W}} \tag{7}$$

where η is the learning rate, and $\mathbf{f}(\bar{\mathbf{y}}) = -\nabla_{\bar{\mathbf{y}}} \log p(\bar{\mathbf{y}})$ is the ML activation function, using $p(\bar{\mathbf{y}}) = \prod_{p=1}^{P} p(\bar{y}_p)$ for some prior $p(\bar{y}_p)$. We use the generalized exponential prior $p(\bar{y}_p) \propto \exp(-|\bar{y}_p|^\alpha)$ where the exponent α is estimated through maximum likelihood [30].

Given a learned unmixing matrix $\bar{\mathbf{W}}$, we can consider the (reshaped) observation vectors $\bar{\mathbf{x}}$ to be represented by scalar combinations of basis vectors $\bar{\mathbf{a}}_k$ which are the columns of the inverse unmixing matrix $\bar{\mathbf{A}} = \bar{\mathbf{W}}^{-1}$. To give a direct physical interpretation in terms of the two stereo microphone channels we de-interleave the basis vectors $\bar{\mathbf{a}}_k$ to extract the stereo basis vector pairs $\mathbf{a}_k^{(1)}, \mathbf{a}_k^{(2)}$ using

$$[\mathbf{a}_k^{(1)}]_n = [\bar{\mathbf{a}}_k]_{2n-1}, \qquad [\mathbf{a}_k^{(2)}]_n = [\bar{\mathbf{a}}_k]_{2n}, \qquad n = 1, \ldots, K/2. \tag{8}$$

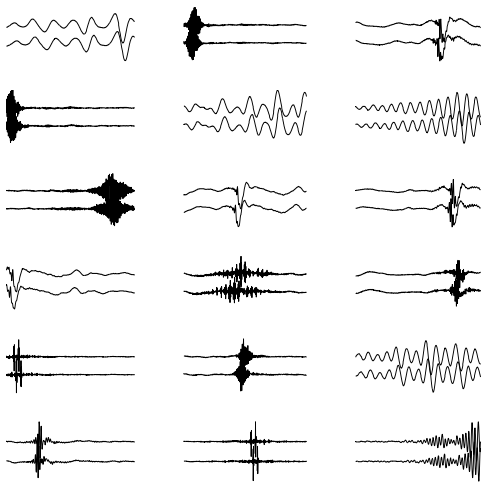

Fig. 2. Examples of stereo basis vector pairs extracted with the adaptive stereo basis algorithm.

Figure 2 shows some of the basis vector pairs obtained from a stereo mixture generated when two male speech signals were synthetically mixed using a source image technique, in low noise and low reverberation conditions (see Section 4). This figure illustrates that the basis vector pairs encode how the extracted features are received at the microphones. Many of the basis vectors are localised in time, and they seem to capture information about time-delay and amplitude differences that characterise the mixing channel. This observation, together with measurements of the relative time delay (see Fig. 3 below), suggests that the convolutive nature of the mixing process has been captured by the algorithm, and that each basis vector pair relates to a particular source.

3.3 Grouping the basis components

Having extracted a set of basis vectors $\bar{\mathbf{a}}_k$, we now need to group these together into subsets that correspond to each source we wish to extract. As for FD-ICA, we could use a variety of methods to perform this grouping. In earlier work we

used a higher-order correlation (F-correlation) between component activities to perform this grouping [12]. However, in Figure 2 we observe that the stereo basis vector pairs tend to be relatively wideband, and exhibit a clear relative time delay between the left and right channels. In this article we therefore propose to group the basis vectors into subsets based on their time delay, or direction of arrival (DOA), as we saw has already been used for FD-ICA.

For each basis pair k we find the time delay τ_k between the vectors in the pair, using the generalised cross-correlation with phase transform (GCC-PHAT) algorithm [31]

$$R_k(\tau) = \int_{-\infty}^{\infty} \frac{A_k^{(1)}(\omega) A_k^{(2)}(\omega)^*}{|A_k^{(1)}(\omega) A_k^{(2)}(\omega)^*|} e^{j\omega\tau} d\omega \tag{9}$$

where $A_k^{(1)}(\omega), A_k^{(2)}(\omega)$ are the Fourier transforms of the stereo basis vector pairs $\mathbf{a}_k^{(1)}$ and $\mathbf{a}_k^{(2)}$ respectively. We have observed that the function $R_k(\tau)$ typically exhibits a single sharp peak at the lag corresponding to the time delay between the two signals. In contrast to the STFT bases used in the FD-ICA algorithm, which exhibits multiple peaks leading to the spatial aliasing problem, this single peak is consistent with the ASB basis vectors being relatively wideband, and with a dominant DOA, hence avoiding the spatial aliasing problem.

The upper plot in Figure 3 illustrates the time-delay estimates obtained with GCC-PHAT, for all basis vector pairs from which those shown in Figure 2 were selected. The histogram of the estimated time-delays is shown in the lower plot of Figure 3. The figure shows that the directions of the two sources (corresponding to a delay of about 9 and -9 samples) are clearly visible, and most basis functions have time delays closely associated with one of the two directions of arrival.

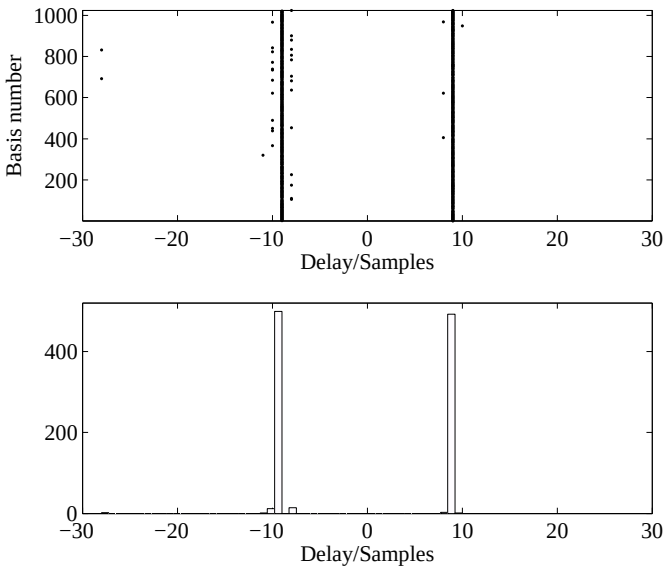

Fig. 3. Plot of the time delays estimated for all basis vectors (upper plot), and its histogram (lower plot).

To group the basis vectors, we use the K-means clustering algorithm to find the time delay 'centroid' T_p, $p = 1, \ldots, P$ corresponding to each of the P sources.

We then construct the index sets $\gamma_p = \{k \mid (T_p - \Delta) \leq \tau_k \leq (T_p + \Delta)\}$ corresponding to basis vectors with delays within some threshold Δ of the cluster centroid, reserving a 'discard' cluster $\gamma_0 = \{k \mid k \notin \gamma_p, p = 1, \ldots, P\}$ for 'noise' basis vectors which will not be associated with any of the P sources. In our 'reshaped' space of vectors $\bar{\mathbf{x}}$, we therefore have a subspace $E_p = \text{span}\{\bar{\mathbf{a}}_k, k \in \gamma_p\}$ corresponding to each source.

3.4 Extracting the source image estimates

To extract the separate source estimates, we project the reshaped vector sequence $\bar{\mathbf{x}}(n)$ into each of the P subspaces E_p as follows. We construct a set

of mask matrices $\mathbf{H}^{(p)} = \mathrm{diag}(h_1^{(p)}, \ldots, h_K^{(p)})$ for $p = 1, \ldots, P$, with the mask values given by

$$h_k^{(p)} = \begin{cases} 1 & \text{if } k \in \gamma_p \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, \ldots, K$. Thus the diagonal elements of $\mathbf{H}^{(p)}$ are one or zero depending on whether or not a transform component is considered to belong to the subspace E_p corresponding to the p-th source. Note that, in contrast to the time-frequency mask used in the DUET algorithm [11], which depends both on the frequency bin index f and the time frame index t, the ASB masking matrix $\mathbf{H}^{(p)}$ operates across basis pair indices k only and is independent of the time frame.

We then form the orthogonal projection matrices $\mathbf{P}_p = \bar{\mathbf{A}}\mathbf{H}_p\bar{\mathbf{W}} = \bar{\mathbf{W}}^{-1}\mathbf{H}_p\bar{\mathbf{W}}$ which is clearly a projection since $\mathbf{P}_p^2 = \mathbf{P}_p$, and where the column span of $\mathbf{P}_p$ is the subspace E_p. The estimated (reshaped) image $\hat{\bar{\mathbf{x}}}_p$ of the p-th source is given by

$$\hat{\bar{\mathbf{x}}}_p = \mathbf{P}_p\bar{\mathbf{x}} = \bar{\mathbf{A}}\mathbf{H}_p\bar{\mathbf{W}}\bar{\mathbf{x}}. \tag{10}$$

Finally, we de-interleave $\hat{\bar{\mathbf{x}}}_p$, using the reverse of the process described in Section 3.1. This de-interleaving process involves overlapping blocks, similar to overlapping windows in the inverse STFT, so we take averages of the overlapping blocks, which reduces blocking artefacts. De-interleaving $\hat{\bar{\mathbf{x}}}_p$ in this way yields the source image $\hat{\mathbf{x}}_p = [\hat{x}_{1p}, \hat{x}_{2p}, \ldots, \hat{x}_{Qp}]^T$, *i.e.* the vector of images of the p-th source at all Q microphones.

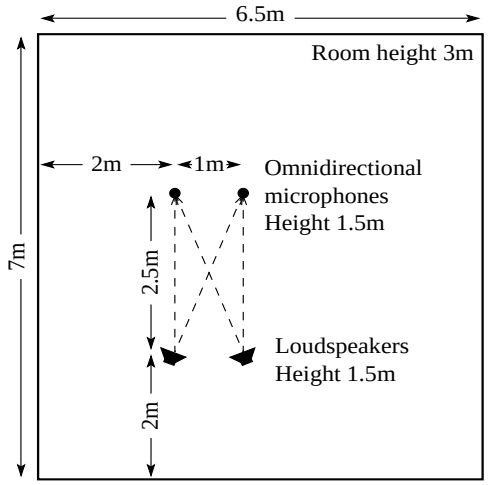

Fig. 4. Experimental setup for simulated speech recordings. The reverberation times were set to either 20 ms, 80 ms or 320 ms.

4 Evaluation

4.1 Experimental setup

We evaluated the proposed ASB algorithm and compared with FD-ICA and DUET on several mixtures of two male speech sources. The speech sources were sampled at 16 kHz with a duration of 1 minute each.

To allow us to control the room Reverberation Time (RT) and the Input Signal-to-Noise Ratio (ISNR), the sources were mixed using simulated room impulse responses, determined by the image technique [32] using McGovern's RIR Matlab function [1] . The positions of the microphones and the loudspeakers are illustrated in Figure 4. Six different mixing conditions were obtained by varying RT between 20 ms (320 samples), 80 ms (1280 samples) and 320 ms (5120 samples), and adding white noise to the mixture with ISNRs of 40 dB and 20 dB.

[1] `http://2pi.us/code/rir.m`

We chose the STFT frame lengths separately for each algorithm, but fixed for all the reverberation times tested. We used the FD-ICA algorithm with the MuSIC-based permutation alignment algorithm described by Mitianoudis and Davies [9], setting the STFT frame size to 2048 samples, which was previously found to be appropriate for this algorithm at a 16kHz sampling rate [9,33]. For the DUET algorithm we used an STFT frame size of 1024 samples, which was found by Yilmaz and Rickard [11] to give the best separation performance at 16 kHz. For the proposed adaptive stereo basis algorithm, we used an adaptive basis frame size of 512 samples, to be consistent with preliminary experiments which indicated that this would be sufficient for separation at a 16 kHz sampling rate with reasonable room reverberation times [33]. Excerpts of the original mixture and source signals and of the estimated source signals are available for listening on our demo web page [2].

4.2 Objective evaluation

We evaluated the performance of each method using the objective criteria of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Noise Ratio (SNR) and Signal-to-Artefacts Ratio (SAR) as defined in [34]. SDR measures the difference between an estimated source and a target source allowing for possible linear filtering between the estimated and target source. We allowed for time-invariant filtering of filter length 1024 samples when calculating SDR. SIR, SNR and SAR provide a more detailed diagnosis of the performance by distinguishing between the elements of the total distortion which are due to unwanted interfering sources (SIR), remaining mixing noise

[2] `http://www.elec.qmul.ac.uk/people/mariaj/asb_demo/`

(SNR) and other artefacts (SAR). Additive noise will be included within the SNR measure.

The SDR, SIR, SNR and SAR criteria are defined in [34] on a per-source basis. To gain a single figure for all sources, we averaged the criteria across all microphones and all sources. The results are presented in Table 1.

In an earlier preliminary investigation [33], we found that the objective SDR measures did not always correspond to our perceived quality of the separation. This difference may be due to the calculation of the objective criteria requires a reconstruction filter to be estimated, which is non-trivial for convolutive mixing or to distortions which are perceptually minor but which are not allowed for by the (linear, time-invariant) filter [34]. For the present study, we therefore conducted a formal subjective listening test to give a more definitive comparison of the relative performance of the three algorithms.

4.3 Evaluation using listening tests

Listening tests are common in audio coding, with standardized test procedures such as MUSHRA (MUltiple Stimulus test with Hidden Reference and Anchors) [35], but have not yet found widespread use in the source separation community.

For the listening test conducted here, we adapted the MUSHRA standard and built a Matlab graphical interface to allow subjects to listen to the stimuli and input their scores [36]. Subjects were asked to assess the *basic quality* of each stimulus, a term used to mean the overall perceived quality of the sound, including all possible types of distortion. Each subject was asked to grade the

Table 1

Objective performance of FD-ICA, DUET and ASB with default frame sizes on simulated speech recordings. All values are expressed in decibels (dB). Bold numbers indicate the best SDR for each mixing condition. See text for comments.

Mixing conditions	ISNR	40 dB			20 dB		
	RT	20 ms	80 ms	320 ms	20 ms	80 ms	320 ms
FD-ICA	SDR	7.0	**11.2**	**6.3**	6.2	6.5	**4.2**
	SIR	10.4	16.1	9.1	12.3	14.0	9.1
	SNR	19.1	19.9	28.9	26.7	10.7	25.8
	SAR	11.1	14.2	10.3	7.7	11.4	7.0
DUET	SDR	7.9	8.2	5.3	6.3	5.7	3.5
	SIR	13.4	13.8	10.0	14.7	12.7	8.9
	SNR	21.0	21.0	20.3	11.8	11.8	11.5
	SAR	10.3	10.2	7.9	9.3	9.0	7.3
ASB	SDR	**15.4**	7.7	1.3	**8.3**	**6.8**	-4.2
	SIR	25.7	16.3	8.9	19.7	17.8	7.4
	SNR	20.2	28.0	22.9	12.5	26.3	16.9
	SAR	18.2	9.8	4.2	12.6	7.5	-2.1

basic quality of the estimated sources compared to a given target source on a scale between 0 and 100, where 100 corresponded to the target source and 0 to the worst estimated source over all conditions. For more details on the

listening test procedure, see [36].

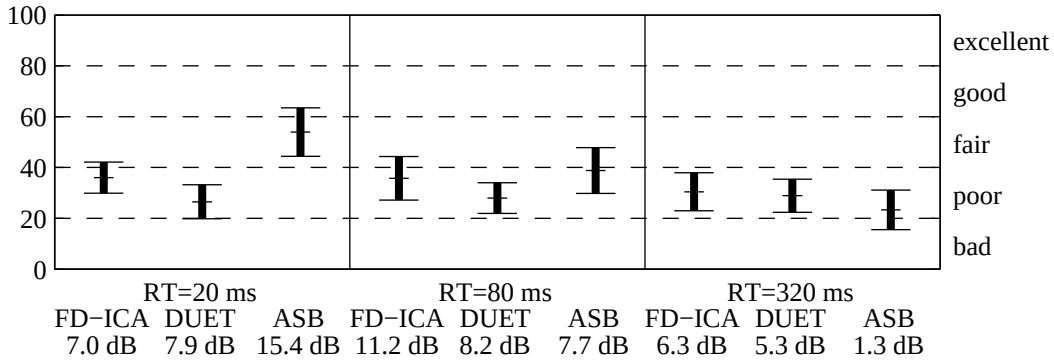

Fig. 5. Subjective performance of FD-ICA, DUET and ASB with default frame sizes on simulated speech recordings with ISNR=40 dB. Bars indicate 95% confidence intervals. SDR values are displayed below for comparison. See text for comments.

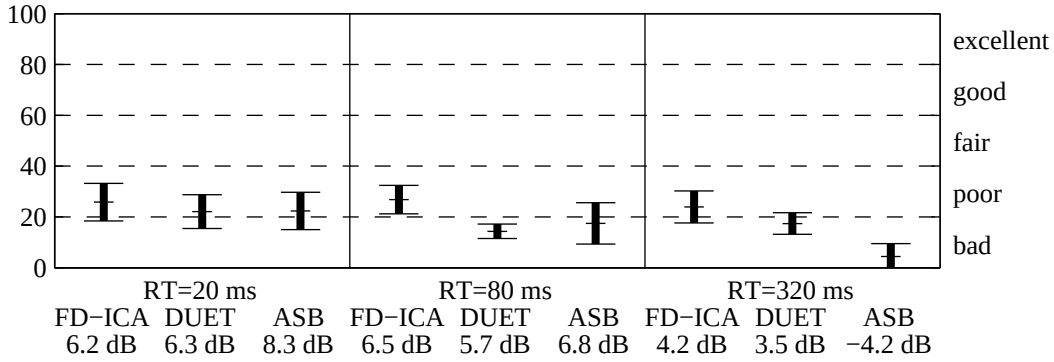

Fig. 6. Subjective performance of FD-ICA, DUET and ASB with default frame sizes on simulated speech recordings with ISNR=20 dB. Bars indicate 95% confidence intervals. SDR values are displayed below for comparison. See text for comments.

Eight subjects took part in the listening tests, and each complete listening test took between about 1 and 2 hours, including breaks. The algorithm developers who had already heard the stimuli were excluded from the listening test. Listeners were not pre-screened for auditory losses, but there was no evidence for any listener exhibiting a response significantly different from those of the other listeners. The results of all listeners were retained for the statistical calculations. The test results are shown in Figures 5 and 6.

4.4 Analysis of results

In the objective comparison (Table 1), we see that with short reverberation times (RT=20 ms) our proposed method outperforms both FD-ICA and DUET by more than 7 dB SDR in relatively clean conditions (ISNR=40 dB) and by about 2 dB SDR in more noisy conditions (ISNR=20 dB). The results of the listening tests (Figures 5 and 6) are generally consistent with the objective criteria, confirming that the proposed ASB algorithm performs significantly better than FD-ICA and DUET in clean, less reverberant, conditions (ISNR=40 dB, RT=20 ms).

For intermediate reverberation times (RT=80 ms), all algorithms show comparable objective performance, although with FD-ICA exhibiting higher objective performance in less noisy conditions (ISNR=40 dB). However, the listening tests indicate that ASB and FD-ICA have similar subjective performance, even though the frame size for ASB (512 samples) is much smaller than for FD-ICA (2048 samples) with DUET giving slightly lower performance than the other algorithms.

In the most reverberant conditions tested, the FD-ICA algorithm gave highest performance. Further investigation of the reason for the negative SDR for the ASB algorithm in noisy reverberant conditions (ISNR=20 dB, RT=320 ms), indicated that the unsupervised K-means clustering algorithm used in the proposed algorithm failed to find one of the source clusters. Supervised clustering based on the true source directions improved the SDR to -0.6 dB, but this remained lower than with FD-ICA and DUET in this case. Supervised clustering did not change the performance of the proposed algorithm

significantly in other conditions.

5 Discussion

5.1 Algorithm comparison

FD-ICA with beamforming-based source matching, DUET and the proposed adaptive stereo basis (ASB) algorithms are based on an essentially similar approach. A transformation is applied on the observed data in order to find a set of basis vectors, followed by direction-based clustering to associate each vector with a source. However, they exhibit some differences that become important when applied to realistic mixtures. We summarize their respective advantages and limitations below.

The main characteristic of ASB is that it is based on an adaptive transform of the observed data, where the basis vectors are estimated from the data. Conversely, FD-ICA and the DUET algorithm use the STFT, a fixed time-frequency transform. Thus we believe that ASB has the potential to provide a sparser representation of the data, and hence improve performance.

DUET and ASB achieve separation by clustering the dictionary elements, the former according to phase (delay) and amplitude information, and the latter according to phase only. FD-ICA with beamforming also uses phase information to align the permutations across all frequencies. Both FD-ICA and DUET suffer from phase ambiguities in the upper frequencies. To avoid this problem, DUET was designed under the assumption that the microphone separation, d, is small enough so that phase ambiguities do not arise [11]. Clearly, this

assumption cannot always be satisfied, particularly when the problem is truly blind (i.e. the microphone separation is not known, and cannot be controlled), or for certain applications, such as for CD recordings where phase ambiguities would arise with a sensor spacing of less than 1cm at 44.1 kHz [37]. To help select the correct phase difference between the two sensors where phase ambiguities are possible, a modified version of DUET has been proposed which uses amplitude differences in the high frequency range [37]. In the ASB algorithm we found experimentally that the basis vectors learned by the algorithm are typically time-localized rather than narrowband. It is therefore possible to identify a unique time delay between the left and right channels, using in our case the GCC-PHAT algorithm, and the phase ambiguity problem does not arise.

DUET was developed for anechoic mixing, and can have difficulties dealing with echoic (convolutive) mixing. Histograms obtained from anechoic mixtures are typically well localised, with distinct peak regions corresponding to the sources, while they are more spread out for echoic mixtures [11]. Conversely, ASB does not make any specific assumptions regarding the mixing channel. The learned basis pairs should automatically capture the nature of the channel, so we would expect the method to be able to deal with reverberation. However, the performance of the ASB algorithm does degrade with longer reverberation times (RT = 80 ms and above), perhaps due to the current frame size limit: 80 ms is equivalent to 1280 samples, compared to the currently feasible frame size of 512 samples in the ASB algorithm.

In comparison to methods that used a fixed basis, the adaptive stereo basis algorithm requires the fitting of an ICA model to the frames of stereo data. This involves additional computational expense, and also leads to a potential problem of *overfitting* due to the large effective dimensionality of the model. The first problem, that of computational expense, is partly due to the use of a stochastic gradient optimisation in the current implementation. We expect that some reduction in computation time would be possible through the use of second-order derivatives (*i.e.* curvature) to improve the convergence of ICA [16]. Note also that it is only the *system identification* stage which requires this computational expense; the separation step is relatively straightforward.

The second problem, that of overfitting, is potentially more serious as it is an intrinsic limitation of the model in its present form. For example, in our experiments, the ICA weight matrix had 512×512 entries and thus required the optimisation of 262144 parameters. At 16 kHz, a two-channel signal requires approximately 8.2 s to deliver this many samples. Our one-minute signals supplied less than 8 times as much data as there were parameters to be optimised, which is rather low and may lead to overfitting.

In applications where the mixing system is known to be stable for long periods, sufficient training data could be collected to avoid overfitting. However, this would of course bring us back to the computational expense of fitting an ICA model to such a large amount of data. Alternatively, there are several structural aspects of the system that could potentially be exploited to regularise or constrain the ICA model [38]. For a further possibility, since the frames used

to train the model are extracted from a longer signal which is assumed to be stationary, there should be no privileged times within the frame. This type of shift invariance has been exploited in single-channel sparse coding [39] and could possibly be adapted for use here.

6 Conclusions

We have considered the problem of convolutive blind audio source separation, and we have presented a stereo coding method. The method is based on the identification of stereo basis vectors adapted to the data. The basis functions are mostly temporally localized, and can be clustered according to directions of arrival (DOA). Separation can then be performed using binary masking on the resulting basis components.

The performance of the algorithm was compared to that of frequency domain ICA (FD-ICA) and the DUET algorithm, using speech signals mixed in a simulated room. Evaluation was performed using both objective measures and subjective listening tests.

The results of both the objective SDR comparison and the formal listening tests indicate that the proposed stereo coding method is competitive with both FD-ICA and the DUET algorithm at short and intermediate reverberation times, and significantly outperforms either of the other algorithms with low noise and short reverberation times (RT = 20 ms or 320 samples) of the same order as the frame size used in the ASB algorithm (512 samples). However, the performance of ASB on more echoic rooms (RT above 80 ms) indicates there is still more work to be done. The adaptive basis means that method is

currently computationally intensive, limiting the frame size and hence limiting its performance for long reverberation times.

In future work, we plan to explore frame sizes longer than 512 samples. To ameliorate the increased computation time involved, we plan to investigate ways to partially structure the ICA bases to allow faster and more robust learning. Other methods may prove useful to learn the basis vector sets, such as the recent K-SVD algorithm [40]. We believe the proposed adaptive stereo basis method is interesting and promising, although further investigation is required in order to reduce the computation cost and improve its robustness to noise and reverberation.

Acknowledgements

The authors wish to thank Nikolaos Mitianoudis and Scott Rickard for providing implementations of the FD-ICA and DUET algorithms, and all the subjects who participated in the listening tests. We would also like to thank two anonymous referees whose comments helped to improve the article before publication.

References

[1] S.-I. Amari, S. C. Douglas, A. Cichocki, H. H. Yang, Multichannel blind deconvolution and equalization using the natural gradient, in: Proceedings of the First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Paris, France, 1997, pp. 101–104.

[2] K. Torkkola, Blind separation of convolved sources based on information maximization, in: Proc. of the IEEE Workshop on Neural Networks for Signal Processing (NNSP), 1996, pp. 423–432.

[3] S. C. Douglas, X. Sun, Convolutive blind separation of speech mixtures using the natural gradient, Speech Communication 39 (2003) 65–78.

[4] S. Makino, H. Sawada, R. Mukai, S. Araki, Blind source separation of convolutive mixtures of speech in frequency domain, IEICE Trans. Fundamentals E88 (2005) 1640–1655.

[5] P. Smaragdis, Blind separation of convolved mixtures in the frequency domain, Neurocomputing 22 (1998) 21–34.

[6] S. Ikeda, N. Murata, A method of ICA in time-frequency domain, in: Proc. of the International Conference on Independent Component Analysis and Blind Source Separation (ICA99), Aussois, France, 1999, pp. 365–371.

[7] L. C. Parra, C. Spence, Convolutive blind separation of non-stationary sources, IEEE Trans. on Speech and Audio Processing 8 (2000) 320–327.

[8] M. E. Davies, Audio source separation, in: J. G. McWhirter, I. K. Proudler (Eds.), Mathematics of Signal Processing, Oxford University Press, 2002, pp. 57–68.

[9] N. Mitianoudis, M. E. Davies, Audio source separation of convolutive mixtures, IEEE Trans. on Speech and Audio Processing 11 (2003) 489–497.

[10] A. Jourjine, S. Rickard, Ö. Yilmaz, Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures, in: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 5, 2000, pp. 2985–2988.

[11] Ö. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Trans. on Signal Processing 52 (2004) 1830–1847.

[12] S. A. Abdallah, M. D. Plumbley, Application of geometric dependency analysis to the separation of convolved mixtures, in: Proc. of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004), Granada, Spain, 2004, pp. 22–24.

[13] H. Sawada, R. Mukai, S. Araki, S. Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation, IEEE Trans. on Speech and Audio Processing 12 (2004) 530–538.

[14] J.-F. Cardoso, B. Laheld, Equivariant adaptive source separation, IEEE Trans. on Signal Processing 44 (1996) 3017–3030.

[15] S. Amari, A. Cichocki, Adaptive blind signal processing - neural network approaches, Proceedings of the IEEE 86 (1998) 2026–2048.

[16] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. on Neural Networks 86 (1999) 626–634.

[17] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, F. Itakura, Evaluation of blind signal separation method using directivity pattern under reverberant conditions, in: Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Vol. 5, 2000, pp. 3140–3143.

[18] H. Saruwatari, S. Kurita, K. Takeda, Blind source separation combining frequency-domain ICA and beamforming, in: Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Vol. 5, 2001, pp. 2733–2736.

[19] S. Araki, S. Makino, R. Mukai, Y. Hinamoto, T. Nishikawa, H. Saruwatari, Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming, in: Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'02), Vol. 2, 2002, pp. 1785–1788.

[20] M. Z. Ikram, D. R. Morgan, A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation, in: Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Vol. 1, 2002, pp. 881–884.

[21] R. O. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Trans. on Antennas and Propagation 34 (1986) 276–280.

[22] N. Mitianoudis, M. E. Davies, Permutation alignment for frequency domain ICA using subspace beamforming methods, in: Proc. of the International Conference on Independent Component Analysis and Blind Source Separation (ICA 2004), Granada, Spain, 2004, pp. 669–676.

[23] P. D. O'Grady, B. A. Pearlmutter, S. T. Rickard, Survey of sparse and non-sparse methods in source separation, International Journal of Imaging Systems and Technology 15 (2005) 18–33.

[24] J.-F. Cardoso, Blind signal separation: Statistical principles, Proceedings of the IEEE 86 (1998) 2009–2025.

[25] A. Bell, T. Sejnowski, Learning the higher-order structure of a natural sound, Computation in Neural Systems 7 (1996) 261–266.

[26] S. A. Abdallah, M. D. Plumbley, If edges are the independent components of natural images, what are the independent components of natural sounds?, in: Proc. of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001), San Diego, California, 2001, pp. 534–539.

[27] M. S. Lewicki, Efficient coding of natural sounds, Nature Neuroscience 5 (2002) 356–363.

[28] J.-F. Cardoso, Multidimensional independent component analysis, in: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98), Vol. 4, 1998, pp. IV–1941–1944.

[29] M. E. Davies, M. G. Jafari, S. A. Abdallah, E. Vincent, M. D. Plumbley, Blind speech separation using space-time ICA filters, to appear (2007).

[30] L. Zhang, A. Cichocki, S.-I. Amari, Self-adaptive blind source separation based on activation functions adaptation, IEEE Transactions on Neural Networks 15 (2) (2004) 233–244.

[31] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. on Acoustic, Speech, and Signal Processing 24 (1976) 320–327.

[32] S. McGovern, A model for room acoustics, Available at: http://2pi.us/rir.html (2003).

[33] M. G. Jafari, S. A. Abdallah, M. D. Plumbley, M. E. Davies, Sparse coding for convolutive blind audio separation, in: Proc. of the International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006), Charleston, SC, USA, Springer-Verlag, Berlin, 2006, pp. 132–139.

[34] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE Trans. on Audio, Speech and Language Processing 14 (4) (2006) 1462–1469.

[35] International Telecommunication Union, Recommendation ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems (2003).

[36] E. Vincent, M. G. Jafari, M. D. Plumbley, Preliminary guidelines for subjective evaluation of audio source separation algorithms, in: A. K. Nandi, X. Zhu (Eds.), Proc. of the ICA Research Network International Workshop, Liverpool, UK, 2006, pp. 93–96.

[37] H. Viste, G. Evangelista, On the use of spatial cues to improve binaural source separation, in: Proceedings of the International Conference on Digital Audio

Effects (DAFx-03), London, UK, 2003, pp. 209–213.

[38] Y. Matsuda, K. Yamaguchi, Linear multilayer ICA integrating small local modules, in: 4th Intl. Symp. on Independent Component Analysis and Signal Separation (ICA2003), Nara, Japan, 2003, pp. 403–408.

[39] T. Blumensath, M. Davies, Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music, in: Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2004), Vol. 5, Montreal, Canada, 2004, pp. V–497–500.

[40] M. Aharon, M. Elad, A. Bruckstein, K-SVD: Design of dictionaries for sparse representation, in: Proceedings of the Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS'05), Rennes, France, 2005, pp. 9–12.

A Constructing source image estimates for the DUET algorithm

The DUET algorithm [10,11] performs separation of stereo sources in the time-frequency domain. Using estimates of relative amplitude and delay parameters, a set of binary time-frequency masks $\mathbf{M}_p(f,t),\ p = 1,\ldots,P$ is then constructed to perform separation of the sources s_p, either by masking one of the microphones, or via maximum likelihood (ML) source estimation [11].

In this article, we wish to measure separation performance on the images of the sources at the microphones as in Equation (3). For the evaluation in Section 4, we directly calculate the image $\hat{\tilde{\mathbf{x}}}_{qp}(f,t)$ of the p-th estimated source observed at the q-th microphone using

$$\hat{\tilde{\mathbf{x}}}_{qp}(f,t) = \mathbf{M}_p(f,t)\tilde{x}_q(f,t),\ \forall f,t. \tag{A.1}$$

The time-domain estimate $\hat{\mathbf{x}}_{qp}(n)$ is obtained by inverting the STFT for each source/microphone pair.

Conceptually this approach uses DUET time-frequency masking to directly calculate an estimate of the image x_{qp} of source s_p at the q-th microphone, without calculating a single source estimate as an intermediate stage. We have observed that attempting to construct source images from a single estimated source can produce poor results for echoic convolutive mixtures, perhaps due to inaccurate estimates of the mixing delays in such situations.