

BLIND SOURCE SEPARATION OF CONVOLUTIVE AUDIO USING AN ADAPTIVE STEREO BASIS

Maria G. Jafari ^a, Emmanuel Vincent ^a, Samer A. Abdallah ^a, Mark D. Plumbley ^a, and Mike Davies ^b

^a Centre for Digital Music, Queen Mary University of London, UK

^b Institute for Digital Communication, University of Edinburgh, UK

{maria.jafari,emmanuel.vincent,samer.abdallah,mark.plumbley}@elec.qmul.ac.uk

mike.davies@ed.ac.uk

ABSTRACT

We consider the problem of convolutive blind source separation of audio mixtures. We propose an Adaptive Stereo Basis (ASB) method based on learning a set of basis vectors pairs from the time-domain stereo mixtures. The basis vector pairs are clustered using estimated directions of arrival (DOAs) such that each basis vector pair is associated with one source. The ASB method is compared with the DUET algorithm on convolutive speech mixtures at different reverberation times and noise levels.

1 INTRODUCTION

The convolutive blind audio source separation problem arises when an array of sensor microphones is placed in a room, so that as well as recording a mixture of the source signals, multipath copies of the sources are also present. Many methods have been proposed for convolutive source separation, including time-domain deconvolution and frequency-domain ICA [8].

One approach that has been found to be successful in practical blind audio source separation applications is the degenerate unmixing estimation technique (DUET) [5]. DUET is a time-frequency (TF) masking method designed to address the underdetermined blind source separation (BSS) problem, where there are fewer mixtures than sources. It separates an arbitrary number of source signals from two mixtures [5], under the assumption that in the time-frequency domain each time-frequency point of a mixture signal is due only to one of the sources, a property denoted as W-disjoint orthogonality [10]. To estimate the dominating source at each time-frequency point, DUET assumes anechoic mixing, *i.e.* that only delays and attenuations are present in the mixture, with no echoes.

We proposed an approach to convolutive audio BSS [2], using ideas of signal transforms and masking similar to that used in DUET, but instead of using a fixed trans-

form we use an adaptive dictionary of stereo basis vectors. The method applies independent component analysis (ICA) to the observed time-domain data to find a set of basis functions (dictionary elements), and then assigns each basis function to one of the sources present in the sound field using a dependency analysis. In [4], this approach is modified, so that clustering is performed based on the estimated direction of arrival (DOA) of the sources. We call this the Adaptive Stereo Basis (ASB) method.

In this paper, our proposed ASB approach [4] is compared to the DUET algorithm, and computer simulations are carried out to assess their performance. The convolutive BSS problem is described in section 2, and the DUET is discussed in section 3, while the ASB method is summarised in section 4. The performance of ASB and DUET algorithm are compared in section 5, followed by discussion and conclusions.

2 PROBLEM FORMULATION

The convolutive blind source separation problem arises when a set of observations each contain mixtures of the original source signals, at several delay times and amplitude levels, as well as multipath copies of the sources, distorted by the environment [8]. In audio, this is the typical situation for recordings in an echoic room. This mixing system can be modelled by

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

where $x_q(n)$ is the signal recorded at the q -th microphone at time n , $s_p(n)$ is the p -th source signal, $a_{qp}(l)$ denotes the impulse response from source p to sensor q , and L_m is the maximum length of all impulse responses. The aim of convolutive blind source separation is to recover the original source signals $s_p(n)$ given only the mixtures $x_q(n)$.

3 DUET ALGORITHM

DUET uses the short-time Fourier transform (STFT) to separate an arbitrary number of source signals from two mixtures [5]. DUET assumes that the sources are W-disjoint orthogonal, *i.e.* that each STFT time-frequency point of a mixture signal is due only to one of the sources

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

[10]. An anechoic mixing model is assumed, given by

$$x_q(n) = \sum_{p=1}^P \alpha_{qp} s_p(n - \delta_{qp}), \quad q = 1, 2 \quad (2)$$

where α_{qp} and δ_{qp} are the attenuation and time delay coefficients associated with the path between the p -th source and the q -th sensor [14]. Fixing $\alpha_{1p} = 1$ and $\delta_{1p} = 0$, $\forall p$, for the first mixture, the subscript q ($= 2$) for the remaining parameters can be dropped [5]. The parameters for the dominating source in time-frequency bin (f, t) can be estimated by $\hat{\alpha}(f, t) = \left| \frac{x_2(f, t)}{x_1(f, t)} \right|$ and $\hat{\delta}(f, t) = -\frac{1}{2\pi f} \angle \left(\frac{x_2(f, t)}{x_1(f, t)} \right)$, where $\angle(\cdot)$ denotes the phase of a complex number taken between $-\pi$ and π . A weighted two-dimensional histogram is then produced, with the number of sources estimated from the number of peaks, and the position of the peaks determining the parameters for each source. Binary masks are used to perform separation, based on the proximity of each time-frequency point to the peak corresponding to that source [14].

With microphone spacing larger than half the wavelength of the maximum audio frequency, the estimation of $\delta(f, t)$ has a phase ambiguity [14], due to equivalent phase differences of multiples of 2π . Thus, the original algorithm was designed under the assumption that sensor spacing is small enough not to introduce these ambiguities.

4 ADAPTIVE STEREO BASIS APPROACH

Studies of time-domain analysis of sounds using independent component analysis (ICA) in [1] and [7] reported that features (basis vectors) learned from speech signals are mostly well localised in time and frequency, yielding representations that exhibit both wavelet- and Fourier-like bases, depending on the characteristics of the data. In [2] this analysis was extended to stereo audio observations, where it was observed that each learned basis vector pair appeared to correspond to a single source.

In our proposed ASB method, we take frames of the observed vectors $\mathbf{x}(n)$ and reshaped into a matrix $\tilde{\mathbf{X}}$ where each column is a stereo frame pair [4]. The basis vectors are learned with the ICA algorithm

$$\Delta \mathbf{B} = \eta (\mathbf{I} - E\{\mathbf{f}(\mathbf{u})\mathbf{u}^T\}) \mathbf{B} \quad (3)$$

where $\mathbf{B} \in \mathbb{R}^{K \times K}$ is a separating matrix giving $\mathbf{u} = \mathbf{B}\tilde{\mathbf{X}}$, η is the learning rate, and $\mathbf{f}(\mathbf{u}) = -\Delta_u \log p(\mathbf{u})$ is the activation function, with prior $p(\mathbf{u}) = \prod_{p=1}^P p(u_p)$. We use a generalized exponential prior $p(c_p) \propto \exp(-|c_p|^\alpha)$, with α estimated through maximum likelihood [2]. The basis vector pairs are extracted from the columns of \mathbf{B}^{-1} . For details, see [4].

Figure 1 shows some of the basis vectors (columns of \mathbf{B}^{-1}) obtained from two mixtures of male speech signals, as described in more detail in section 5.

Most of the basis vectors are localised in time, allowing the relative delay to be estimated (Fig. 2). We esti-

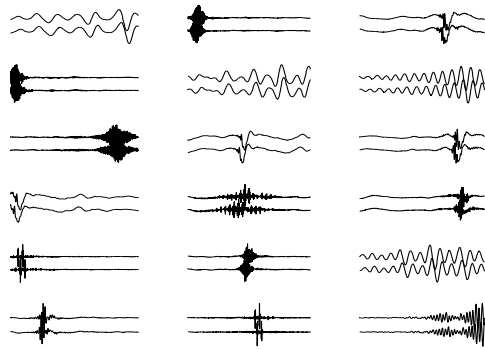


Figure 1: Examples of basis vectors extracted with the stereo sparse coding algorithm.

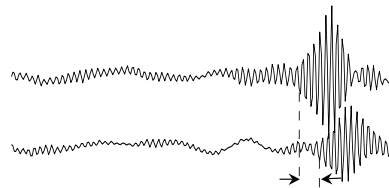


Figure 2: Basis vector pair showing relative time delay.

mate the relative time delay using the GCC-PHAT algorithm [6], corresponding to a direction of arrival (DOA) for a particular source [4]. Figure 3 depicts the time-delay estimates obtained with this method. It shows that ASB correctly identifies the directions of the two sources (corresponding to a delay of about 9 and -9 samples), and most basis functions are associated with one of the directions of arrival. We cluster the basis functions according to the associated direction of arrival. Finally we separate the source by selecting only those output components \mathbf{u} corresponding to the source of interest, and invert the transform to recover the source at the microphones [4].

5 EVALUATION

We evaluated DUET and ASB on several signals containing two male speech signals, sampled at 16 kHz and with a duration of 1 minute. To allow precise control of the Reverberation Time (RT) and the Input Signal-to-Noise Ratios (ISNR), we used simulated room impulse responses. These were determined by the image technique [11] using the Rir Matlab function¹.

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 (1300 samples), and 320 ms (5100 samples), and adding white noise to the mixture with ISNRs of 40 dB and 20 dB. The frame length was set to 1024 samples for DUET and 512 samples for ASB. Excerpts of the original mixture and source signals and of the estimated source signals are available for listening on our demo webpage².

¹<http://2pi.us/code/rir.m>

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

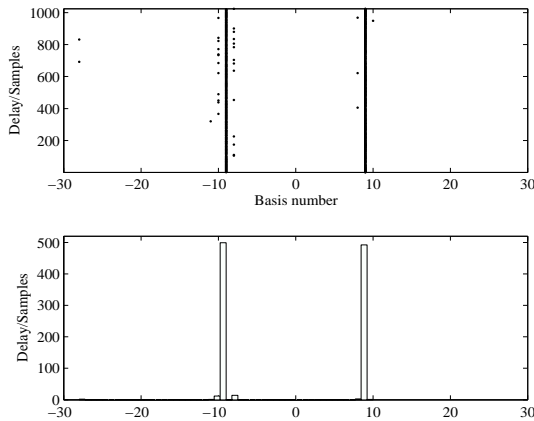


Figure 3: Time delays estimated for all basis vectors (upper plot), and histogram (lower plot).

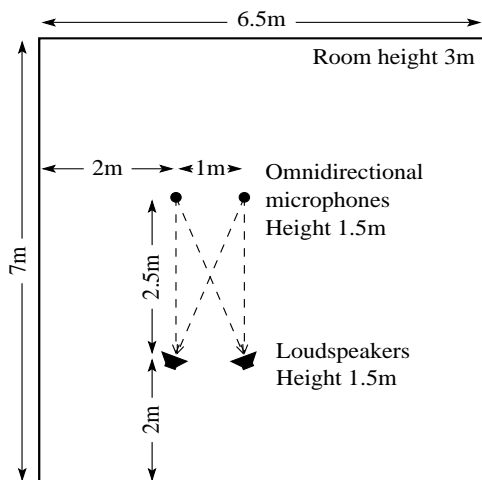


Figure 4: Experimental setup for simulated speech recordings. The reverberation times were set to 20 ms, 80 ms and 320 ms, respectively.

We evaluated the performance of each method using the criteria of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Noise Ratio (SNR) and Signal-to-Artifacts Ratio (SAR) defined in [12]. SDR measures the difference between an estimated source and a target source in terms of a power ratio, allowing limited linear distortion. In our evaluations, we allowed for time-invariant filter distortions of length 1024 samples. SIR, SNR and SAR provide a more detailed diagnosis of the performance by distinguishing difference due to interfering sources, remaining noise and other distortions. These criteria were measured on source images at the microphones, to avoid problems with spectral shape ambiguity [9], and we averaged across all sources and microphones to obtain single performance figures.

Results are presented in Table 1. At low reverberation times (RT=20 ms) we can see that ASB outperforms DUET by more than 7 dB SDR in clean conditions (ISNR=40 dB), and by about 2 dB SDR in more noisy conditions (ISNR=20 dB). However, the performance of ASB degrades faster than the DUET algorithm in the pres-

ence of reverberation (RT=320 ms). In noisy reverberant conditions (ISNR=20 dB, RT=320 ms), the clustering algorithm associated with ASB failed to find both source directions, resulting in a negative SDR. We investigated using supervised clustering for ASB, using the known delays. This improved the performance of ASB slightly in noisy reverberant conditions, but did not change the performance significantly in other conditions. The general degradation of the performance of ASB in the presence of reverberation may be due to the relatively short frame size of 512 samples, instead of 1024 samples for DUET.

6 DISCUSSION

ASB and DUET are based on a similar approach. A transformation is applied on the observed data in order to find a set of basis vectors, followed by clustering to associate each vector with a source. Both methods exploit directional information to perform separation. However, ASB is based on an adaptive transform, such that the basis vectors are estimated from the data, while DUET uses a fixed transform (the STFT). Thus, ASB has the potential to provide a sparser representation of the data, which may help separation. DUET and ASB achieve separation by explicitly clustering the dictionary elements, the former according to phase and amplitude information, and the latter according to phase only. In DUET the separation is achieved by time-frequency masking, *i.e.* masking each (f, t) point separately, while in ASB the masking depends on the basis vector only, not on the time frame.

DUET was designed under the assumption that the microphone separation, d , is small enough so that ambiguities do not arise in its narrowband phase estimation method [14]. However, this is an assumption that cannot always be satisfied [13]. ASB does not appear to suffer from this phase ambiguity problem, since the basis vectors are typically more wideband, and the GCC-PHAT algorithm [6] we use to estimate the phase delays normally produces a unique and sharp delay estimate.

It was found in [14] that, at a sampling rate of 16kHz, a window length of 1024 provided the best performance with DUET. On the other hand, ASB was found to provide good separation with shorter frame sizes (*e.g.* 512 samples, as used here), at least in less reverberant conditions. Learning basis sets in much longer frame sizes in ASB is currently very computationally expensive.

ASB does not use any assumptions regarding the mixing channel, but relies on the learned basis pairs capturing the nature of the channel. While this should make it robust to dealing with reverberation, its performance on longer reverberation times is currently poor. This seems to be due to the current frame size limitations, in that performance appears to decrease when the reverberation time significantly exceeds the frame size. This emphasizes the need for more efficient methods of adapting the stereo basis for larger frame sizes. In future work we plan to investigate alternative dictionary learning methods such as the K-SVD [3].

Table 1: Performance of 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 RT	40 dB			20 dB		
		20 ms	80 ms	320 ms	20 ms	80 ms	320 ms
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

7 CONCLUSIONS

The performance of our proposed adaptive stereo basis (ASB) algorithm was compared to that of the DUET algorithm, for speech signals mixed in simulated rooms. The proposed ASB method performs well in small to medium reverberation times. However, its performance degrades significantly in more reverberant conditions (RT=320 ms), most likely due to the frame size used (512 samples = 32 ms at 16 kHz).

In future we plan to investigate the use of longer frame sizes in ASB on longer frame sizes, to attempt to improve its performance with longer reverberation times. For computational reasons this will require more efficient dictionary learning methods, either through introducing some structure into the basis, or through alternative methods such as K-SVD [3]. We will also investigate the algorithm performance on overcomplete mixtures, since the basis set clustering is not theoretically limited to the same number of sources as sensors.

References

- [1] S. A. Abdallah and M. D. Plumbley. If the independent components of natural images are edges, what are the independent components of natural sounds? In *Proc. of the International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, pages 9–13, 2001.
- [2] S. A. Abdallah and 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 Source Separation (ICA)*, pages 22–24, 2004.
- [3] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: Design of dictionaries for sparse representation. In *Proceedings of SPARSE'05, Rennes, France*, pages 9–12, November 2005.
- [4] M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and 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)*, pages 132–139, 2006.
- [5] A. Jourjine, S. Rickard, and 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)*, pages 2985–2988, 2000.
- [6] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustic, Speech, and Signal Processing*, 24:320–327, 1976.
- [7] M. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5:356–363, 2002.
- [8] S. Makino, H. Sawada, R. Mukai, and S. Araki. Blind source separation of convolutive mixtures of speech in frequency domain. *IEICE Trans. Fundamentals*, E88:1640–1655, 2005.
- [9] N. Mitianoudis and M. E. Davies. Audio source separation of convolutive mixtures. *IEEE Trans. on Speech and Audio Processing*, 11:489–497, 2003.
- [10] P. D. O’Grady, B. A. Pearlmutter, and S. T. Richard. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, 15:18–33, 2005.
- [11] S. McGovern. A model for room acoustics, 2003. Available at: <http://2pi.us/rir.html>.
- [12] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4), 2006. to appear.
- [13] H. Viste and G. Evangelista. On the use of spatial cues to improve binaural source separation. In *Proc. of the International Conference on Digital Audio Effects (DAFx)*, 2003.
- [14] Ö. Yilmaz and S. Richard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*, 52:1830–1847, 2004.