SPEECH SEPARATION USING AN ADAPTIVE SPARSE DICTIONARY ALGORITHM

*Maria G. Jafari, Mark D. Plumbley**

Department of Electronic Engineering,
Queen Mary University of London,
London, UK

Mike E. Davies

IDCoM & Joint Res. Inst. for Signal & Image Process.,
Edinburgh University,
Edinburgh, UK

ABSTRACT

We present a greedy adaptive algorithm that builds a sparse orthogonal dictionary from the observed data. In this paper, the algorithm is used to separate stereo speech signals, and the phase information that is inherent to the extracted atom pairs is used for clustering and identification of the original sources. The performance of the algorithm is compared to that of the adaptive stereo basis algorithm, when the sources are mixed in echoic and anechoic environments. We find that the algorithm correctly separates the sources, and can do this even with a relatively small number of atoms.

***Index Terms*—** Orthogonal transform, sparse dictionary, adaptive dictionary, source separation.

1. INTRODUCTION

When placed in a real environment, an array of microphones records mixtures of sound sources characterized by time delays or echoes, which are determined by the mixing medium. Source separation techniques are used to learn the original sources, usually based on some statistical properties that might facilitate the separation process, such as source sparsity. Its advantage is that sparse sources will not overlap in a transformed domain, hence making separation much easier to perform. Exploitation of source sparsity has motivated the development of a wide variety of separation methods, including in the time-frequency [1, 2] and wavelet domains [3], as well as the derivation of the more general framework of sparse component analysis. This is a four-step approach focused on dictionary learning, with the aim of finding a sparse signal decomposition from which the mixing can be estimated, and the sources reconstructed [4].

In the time-frequency domain, the problem is typically addressed by performing source separation independently at each frequency bin, resulting in the introduction of the well-known permutation problem [1]. The separated components at each frequency bin must be clustered in order to estimate the original sources, correctly separated, and this is often done using beamforming methods which however suffer from phase ambiguities in the upper frequencies. An alternative approach is the adaptive stereo basis (ASB) method proposed in [5]. The algorithm learns a dictionary from the observed stereo data, simultaneously across the two channels, under the assumption that the sources are sparse. The basis vectors are then clustered according to the relative time-delays between the left and right channels of the basis pairs, corresponding to the directions of arrival (DOAs) of the sources.

In this article, we propose to separate stereo speech signals using an approach similar to that of ASB, comprising a dictionary learning stage, followed by clustering and source reconstruction. An orthogonal dictionary is learned directly from the data, using a greedy adaptive sparse dictionary (GASD) algorithm, which extracts dictionary elements from regions of the observed data where the energy is maximum, while maintaining a minimum L1-norm. Source separation is also performed with the ASB algorithm, for comparison purposes. The paper is organized as follows: Section 2 introduces the source separation problem, the GASD algorithm is presented in Section 3, while experimental results and conclusions are given, respectively, in Sections 4 and 5.

2. SOURCE SEPARATION

We address the source separation problem for two convolutive mixtures, $\mathbf{x}(n)$, of two sampled real-valued speech signals, $\mathbf{s}(n)$. The q-th microphone records a mixture, $x_q(n)$, of the source signals, $s_p(n)$, $p = 1, 2$, convolved with the impulse response between each source and sensor, as follows

$$x_q(n) = \sum_{p=1}^{2} \sum_{l=1}^{L} a_{qp}(l) s_p(n-l), \quad q = 1, 2 \qquad (1)$$

where $a_{qp}(l)$ is the impulse response from source p to sensor q, and L is the maximum length of all impulse responses. The aim is to find estimates for the unmixing filters $w_{qp}(l)$, using only the sensor measurements, and to reconstruct the sources from

$$y_p(n) = \sum_{q=1}^{2} \sum_{l=1}^{L} w_{qp}(l) x_q(n-l), \quad p = 1, 2 \qquad (2)$$

*This work was funded by EPSRC grant GR/S85900/01.

where $y_p(n)$ is the p-th recovered source. In matrix form, the mixing and separating models in (1) and (2) become, respectively $\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n)$ and $\mathbf{y}(n) = \mathbf{W}(n)\mathbf{x}(n)$.

The source separation problem is addressed here by leaning a dictionary from the observed data, using a greedy adaptive sparse dictionary algorithm, operating across the two channels. This is then followed by DOA based clustering, and source reconstruction.

3. GREEDY ADAPTIVE SPARSE DICTIONARY ALGORITHM

The GASD algorithm adaptively learns a data dependent dictionary by sequentially extracting the columns of the matrix $\mathbf{X}$, which is generated by taking overlapping data frames from the observed data. The method is inspired by the idea of reducing the L2-norm of the data by a maximum amount, across all frames, while ensuring that the L1-norm is reduced by a minimum amount. This is achieved by setting each new atom equal to the column of $\mathbf{X}$ that satisfies:

$$\max_k \frac{||\mathbf{x}_k||_2}{||\mathbf{x}_k||_1} \tag{3}$$

where $\mathbf{x}_k$ the k-th column of $\mathbf{X}$. In practice, the L1-norm is not re-normalized at each step, and therefore 3 is strictly achieved only for the first atom. The GASD algorithm learns the dictionary atoms according to the steps outlined below.
At iteration $j = 1$,

- ensure that the columns of $\mathbf{X}$ have unit L1-norm $\tilde{\mathbf{x}}_k = \frac{\mathbf{x}_k}{||\mathbf{x}_k||_1}$. This leads to a new data matrix $\tilde{\mathbf{X}}$, whose columns now have unit L1-norm;
- set the residual matrix $\mathbf{R}^0 = \tilde{\mathbf{X}}$ where $\mathbf{R}^j = [\mathbf{r}_1^j, \ldots, \mathbf{r}_{k_{\max}}^j]$ and $\mathbf{r}_k^j \in \mathbb{R}^{k_{\max}}$ is a residual column vector corresponding to the k-th column of $\mathbf{R}^j$.

Repeat, for all atoms to be extracted:

1. Compute the L2-norm of each frame
$$E_k = ||\mathbf{r}_k^j||_2 = \sum |\mathbf{r}_k^j|^2. \tag{4}$$
3. Set the j-th dictionary element $\boldsymbol{\psi}^j$ to be equal to the residual vector with largest L2-norm $\mathbf{r}^j{}_{\hat{k}}$
$$\boldsymbol{\psi}^j = \mathbf{r}_{\hat{k}}^j. \tag{5}$$
where
$$\hat{k} = \arg\max_{k \in \mathbb{K}}(E_k) \tag{6}$$
is the index corresponding to the signal block with largest L2-norm.
4. Evaluate the coefficients of expansion, given by the inner product between the residual vector $\mathbf{r}_k^j$, and the atom $\boldsymbol{\psi}^j$
$$\alpha_k^j = \langle \mathbf{r}_k^j, \boldsymbol{\psi}^j \rangle. \tag{7}$$
5. Compute the new residual, by removing the component along the chosen atom, for each element k in $\mathbf{r}_k^j$
$$\mathbf{r}_k^j = \mathbf{r}_k^{j-1} - \frac{\alpha_k^j}{\langle \boldsymbol{\psi}^j, \boldsymbol{\psi}^j \rangle} \boldsymbol{\psi}^j. \tag{8}$$

The last step ensures that the transform is orthogonal, by removing the contribution of the atom from each residual vector.

3.1. Applying GASD to source separation

Since we seek to separate a stereo mixture, prior to generating the matrix $\mathbf{X}$, the samples from the observed stereo signal are interleaved, as discussed in [5]. This emphasizes the correlations between the original source signals at the two microphones, leading to basis pairs that encode information about the mixing channel. This is followed by learning the dictionary with GASD, yielding a set of basis vector pairs, $\boldsymbol{\psi}_l^{(i)}(n),\ i = 1, 2;\ l = \{1, \ldots, L\}$, rather than individual atoms. To obtain the separated sources, the atom pairs must be clustered into subsets corresponding to each original signal, followed by source reconstruction. Clustering is done by finding the time delay, or direction of arrival (DOA), between the atoms in each pair with the generalized cross-correlation with phase transform (GCC-PHAT) algorithm [6],

$$R_l(\tau) = \int_{-\infty}^{\infty} \Psi_l^{(1)}(\omega)\Psi_l^{(2)}(\omega)^* / (|\Psi_l^{(1)}(\omega)\Psi_l^{(2)}(\omega)^*|) e^{j\omega\tau} d\omega \tag{9}$$

where $\Psi_l^{(1)}(\omega)$, $\Psi_l^{(2)}(\omega)$ are the Fourier transforms of the basis vectors. The atoms are then grouped using the K-means clustering algorithm, with the cluster centers corresponding to the time delays, T_i, $i = 1, 2$, for each source. This allows us to define a set of indices

$$\gamma_i = \{l \mid (T_i - \Delta) \leq \tau_l \leq (T_i + \Delta)\} \tag{10}$$

corresponding to the atoms with delays within some threshold Δ of the cluster center, and reserving a 'discard' cluster $\gamma_0 = \{l \mid l \notin \gamma_i, i = 1, 2\}$ for atoms that will not be associated with any of the sources. Then, to reconstruct the original sources, two mask matrices $\mathbf{H}^{(i)},\ i = 1, 2$, with their diagonal elements given by

$$h_l^{(i)} = \begin{cases} 1 & \text{if } l \in \gamma_i \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

for $l = 1, \ldots, L$. Then, the estimated image $\hat{\hat{X}}^{(i)}$ of the i-th source at both microphones is given by

$$\hat{\hat{X}}^{(i)} = \mathbf{D}^T \mathbf{H}^{(i)} \mathbf{D} \mathbf{X}^{(i)} \tag{12}$$

where $\mathbf{D}$ is the orthogonal dictionary matrix obtained with the GASD algorithm. It should be noted that in a general framework, the right-hand side of equation (12) is $\mathbf{A}\mathbf{H}^{(i)}\mathbf{W}\mathbf{X}^{(i)}$, where $\mathbf{A}$ is the dictionary matrix, and $\mathbf{W} = \mathbf{A}^{-1}$. Hence, since the proposed GASD algorithm results in an orthogonal dictionary matrix, it has the advantage that the source reconstruction step avoids matrix inversion, which is replaced by matrix transposition.

Finally, we reverse the reshaping process to find the source image $\hat{\mathbf{x}}^{(i)}(n) = \left[\hat{x}_1^{(i)}(n), \hat{x}_2^{(i)}(n)\right]^T$, that is, the vector of images of the i-th source at both microphones.

4. EXPERIMENTAL RESULTS

The GASD algorithm was used to separate the components from a stereo mixture generated when two male speech signals were synthetically mixed according to the mixing model in equation (1). We consider two different mixing conditions, obtained with the reverberation time set to 0 ms (anechoic mixing), and 320 ms (echoic mixing). In the anechoic mixing case, the position of the sources was such that time-delays of approximately -16 and 23 samples resulted, while in the echoic case, the time-delays were approximately -9 and 9 samples. For comparisson purposes, separation was also performed with the ASB algorithm in [5], which learns the dictionary atoms using an independent component analysis algorithm with sparse prior (see [5] for more details), while atom pair clustering and source reconstruction is performed as outlined in Section 3.1. The upper plots in figure 1 show some of the atom pairs obtained with ASB from the anechoic mixtures, while examples of the dictionary elements extracted with GASD are shown in the lower plots. Comparing these, we see that the former extracts much more elementary signal features, which can be used to describe most speech signals, while the latter yields more complex atoms that capture information unique to the analyzed signal. In a similar fashion to the ASB basis pairs, the GASD atom pairs encode how the extracted features are received at the microphone, that is, they capture information about time-delays and amplitude differences. This can be seen especially from atom pair in the top-left ($l = 39$ from the ASB atoms), and from the bottom-middle plot ($l = 9$ from the GASD atoms).

Time-delays estimates obtained with the two algorithms from all basis vector pairs, in the case of anechoic and echoic mixtures, are depicted in figures 2 and 3 respectively (upper plots), which also show their histograms. In all cases, the time-delays of the two sources are clearly visible, and correctly identified as -16 and 23 samples in the anechoic case, and -9 and 9 samples in the echoic case. It is interesting to see how the GASD algorithm seems to correctly identify the source directions in those atoms that are extracted first (with higher L2-norm), while picking up more noise as the L2-norm decreases. This can be seen in the upper-left plots of figures 2 and 3, where in the higher atom numbers, the time-delay plots become noisy, resulting in less accurate DOA estimates. This would suggest that source reconstruction with GASD does not require the use of all the atoms extracted. To test this hypothesis, the performance of the two algorithms was evaluated using the objective criteria of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artefacts Ratio (SAR) measuring, respectively, the difference between the estimated and target source, allowing for possible linear filtering between them, the distortion due to interfering sources and to other artefacts [7]. Table 1 shows the criteria obtained for the anechoic mixing case; similar figures were obtained for the echoic case, but are not reported here due to lack of space. The single figures were obtained by averaging the criteria across all sources and microphones. Negative SIR

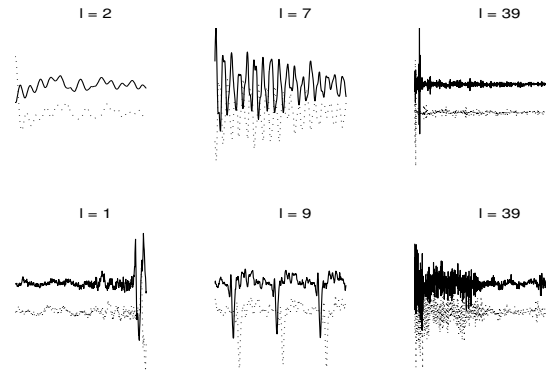

Fig. 1. Examples of the atoms pairs learned with the ASB (upper plots) and GASD algorithm (lower plots). The value l denotes the position of the atom within the dictionary.

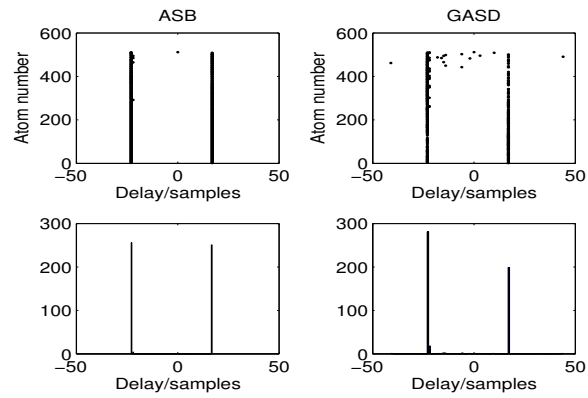

Fig. 2. Time-delays (upper plots) and their histograms (lower plots) estimated for ASB and GASD, under anechoic mixing.

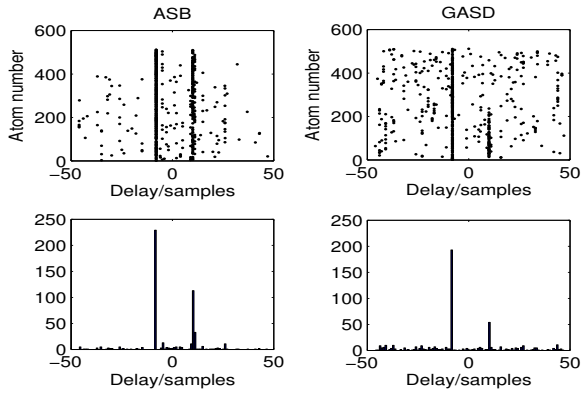

Fig. 3. Time-delays (upper plots) and their histograms (lower plots) estimated for ASB and GASD, under echoic mixing, with a reverberation time of 320 ms.

Number of Atoms	Method	SDR	SIR	SAR
512	GASD	0.4	5.4	3.0
	ASB	-37.8	-0.4	-2.9
400	GASD	0.4	5.4	3.0
	ASB	-28.9	1.1	-2.2
200	GASD	0.4	5.7	2.6
	ASB	-16.9	-.6	-2.7
100	GASD	0.3	5.9	1.6
	ASB	-9.1	-2.1	-3.4
50	GASD	0.2	5.5	1.1
	ASB	-4.8	-2.4	-3.7
30	GASD	0.1	4.1	1.3
	ASB	-2.6	-2.8	-3.6

Table 1. Objective performance of GASD and ASB. All values are expressed in decibels (dB).

values indicate that the algorithm has failed to recover the target, while negative SAR values, indicate that large artifacts are present; together they result in negative SDR values. The results suggest that the sources recovered with GASD remain of similar quality even when the number of atoms used in the reconstruction where reduced by a fifth, and performance did not deteriorate drastically when even fewer than 100 atoms were used. An informal listening test was also conducted, and it supported these results. In contrast, the performance of ASB worsens as the number of atoms are reduced, with the SIR consistently falling when fewer than 400 atoms are used for reconstruction.

5. CONCLUSIONS

A greedy adaptive sparse dictionary algorithm that learns an orthogonal dictionary from the data has been presented. The algorithm was used to separate anechoic and echoic stereo mixtures of speech signals, hence yielding basis vector pairs, which capture spatial information about the mixing channel. This was exploited to cluster the atom pairs, and thus reconstruct the original source signals. The method was shown to correctly identify the time-delays corresponding to each source, both in the anechoic and echoic situations, and to reconstruct the source signals with fewer atoms than those extracted.

6. REFERENCES

[1] H. Sawada, R. Mukai, S. Araki, and 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*, vol. 12, pp. 530–538, 2004.

[2] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[3] M. G. Jafari and J. A. Chambers, "Fetal electrocardiogram extraction by sequential source separation in the wavelet domain," *IEEE Trans. on Biomedical Engineering*, vol. 52, pp. 390–400, 2005.

[4] R. Gribonval and S. Lesage, "A survey of sparse component analysis for blind source separation: principles, perspectives, and new challanges," in *Proceedings of the 2006 European Symposium on Artificial Neural Networks (ESANN '06)*, 2006, pp. 323–330.

[5] M. G. Jafari, E. Vincent, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "An adaptive stereo basis method for convolutive blind audio source separation," *Neurocomputing*, 2008, To appear.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustic, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[7] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide," Tech. Rep. 1706, IRISA, http://www.irisa.fr/metiss/bss_eval/, 2005.