SOLVING THE INDETERMINATIONS OF BLIND SOURCE SEPARATION OF CONVOLUTIVE SPEECH MIXTURES

Bertrand Rivet, Laurent Girin

Speech Communication Institute
Grenoble National Polytechnic Institute
Grenoble, France

Christian Jutten

Image and Signal Processing Laboratory
Grenoble National Polytechnic Institute
Grenoble, France

ABSTRACT

Looking at the speaker's face seems useful to better hear a speech signal and extract it from competing sources before identification. In this paper, we present a novel algorithm plugging audiovisual coherence of speech signals, estimated by statistical tools, on audio blind source separation (BSS) algorithms in the difficult case of convolutive mixtures. The algorithm mainly works in the frequency (transform) domain, where the convolutive mixture becomes an additive mixture for each frequency channel. Frequency by frequency separation is made by an audio BSS algorithm, and the audiovisual information is used to solve the standard source permutation and scale factor problems at the output of the separation stage, for each frequency. The proposed method is shown to be efficient in the case of 2×2 convolutive mixtures.

1. INTRODUCTION

Looking at the speaker's face seems useful to better hear a speech signal and to extract it from competing sources before identification [1]. Schwartz et *al.* [2] attempted to show that vision may enhance audio speech in noise and therefore provide what they called a "very early" contribution to speech intelligibility, different and complementary to the classical lipreading effect. This suggests to elaborate new speech enhancement or extraction techniques exploiting the audiovisual coherence of speech stimuli. Girin et *al.* [3] developed a technological implementation of this idea: a first system for automatically enhancing audio speech embedded in white noise by using filters which parameters were partly estimated from the video input. Then Sodoyer et *al.* [4] have developed an approach exploring the link between two signal processing streams that were completely separated: sensor fusion in audiovisual speech processing on the one hand, and blind source separation (BSS) techniques [5, 6] on the other hand. They have proposed to use a statistical model of audiovisual coherence to estimate the separating matrix in the case of a simple additive mixture.

In this study, we focus on the more complex problem of convolutive mixtures of speech signals, where the permutation and scale factor indeterminations arise for each frequency bin. In a recent paper [7], we proposed an approach to the permutation problem based on a statistical model of the audiovisual coherence of speech signals. In this paper, we complete and thus improve the estimation of the sources by a complementary approach to solve the scale permutation problem exploiting the marginal audio model extracted from the audiovisual one.

This paper is organized as follows. Section 2 introduces the BSS problem of convolutive mixtures. Section 3 explains the audiovisual approach to improve the estimation of the speech sources. Section 4 proposes numerical experiments before conclusions and perspectives in section 5.

2. BSS OF CONVOLUTIVE MIXTURES

Let us consider the case of a stationary convolutive mixture of N sources $\mathbf{s}(m) = [s_1(m), \cdots, s_N(m)]^T$ to be separated from P observations $\mathbf{x}(m) = [x_1(m), \cdots, x_P(m)]^T$ (^{T} denoting the tranpose):

$$x_p(k) = \sum_{n=1}^{N} \sum_{m=-\infty}^{\infty} h_{p,n}(m) s_n(k - m) \qquad (1)$$

The filters $\{h_{p,n}(m)\}$, that model the impulse response between each source $s_n(k)$ and the pth sensor, are entries of the mixing filter matrix $\{\mathcal{H}(m)\}$. The goal of the BSS is to recover the sources by using a dual filtering process:

$$\hat{s}_n(k) = \sum_{p=1}^{P} \sum_{m=-\infty}^{\infty} g_{n,p}(m) x_p(k - m) \qquad (2)$$

where $\{g_{n,p}(m)\}$ are entries of the demixing filter matrix $\{\mathcal{G}(m)\}$ which are estimated such that the components of the output signals vector $\hat{\mathbf{s}}(m) = [\hat{s}_1(m), \cdots, \hat{s}_N(m)]^T$ are as mutually independent as possible. This problem is generally considered in the dual frequency domain where the
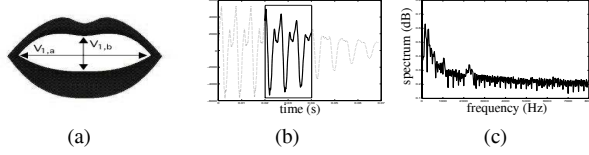
Fig. 1. Audio and video parameters. Fig. 1(a) shows the two video parameters. Fig. 1(b) displays the temporal windowed signal and Fig. 1(c) its spectral local characteristics (obtained by the FFT of the segment in Fig. 1(b)).

equations (1) and (2) lead to

$$S_x(m, f) = \mathcal{H}(f)S_s(m, f)\mathcal{H}^*(f) \qquad (3)$$

$$S_{\hat{s}}(m, f) = \mathcal{G}(f)S_x(m, f)\mathcal{G}^*(f) \qquad (4)$$

where $S_s(m, f)$, $S_x(m, f)$ and $S_{\hat{s}}(m, f)$ are the time varying power spectrum density matrices of respectively the sources $\mathbf{s}(m)$, the observations $\mathbf{x}(m)$ and the output $\hat{\mathbf{s}}(m)$. $\mathcal{H}(f)$ and $\mathcal{G}(f)$ are the frequency response matrices of the mixing and demixing filter matrices (* denoting the conjugated transpose).

If we assume that the sources are mutually independent (or at least decorrelated) $S_s(m, f)$ is a diagonal matrix and an efficient separation must lead to a diagonal matrix $S_{\hat{s}}(m, f)$. Thus, a basic criterion for BSS is to adjust the matrix $\mathcal{G}(f)$ so that $S_{\hat{s}}(m, f)$ is as diagonal as possible. This can be done by the joint diagonalization process described in [8], and in the following we use this method. The well-known crucial limitation of the BSS problem is that for each frequency bin, $\mathcal{G}(f)$ can only be provided up to a scale factor and a permutation between the sources:

$$\mathcal{G}(f) = \mathcal{P}(f)\mathcal{D}(f)\hat{\mathcal{H}}^{-1}(f) \qquad (5)$$

where $\mathcal{P}(f)$ and $\mathcal{D}(f)$ are arbitrary permutation and diagonal matrices. Pham *et al.* [9] proposed to reconstruct the frequency response $\{\mathcal{G}(f)\}$ by exploiting the continuity between consecutive frequency bins. They select the permutations that assume a smooth reconstruction of the frequency response and they do not solve the scale factor problem.

3. SOLVING INDETERMINATION PROBLEMS

In this section, we first present the statistical model of the audiovisual coherence of speech signals. Then we present how this model can be used to successively solve the permutation and the scale factor indeterminations.

3.1. Modeling the audiovisual coherence of speech

We assume that we want to extract a particular speech source, say $s_1(t)$, from the audio mixtures $\mathbf{x}(t)$ and we exploit additional observations, which consist of a video signal $\mathbf{v}_1(t)$

extracted from the speaker's face and synchronous with the acoustic signal $s_1(t)$. This video signal consists of the trajectory of basic geometric lip shape parameters. The speaker's lip parameters and the local spectral characteristics of the acoustic signal are related by a complex relationship which can be described in statistical terms. Hence, we assume that we can build a statistical model providing the joint probability $p_{AV}(\mathbf{A}_1(t), \mathbf{v}_1(t))$ of a video vector $\mathbf{v}_1(t) = [V_{1,a}(t), V_{1,b}(t)]^T$ containing the lip internal width and height (Fig 1(a)) and an audio vector $\mathbf{A}_1(t) = [A_1(t, f_1), \cdots, A_1(t, f_L)]^T$ containing local spectral characteristics (Fig. 1(c)). These video and audio vectors represent the useful information of a signal frame (Fig. 1).

$$(\mathbf{v}_1(t), \mathbf{A}_1(t)) \sim \sum_{i=1}^{I} \omega_i \mathcal{N}\left(\mu_i^{AV}, \Gamma_i^{AV}\right) \qquad (6)$$

where $\mathcal{N}(\mu, \Gamma)$ is the Gaussian distribution of mean vector μ and covariance matrix Γ. $\{\omega_i, \mu_i^{AV}, \Gamma_i^{AV}\}$ are the parameters of the ith Gaussian kernel.

3.2. Permutation ambiguity

Regularizing the permutation problem of frequency domain BSS consists in searching the permutations set $\{\hat{\mathcal{P}}(f)\}$ that assume $\hat{\mathbf{A}}_{1, \{\hat{\mathcal{P}}(f)\}}(t) \simeq \mathbf{A}_1(t)$, where $\hat{\mathbf{A}}_{1, \{\mathcal{P}(f)\}}(t)$ are the estimated audio coefficients of the real audio parameters $\mathbf{A}_1(t)$ given by (5) from the BSS algorithm of [9] up to the permutation matrices $\{\mathcal{P}(f)\}$. To estimate $\{\mathcal{P}(f)\}$, we propose to minimize the audiovisual criterion $J_{AV}(\{\mathcal{P}(f)\}, t)$ between the audio spectrum output on channel 1 and the visual information $\mathbf{v}_1$:

$$\{\hat{\mathcal{P}}(f)\} = \arg \min_{\{\mathcal{P}(f)\}} J_{AV}(\{\mathcal{P}(f)\}, t) \qquad (7)$$

with

$$J_{AV}(\{\mathcal{P}(f)\}, t) = -\log\left[p_{AV}\left(\hat{\mathbf{A}}_{1, \{\mathcal{P}(f)\}}(t), \mathbf{v}_1(t)\right)\right] \qquad (8)$$

To improve the criterion, we introduce the possibility to cumulate the probabilities over time. For this purpose, we assume that the values of audio and visual characteristics at several consecutive time frames are independent from each other and we define an integrated audiovisual criterion by:

$$J_{AV}^T(\{\mathcal{P}(f)\}) = \sum_{t=0}^{T-1} J_{AV}(\{\mathcal{P}(f)\}, t) \qquad (9)$$

Since there are $(N!)^L$ possible permutation matrices if the short term Fourier transform is calculated over L frequencies, it is not possible to attempt an exhaustive research, because of huge computational load. To overcome this, we already proposed an original algorithm in two stages [7]. First, we use a dichotomic scheme during which we simplify the criterion (9) by marginalizing the audiovisual probability $p_{AV}(\mathbf{A}_1(t), \mathbf{v}_1(t))$ regarding sets of frequencies.

Then, we use the previous estimation of the permutation matrices as the initialization of the second stage, in which we exploit the joint criterion (9) using a recursive scheme.

3.3. Scale factor ambiguity

To solve the scale factor problem, we propose a novel process exploiting the audio model achieved by marginalizing the audiovisual model (6) regarding the video parameters. Resulting from the audiovisual probability, the model of the audio parameters is also a mixture of Gaussian kernels

$$\mathbf{A}_1(t) \sim \sum_{i=1}^{I} \omega_i \mathcal{N}\left(\mu_i^A, \Gamma_i^A\right) \tag{10}$$

Now, regularizing the scale factor ambiguity consists in searching, for each frequency bin f, the parameter $\alpha(f)$ which leads to $\alpha(f)A_1^\dagger(t,f) = A_1(t,f)$ where $A_1^\dagger(t,f)$ is the estimated audio parameter reached after our permutation cancellation (*i.e.* $A_1^\dagger(t,f) = \hat{A}_{1,\hat{P}(f)}(t,f)$). To estimate $\alpha(f)$ we propose to exploit the audio model (10). Indeed, the variance of $\alpha(f)A_1^\dagger(t,f)$ verifies

$$\mathrm{Var}\left(\alpha(f)A_1^\dagger(t,f)\right) = \mathrm{Var}(A_1(t,f)) \tag{11}$$

where $\mathrm{Var}(\cdot)$ is the variance operator. Moreover

$$\mathrm{Var}(A_1(t,f)) = \sum_{i=1}^{I} \omega_i \gamma_i^A(f) + \sum_{i=1}^{I} \omega_i \left(\mu_i^A(f)\right)^2 \cdots$$
$$\cdots - \left(\sum_{i=1}^{I} \omega_i \mu_i^A(f)\right)^2 \tag{12}$$

where $\gamma_i^A(f)$ is the fth diagonal coefficient of the covariance matrix Γ_i^A and $\mu_i^A(f)$ the fth coefficient of the mean vector μ_i^A. Thus we propose to estimate $\alpha(f)$ thanks to

$$\hat{\alpha}(f) = \sqrt{\frac{\mathrm{Var}\left(A_1(t,f)\right)}{\mathrm{Var}\left(A_1^\dagger(t,f)\right)}} \tag{13}$$

where $\mathrm{Var}(A_1(t,f))$ is defined by (12) and $\mathrm{Var}(A_1^\dagger(t,f))$ is estimated by a classical variance estimator.

4. NUMERICAL EXPERIMENTS

In the following, we consider the case of two sources mixed by 2×2 matrices of filters. All mixing filters are artificial finite impulse response filters up to 64 lags with 3 significant echos. They fit a simplified acoustic model of a room impulse response. The local audio parameters $\mathbf{A}_1(t)$ are obtained by the short term Fourier transform. Moreover, we need synchronous video parameters $\mathbf{v}_1(t)$ and local audio parameters $\mathbf{A}_1(t)$. Since the video channel is sampled
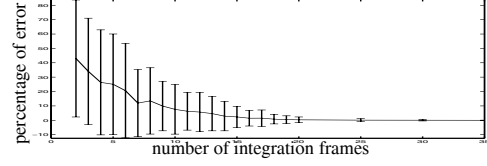

Fig. 2. Percentage of detection errors versus number of integration frames.

at 50Hz, we choose the length of the temporal block equal to 20ms and the audio signals are sampled at 16kHz. In this study we choose a limited number of audio parameters to the permutation cancellation so we subdivide the spectrum into 32 consecutive bands of 250Hz and we calculate the energy of the signal in these bands. Furthermore, the joint statistical model of the audiovisual information consists in a mixture of 16 Gaussian kernels estimated during a training phase using the EM algorithm [10]. The corpus consisted in 110 phonetically well-balanced sentences in French. The audiovisual model was trained by using the 80 first sentences, representing about 7000 audiovisual vectors for each speaker (different male and female speakers were tested with one audiovisual model for each one of them). The overall process was tested with the 30 sentences not used for training.

Fig. 2 shows the percentage of detection error[1] versus the number of integration frames: the solid line is the mean and the error bars are the standard deviations (this simulation is repeated over 40 different speech sentences). This stresses the importance of frames integration for the criteria. Choosing around 20 frames of integration seems to be a good trade-off between computation time and detection error.

Fig. 3 presents an example of the achieved separation: Fig 3(a) shows the two sources, Fig 3(e) the mixtures and Fig 3(b) the estimated sources given by the BSS algorithm, without permutation cancellation. Fig. 3(f) shows the spectrum $|\mathcal{G}(f)\mathcal{H}(f)|$ of the global filter $(\mathcal{G} * \mathcal{H})(n)$. Fig. 3(c) (resp. Fig. 3(g)) displays the estimated sources (resp. the spectrum of the global filter) after our permutation cancellation algorithm (*cf.* 3.2). One can see that, for all frequencies f, $|(\mathcal{G}\mathcal{H})_{12}(f)|$ (resp. $|(\mathcal{G}\mathcal{H})_{21}(f)|$) is much smaller than $|(\mathcal{G}\mathcal{H})_{11}(f)|$ (resp.$|(\mathcal{G}\mathcal{H})_{22}(f)|$). This first means that our algorithm found all the permutations and also that the two sources are nearly separated. However we can see that the diagonal terms of the global filter are not equal to the identity. Thus, the estimated sources are the original sources modified by these filters. Using our estimation of the scale factor (*cf* 3.3), we achieve the source estimation as shown Fig. 3(h): this global filter is closer to the identity than the

---

[1]The permutation errors contain both the unsolved permutations (actual permutations undetected by our algorithm) and the wrong permutations (bad decision of the algorithm).
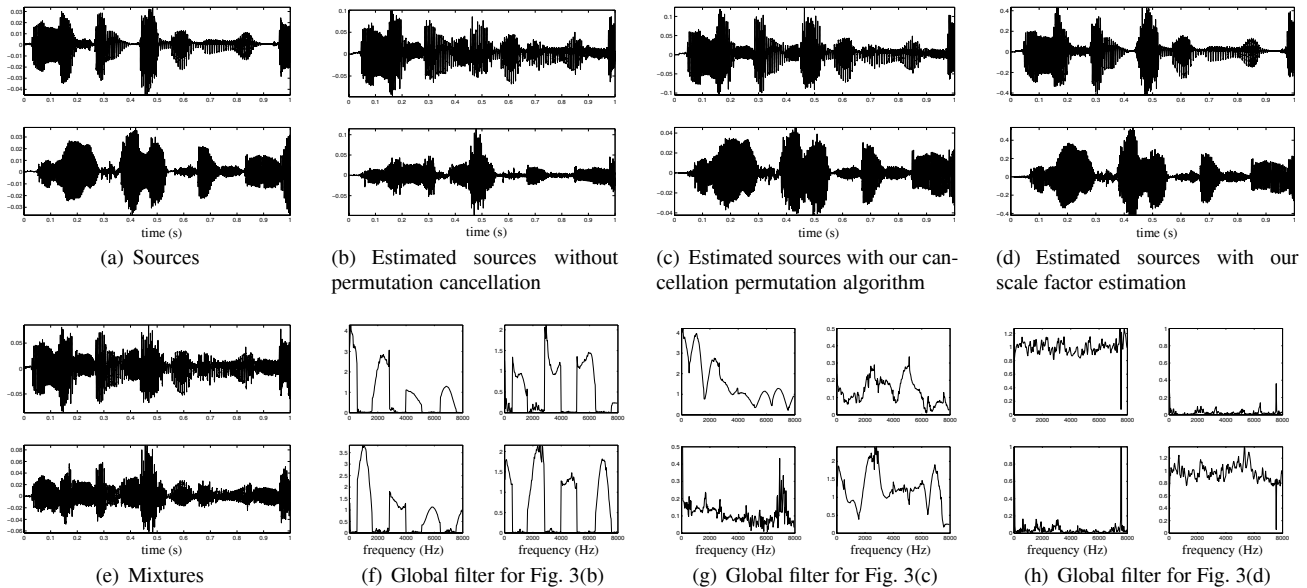
(a) Sources

(b) Estimated sources without permutation cancellation

(c) Estimated sources with our cancellation permutation algorithm

(d) Estimated sources with our scale factor estimation

(e) Mixtures

(f) Global filter for Fig. 3(b)

(g) Global filter for Fig. 3(c)

(h) Global filter for Fig. 3(d)

Fig. 3. Sources, mixtures, estimated sources and global filters magnitude spectrum.

Fig. 3(g) one. So the estimated sources (Fig. 3(d)) are much closer to the original sources.

5. CONCLUSIONS AND PERSPECTIVES

The BBS problem of convolutive speech mixtures can be achieved by using a joint diagonalization process in the time-frequency domain [9]. However, this only gives a solution up to a permutation and a scale factor. In this paper, we proposed a new method to overcome these problems exploiting the audiovisual coherence of speech and the audio model resulting from it. We also showed the importance to cumulate the probabilities on consecutive frames in order to adequately exploit this coherence.

As a further step, we will also extend the permutation cancellation to use directly all the frequency spectrum coefficients instead of a limited number of audio parameters. It is important to note that although the presented results concerned the mixing of two speech sources, our algorithm can be used to extract a speech signal corrupted by any kind of noisy environment. This point is part of our future works.

6. REFERENCES

[1] K. Grant and P. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences." *J. Acoust. Soc. Am.*, vol. 108, pp. 1197–1208, 2000.

[2] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Audio-visual scene analysis; evidence for a "very-early" integration process in audio-visual speech perception." in *Proc. ICSLP'2002*, 2002, pp. 1937–1940.

[3] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. of Am.*, vol. 109, no. 6, pp. 3007–3020, June 2001.

[4] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources." in *Eurasip JASP*, 2002, pp. 1164–1173.

[5] J.-F. Cardoso, "Blind signal separation : statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, October 1998.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[7] B. Rivet, L. Girin, C. Jutten, and J.-L. Schwartz, "Using audiovisual speech processing to improve the robustness of the separation of convolutive speech mixtures." in *MMSP 2004*, Siene, Italy, October 2004.

[8] D.-T. Pham, "Joint approximate diagonalization of positive definite matrices," *SIAM J. Matrix Anal. And Appl.*, vol. 22, no. 4, pp. 1136–1152, 2001.

[9] D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of convolutive audio mixtures using non-stationary," in *ICA 2003*, Nara, Japan, April 2003.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Ser. B.*, vol. 39, 1977.