BLIND SPEECH SEPARATION FOR CONVOLUTIVE MIXTURES USING AN ORIENTED PRINCIPAL COMPONENTS ANALYSIS METHOD

Y. Benabderrahmane[1], S. A. Selouani[2], and D. O'Shaughnessy[1]

[1] INRS-EMT, 800 de la Gauchetière O, H5A 1K6, Montréal, QC, Canada,

[2]Université de Moncton, campus de Shippagan E8S 1P6 NB, Canada

ABSTRACT

This paper deals with blind speech separation of convolutive mixtures of sources. The separation criterion is based on the Oriented Principal Components Analysis (OPCA) method. OPCA is a (second order) extension of standard Principal Component Analysis (PCA) aiming at maximizing the power ratio of a pair of signals. The convolutive mixing is obtained by modeling the Head Related Transfer Function (HRTF). Experimental results show the efficiency of the proposed approach in terms of subjective and objective evaluation, when compared to the widely used CFICA (Convolutive Fast-ICA) algorithm.

Index Terms— Blind source separation (BSS), convolutive mixture, speech signals, Oriented Principal Component Analysis

1. INTRODUCTION

The objective of Blind Source Separation (BSS) is to extract the original source signals from their mixtures and possibly to estimate the unknown mixing channel using only the information of the observed signal with no, or very limited, knowledge about the source signals and the mixing channel. Methods for this problem can be divided into methods using second-order [1] or higher-order statistics [2], the maximum likelihood principle [3], the Kullback-Liebler distance [4] PCA methods, non-linear PCA [5], and ICA methods [2], [4], [6]. Further information on these methods and some applications of ICA can be found in [7]. Most approaches to BSS assume the sources are statistically independent and thus often seek solutions of separation criteria using higher-order statistical information [2] or using only second-order statistical information in cases where the sources have temporal coherency [3], are non-stationary [4], or eventually are cyclo-stationary. We must note that second-order methods do not actually replace higher-order ones since each approach is based on different assumptions. For example, second-order methods assume that the sources are temporally coloured whereas higher-order methods assume white sources. Another difference is that higher-order methods do not apply to Gaussian signals but second-order methods do not have any such constraint.

This paper is organized as follows: in Section 2, we present the mixing model. In section 3, we present the separation model. In Section 4 we briefly describe the implementation of the OPCA method that we propose for the separation of mixed speech signals. Section 5 presents the experimental results and discusses them. Finally, Section 6 concludes and gives a perspective of our work.

2. THE MIXING MODEL

At the discrete time index t, a set of M source signals $s(t) = (s_1(t), \ldots, s_M(t))$ is received at an array of N sensors. The received signals are denoted $x(t) = (x_1(t), \ldots, x_N(t))$. In many real-world applications the sources are said to be *convolutively* (or dynamically) mixed. The convolutive model introduces the following relation between the n'th mixed signal and the original source signals.

The real convolutive mixing process (including delays) can be assumed as:

$$x_m(t) = \sum_{n=1}^{N} \sum_{k=0}^{K-1} a_{mnk} s_n(t-k), \qquad (1)$$

The mixed signal is a linear mixture of filtered versions of each of the source signals, and a_{mnk} represents the corresponding mixing filter coefficients. In practice, these coefficients may also change in time, but for simplicity the mixing model is often assumed stationary.

In matrix form, the convolutive model can be written as:

$$x(t) = \sum_{k=0}^{K-1} A_k s(t-k), \qquad (2)$$

where A_k is an $M \times N$ matrix which contains the k'th filter coefficients.

The convolutive mixing process in eq. (2) can be simplified by transforming the mixtures into the frequency domain. The linear convolution in the time domain can be written in the frequency domain as separate multiplications for each frequency:

$$X(f) = A(f)S(f). \qquad (3)$$

At each frequency f, $A(f)$ is a complex $M \times N$ matrix, $X(f)$ is complex $M \times 1$ vector, and similarly $S(f)$ is a complex $N \times 1$

 ISSN 2076-1465

vector. The frequency transformation is typically computed using a discrete Fourier transform (DFT) within a time frame of size T starting at time t:

$$X(f,t) = DFT([x(t),....x(t+T-1)]), \quad (4)$$

and correspondingly for $S(f, t)$. Often a windowed discrete Fourier transform is used:

$$X(f,t) = \sum_{\tau=0}^{T-1} w(\tau)x(t+\tau)e^{-j2\pi f\tau/T} \quad (5)$$

where the window function $\boldsymbol{w(\tau)}$ is chosen to minimize band-overlap. By using the fast Fourier transform (FFT) convolutions can be implemented efficiently in the discrete Fourier domain.

3. THE SEPARATION MODEL

The objective of blind source separation is to find an estimate, $\hat{s}(t)$, which is a model of the original source signals $s(t)$. For this, it may not be necessary to identify the mixing filters A_k explicitly. Instead, it is often sufficient to estimate separation filters W that remove the cross-talk introduced by the mixing process (figure 1).

The goal in source separation is not necessarily to recover identical copies of the original sources. Instead, the aim is to recover model sources without interferences from other sources; each separated signal $\hat{s}_n(t)$ should contain signals originating from a single source only. Therefore, each model source signal can be a filtered version of the original source signals,

$$\hat{S}(f,t) = W(f)A(f)S(f,t) \quad (6)$$

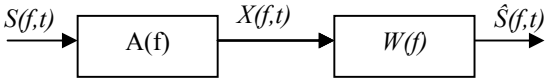

Figure 1: Source separation system

The criterion for separation is satisfied if the recovered signals are permuted, and possibly scaled and filtered, versions of the original signals,

$$W(f)A(f) = P(f)D(f), \quad (7)$$

where P is a permutation matrix and $D(f)$ is a diagonal matrix with scaling filters on its diagonal. If one can identify $A(f)$ exactly and choose $W(f)$ to be its inverse, then $D(f)$ is an identity matrix, and one recovers the sources exactly.

A survey of frequency-domain BSS is provided in [8]. An advantage of blind source separation in the frequency domain is that the separation problem can be decomposed into smaller problems for each frequency bin in addition to the significant gains in computational efficiency [9]. The convolutive mixture problem is reduced to "instantaneous" mixtures for each frequency. Another problem that arises in the frequency domain is the permutation and scaling ambiguity. If the convolutive problem is treated for each frequency as a separate problem, the source signals in each frequency bin may be estimated with an arbitrary permutation and scaling,

$$\hat{S}(f,t) = P(f)D(f)S(f,t). \quad (8)$$

4. OPCA METHOD

The OPCA algorithm was previously proposed by Diamantaras and Papadimitriou [10], specifically for separating four multilevel PAM (Pulse Amplitude Modulation) signals filtered by an ARMA (Auto-Regressive Moving Average) coloring filter. In this work we aim to use OPCA to perform a BSS on a convolutive mixture of speech signals according to the model illustrated in Figure 2.

OPCA can be considered as a generalization of PCA. It corresponds to the generalized eigenvalue decomposition of a pair of covariance matrices in the same way that PCA corresponds to the eigenvalue decomposition of a single covariance matrix. Oriented PCA (OPCA) describes an extension of PCA involving two signals $u(k)$ and $v(k)$. The aim is to identify the so-called oriented principal directions $e_1,...,e_n$ that maximize the signal-to-signal power ratio $E(e_i^T u)^2/E(e_i^T v)^2$ under the orthogonality constraint: $e_i^T R_u e_j = 0$, $i \neq j$. OPCA is a second-order statistics method, which reduces to standard PCA if the second signal is spatially white $R_v = I$. The solution of OPCA, as shown in Figure 2, is a generalized eigenvalue decomposition of the matrix pencil $[R_u, R_v]$. Subsequently, we shall relate the BSS problem with the OPCA analysis of the observed signal x and almost any filtered version of it. Note that the 0-lag covariance matrix of $x(k)$ is:

$$R_x(0) = AR_S(0)A^T = AA^T \quad (9)$$

Now, consider a scalar, linear filter having $h=[h_0,...,h_M]$ (referred to as J-Filter in Figure 2) operating on $X(f,t)$:

$$Y(f,t) = \sum_{m=0}^{M} H_m X(f, t-l_m). \quad (10)$$

The 0-lag covariance matrix of Y is expressed as:

$$R_Y(0) = E\{Y(f,t)Y(f,t)^T\} = \sum_{p,q}^{M} H_p H_q R_X(l_p - l_q). \quad (11)$$

From Eq. (1) it follows that:

$$R_X(l_m) = AR_S(l_m)A^T. \quad (12)$$

So

$$R_Y(0) = ADA^T, \quad (13)$$

with

$$D = \sum_{p,q=0}^{M} H_p H_q R_S(l_p - l_q). \quad (14)$$

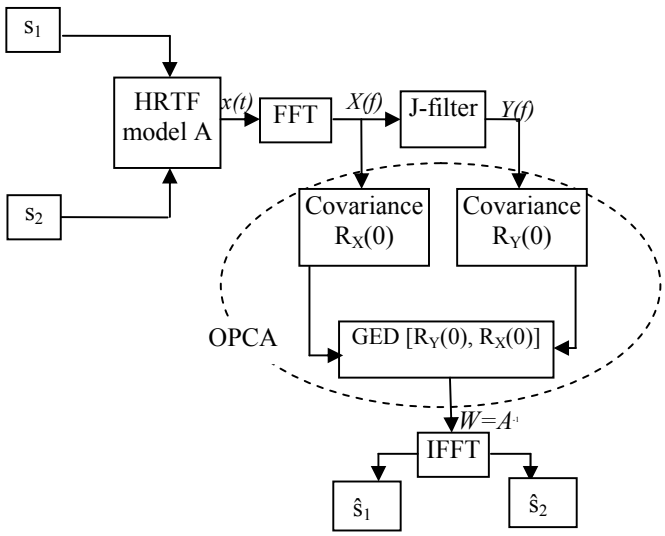

Figure 2: Block diagram of BSS for convolutive mixtures using the OPCA method.

Provided that A is square and invertible we can write:

$$R_Y(0)A^{-T} = AD = AA^T A^{-T} D = R_X(0)A^{-T} D\,. \qquad (15)$$

Eq. (15) expresses a Generalized Eigenvalue Decomposition problem for the matrix pencil $[R_Y(0), R_X(0)]$. This is equivalent to the OPCA problem for the pair of signals $[Y(f,t), X(f,t)]$. The generalized eigenvalues for this problem are the diagonal elements of D. The columns of the matrix A^{-T} are the generalized eigenvectors. The eigenvectors are unique up to a permutation and scale provided that the eigenvalues are distinct (this is true in general). In this case, for any generalized eigenmatrix W we have $W = A^{-T}P$ with P being a scaled permutation matrix; each row and each column contains exactly one non-zero element. Then the sources can be estimated as:

$$\hat{S}(f,t) = W^T X(f,t)\,, \qquad (16)$$

which can be written as:

$$\hat{S}(f,t) = P^T A^{-1} AS(f,t) = P^T S(f,t)\,, \qquad (17)$$

where $\hat{S}(f,t)=[\hat{S}_1(f,t), \hat{S}_2(f,t)]^T$ is the estimated source signal vector and $W(f)$ represents an unmixing matrix at frequency bin f. The unmixing matrix $W(f)$ is determined so that $\hat{S}_1(f,t)$ and $\hat{S}_2(f,t)$ become mutually uncorrelated, because the source signals $S_1(f,t)$ and $S_2(f,t)$ are assumed to be zero mean and mutually uncorrelated. The estimated sources are equal to the true ones except for the (unobservable) arbitrary order and scale.

Then we apply the IFFT of $\hat{S}(f,t)$ for recovering the estimated signals in time domain.

$$\hat{s}(t) = IFFT\left(\hat{S}(f,t)\right)$$

The J-filter mentioned in Figure 2 is expressed as:

$$h = [h_0, h_1, h_2] = [1, \alpha, \beta], \qquad (18)$$

where α and β are parameters to be fixed. These parameters are optimized by re-formulating the D matrix of eq. (15) as the following [10]:

$$D = (1+\alpha^2+\beta^2)I + 2\alpha Rs(I_\alpha) + 2\beta R_s(I_\beta) + 2\alpha\beta Rs(I_\alpha - I_\beta). \qquad (19)$$

Note that the optimality criterion of the J-filter is related to the eigenvalue spread [10]. The maximization criterion used to find α and β is given by:

$$J(\alpha,\beta) = \min_i \left[\min_{j\neq i} \frac{(d_i - d_j)^2}{\max_k d_k^2} \right], \qquad (20)$$

where $d_{i,j}$ represents the diagonal elements of D. In our experiments, the J-filter order of 3 was chosen. The search of the optimal filter is transformed into the search for the filter that spreads the eigenvalues as much as possible [10]. The search is exhaustive and is performed for values of α and β varying within a given interval of h ($\forall\ \alpha, \beta \in [h_{min}, h_{max}]$). In the experiments we fixed h_{min} = -5, h_{max} = 5, while the increasing step was 0.2.

5. EXPERIMENTS AND RESULTS

In the following experiments the TIMIT database was used. The TIMIT corpus contains broadband recordings of a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States, each reading 10 phonetically rich sentences [11]. Some sentences of the TIMIT database were chosen to evaluate our BSS methods. We tested OPCA using a filter of order 3, as mentioned earlier. The use of more correlation matrices increases the information input in the estimation process and then improves the separation quality. We consider a two-input, two-output convolutive BSS problem, so we mixed in convolution two speech signals: $s_1(n)$ and $s_2(n)$, that respectively pronounced by a man and a woman.

In the experiment a dummy head with two microphones (one in each ear) was used instead of the microphone array. This kind of recording was used to investigate how effective the BSS is during a more natural configuration of the sources. This situation takes into account all the changes in an acoustic field connected with the head, i.e., the Head Related Transfer Function. The HRTF influences both sound pressure level and spectra of the source signals reaching the ears.

We tested our overall framework with a mixing filter measured at the ears of a dummy head. We selected impulse responses associated with source positions defined by 30 and -80-degree angles in relation to the dummy head as we can see in figure 3.

To evaluate our approach in the convolutive case, we compared it with the well-known C-FICA and DUET techniques.

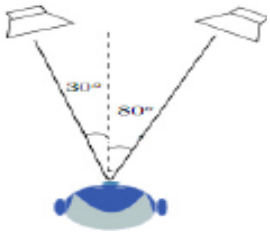

Figure 3: The convolutive (HRTF) model with source positions at 30-and 80-degree angles in relation to the dummy head

- The C-FICA algorithm (Convolutive extension of Fast-ICA) [7] is a time-domain fast fixed-point algorithm that realizes blind source separation of convolutive mixtures. It is based on a convolutive sphering process (or spatio-temporal sphering) that lets the use of the classical Fast-ICA updates extract iteratively the innovation processes of the sources in a deflation procedure.
- DUET (Degenerate Unmixing and Estimation Technique) is a method that applies when sources are W-disjoint orthogonal, that is, when the time-frequency representations of any two signals in the mixtures are disjoint sets. The method uses an online algorithm to perform gradient search for the mixing parameters and simultaneously construct binary time-frequency masks that are used to partition one of the mixtures to recover the original source signals [12].

Through this comparison, we aim to demonstrate the effectiveness of the proposed separation technique based on the OPCA method. The OPCA method is effective, as can be seen in the time domain, where we note that the original signals (Figure 5) and estimated signals by OPCA (Figure 7) are very close. The OPCA method has the advantage that the time processing is less than with the C-FICA algorithm. With our experimental setup, the OPCA method takes 45 sec while the C-FICA technique takes 55 sec to be performed. These results were achieved with a computer whose specifications are: Processor: Intel (R) Core(TM) 2 Quad CPU Q9550 @ 2.83 GHz, RAM: 4 GB, OS: Windows 7 professional 64-bit. The method was verified subjectively by listening to the original, mixed and separated signals. We obtained a very good separation.

To measure the speech quality, one of the reliable methods is the Perceptual Evaluation of Speech Quality (PESQ). This method is standardized in ITU-T recommendation P.862 [13]. PESQ measurement provides an objective and automated method for speech quality assessment. As illustrated in Figure 4 [14], the measure is performed by using an algorithm comparing a reference speech sample to the speech sample processed by a system. Theoretically, the results can be mapped to relevant mean opinion scores (MOS) based on degradation of the sample [15]. The PESQ Algorithm is designed to predict subjective opinion scores of a degraded

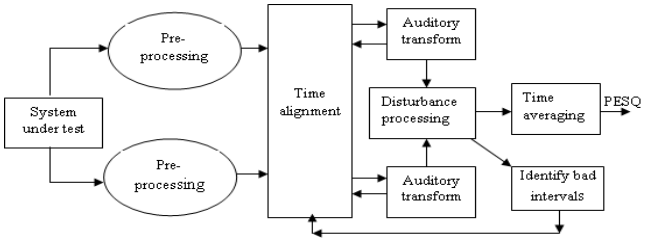

Figure 4: Block diagram of the PESQ measure computation

speech sample. PESQ returns a score from 0.5 to 4.5, with higher scores indicating better quality. For our experiments we used the code provided by Loizou in [14]. This technique is generally used to evaluate speech enhancement systems. Usually, the reference signal refers to an original (clean) signal and the degraded signal refers to the same utterance pronounced by the same speaker as in the original signal but submitted to diverse adverse conditions. In the PESQ algorithm, the reference and degraded signals are level-equalized to a standard listening level thanks to the preprocessing stage. The gain of the two signals is not known a priori and may vary considerably. In the original PESQ algorithm, the gains of the reference, degraded and corrected signals are computed based on the root mean square values of band-passed-filtered (350-3250 Hz) speech. The full frequency band is kept in our scaled version of normalized signals. The filter with a response similar to that of a telephone handset, existing in the original PESQ algorithm, is also removed. The PESQ method is used throughout all our experiments to evaluate the OPCA estimated speech. It has the advantage to be independent of listeners and number of listeners.

For PESQ evaluation, OPCA was the best one in comparison with C-FICA and DUET approach, which we can see in Table 1. We note the very good improvement in PESQ of OPCA method compared to mixed signals.

Table1: Comparison of PESQ for the C-FICA, DUET, OPCA methods and convolved mixed signals without any processing

PESQ	PESQ (female speech)	PESQ (male speech)
Mixed Signals	0.78	1.2
C-FICA	1.686	1.734
DUET	0.417	1.002
OPCA	3.599	3.696

In frequency domain algorithms, the challenge is to solve the permutation ambiguity, i.e., to make the permutation matrix *P(f)* independent of frequency. Especially when the number of sources and sensors is large, recovering consistent permutations is a severe problem. With N model sources there are N! possible permutations in each frequency bin [8]. Many frequency domain algorithms provide *ad hoc* solutions, which solve the permutation ambiguity only partially, thus requiring a combination of different methods. The problem is not very severe in our case, because we work with two sources.

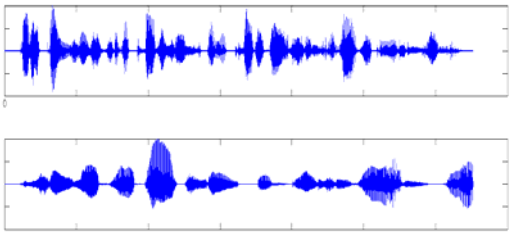

Figure 5: (a): Original signals: Male sentence: "This brings us to the question of accreditation of art schools in general", (b): female sentence: "She had your dark suit in greasy wash water all year".

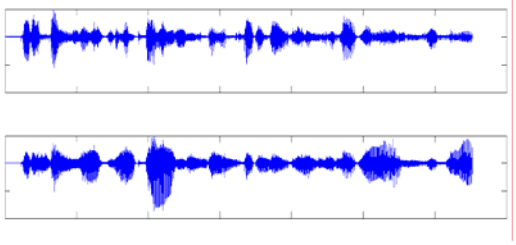

Figure 6: Signals mixed by convolution.

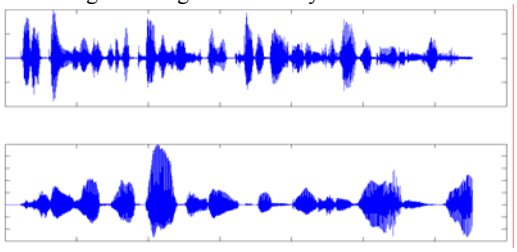

Figure 7: Estimated signals by the OPCA method.

6. CONCLUSION

We have presented a blind speech separation technique of convolution mixtures using an oriented principal component analysis method. All earlier approaches have consistently used two steps: one pre-processing (sphering) step followed by a second-order analysis method such as PCA .The OPCA approach has the advantage that no pre-processing step is required as sphering is implicitly incorporated in the signal-to-signal ratio criterion which is optimized by OPCA [10].

The proposed separation technique of mixed observations into source estimates is effective, as shown in the time domain. Subjective evaluation is performed through listening to the estimated signals before and after mixing and after separation was used. The results are very satisfactory; we obtained a very good separation. We tested the method with other speech signals from the TIMIT, Noizeus and AURORA databases. We experimented also with other types of mixtures (e.g. like anechoic) and the results were similar. These results confirm the efficiency of the OPCA method that we previously used for the first time, in the separation of speech signals in an instantaneous mixing case [16]. For future work, we will use mixtures of more than two sources, and also we are continuing our research efforts by implementing a combination of OPCA and different methods, for resolving the problem of permutation ambiguity and applying it in a mobile communication framework.

6. REFERENCES

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A Blind Source Separation Technique Using Second-Order Statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[2] J.-F. Cardoso, "Source separation using higher order moments," in *Proc. IEEE ICASSP,* Glasgow, U.K., 1989, vol. 4, pp. 2109–2112.

[3] J. Basak and S. Amari, "Blind separation of uniformly distributed signals: A general approach," *IEEE Trans. Neural Networks*, vol. 10, pp. 1173–1185, September 1999.

[4] D.T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. Signal Processing*, vol. 44, pp. 2768–2779, November 1996.

[5] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, vol. 7, pp. 113–127, 1994.

[6] A. Hyvarinen and E. Oja, "A Fast Fixed-point Algorithm for Independent Component Analysis", *Neural Computation*, Vol. 9, No. 6, pp. 1483-92, 1997.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, "*Independent Component Analysis"*, John Wiley, NY, 2001.

[8] M. S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra, "A Survey of Convolutive Blind Source Separation Methods", Springer Handbook on Speech Processing and Speech Communication, 2007.

[9] S. Makino, H. Sawada, R. Mukai, and S. Araki,"Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1640–1655, Jul 2005.

[10] K. I. Diamantaras, Th. Papadimitriou, "Oriented PCA and Blind Signal Separation", *4th International Symposium on Independent Component Analysis and Blind Signal Separation* (ICA 2003), pp 609-613, April 2003, Nara, Japan.

[11] W. Fisher, G. Dodington, & K. Goudie-Marshall, The TIMIT-DARPA speech recognition research database: Specification and status, *DARPA Workshop on Speech Recognition,* 1986.

[12] S. Makino, T.W. Lee, H. Sawada, "Blind Speech Separation", *Signals and Communication Technology*, published by Springer, 2007.

[13] ITU, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs", *ITU-T Recommendation* 862, 2000.

[14] P Loizou, 2007. "Speech Enhancement: Theory and Practice". *CRC Press LLC, Boca Raton*, FL, 2007.

[15] ITU-T Recommendation P.800, "Methods for Subjective Determination of Speech Quality", *Intern. Telecommunication Union,* Geneva, 2003.

[16] Y. Benabderrahmane, S.A. Selouani, and D. O'Shaughnessy, H. Hamam, "A Comparative Study of Blind Speech Separation using Subspace Methods and Higher Order Statistics", *Lecture Notes in Computer Science, Springer eds*., pp. 117-124, 2009.