Blind Non-stationnary Sources Separation by Sparsity in a Linear Instantaneous Mixture

Bertrand Rivet

GIPSA-lab, CNRS UMR-5216, Grenoble INP,
Domaine Universitaire, BP 46, 38402 Saint Martin d'Hères cedex, France
bertrand.rivet@gipsa-lab.inpg.fr

Abstract. In the case of a determined linear instantaneous mixture, a method to estimate non-stationnary sources with non activity periods is proposed. The method is based on the assumption that speech signals are inactive in some unknown temporal periods. Such silence periods allow to estimate the rows of the demixing matrix by a new algorithm called Direction Estimation of Separating Matrix (DESM). The periods of sources inactivity are estimated by a generalised eigen decomposition of covariance matrices of the mixtures, and the separating matrix is then estimated by a kernel principal component analysis. Experiments are provided with determined mixtures, and shown to be efficient.

1 Introduction

Blind source separation consists of estimating unknown signals (denoted sources) from mixtures of them without prior knowledge neither about the nature of mixing function nor about the sources. When involved sources have specific properties, the source separation can be based on them leading thus to semi-blind source separation methods. For instance when speech signals are present among the sources, non-stationarity [10,13], or sparse decomposition in a specific basis [17,1,2] have been exploited. In parallel, the bi-modal (audio-visual) nature of speech was used [14,16,11,12]. Audiovisual speech source separation is based on the strong links which exist between the sound produced by a speaker and visual speech signals, in particular speaker's lips movement: these methods exploit the complementarity and the redundancy of these modalities.

The proposed method is based on the "sparsity" of speech signals. Indeed, they are highly non-sationnary: there are a lot of lapses of time during which the signal power is negligible compared its averaged power, for instance time between words. The proposed method draws from [12] where the silent moments of a speaker (estimated by a purely visual voice activity detection [3]) are used to identify the function which allows to extract this specific speech signal. Even if this audiovisual approach is efficient (even in the convolutive case), it requires a specific device to record simultaneously audio and video signals by microphones and camera, respectively. In the present study, the instantaneous mixing case is addressed by a purely acoustic method which estimates jointly the voice non-activity periods and the separation matrix.

T. Adali et al. (Eds.): ICA 2009, LNCS 5441, pp. 314–321, 2009.

This paper is organised as follows. Section 2 presents the proposed approach to exploit the natural speech sparsity, while Section 3 describes DESM algorithm to estimate sources. Numerical experiments and results are given in Section 4 before conclusions and perspectives in Section 5.

2 Exploitation of Natural Speech "Sparsity"

In this section, the general framework of source separation with instantaneous mixture is recalled before introducing the proposed method to exploit the natural sparsity of speech.

Let $\mathbf{s}(t) \in \mathbb{R}^{N_s}$ denotes the N_s dimensional column vector of source signals whose j-th component is $s_j(t)$. With instantaneous mixtures, observations $x_i(t)$ are expressed as a linear combination of sources $s_j(t)$: $x_i(t) = \sum_j a_{i,j} s_j(t)$, or with matrix notation

$$\mathbf{x}(t) = A\,\mathbf{s}(t), \tag{1}$$

where $A \in \mathbb{R}^{N_m \times N_s}$ denotes the mixing matrix whose (i, j)-th entry is $a_{i,j}$ and $\mathbf{x}(t) \in \mathbb{R}^{N_m}$ is the N_m dimensional column vector of the observations. In this study, the determined case is considered: the number of mixtures N_m is so equal to the number of sources N_s. The source estimation problem is then equivalent to estimate a separating matrix $B \in \mathbb{R}^{N_s \times N_s}$ such that

$$\mathbf{y}(t) = B\,\mathbf{x}(t) \tag{2}$$

is a vector whose components are the estimate of sources $s_i(t)$.

The independant component analysis (ICA) [6,4], which exploits the mutual independance between the sources, was widely used to solve this problem. Recently sparsity was introduced in source separation [8]. For instance, methods proposed in [17,1,2] are based on the assumption that, in some basis, there exist some parts where at most one source is present at the time allowing thus to estimate the mixing matrix. Indeed, if at time index τ, only source $s_n(\tau)$ is active (*i.e.*, $\forall i \neq n, s_i(\tau) = 0$) then $\mathbf{x}(\tau) = \mathbf{a}_n s_n(\tau)$. In other words, mixtures $\mathbf{x}(\tau)$ are proportional to n-th column $\mathbf{a}_n$ of mixing matrix A. It is thus possible to estimate all the columns of the mixing matrix, and to express the separating matrix B as the inverse of the estimated mixing matrix.

The proposed method is quite different since it is based on the assumption that there exist some time indexes where at least one source is inactive: *i.e.* for $t = \tau$, $\exists\, n \;/\; s_n(\tau) = 0$. Let suppose, in this section, that all the sources are stationnary excepted one, let say $s_1(t)$ without any loose of generality. Let R_1 denotes the covariance matrix of observations $\mathbf{x}(t)$ computed for all time indexes t, and let R_2 denotes covariance matrix of observations computed during an inactivity period of source $s_1(t)$. The proposed method is based on the general eigenvalue decomposition of couple (R_2, R_1) [15]. It is easy to check that (R_2, R_1) admits only two disctint generalised eigenvalues: 1 degenerated $N_s - 1$ times (whose eigensubspace $\mathcal{E}$ is a hyperplan complementary to $\mathbf{a}_1$), and 0 whose generalised eigenvector $\mathbf{v}$ is orthogonal to $\mathcal{E}$. Thus the projection of observations

$\mathbf{x}(t)$ on generalised eigenvector $\mathbf{v}$ allows to extract source $s_1(t)$ by cancelling the contribution of other sources: $\forall i \neq 1,\ \mathbf{v}^T \mathbf{a}_i = 0$, since $\mathbf{a}_i$ for $i \neq 1$ are in hyperplan $\mathcal{E}$.

This method allows first to detect if source $s_1(t)$ vanishes by testing generalised eigenvalues and then to extract source $s_1(t)$ when it is active by projecting the observations on the generalised eigenvector associated with the generalised eigenvalue equal to zero.

3 DESM Algorithm

In the previous section, only one source was considered to be non-stationnary with non active periods. However, several sources can be non active (possibly in different periods): for instance if the mixtures contain several speech sources. Moreover, the inactivity periods are unknown. In this section, the proposed Direction Estimation of Separating Matrix (DESM) algorithm is presented: it extends the previously proposed method (Section 2) to extract from the mixture all the sources with non active periods.

To detect time periods where at least one source is non active, we proposed to compute the generalised decomposition of couple $\{(R_2(\tau), R_1)\}_\tau$ where R_1 is the covariance matrix of observations $\mathbf{x}(t)$ estimated with all time samples and $R_2(\tau)$ is the covariance of observations $\mathbf{x}(t)$ estimated on windowed samples around τ (typically, this window is about 100 milliseconds). The generalised eigen decompositions of $\{(R_2(\tau), R_1)\}_\tau$ provide

$$R_2(\tau)\, \Phi(\tau) = R_1\, \Phi(\tau)\, \Lambda(\tau), \tag{3}$$

where $\Lambda(\tau)$ is a diagonal matrix whose diagonal terms $\lambda_1(\tau) \leq \cdots \leq \lambda_{N_s}(\tau)$ are the generalised eigenvalues and $\Phi(\tau)$ is an orthonormal matrix whose columns $\phi_i(\tau)$ are the generalised eigenvectors. Thus at τ time, if N sources are inactive then N generalised eigenvalues are null whose associated generalised eigenvectors defined a subspace orthogonal to the subspace spanned by the $N_s - N$ active sources.

The proposed DESM algorithm can thus be decomposed in two steps:

1. the first one is to detect the periods where at least one source is inactive by testing the generalised eigenvalues $\{\lambda_1(\tau)\}_\tau$: if $\lambda_1(\tau) \leq \eta$, where η is a threshold chosen *a priori*, then the algorithm decides that at least one source was inactive during time window centred on τ. Let $\Theta = \{\tau \mid \lambda_1(\tau) \leq \eta\}$ be the set of time indexes where at least one source is inactive (the cardinal of Θ is N_τ). This provides a set of vectors $\{\phi_1(\tau)\}_{\tau \in \Theta}$ defined as the set of the first generalised eigenvector with $\tau\ \in \Theta$. These vectors are mainly aligned in the directions which allows to extract corresponding sources (Fig. 2). These direction are the rows of the separating matrix.
2. then the DESM algorithm estimates these directions thanks to a kernel principal component analysis (kernel PCA) [9,7], where the kernel is chosen as

$$k\big(\phi_1(t), \phi_1(t')\big) = k_{t,t'} \triangleq \begin{cases} \frac{\phi_1^T(t)\phi_1(t') - \cos\theta_0}{1-\cos\theta_0}, & \text{if } \phi_1^T(t)\phi_1(t') \geq \cos\theta_0 \\ 0, & \text{else} \end{cases} \tag{4}$$

with t and t' in Θ, and where θ_0 is an angle which is chosen *a priori*. Kernel PCA consists in performing an eigen decomposition of matrix $K \in \mathbb{R}^{N_\tau \times N_\tau}$ whose (i, j)-th entry is $k_{i,j}$:

$$K = \Psi \Delta \Psi^T, \tag{5}$$

where Δ is a diagonal matrix of eigenvalues of K, and Ψ is an orthonormal matrix whose columns are eigenvectors of K. Let $W = [\psi_1, \cdots, \psi_{N_s}]$ be the matrix composed by the concatenation of N_s eigenvectors ψ_i associated with the N_s largest eigenvalues.

The separation matrix is then obtained by

$$B = W^T\, K\, V, \tag{6}$$

where $V = [\phi_1(t \in \Theta)]$ is the matrix obtained by the concatenation of generalised eigenvector associated with the smallest generalised eigenvalue $\lambda_1(t)$ (3) with $t \in \Theta$. The sources are finally estimated thanks to

$$\hat{s}(t) = B\mathbf{x}(t), \tag{7}$$

for all time indexes t, including those when sources are active.

Finally, DESM algorithm which allows to extract non-stationnary sources with inactive periods, is summarised in Algorithm 1.

Algorithm 1. DESM algorithm

```
 1: Compute covariance matrix $R_1$ from all time samples
 2: for each $\tau$ do
 3:     Compute covariance matrix $R_2(\tau)$ with time window centred on $\tau$
 4:     Compute generalised eigen decomposition (3) of couple $(R_2(\tau), R_1)$
        $\Rightarrow \ (\Phi(\tau), \Lambda(\tau))$
 5: end for
 6: Estimate $\Theta = \{\tau \mid \lambda_1(\tau) \leq \eta\}$
 7: Compute matrix $K$ defined by (4)
 8: Perform eigen decomposition (5) of $K \ \Rightarrow \ (\Psi, \Delta)$
 9: Compute $W = [\psi_1, \cdots, \psi_{N_s}]$ and $V = [\phi_1(t \in \Theta)]$
10: Compute $B = W^T\, K\, V$ (6)
11: Estimate sources by $\hat{\mathbf{s}}(t) = B\mathbf{x}(t)$
```

Note that using generalised eigenvalues of couple $(R_1, R_2(\tau))$ in stage 4, instead of using simple eigenvalues of $R_2(\tau)$, overcomes the problem of relative power of sources. In particular when some of the sources are definitely less powerful than others, using eigenvalues can lead to consider that these sources are inactive.

4 Numerical Experiments

In this section, the principle of the proposed DESM algorithm is illustrated to extract speech signals from linear instantaneous mixtures of audio sources (*i.e.* speech and musical sources). The sources are from two databases: the first one is composed of 18 French sentences read by French speakers (males and females), the second one is composed of music signals. All signals were sampled at 16kHz. In the different tested configurations, the sources are randomly chosen and the entries of the mixing matrix are randomly chosen from a uniform random variable distributed from -1 to 1. For each configuration (*i.e.* for each number of sources) 100 different mixtures were tested.

In the first experiment, the extraction of two speech signals from three mixtures is illustrated (Fig. 3). One of the source is thus a musical signal without inactive periods, the second in this example. First of all, the estimation of non-activity periods thanks to generalised eigen decomposition is illustrated on Fig. 1. As one can see on the top plot, which represents the power of the three sources computed with a time-sliding window of 100ms, the two speech sources have (possibly overlapped) non-activity periods while the musical source has its short term power almost constant. It is quite interesting to note that the smallest generalised eigenvalue $\lambda_1(t)$ (bottom plot) allows to detect these inactivity periods, without labbeling which speech signals are inactive. Moreover, generalised eigenvectors $\phi_1(t)$ (Fig. 2) are mainly in two directions corresponding to the rows of the separating matrix that extract the two speech signals. Fig. 3 shows that the proposed DESM algorithm is efficient to extract speech signals ($\hat{s}_1(t)$ et $\hat{s}_2(t)$). Moreover, the third estimated source is still a mixture of the three sources since kernel PCA of matrix K_1 only presents two significant eigenvalues. More generaly, the number of significant eigenvalues of matrix K_1 could be used to estimate the number of speech sources to only extract these sources.

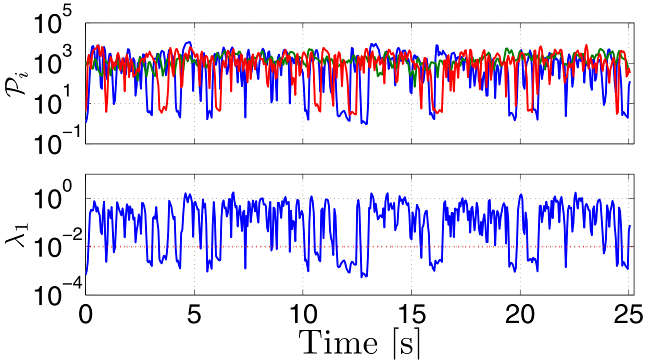


Fig. 1. Estimation of non-activity periods thanks to DESM algorithm. Top figure shows the power of the three sources on a time-sliding window of 100ms (blue, green and red curves for the 1st, 2nd and 3rd source, respectively). Bottom figure shows the smallest generalised eigenvalue (3) $\lambda_1(\tau)$ (blue curve) as well as chosen threshold η (red dotted curve). Plots are in logarithm scale.

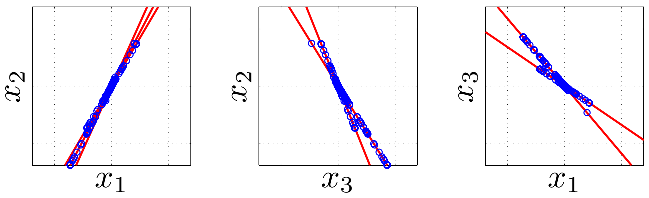
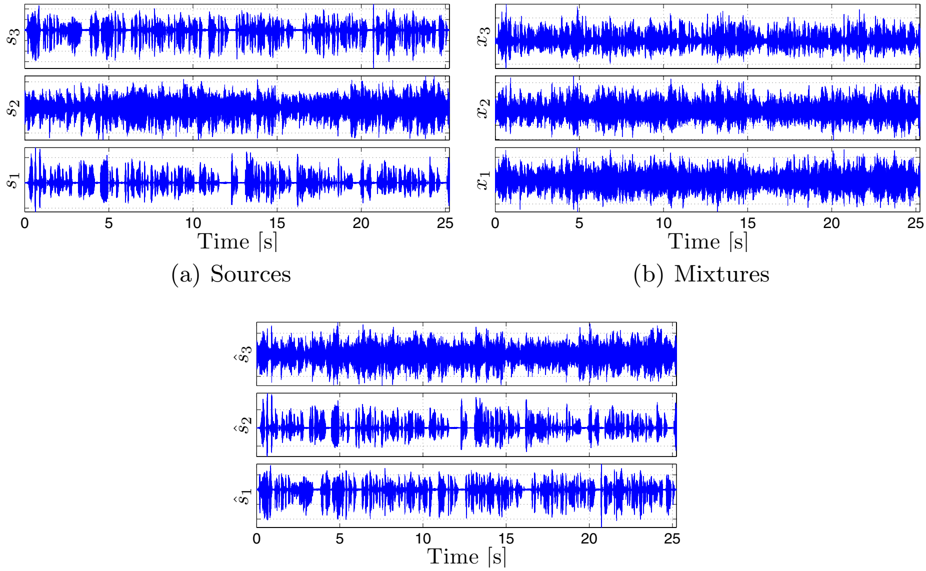


Fig. 2. Estimation of separating matrix B (6) thanks to DESM algorithm with three sources. Projections of estimated rows (red curves) and generalised eigenvectors $\phi_1(t)$ with $t \in \Theta$ (blue points) on (x_1, x_2), (x_2, x_3) et (x_1, x_3). Generalised eigenvectors are multiplied by $1/\lambda_1(t)$.



(a) Sources

(b) Mixtures

(c) Estimated sources

Fig. 3. Illustration of DESM algorithm

In the second experiment (Fig. 4), the performance of the proposed algorithm was estimated. To evaluate the estimation of the rows of the separating matrix, we use the performance index defined as

$$PI = \sum_{i \in \mathcal{S}} \sum_{j} \left| \frac{C_{i,j}}{\max_k |C_{i,k}|} \right| - 1, \text{ with } C = B\,A, \tag{8}$$

where $\mathcal{S}$ denotes the set of speech sources. So the smaller the performance index is, the better the extraction is. In these experiments, only two sources are speech sources, all the other sources are musical signal. Figure 4 shows the median performance index versus the number of sources. The performance achieved by

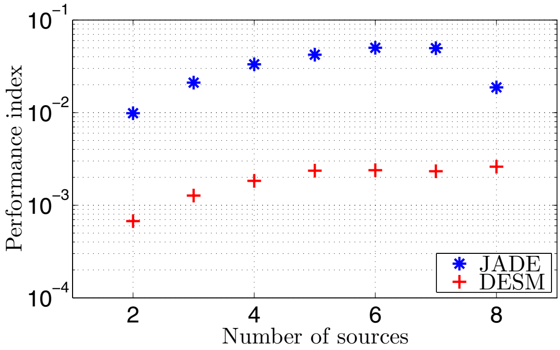


Fig. 4. Performance index

the proposed DESM algorithm are compared with the performance provided by the JADE algorithm [5]. As one can see, the DESM algorithm compares favourably with the JADE algorithm, even with numerous sources.

5 Conclusions and Perspectives

In this paper, a new algorithm denoted DESM (Direction Estimation of Separating Matrix) is proposed to extract the sources with non active periods from a linear instantaneous mixture. The detection of these inactive periods allows to estimate the separating matrix which is then used to extract these sources when they are active. The proposed algorithm was tested with different configurations and shown to be efficient at a low computational cost. Even if in this study, the purpose was to extract speech sources, the DESM algorithm can be used in a more general context (*i.e.* to extract any "sparse" sources). In perspective, this methods could be used with convolutive mixtures in the frequency domain. However, this leads to the classical permutation problem [4] which could be fixed by one of the numerous methods proposed in the literature.

References

1. Abrard, F., Deville, Y.: A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. Signal Processing 85(7), 1389–1403 (2005)
2. Arberet, S., Gribonval, R., Bimbot, F.: A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture. In: Proc. ICA, Charleston, USA, pp. 536–543 (March 2006)
3. Aubrey, A., Rivet, B., Hicks, Y., Girin, L., Chambers, J., Jutten, C.: Two novel visual voice activity detectors based on appearance models and retinal filltering. In: Proc. European Signal Processing Conference (EUSIPCO), Poznan, Poland, pp. 2409–2413 (September 2007)

4. Cardoso, J.-F.: Blind signal separation: statistical principles. Proceedings of the IEEE 86(10), 2009–2025 (1998)
5. Cardoso, J.-F., Souloumiac, A.: Blind beamforming for non Gaussian signals. IEE Proceedings-F 140(6), 362–370 (1993)
6. Comon, P.: Independent component analysis, a new concept? Signal Processing 36(3), 287–314 (1994)
7. Desobry, F., Févotte, C.: Kernel PCA based estimation of the mixing matrix in linear instantaneous mixtures of sparse sources. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, vol. 5, pp. 669–672 (May 2006)
8. Gribonval, R., Lesage, S.: A survey of Sparse Component Analysis for Blind Source Separation: principles, perspectives, and new challenges. In: Proc. European Symposium on Artificial Neural Networks (ESANN), Bruges, pp. 323–330 (April 2006)
9. Muller, K.-R., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B.: An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks 12(2), 181–201 (2001)
10. Parra, L., Spence, C.: Convolutive blind separation of non stationary sources. IEEE Transactions on Speech and Audio Processing 8(3), 320–327 (2000)
11. Rivet, B., Girin, L., Jutten, C.: Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. IEEE Transactions on Speech and Audio Processing 15(1), 96–108 (2007)
12. Rivet, B., Girin, L., Jutten, C.: Visual voice activity detection as a help for speech source separation from convolutive mixtures. Speech Communication 49(7-8), 667–677 (2007)
13. Servière, C., Pham, D.-T.A.: A novel method for permutation correction in frequency-domain in blind separation of speech mixtures. In: Puntonet, C.G., Prieto, A.G. (eds.) ICA 2004. LNCS, vol. 3195, pp. 807–815. Springer, Heidelberg (2004)
14. Sodoyer, D., Girin, L., Jutten, C., Schwartz, J.-L.: Developing an audio-visual speech source separation algorithm. Speech Communication 44(1–4), 113–125 (2004)
15. Souloumiac, A.: Blind source detection and separation using second order non-stationarity. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Detroit, USA, vol. 3, pp. 1912–1915 (May 1995)
16. Wang, W., Cosker, D., Hicks, Y., Sanei, S., Chambers, J.A.: Video assisted speech source separation. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, USA (March 2005)
17. Yilmaz, Ö., Rickard, S.: Blind Separation of Speech Mixtures via Time-Frequency Masking. IEEE Transactions on Signal Processing 52(7), 1830–1847 (2004)