Open access • Journal Article • DOI:10.1109/TASL.2006.872619

Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals From Convolutive Mixtures — Source link

Bertrand Rivet, Laurent Girin, C. Jutten

Institutions: École Normale Supérieure

Published on: 01 Jan 2007 - IEEE Transactions on Audio, Speech, and Language Processing (IEEE)

Topics: Speech enhancement, Speech processing, Speech coding, Blind signal separation and Audio signal processing

Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals From Convolutive Mixtures

Bertrand Rivet, Laurent Girin, Christian Jutten

Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals From Convolutive Mixtures

Bertrand Rivet, Laurent Girin, and Christian Jutten, *Member, IEEE*

***Abstract*—Looking at the speaker's face can be useful to better hear a speech signal in noisy environment and extract it from competing sources before identification. This suggests that the visual signals of speech (movements of visible articulators) could be used in speech enhancement or extraction systems. In this paper, we present a novel algorithm plugging audiovisual coherence of speech signals, estimated by statistical tools, on audio blind source separation (BSS) techniques. This algorithm is applied to the difficult and realistic 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. The audio and visual informations are modeled by a newly proposed statistical model. This model is then used to solve the standard source permutation and scale factor ambiguities encountered for each frequency after the audio blind separation stage. The proposed method is shown to be efficient in the case of 2×2 convolutive mixtures and offers promising perspectives for extracting a particular speech source of interest from complex mixtures.**

***Index Terms*—Audiovisual coherence, blind source separation, convolutive mixture, speech enhancement, statistical modeling.**

I. Introduction

FOR understanding speech, "two senses are better than one" [1]: we know, since [2], that lip-reading improves speech identification in noise since there exists an intrinsic coherence between audition and vision for speech perception. Indeed, they are both consequences of the articulatory gestures. This coherence can be exploited in adverse environments, where audio and visual signals of speech are complementary [3]. Indeed, acoustic speech features that are robust in noise are generally poorly visible (e.g., voicing/unvoicing). On the contrary, the phonetic contrasts less robust in noisy auditory perception are the most visible ones, both for consonants [3] and vowels [4]. Thus, visual cues can compensate to a certain extent the deficiency of the auditory ones. This explains that the fusion of auditory and visual informations meets a great success in several speech applications, mainly in speech recognition in noisy environments, from the pioneer works of Petajan [5] to more recent studies, e.g., [6].

The recent discovery by Grant and Seitz [7], confirmed in [8] and [9], that vision of the speaker's face intervenes in the audio *detection* of speech in noise, suggests that for *hearing* speech also, two senses are better than one. On this basis, Schwartz *et al.* [10] 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 lip-reading effect. In parallel, Girin *et al.* developed in [11] a technological counterpart of this idea: a first audiovisual system for automatically enhancing audio speech embedded in white noise by using filters whose parameters were partly estimated from the video input. Deligne *et al.* [12] and Goecke *et al.* [13] provided an extension of this work using more powerful techniques. Also, their audio-visual speech enhancement system was applied to speech recognition in adverse environments.

The extension of the speech enhancement problem to the separation of multiple simultaneous speech/audio sources is a major issue of interest in the speech/audio processing area. This problem is sometimes referred to as the "cocktail-party" problem. It is a quite difficult problem to address, even when several sensors are used, since both the signals and the recorded mixtures are complex: e.g., signals are nonstationary in time and space, mixtures are convolutive since signals are reflected and attenuated along the way to the sensors. To deal with this problem in the case of speech signals, Girin *et al.* [14] and then Sodoyer *et al.* [15], [16] have extended the previous work of [11] to a more general and hopefully more powerful approach. They began to explore the link between two signal processing streams that were completely separated: sensor fusion in audio-visual speech processing and blind source separation (BSS) techniques.

The problem generally labeled under the denomination "source separation" consists in recovering signals $\mathbf{s}(t)$, also called sources, from mixtures of them $\mathbf{x}(t)$, typically signals recorded by a sensor array. In the blind context, both the sources $\mathbf{s}(t)$ and the mixing process $\mathcal{H}$ are unknown: this situation is called the blind source separation (BSS) [17]–[19]

$$\mathbf{x}(t) = \mathcal{H}\left(\mathbf{s}(t)\right). \tag{1}$$

The lack of prior knowledge about the sources and the mixing process is generally overcame by a statistically strong assumption: the independence between the sources. Hence, the method

Manuscript received January 31, 2005; revised October 31, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shoji Makino.

B. Rivet and L. Girin are with the Institut de la Communication Parlée, Ecole Nationale d'Electronique et de Radioélectricité, Grenoble 38000, France (e-mail: rivet@icp.inpg.fr; rivet@lis.inpg.fr; girin@icp.inpg.fr).

C. Jutten is with the Images and Signals Laboratory, Ecole Nationale d'Electronique et de Radioélectricité, Grenoble 38000, France (e-mail: Christian.Jutten@lis.inpg.fr).

Digital Object Identifier 10.1109/TASL.2006.872619

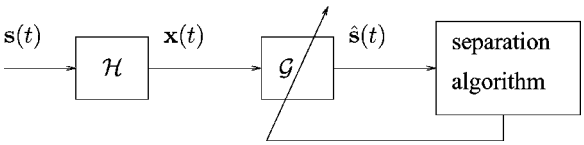

Fig. 1. BSS principle.

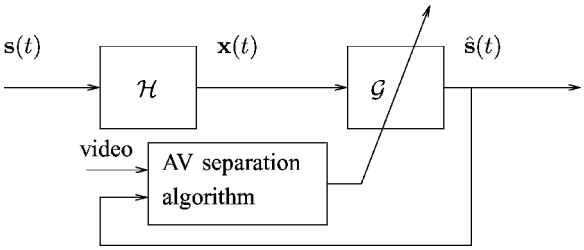

Fig. 2. BSS using audiovisual bimodality of speech.

for solving BSS using independence is called the independent component analysis (ICA) [20]. In this case, a demixing process $\mathcal{G}$ is estimated so that the reconstructed sources $\hat{\mathbf{s}}(t)$ are as independent as possible (Fig. 1)

$$\hat{\mathbf{s}}(t) = \mathcal{G}\left(\mathbf{x}(t)\right). \tag{2}$$

Introduced first by Hérault and Jutten [21], [22] in the middle of the 1980s, the BSS became an attractive field of signal processing due to the few prior knowledge and its wide range of applications. In the 1990s, many efficient algorithms have been developed, especially JADE [23], FAST-ICA [24], Infomax [25], and the theoretical statistical frameworks of ICA and BSS have been proposed especially by Cardoso [17] and Comon [20]. Consequently, a large number of application fields exploited the BSS model: in biomedical signal processing (e.g., extraction of the fetal ECG [26], [27]), communication, audio (see [19] for more references).

Recently, as mentioned above, Sodoyer *et al.* have proposed in [15] and [16] to introduce the audiovisual bimodality of speech in BSS in order to improve the separation of the acoustic signal thanks to visual information (Fig. 2). The principle of their study was the following. Instead of estimating the demixing process using a criterion based on the independence of the source, they proposed to use a criterion based on the audiovisual (AV) coherence: one speech source of interest is extracted using the visual information simultaneously recorded from the speaker's face by video processing. The proposed system was shown to efficiently estimate the separating matrix in the case of a simple instantaneous additive mixture. Later, Dansereau [28] and Rajaram *et al.* [29] also proposed an audiovisual speech source separation system, respectively, plugging the visual information in a 2×2 decorrelation system with first-order filters and in the Bayesian framework for a $2\times$ 2 linear mixture. Unfortunately, as mentioned above, real speech/audio mixtures are generally more complex, and better described in terms of convolutive mixtures. Therefore, the aim of this paper is to explore the audiovisual speech source separation problem in the more realistic case of convolutive mixtures.

In this paper, we address the problem in the dual frequency domain, as already proposed in [30]–[33] for the classical acoustic convolutive mixture problem. However, the audiovisual coherence is not used directly to estimate demixing matrices (this task is achieved by using pure audio techniques), but to solve the indeterminacies[1]generally encountered by separation techniques based on the independence assumption: permutation and scale ambiguities that are described in the following. The preliminary works [34] and [35], in which we showed how audiovisual processing could be used to estimate the indeterminacies, are the basis of this study. In these papers, we only exploited a limited number of power audio parameters whereas in the present study, we extend the AV model to the log-modulus of the coefficients of the short-time Fourier transform, and to all the frequency bins, using the statistical modeling [36]. Moreover, we present a new AV process to estimate the scale indeterminacy, as well as a new bootstrap algorithm to improve both the scale and permutation cancellation. Finally, we show that the new AV model is efficient.

This paper is organized as follows. Section II introduces the BSS problem for speech convolutive mixtures. In Section III, we present the audiovisual model that is plugged into the presented BSS system, together with the audiovisual data that feed this model. Section IV presents how this audiovisual model is used to solve the permutation and scale indeterminacies. Experiments on speech signals are presented in Section V, and the results and possible extensions of this work are discussed in Section VI.

II. BSS of Convolutive Mixtures

In the case of a stationary convolutive mixing process [31], [37], $\mathcal{H}$ is a linear filter matrix. Thus, the P observations $\mathbf{x}(t) = [x_1(t), \ldots, x_P(t)]^T$ are the sum of the contributions of the N sources $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$, each of them being filtered by a row of $\mathcal{H}(t)$ (^{T} is the transpose operator)

$$x_p(t) = \sum_{n=1}^{N} \sum_{m=-\infty}^{+\infty} h_{p,n}(m) s_n(t-m), \quad 1 \le p \le P \tag{3}$$

where the entries of the mixing filter matrix[2] $\{\mathcal{H}(t)\}_t$ are the filters $\{h_{p,n}(t)\}_t$ which model the impulse response between the source $s_n(t)$ and the pth sensor. So rewriting (3) using matrix form, leads to

$$\mathbf{x}(t) = \mathcal{H}(t) * \mathbf{s}(t). \tag{4}$$

[1]Typically, for instantaneous linear mixtures, $\mathcal{H}$ and $\mathcal{G}$ are modeled by matrices and there are two indeterminacies, permutation and scale, which are easy to understand. In fact, each sensor observes a mixture $x_j(t) = \sum_{i-1}^{N} h_{j,i} s_i(t)$, and it is possible to observe the same mixture with different sources (new amplitude and new order): $x_j(t) = \sum_{i-1}^{N} (h_{j,i}/a_i)(a_i s_i(t))$.

[2]In this paper, $\{a(b)\}_b$ denotes the set of elements $a(b)$, for all b.

In the convolutive case, the demixing process $\mathcal{G}$ is chosen as a linear filter matrix with entries $\{g_{n,p}(t)\}_t$, so that we have

$$\hat{s}_n(t) = \sum_{p=1}^{P} \sum_{m=-\infty}^{\infty} g_{n,p}(m) x_p(t-m), \quad 1 \leq n \leq N. \quad (5)$$

The demixing filters of $\mathcal{G}$ are estimated so that the components of the output signal vector $\hat{\mathbf{s}}(t) = [\hat{s}_1(t), \ldots, \hat{s}_N(t)]^T$ are as mutually independent as possible. Rewriting (5) in a matrix form leads to

$$\hat{\mathbf{s}}(t) = \mathcal{G}(t) * \mathbf{x}(t). \quad (6)$$

Several authors [30]–[33] have proposed to consider the problem in the dual frequency domain where convolutions become multiplications. Indeed, the mixing process (4) leads to

$$\forall f, \quad S_x(t,f) = \mathcal{H}(f) S_s(t,f) \mathcal{H}^H(f) \quad (7)$$

and the demixing process (6) leads to

$$\forall f, \quad S_{\hat{s}}(t,f) = \mathcal{G}(f) S_x(t,f) \mathcal{G}^H(f) \quad (8)$$

where $S_s(t,f)$, $S_x(t,f)$, and $S_{\hat{s}}(t,f)$ are the time-varying power spectrum density (psd) matrices of the sources $\mathbf{s}(t)$, the observations $\mathbf{x}(t)$, and the output $\hat{\mathbf{s}}(t)$, respectively. $\mathcal{H}(f)$ and $\mathcal{G}(f)$ are the frequency response matrices of the mixing and demixing filter matrices (^{H} denoting the conjugated transpose). Since the mixing process is assumed to be stationary, $\mathcal{H}(f)$ and $\mathcal{G}(f)$ are not time-dependent, although the signals (i.e., sources, mixtures) may be nonstationary. If we assume that the sources are mutually independent (or at least decorrelated), for all frequency bins f, $S_s(t,f)$ is a diagonal matrix and thus an efficient separation must lead to a diagonal matrix $S_{\hat{s}}(t,f)$. Therefore, a basic criterion for BSS is to adjust the matrices $\mathcal{G}(f)$ so that $S_{\hat{s}}(t,f)$ is as diagonal as possible for every frequency [31], [33]. This can be done by the joint diagonalization method described in [38] that exploits the nonstationarity of signals: at each frequency bin f, $\mathcal{G}(f)$ is the matrix that jointly diagonalizes a set of psd matrices $\{S_{\hat{s}}(t,f)\}_t$ for several time indexes t.

Now, any blind source separation based on the independence assumption between sources is faced with two crucial limitations. Indeed, since the order and the amplitude of the estimated sources do not influence their mutual independence, an independence-based criterion can only provide the estimated sources up to a scale factor and a permutation. In frequency domain separation, these limitations are encountered for each frequency bin. In other words, $\mathcal{G}(f)$ can only be estimated up to a scale factor and a permutation between the estimated sources, and the scale and permutation can be different for each frequency bin. Therefore, in the specific square case, where there are as many sources as observations ($N = P$), the separating filter matrices can be expressed as

$$\forall f, \quad \mathcal{G}(f) = \mathcal{D}(f) \mathcal{P}(f) \mathcal{H}^{-1}(f) \quad (9)$$

where $\mathcal{D}(f)$ (resp. $\mathcal{P}(f)$) is the arbitrary diagonal (resp. permutation) matrix which represents the scale (resp. permutation) indeterminacy. As a result, even if the separation matrices $\mathcal{G}(f)$ are well estimated for each frequency bin (i.e., $S_{\hat{s}}(t,f)$ are diagonal), this does not ensure that the estimated sources $\hat{\mathbf{s}}(t) = \mathcal{G}(t) * \mathbf{x}(t)$ are well reconstructed due to the permutation $\mathcal{P}(f)$ and diagonal $\mathcal{D}(f)$ matrices. In order to have a good reconstruction of the sources, it is necessary to have the same scale factor and the same permutation for all frequency bins f

$$\forall (f_1, f_2), \quad \begin{cases} \mathcal{P}(f_1) = \mathcal{P}(f_2) & (10a) \\ \mathcal{D}(f_1) = \mathcal{D}(f_2). & (10b) \end{cases}$$

These two equations have distinct influence on the quality of the reconstructed sources. First, (10a) ensures that there is no interference between the estimated sources $\mathbf{s}(t)$. Thus, each estimated source is obtained from only one source: $\forall n$, $\hat{s}_n(t) = g(s_n(t))$ (without any loss of general information, we suppose that $\mathcal{P}(f)$ is the identity matrix). Second, (10b) ensures that the shape of the power spectrum density of the estimated sources across frequency bins is correct. So, verifying (10a) without (10b) only provides the sources up to a filter.

Among the numerous methods proposed to control the permutation indeterminacy, Pham *et al.* [33] proposed to equalize the permutation matrices across frequency bins by exploiting the continuity between consecutive frequency bins of $\{\mathcal{G}(f)\}_f$: they select the permutation that provides a smooth reconstruction of the frequency response. The method presents a major drawback: it selects the permutation at the frequency bin f_i based on the value of $\mathcal{G}(f_{i-1})$. Thus, if a wrong decision is done at a frequency bin, then the following frequency bins are also wrong unless another "lucky" wrong decision eventually corrects the process. Moreover, in this study, Pham *et al.* do not solve the scale factor ambiguity. Other proposed methods (e.g., [39]) ensure that the estimated sources do not depend on the rescaling of the separation matrices. However, they do not ensure that these estimated sources are the original ones (i.e., (10b) is not satisfied, which means that the sources can be estimated up to a nonflat filter).

In this paper, we propose a method, which exploits the coherence between the acoustic speech signal and the speaker's lips, to solve both the permutation (10a) and the scale factor (10b) ambiguities.

III. Audiovisual Model

In this section, we explain the audiovisual model used for modeling the relationship between audio and visual speech signals. We first give the description of the video and audio parameters used in this study, then we present the statistical audiovisual model.

A. Video Parameters

At each time index t, the video signal $\mathbf{v}(t)$ consists of a vector of basic lip shape parameters: internal width and height of the lips (Fig. 3). Indeed, several studies have shown that the basic facial lip edge parameters contain most of the visual information, according to both intelligibility criterion [40], [41] and statistical analysis: the internal width and height represent 85% of

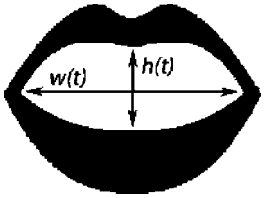

Fig. 3. Video parameters: internal width w and internal height h.

the variance of the visual data used in [42]. The video vector is then

$$\mathbf{v}(t) = \begin{pmatrix} w(t) \\ h(t) \end{pmatrix} \tag{11}$$

where w (resp. h) refers to the internal width (resp. height). The video signal is sampled at 50 Hz and then the video parameters are automatically extracted every 20 ms by a device and an algorithm developed at the ICP [43].

B. Audio Parameters

Let $s(t)$ denote the corresponding acoustic speech signal. The vector of audio parameters must contain local spectral characteristics of the speech signal $s(t)$ (Fig. 4). Thus, $s(t)$ is divided into consecutive frames, synchronous with the video signal. The space between two audio frames is 20 ms since the statistical model aims at associating each spectrum with one set of video parameters. Each audio frame is centered, normalized, and multiplied by a Hamming window, and we calculate its short-term Fourier transform $\mathbf{S}(t) = [S(t, f_1), \ldots, S(t, f_L)]^T$. These audio parameter are complex, $\mathbf{S}(t) \in \mathbb{C}^L$, and can be seen as zero-mean circular complex Gaussian random variables [44], [45]. In [46], the diagonal elements of the covariance matrix are interpreted as the power spectrum density. In speech processing, taking into account perceptual properties of human hearing, it is usual to consider the log-modulus of discrete Fourier transform (DFT) coefficients, so we define another audio vector $\mathbf{A}(t)$ such that

$$\mathbf{A}(t) = \log |\mathbf{S}(t)| \tag{12}$$

where $\log(\cdot)$ denotes here the component-wise decimal logarithm. These coefficients $\mathbf{A}(t)$ were shown to follow a Log–Rayleigh distribution [36] (see also Appendix I).

C. Audiovisual Model

In the audiovisual speech processing community, it is usual to consider that the relationship between the audio and video parameters is complex and can be expressed in statistical term [15], [47]. So we choose to model the audiovisual data by a mixture of kernels

$$p_{AV}\left(\mathbf{v}(t), \mathbf{A}(t)\right) = \sum_{i=1}^{I} \omega_i^{AV} p\left(\mathbf{v}(t), \mathbf{A}(t) | \mu_i^{AV}, \Sigma_i^{AV}\right) \tag{13}$$

where $\{\omega_i^{AV}, \mu_i^{AV}, \Sigma_i^{AV}\}$ is the parameter set {weight, mean value vector, covariance matrix} of the ith kernel, and I is the number of kernels. In this model, we choose, for each kernel, a separable model[3] with diagonal covariance matrices

$$p\left(\mathbf{v}(t), \mathbf{A}(t) | \mu_i^{AV}, \Sigma_i^{AV}\right) = p_G\left(\mathbf{v}(t) | \mu_i^V, \Sigma_i^V\right) p_{LR}\left(\mathbf{A}(t) \left| \frac{\Sigma_i^A}{2} \right.\right) \tag{14}$$

where $p_G(\mathbf{x}|\mu, \Sigma)$ denotes the value of the Gaussian distribution of $\mathbf{x}$ given the mean value vector μ and the covariance matrix Σ and $p_{LR}(\cdot|\beta^2)$ denotes the Log–Rayleigh probability density function (42) given the localization parameter β^2. Moreover

$$\mu_i^{AV} = \begin{pmatrix} \mu_i^V \\ 0 \end{pmatrix}$$

and

$$\Sigma_i^{AV} = \begin{pmatrix} \Sigma_i^V & 0 \\ 0 & \Sigma_i^A \end{pmatrix}$$

with

$$\Sigma_i^V = \operatorname{diag}\left(\left(\sigma_i^V(w)\right)^2, \left(\sigma_i^V(h)\right)^2\right)$$
$$\Sigma_i^A = \operatorname{diag}\left(\left(\sigma_i^A(f_1)\right)^2, \ldots, \left(\sigma_i^A(f_L)\right)^2\right)$$

where $\operatorname{diag}(\cdot)$ denotes the diagonal matrix whose diagonal elements are its arguments.

The sets of parameters $\Theta = \{\omega_i^{AV}, \mu_i^V, \Sigma_i^V, \Sigma_i^A\}_{1 \le i \le I}$ have to be estimated from training data in a previous stage. By using the penalized version [48], [49] of the widely used expectation-maximization (EM) algorithm [50] (see Appendix II), we obtain the results on the audiovisual model presented in Section V-A.

IV. Solving Indeterminacy Problems

In this section, we propose an approach to the indeterminacy problems (see Section II) exploiting the audiovisual coherence of speech signals through the AV model presented in Section III. We explain how to estimate first the permutations (IV-B), then the scale factors (IV-C), and finally in IV-D we show how to plug together the scale factor estimation and the permutation cancellation.

In this paper, we want to extract only one particular speech source of interest, say $s_1(t)$, from the mixture $\mathbf{x}(t)$. For this purpose, we exploit additional observations, which consist in the video signal $\mathbf{v}_1(t)$ extracted from the speaker's face. In the following, we suppose, without any loss of general information, that $s_1(t)$ is the first component of $\mathbf{s}(t)$.

A. Notations

The sources $S_n(t, f)$ can be seen as a three-dimensional array: (index of the nth source $\times$ time $t \times$ frequency bin f).

[3]Note that, even if each audiovisual kernel (14) models the audiovisual information in an independent way, the global audiovisual model (13) does not, since it is a sum of several kernels.

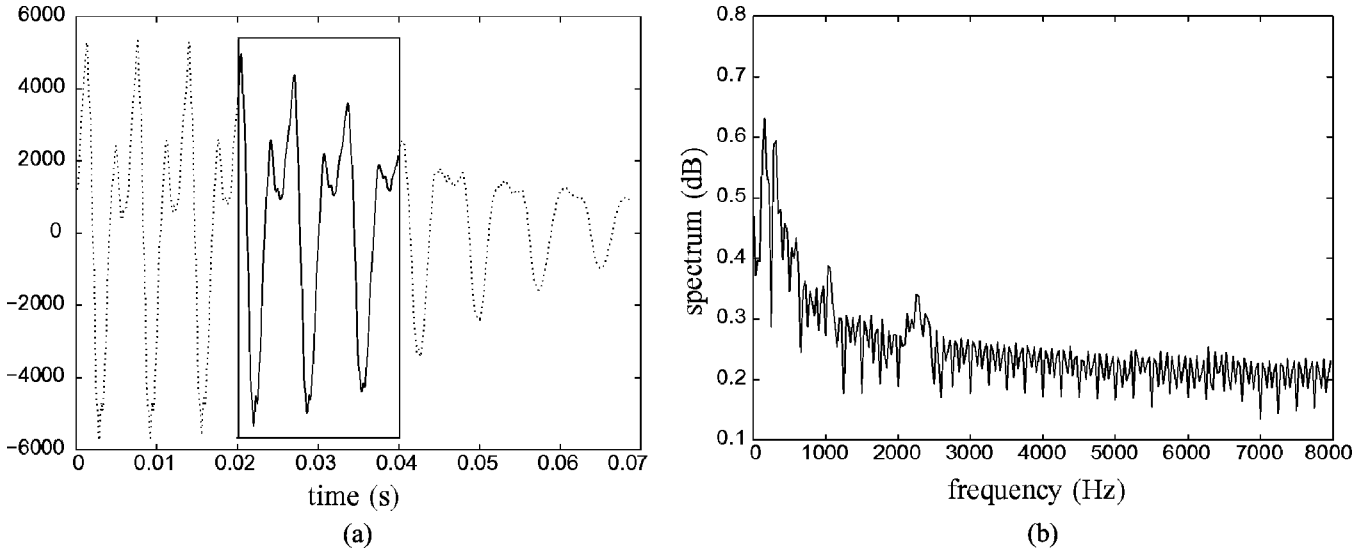

Fig. 4. Audio parameters. (a) Speech signal and the frame. (b) Spectrum of the frame.

Thus, first, let $\mathbf{S}_n(t,:)$ denote the audio column vector of the nth source at time t : $\mathbf{S}_n(t,:) = [S_n(t,f_1),\ldots,S_n(t,f_L)]^T$.

Second, let $\mathbf{S}_{:}^{\dagger}(t,f) = [S_1^{\dagger}(t,f),\ldots,S_N^{\dagger}(t,f)]^T$ denote the estimated audio column vector provided by the demixing process (9) at time t and frequency f. So, this vector $\mathbf{S}_{:}^{\dagger}(t,f)$ is equal to $\mathbf{S}_{:}(t,f) = [S_1(t,f),\ldots,S_N(t,f)]^T$ up to the permutation $\mathcal{P}(f)$ and the scale factor $\mathcal{D}(f)$

$$\mathbf{S}_{:}^{\dagger}(t,f) = \mathcal{D}(f)\mathcal{P}(f)\mathbf{S}_{:}(t,f). \tag{15}$$

Moreover, let $\mathbf{A}_{:}^{\dagger}(t,f) = [A_1^{\dagger}(t,f),\ldots,A_N^{\dagger}(t,f)]^T$ denote the component-wise log-modulus of $\mathbf{S}_{:}^{\dagger}(t,f)$:

$$\mathbf{A}_{:}^{\dagger}(t,f) = \log\left|\mathbf{S}_{:}^{\dagger}(t,f)\right|. \tag{16}$$

Note that, if $\mathcal{P}$ is a permutation matrix and $\mathbf{y}$ a vector, then

$$\log|\mathcal{P}\mathbf{y}| = \mathcal{P}\log|\mathbf{y}|. \tag{17}$$

Finally, for sake of simplicity, let $\mathcal{P}$ (resp. $\mathcal{D}$) denote the permutation (resp. diagonal) matrices set $\{\mathcal{P}(f)\}_f$ (resp. $\{\mathcal{D}(f)\}_f$).

B. Permutation Ambiguity

Now, regularizing the permutation problem (10a) in frequency domain BSS, consists in searching a permutation set $\widehat{\mathcal{P}^*}$ such that

$$\mathbf{A}^{\dagger}_{1,\widehat{\mathcal{P}^*}}(t,:) \simeq \mathbf{A}_1(t,:) \tag{18}$$

where $\mathbf{A}^{\dagger}_{1,\mathcal{P}^*}(t,:)$ are the estimated audio coefficients of $s_1(t)$ up to the permutation set $\mathcal{P}^*$

$$\forall f, \quad A^{\dagger}_{1,\mathcal{P}^*}(t,f) = \left(\mathcal{P}^*(f)\mathbf{A}_{:}^{\dagger}(t,f)\right)_1 \tag{19}$$

where $(\mathbf{y})_i$ is the ith component of vector $\mathbf{y}$.

To estimate $\mathcal{P}^*$, we propose to minimize the audiovisual criterion $j_{AV}(\mathcal{P}^*,t)$ between the audio spectrum output on channel 1 and the visual information $\mathbf{v}_1$

$$\widehat{\mathcal{P}^*} = \arg\min_{\mathcal{P}^*} j_{AV}(\mathcal{P}^*,t) \tag{20}$$

with

$$j_{AV}(\mathcal{P}^*,t) = -\log\left[p_{AV}\left(\mathbf{v}_1(t),\mathbf{A}^{\dagger}_{1,\mathcal{P}^*}(t,:)\right)\right]. \tag{21}$$

Note that, even if the optimal solution $\hat{\mathcal{P}}^*(f)$ would be $\mathcal{P}^{-1}(f)$ for all f, since the criterion (20) only considers channel 1 (i.e., $\hat{s}_1(t)$), it provides at best $(\hat{\mathcal{P}}^*(f)\mathcal{P}(f))_{1,1} = 1$ for all f, leaving the other terms of $(\hat{\mathcal{P}}^*(f)\mathcal{P}(f))$ unspecified. In other words, this means that, at best, the proposed method will only ensure that the components $\hat{S}_1(t,f)$ actually correspond to the source of interest, without any constraint on the other estimated sources. This is not a problem because we are only interested in extracting $s_1(t)$ using $v_1(t)$. Extracting other sources with our method would require additional video information about the other sources to extract.

In [15] and [16], it was shown that temporal integration is necessary to efficiently exploit the audiovisual information, since the audiovisual coherence is mainly expressed in the time-dynamic of speech. So 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

$$J_{AV}(\mathcal{P}^*) = \sum_{t=0}^{T-1} j_{AV}(\mathcal{P}^*,t). \tag{22}$$

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 the huge computational load.

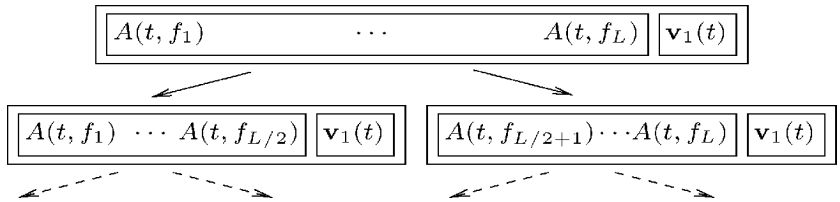

Fig. 5. Marginal recursive scheme: at each step, test, with the marginal criterion (24), a permutation of audio parameters set decreased by a factor 2.

In [34], we already proposed an algorithm for minimizing criterion (22). In this paper, this algorithm is used as a basis idea, but as shown in the following, it had to be modified. To simplify, we present our algorithm for two sources and two mixtures, but it can be easily extended to larger cases. First, we use a dichotomic scheme during which we simplify criterion (22) by marginalizing the audiovisual probability $p_{AV}(\mathbf{v}(t), \mathbf{A}(t))$ regarding subsets of contiguous frequencies. Thus, let $\mathcal{F}$ denote an arbitrary subset of frequencies f_j and $p_{AV}^{\mathcal{F}}(\mathbf{v}(t), \mathbf{A}(t))$ the marginal audiovisual probability regarding the frequency set $\mathcal{F}$

$$p_{AV}^{\mathcal{F}}(\mathbf{v}(t), \mathbf{A}(t)) = \int p_{AV}(\mathbf{v}(t), \mathbf{A}(t))\, dA(t, f_j)_{f_j \notin \mathcal{F}} \quad (23)$$

and the marginal form of (22) is

$$J_{AV}(\mathcal{P}^*, \mathcal{F}) = \sum_{t=0}^{T-1} j_{AV}(\mathcal{P}^*, t, \mathcal{F}) \quad (24)$$

where

$$j_{AV}(\mathcal{P}^*, t, \mathcal{F}) = -\log\left[p_{AV}^{\mathcal{F}}\left(\mathbf{v}_1(t), \mathbf{A}_{1,\mathcal{P}^*}^{\dagger}(t,:)\right)\right]. \quad (25)$$

Now, exploiting this simplification, we use the following descending dichotomic scheme, denoted marginal AV algorithm (Fig. 5).

1) First, test the permutation (between the two estimated sources) on all audio parameters, $\mathcal{F} = \{f_1, \ldots, f_L\}$, which minimizes J_{AV}

$$J_{AV}(\mathcal{J}, \mathcal{F}) \underset{H_0}{\overset{H_1}{\lessgtr}} J_{AV}(\mathcal{I}, \mathcal{F})$$

where $\mathcal{J}$ is the unitary antidiagonal matrix, and $\mathcal{I}$ is the identity matrix.[4] H_0 means "do nothing" and H_1 means "permute between the two estimated sources the audio coefficients for the frequencies set $\mathcal{F}$."

2) Then, sharpen the estimation of the permutation matrices set by testing separately with (24):
   - permutation on the first half of the audio parameters set $\mathcal{F}_1 = \{f_1, \ldots, f_{L/2}\}$

$$J_{AV}(\mathcal{J}, \mathcal{F}_1) \underset{H_0}{\overset{H_1}{\lessgtr}} J_{AV}(\mathcal{I}, \mathcal{F}_1)$$

[4]Remember that, for sake of clarity but without loss of generality, we only consider two sources. There are then only two possible permutations: $\mathcal{I}$ (no permutation) and $\mathcal{J}$ (permutation).

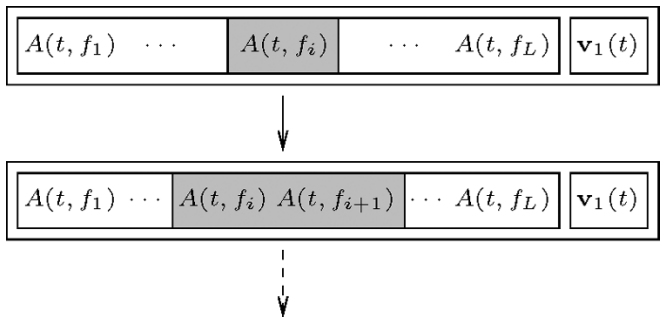

Fig. 6. Joint recursive scheme: at each step, test with the joint criterion (22), a permutation of audio parameters set increased by a factor 2.

   - a permutation on the second half of the audio parameters set $\mathcal{F}_2 = \{f_{L/2+1}, \ldots, f_L\}$

$$J_{AV}(\mathcal{J}, \mathcal{F}_2) \underset{H_0}{\overset{H_1}{\lessgtr}} J_{AV}(\mathcal{I}, \mathcal{F}_2)$$

3) continue with this dichotomic scheme on the next subsets of frequencies.

This initialization scheme, using marginal criterion (24), gives a good estimation of the permutations set $\mathcal{P}^*$, but we observed that this result can be improved. This is not surprising since the marginal AV probability (23) does not take into account all the audiovisual coherence. Thus, we refine the estimation by applying the joint criterion (22) with an ascending recursive scheme, denoted joint AV algorithm (Fig. 6):

1) for all $1 \leq i \leq L$, test with (22) the permutation matrices set $\mathcal{P}_i$ that only permutes, between the two estimated sources, the frequency f_i leaving all the other frequencies $f_j, j \neq i$ unchanged

$$J_{AV}(\mathcal{P}_i) \underset{H_0}{\overset{H_1}{\lessgtr}} J_{AV}(\mathcal{I})$$

2) then, for all $1 \leq i \leq L/2$, similarly test the permutation matrices set $\mathcal{P}_{2i-1,2i}$ that permutes, between the two estimated sources, the couple of frequencies $\{f_{2i-1}, f_{2i}\}$

$$J_{AV}(\mathcal{P}_{2i-1,2i}) \underset{H_0}{\overset{H_1}{\lessgtr}} J_{AV}(\mathcal{I})$$

3) continue with this dichotomic scheme until the permutation matrices set $\mathcal{P}_{1,\ldots,L}$ that permutes the whole frequency set $\{f_1,\ldots,f_L\}$

$$J_{AV}(\mathcal{P}_{1,\ldots,L}) \underset{H_0}{\overset{H_1}{\lessgtr}} J_{AV}(\mathcal{I})$$

4) loop at stage 1 if necessary.

Unfortunately, a very tricky problem appears when using criteria (22) and (24): the scale indeterminacy, i.e., the matrices set $\mathcal{D}$, can dramatically change the value of the probability in (21) and (23). Therefore, it has to be estimated to improve the permutation resolution. Moreover, using the frequency approach, the scale factor estimation is a major challenge of BSS to improve the quality of the reconstructed sources, as mentioned in Section II. That is why we now propose a method to estimate the scale factor set.

C. Scale Factor

Now, regularizing the scale factor ambiguity (10b) of frequency domain BSS consists in searching a scale factor set $\widehat{\mathcal{D}^*}$ that assumes

$$\mathbf{A}^{\dagger}_{1,\widehat{\mathcal{D}^*}}(t,:) \simeq \mathbf{A}_1(t,:) \tag{26}$$

where $\mathbf{A}^{\dagger}_1, \mathcal{D}^*(t,:)$ are the estimated audio coefficients up to the scale factors set $\mathcal{D}^*$

$$\forall f, \quad A^{\dagger}_{1,\mathcal{D}^*}(t,f) = \left(\log\left|\mathcal{D}^*(f)\mathbf{S}^{\dagger}_{:}(t,f)\right|\right)_1$$

so

$$\forall f, \quad A^{\dagger}_{1,\mathcal{D}^*}(t,f) = \log\left|\mathcal{D}^*_{1,1}(f)\right| + A^{\dagger}_1(t,f).$$

Note that, even if the optimal solution $\hat{\mathcal{D}}^*(f)$ would be $\mathcal{D}^{-1}(f)$ for all f, since we are only interested in extracting $s_1(t)$, we will only estimate $\alpha(f) = \mathcal{D}^*_{1,1}(f)$. Thus, regularizing the scale factors consists in searching, for each frequency bin f, the parameter $\alpha(f)$ which leads to

$$\alpha(f)S^{\dagger}_1(t,f) \simeq S_1(t,f). \tag{27}$$

To estimate $\{\alpha(f)\}_f$, we proposed in [35] to exploit the audio model achieved by marginalizing the audiovisual model (13) regarding the video parameters[5]

$$p_A\left(\mathbf{S}_1(t,:)\right) = \sum_{i=1}^{I} \omega_i^A p_G\left(\mathbf{S}_1(t,:)|0,\Sigma_i^A\right) \tag{28}$$

[5]Note that a Log-Rayleigh distribution on $\mathbf{A}(t)$ is equivalent to a complex Gaussian distribution with zero mean value on $\mathbf{S}(t)$. Thus, we return here to this latter distribution since we need to characterize $\mathbf{S}(t)$.

where ω_i^A is the weight of the ith audio kernel derived from the audiovisual one: $\omega_i^A = \omega_i^{AV}$. Since the variance of $\alpha(f)S^{\dagger}_1(t,f)$ verifies

$$\alpha^2(f)\mathrm{Var}\left(S^{\dagger}_1(t,f)\right) = \mathrm{Var}\left(S_1(t,f)\right) \tag{29}$$

where $\mathrm{Var}(\cdot)$ is the variance operator, and since the variance of the marginal audio model verifies

$$\mathrm{Var}\left(S_1(t,f)\right) = \sum_{i=1}^{I} \omega_i^A \left(\sigma_i^A(f)\right)^2 \tag{30}$$

then, we proposed to estimate $\alpha(f)$ thanks to

$$\hat{\alpha}(f) = \sqrt{\frac{\sum_{i=1}^{I} \omega_i^A \left(\sigma_i^A(f)\right)^2}{\mathrm{Var}\left(S^{\dagger}_1(t,f)\right)}} \tag{31}$$

where $(\sigma_i^A(f))^2$ is the fth diagonal element of the covariance matrix Σ_i^A and $\mathrm{Var}(S^{\dagger}_1(t,f))$ is estimated by the classical variance estimator

$$\mathrm{Var}\left(S^{\dagger}_1(t,f)\right) = \frac{1}{T-1}\sum_{t=1}^{T}\left|S^{\dagger}_1(t,f)\right|^2. \tag{32}$$

Nevertheless, this idea can only work if the sampled variance of target source $\mathrm{Var}(S_1(t,f))$ is equal to the variance obtained from the trained model (30). This may not be true since the variance of the trained model is the averaged variance, which can be quite different from the variance of a particular frame of speech. That is why we propose in this study to use a new audiovisual criterion to estimate the scale factor. In (30), the variance of the audio model is the sum of the *a priori* probability ω_i^A of the ith kernel multiplied by the corresponding variance $(\sigma_i^A(f))^2$: it does not take into account that some AV kernel (i.e., some sounds) may not be pronounced. To overcome this, we propose to use the video information by substituting the *a priori* probability by the video *a posteriori* probability

$$\beta_i^{AV} = \frac{1}{T}\sum_{t=1}^{T} p\left(i|\mathbf{v}(t)\right) \tag{33}$$

where $p(i|\mathbf{v}(t))$ is the *a posteriori* probability of the ith kernel given the video vector $\mathbf{v}(t)$

$$p\left(i|\mathbf{v}(t)\right) = \frac{\omega_i^{AV} p\left(\mathbf{v}(t)|\mu_i^V,\Sigma_i^V\right)}{\sum_{k=1}^{I}\omega_k^{AV} p\left(\mathbf{v}(t)|\mu_k^V,\Sigma_k^V\right)}. \tag{34}$$

So, the new model variance verifies

$$\mathrm{Var}\left(S_1(t,f)\right) = \sum_{i=1}^{I} \beta_i^{AV}\left(\sigma_i^A(f)\right)^2 \tag{35}$$

and finally, the scale factor is estimated thanks to

$$\hat{\alpha}_{AV}(f) = \sqrt{\frac{\sum_{i=1}^{I} \beta_i^{AV} \left(\sigma_i^A(f)\right)^2}{\mathrm{Var}\left(S_1^\dagger(t,f)\right)}}. \tag{36}$$

We will now explain how to plug the AV scale factor estimation in the AV permutation cancellation algorithm.

D. Bootstrap Algorithm

In [35], we initially estimated the scale factors set $\mathcal{D}$ after the permutation cancellation algorithm. However, as we explained previously, the unknown scale factors set $\mathcal{D}$ can dramatically change the value of the probabilities (21) and (23). Thus, in order to have good performance, we now propose a new bootstrap algorithm: at each stage of the AV marginal or AV joint algorithms, if H_1 is chosen, then reestimate the scale factors thanks to the AV process (36). If necessary, it is possible to loop several times the joint AV algorithm.

So, this algorithm estimates $\widehat{\mathcal{D}^*}$ and $\widehat{\mathcal{P}^*}$, and the estimated sources $\hat{\mathbf{S}}_{:}(t,f)$ are obtained thanks to

$$\hat{\mathbf{S}}_{:}(t,f) = \widehat{\mathcal{D}^*}(f)\widehat{\mathcal{P}^*}(f)\mathbf{S}_{:}^\dagger(t,f) \tag{37}$$

Finally, the reconstructed source of interest $\hat{s}_1(t)$ is the result of the inverse short-time Fourier transform of $\hat{S}_1(t,f)$.

Note that even if in this paper we only discussed the case of one source of interest, it is easy to extend the criteria (22), (24), and (36), and then to achieve the final estimation (37) to a more significant number of sources of interest, if the corresponding additional video signals are available.

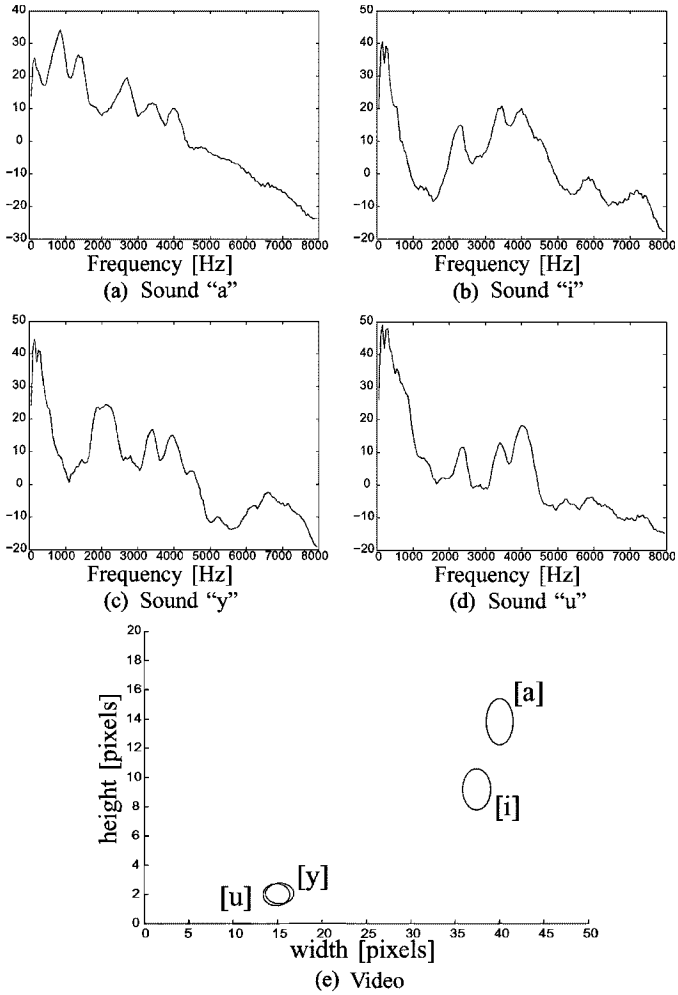

Fig. 7. Illustration of the corpus. Figs. (a), (b), (c), (d) present the audio kernels $[(\sigma_i^A(f_1))^2, \ldots, (\sigma_i^A(f_L))^2]^T$ (in decibels) interpreted as the psd of four French vowels [a], [i], [y], [u], respectively. (e) Corresponding video kernels.

V. Numerical Experiments

In this section, we present first the corpus used in the experiments and the AV model configuration, and then we present the results of the algorithm to cancel the ambiguities.

A. Corpus and Audiovisual Model Configuration

We use two types of corpora for assessing the separation: one is the source of interest, and the second is the disturbing source.

The first corpus, used as the source of interest, consists of French logatoms that are nonsense "V1-C-V2-C-V1" utterances, where "V1" and "V2" are same or different vowels within the set [a],[i],[y],[u], and "C" is a consonant within the plosives set [p],[t],[k],[b],[d],[g] or no plosive [#]. The 112 sequences, representing around 50 s of speech, were pronounced twice by the male speaker: the first time is used for training the AV model and the second time for the test. This corpus is interesting since it groups in a restricted set the basic problems to be addressed by audiovisual studies (Fig. 7): it contains on the first hand, similar lip shapes associated with distinct sounds (such as [y] and [u]), and on the other hand sounds with similar acoustic features and different lip shapes (such as [i] and [y]). Since the video channel is sampled at 50 Hz, we choose the length of the temporal block equal to 20 ms, and the audio signals are sampled at 16 kHz. In this study, the number of the audio parameters L is, therefore, equal to 160 (the first half of the 320 FFT coefficients). Thus, the training data set and the testing data set contain around 2500 audiovisual vectors (two video parameters (11) and 160 audio parameters (12) for each vector).

The second corpus, used as the second source, consists of phonetically well-balanced sentences in French of a female speaker.

In this study, we choose, for the AV model $I = 64$ kernels which parameters $\{\omega_i^{AV}, \mu_i^V, \Sigma_i^V, \Sigma_i^A\}_{1 \le i \le 64}$ are estimated using the training data set with the EM algorithm (see the update equations in Appendix II). The number of kernels (here 64) higher than the number of phonems (here 12) allows the model to fit the AV transitions between the different sounds. Fig. 7 shows several of the resulting AV kernels.

B. Scale Factor Results

In this experiment, we only study the performance of the scale factor audio (31) and audiovisual (36) stages: no mixing or separation were performed.

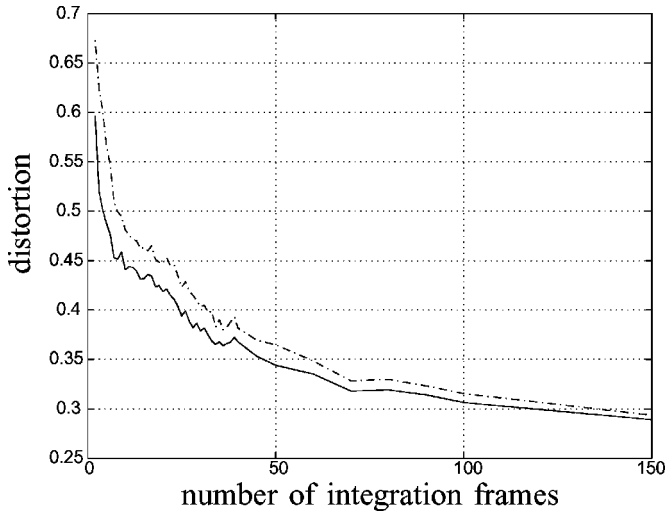

Fig. 8. Distortion (39) versus number of integration frames. Solid line: AV estimation (36) and dashed-dot line: audio estimation (31).

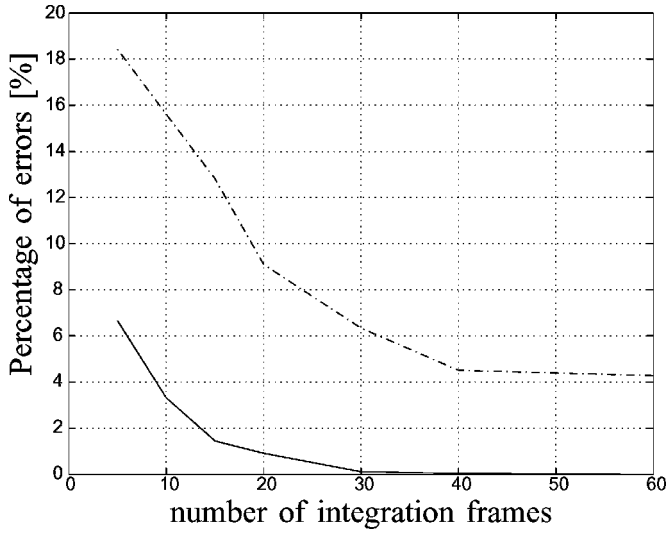

Fig. 9. Percentage of errors versus number of integration frames. Solid line: permutation of 250-Hz bandwidth blocks, dashed-dot line: permutation of 100-Hz bandwidth blocks.

For this purpose, the signal of interest $S_1(t,f)$ [obtained by the Fourier transform of $s_1(t)$] is arbitrary multiplied by a scale factor $\mathcal{D}_{1,1}(f)$ randomly chosen at each frequency bin.

To quantify the performance, we defined the distorsion as

$$d(t) = \sqrt{\frac{1}{L}\sum_{f=f_1}^{f_L} \left|\log|S(t,f)| - \log\left|\hat{S}(t,f)\right|\right|^2}. \quad (38)$$

In the specific case of this experiment, where there is no permutation, the distortion (38) no more depends on time and can be expressed as

$$d = \sqrt{\frac{1}{L}\sum_{f=f_1}^{f_L} \left|\log\left(\widehat{\mathcal{D}^*}_{1,1}(f)\mathcal{D}_{1,1}(f)\right)\right|^2}. \quad (39)$$

Fig. 8 shows the mean distortion versus the number of integration frames for both audio (31) and audiovisual (36) estimations. Each simulation is repeated over 60 different logatoms. Fortunately, for both estimations, the distortion decreases while the number of integration frame increases. This is due to the fact that the variance of a particular section of speech can be different of the variance of the model: the more numerous the integration frames are, the more robust the estimation is. Moreover, the AV estimation (36) is better than the audio one (31), justifying our idea that the video parameters can efficiently improve the estimation of the model variance corresponding of the particular utterance of speech. Thus, for an arbitrary distortion, the number of integration frames is smaller with an AV estimation than with an audio one.

C. Permutation Results

To estimate the performance of our permutation cancellation algorithm, we test permutation detection of blocks of consecutive frequencies. As in the previous section, no mixing or separation were performed. We simply artificially permuted some blocks of consecutive frequencies between the two sources $\mathbf{S}_1(t,:)$ and $\mathbf{S}_2(t,:)$ obtained by the FFT of $s_1(t)$ and $s_2(t)$. Then, we applied our permutation cancellation algorithm (Section IV-B) on these artificially modified signals. First, we test the permutation of 250-Hz bandwidth blocks (corresponding to five consecutive frequencies) for 1, 4, 8, 12, and 16 permuted blocks. In this case, the smallest bandwidth of the block used in our algorithm is equal to 250 Hz. Second, we test the permutation of 100-Hz bandwidth blocks (corresponding to two consecutive frequencies) for 1, 10, 20, 30, and 40 permuted blocks. In this case, the smallest bandwidth of the block used in our algorithm is equal to 100 Hz.

We define the detection error as the sum of the unsolved permutations (actual permutations undetected by our algorithm) and the wrong permutations (bad decision of the algorithm).

For each experimental condition, the simulation is repeated over 60 different logatoms which are randomly chosen but perfectly known. Fig. 9 shows the mean percentage of detection error versus the number of integration frames for the two simulation cases (100- and 250-Hz bandwidth blocks). This figure stresses the importance of integration for the criteria (22) and (24). Indeed, if the number of integration frames is too small, the percentage of errors significantly increases while the computational time decreases. Meanwhile, if the number of integration frames increases, the number of errors decreases toward zero while the computational time increases. It can be noted that, for more than 40 integrated frames, the percentage of error is smaller than 5% for all tested conditions. Also, the mean results in the case of 250-Hz bandwidth block are better than the case of 100-Hz bandwidth block. Thus, for an arbitrary percentage of errors, the resolution of the algorithm can be increased at the price of a larger number of integration frames.

D. Separation Results

In the following, we consider the case of two sources [Fig. 10(a)] and two mixtures [Fig. 10(b)]. All mixing filters are artificial finite impulse response filters up to 320 lags: they fit a simplified acoustic model of a room impulse response (Fig. 11). Even if we plot the sources over four seconds, we only used the first 40 frames in order to estimate the permutation and the

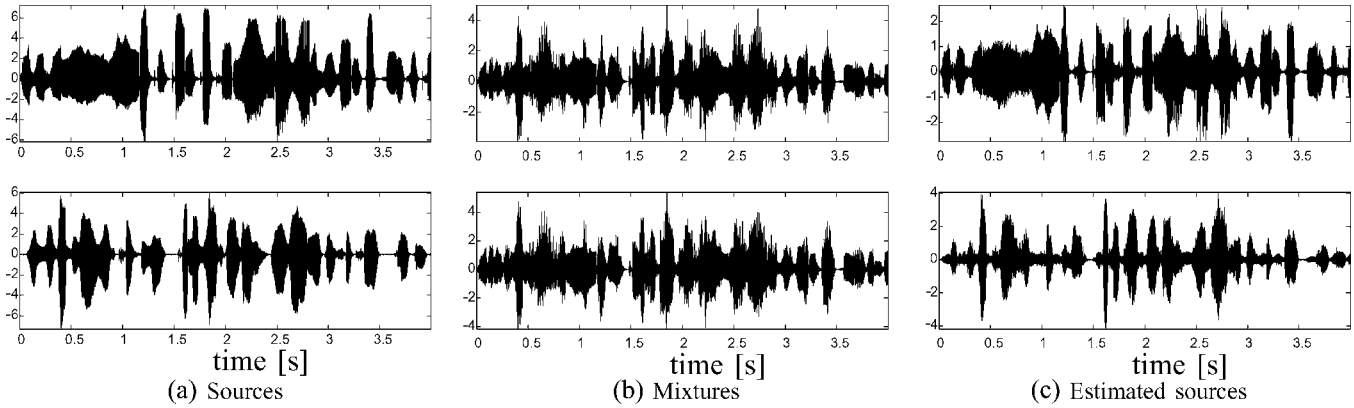

Fig. 10. Sources, mixtures, and estimated sources.

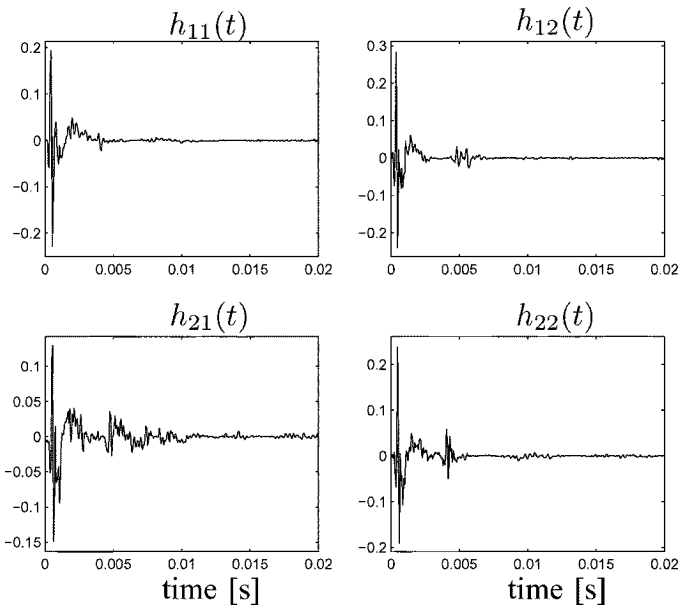

Fig. 11. Mixing filters: impulse response of the four mixing filters.

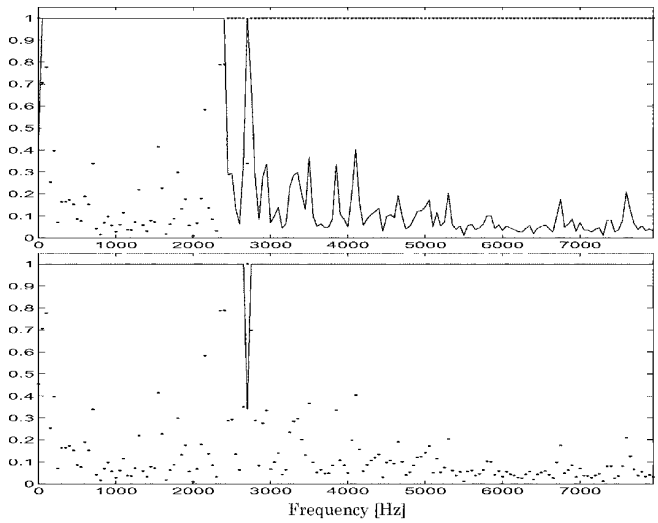

Fig. 12. Performance index $r(f)$ [Eq. (40)] (dots) and its inverse (solid line) truncated at 1, before (upper panel) and after (lower panel) our ambiguities cancellation versus the frequency.

scale factor. This value is coherent with the results obtained in the previous sections.

An indicator of the separation performance is the performance index [33], defined as

$$r(f) = \left| \frac{(\mathcal{GH})_{12}(f)(\mathcal{GH})_{21}(f)}{(\mathcal{GH})_{11}(f)(\mathcal{GH})_{22}(f)} \right|^{\frac{1}{2}} \tag{40}$$

where $(\mathcal{GH})_{i,j}(f)$ is the ij element of the global matrix filter $\mathcal{G}(f)\mathcal{H}(f)$. For a good separation, the index (40) should be close to 0 (or infinity if a permutation has occurred).

Fig. 12 plots $\min(r, 1)$ and $\min(1/r, 1)$ before and after applying our ambiguities detection. One can see that our method corrects all the permutations except one error: the performance index is always smaller than 1 except for one frequency. This is confirmed by the spectrum of the four impulse responses of the global filter $(\mathcal{GH})(f)$ (Fig. 13): for all the frequency bins f (except for the error), $|(\mathcal{GH})_{12}(f)|$ (resp. $|(\mathcal{GH})_{21}(f)|$) is much smaller than $|(\mathcal{GH})_{11}(f)|$ (resp. $|(\mathcal{GH})_{22}(f)|$). This means that the global filter $(\mathcal{G} * \mathcal{H})(t)$ is close to a diagonal filter.

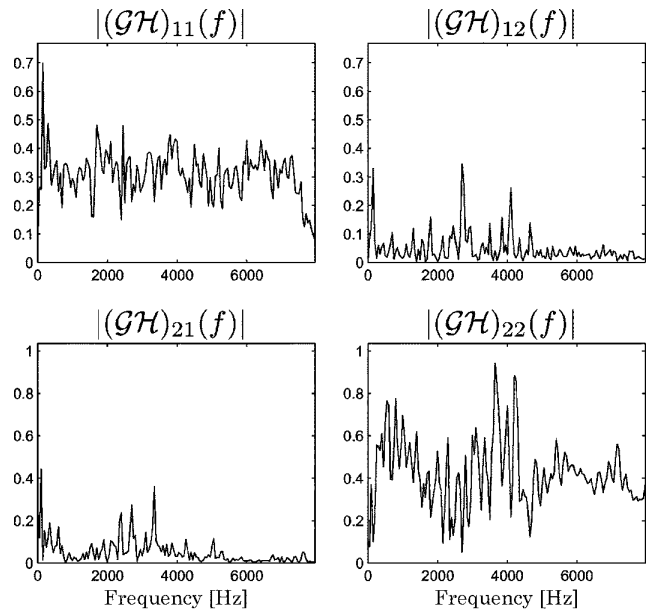

Fig. 13. Global filter: spectrum of the four global filters estimated with our ambiguities cancellation.

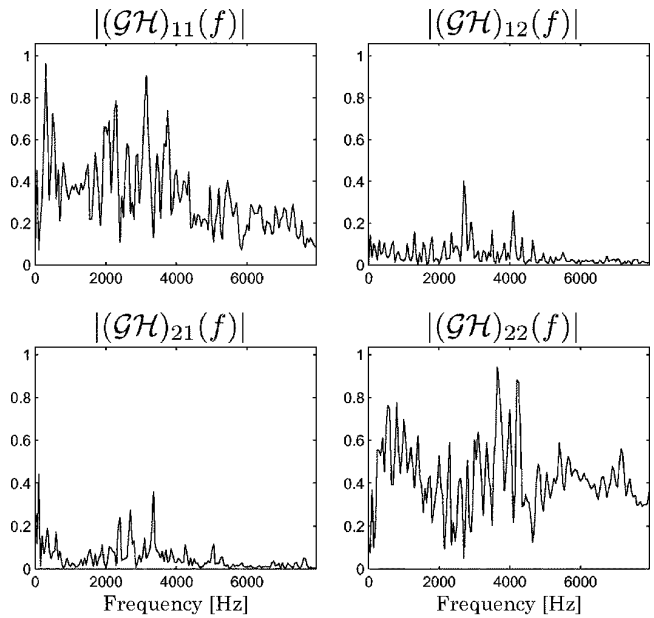

Fig. 14. Global filter without scale factor estimation.

Finally, Fig. 14 shows the spectrum of the global filter where we only made the permutation cancellation: we estimated the permuted frequency bins with our ambiguities cancellation algorithm (Section IV-D) and then the global filter is estimated by $\mathcal{P}(f)\mathcal{G}(f)\mathcal{H}(f)$ (i.e., without scale factor cancellation). One can see that $(\mathcal{GH})_{11}(f)$ is much closer to a constant with the scale factor cancellation (Fig. 13) than without scale factor cancellation (Fig. 14). This provides a better estimation of the spectrum shape of the estimated source of interest. Moreover, note that $(\mathcal{GH})_{21}(f)$ and $(\mathcal{GH})_{22}(f)$ are left unchanged since our criteria only consider channel 1.

VI. Conclusion

The BSS problem of convolutive speech mixtures can be processed by using a pure audio technique like a joint diagonalization process in the time-frequency domain [30]–[33]. However, this only provides a solution up to a permutation and a scale factor at each frequency bin. In this paper, we proposed a new statistical AV model expressing the complex relationship between two basic lip video parameters and acoustic speech parameters which consisted in the log-modulus of the coefficients of the short-time Fourier transform. The series of experiments presented in this paper, using our new AV model, confirm the interest of using AV processing to solve these ambiguities. However, the method required integration over a large number of frames to obtain a good estimation of the permutations. Thus, our method is very efficient in order to estimate large blocks of consecutive permuted frequencies (20 to 40 frames are enough in this case). However, over 100 frames are generally necessary to find isolated permuted frequency. Note that our presented method uses criteria which only consider one source of interest. If we extend the criteria to two sources of interest, the estimation of permutation is quite more efficient. Indeed, if the models of two distinct sources are known, then if a permutation is not detected by one of them, the other may find the permutation. In this paper, we are not only interested in the permutation problem, but also in the scale factor: doing this, we improve both the permutation detection and the estimation of the sources.

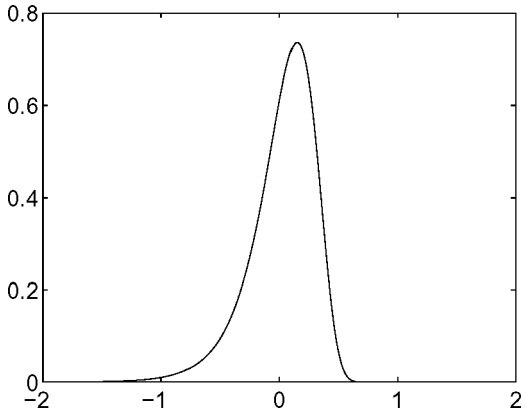

Fig. 15. Probability density function of a Log–Rayleigh of parameter $\beta^2 = 1$.

It is already possible to assert that our AV method seems useful and original to estimate the ambiguities in the difficult and realistic problem of convolutive mixtures. Indeed, we use an additional video information, which is intrinsically robust to acoustic noise. A major strength of our method is that it can extract a speech source from any kind of corrupting signals.

Finally, since the dimension of the parameters vector is large, a further step could be to search for other more efficient data representations and/or associated algorithms. Of course, other developments are still necessary for a complete demonstration of the efficiency of the proposed method: a first step could be to do experiments on a larger and more complex corpus, including continuous speech material and multispeaker AV models. We are currently working on this topic.

Appendix I
Log–Rayleigh Distribution

In this appendix, we briefly recall the definition of the log–Rayleigh distribution (for more details refer to [36]).

Definition 1: Let Z denote a Log–Rayleigh random variable with localization parameter $\beta^2 : Z \sim \mathrm{LogRay}(\beta^2)$. The probability density function of this variable is given by

$$\forall z \in \mathbb{R}, \quad p_{LR}(z|\beta^2) = \ln 10 \frac{(10^z)^2}{\beta^2} \exp\left(-\frac{(10^z)^2}{2\beta^2}\right). \quad (41)$$

This distribution is plotted in Fig. 15.

In the multidimensional case, for sake of simplicity, we note $\mathrm{LogRay}(\Gamma^2)$ the log–Rayleigh distribution where the parameter Γ^2 is the diagonal matrix $\Gamma^2 = \mathrm{diag}(\gamma_1^2, \ldots, \gamma_N^2)$ which is the product of the monodimensional Log–Rayleigh distribution

$$\mathrm{LogRay}(\Gamma^2) = \prod_{i=1}^{N} \mathrm{LogRay}\left(\gamma_i^2\right). \quad (42)$$

Appendix II
Learning the Audiovisual Model

In this appendix, we use the EM algorithm [50] in a penalized version [48], [49] to obtain the update equations for the audiovisual model (13). The EM algorithm is a recursive method

to estimate the parameter set $\Theta = \{\omega_i^{AV}, \mu_i^V, \Sigma_i^V, \Sigma_i^A\}_{1\leq i\leq I}$ thanks to the maximum likelihood. There are two stages.

1) Expectation (E): calculation of the *a posteriori* probability

$$p\left(i|\mathbf{v}(t),\mathbf{A}(t),\Theta^k\right) = \frac{\left(\omega_i^{AV}\right)^k p\left(\mathbf{v}(t),\mathbf{A}(t)|\left(\mu_i^{AV}\right)^k,\left(\Sigma_i^{AV}\right)^k\right)}{\sum_{j=1}^{I}\left(\omega_j^{AV}\right)^k p\left(\mathbf{v}(t),\mathbf{A}(t)|\left(\mu_j^{AV}\right)^k,\left(\Sigma_j^{AV}\right)^k\right)}$$

where $(\cdot)^k$ refers to the parameters at the kth iteration.

2) Maximization (M): update of the parameters
   - weight

$$\left(\omega_i^{AV}\right)^{k+1} = \frac{1}{T}\sum_{t=1}^{T} p\left(i|\mathbf{v}(t),\mathbf{A}(t),\Theta^k\right)$$

   - video parameters

$$\left(\mu_i^V\right)^{k+1} = \frac{\sum_{t=1}^{T}\left[\mathbf{v}(t)p\left(i|\mathbf{v}(t),\mathbf{A}(t),\Theta^k\right)\right]}{\sum_{t=1}^{T} p\left(i|\mathbf{v}(t),\mathbf{A}(t),\Theta^k\right)}$$

   and for d equal to w or h:

$$\left(\left(\sigma_i^V(d)\right)^2\right)^{k+1} = \frac{\sum_{t=1}^{T}\left[\left(v(t,d)-\left(\left(\mu_i^V(d)\right)^{k+1}\right)^2 p\left(i|\mathbf{v}(t),\mathbf{A}(t),\Theta^k\right)\right]+2\alpha_i}{\sum_{t=1}^{T} p\left(i|\mathbf{v}(t),\mathbf{A}(t),\Theta^k\right)+2\beta_i}$$

   where α_i and β_i are the penalized parameters [48],
   - the audio parameters, for all f_l

$$\left(\left(\sigma_i^A(f_l)\right)^2\right)^{k+1} = \frac{\sum_{t=1}^{T}\left[\left(10^{A(t,f_l)}\right)^2 p\left(i|\mathbf{v}(t),\mathbf{A}(t),\Theta^k\right)\right]}{\sum_{t=1}^{T} p\left(i|\mathbf{v}(t),\mathbf{A}(t),\Theta^k\right)}.$$

REFERENCES

[1] L. E. Bernstein and C. Benoît, "For speech perception by humans or machines, three senses are better than one," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 1996, pp. 1477–1480.

[2] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, pp. 212–215, 1954.

[3] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds. Mahwah, NJ: Lawrence Erlbaum, 1987, pp. 3–51.

[4] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, and P. Escudier, "Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise," *J. Acoust. Soc. Amer.*, vol. 103, no. 6, pp. 3677–3689, 1998.

[5] E. D. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D. dissertation, Univ. Illinois, Urbana, 1984.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

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

[8] J. Kim and D. Chris, "Investigating the audio-visual speech detection advantage," *Speech Commun.*, vol. 44, no. 1–4, pp. 19–30, 2004.

[9] L. E. Bernstein, E. T. J. Auer, and S. Takayanagi, "Auditory speech detection in noise enhanced by lipreading," *Speech Commun.*, vol. 44, no. 1–4, pp. 5–18, 2004.

[10] 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. Int. Conf. Spoken Lang. Process. (ICSLP)*, 2002, pp. 1937–1940.

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

[12] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (AudioVisual Codebook Dependent Cepstral Normalization)," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 2002, pp. 1449–1452.

[13] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Orlando, FL, May 2002, pp. 2025–2028.

[14] L. Girin, A. Allard, and J.-L. Schwartz, "Speech signals separation: A new approach exploiting the coherence of audio and visual speech," in *IEEE Int. Workshop Multimedia Signal Process. (MMSP)*, Cannes, France, 2001.

[15] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: a new approach exploiting the audiovisual coherence of speech stimuli," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 11, pp. 1165–1173, 2002.

[16] D. Sodoyer, L. Girin, C. Jutten, and J.-L. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Commun.*, vol. 44, no. 1–4, pp. 113–125, Oct. 2004.

[17] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[18] C. Jutten and A. Taleb, "Source separation: from dusk till dawn," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation (ICA)*, Helsinki, Finland, Jun. 2000, pp. 15–26.

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

[20] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[21] J. Hérault, C. Jutten, and B. Ans, "Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétrique en apprentissage non supervisé," in *Proc. GRETSI*, Nice, France, May 1985, vol. 2, pp. 1017–1020.

[22] C. Jutten and J. Hérault, "Blind separation of sources. Part I: An adaptive algorithm based on a neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, Jul. 1991.

[23] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *Proc. Inst. Elect. Eng. F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[24] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.

[25] A. Bell and T. Sejnowski, "An information-maximization approach to blind source separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[26] L. De Lathauwer, D. Callaerts, B. De Moor, and J. Vandewalle, "Fetal electrocardiogram extraction by source subspace separation," in *Proc. IEEE Workshop HOS*, Girona, Spain, Jun. 12–14, 1995, pp. 134–138.

[27] V. Zarzoso and A. K. Nandi, "Noninvasive fetal electrocardiogram extraction: Blind source separation versus adaptative noise cancellation," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 1, pp. 12–18, Jan. 2001.

[28] R. Dansereau, "Co-channel audiovisual speech separation using spectral matching constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montréal, QC, Canada, 2004.

[29] S. Rajaram, A. V. Nefian, and T. S. Huang, "Bayesian separation of audio-visual speech sources," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Montréal, Canada, 2004, pp. 645–648.

[30] V. Capdevielle, C. Servière, and J.-L. Lacoume, "Blind separation of wide-band sources in the frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Detroit, MI, May 1995, pp. 2080–2083.

[31] L. Para and C. Spence, "Convolutive blind separation of non stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.
[32] A. Dapena, M. F. Bugallo, and L. Castedo, "Separation of convolutive mixtures of temporally-white signals: A novel frequency-domain approach," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation (ICA)*, San Diego, CA, Dec. 2001, pp. 315–320.
[33] D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationary," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation (ICA)*, Nara, Japan, Apr. 2003, pp. 981–986.
[34] 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 *IEEE Int. Workshop Multimedia Signal Process. (MMSP)*, Sienna, Italy, Oct. 2004, pp. 47–50.
[35] B. Rivet, L. Girin, and C. Jutten, "Solving the indeterminations of blind source separation of convolutive speech mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Philadelphia, PA, Mar. 2005, pp. 533–536.
[36] ——, "Log-Rayleigh distribution: A simple and efficient statistical representation of log-spectral coefficients," *IEEE Trans. Audio, Speech, Lang. Process.*, 2006, submitted for publication.
[37] H.-L. Nguyen-Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Process.*, vol. 45, pp. 209–229, 1995.
[38] D.-T. Pham, "Joint approximate diagonalization of positive definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 22, no. 4, pp. 1136–1152, 2001.
[39] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomput.*, vol. 41, no. 1–4, pp. 1–24, Oct. 2001.
[40] B. Le Goff, T. Guiard-Marigny, and C. Benoît, "Read my lips. . . and my jaw! How intelligible are the components of a speaker's face?," in *Proc. Euro. Conf. Speech Communication Technology*, Madrid, Spain, 1995, pp. 291–294.
[41] ——, "Analysis-synthesis and intelligibility of a talking face," in *Progress in Speech Synthesis*J. Van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. New York: Springer-Verlag, 1996, pp. 235–244.
[42] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *Proc. Audio-Visual Speech Processing Workshop (AVSP)*, Aalborg, Denmark, 2001, pp. 90–97.
[43] T. Lallouache, "Un poste visage-parole. Acquisition et traitement des contours labiaux," in *Proc. Journées d'Etude sur la Parole (JEP)* (in French), Montréal, QC, Canada, 1990, pp. 282–286.
[44] B. Picinbono, "Second-order complex random vectors and normal distributions," *IEEE Trans. Signal Process.*, vol. 44, no. 10, pp. 2637–2640, Oct. 1996.
[45] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1293–1302, Jul. 1993.
[46] L. Benaroya, "Séparation de plusieurs sources sonores avec un seul microphone," Ph.D. dissertation, Traitement du signal, Univ. Rennes 1, Rennes, France, Jun. 2003.
[47] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, no. 1, pp. 23–43, 1998.
[48] D. Ormoneit and V. Tresp, "Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates," *IEEE Trans. Neural Netw.*, vol. 9, no. 4, pp. 639–650, Jul. 1998.
[49] H. Snoussi and A. Mohammad-Djafari, "Penalized maximum likelihood for multivariate Gaussian mixture," in *Proc. Bayesian Inference and Maximum Entropy Methods, R. L. Fry, Ed. MaxEnt Workshops*, Aug. 2001, pp. 36–46.
[50] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Ser. B.*, vol. 39, pp. 1–38, 1977.

Bertrand Rivet graduated from the École Normale Supérieure de Cachan, Cachan, France, and received the Agrégation de Physique Appliquée and the Master's degree from the University of Paris-XI, Orsay, France, in 2002 and 2003, respectively. He is currently pursuing the Ph.D. degree in signal processing at the Institut de la Communication Parlée (Speech Communication Laboratory) and at the Laboratoire des Images et des Signaux (Images and Signals Laboratory), Grenoble, France.

His research concerns audiovisual speech and blind source separation.

Laurent Girin received the M.Sc. and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble, Grenoble, France, in 1994 and 1997, respectively.

In 1997, he joined the Ecole Nationale d'Electronique et de Radioélectricité de Grenoble, where he is currently an Associate Professor in electrical engineering and signal processing. His research activity takes place in the Institut de la Communication Parlée (Speech Communication Laboratory) in Grenoble. His current research interests concern audiovisual speech processing with application to speech coding, speech enhancement, and audio/speech source separation, and also acoustic speech analysis, modeling and synthesis.

Christian Jutten received the Ph.D. and the Docteur ès Sciences degrees from the Institut National Polytechnique of Grenoble, Grenoble, France, in 1981 and 1987, respectively.

He was an Associate Professor with the Ecole Nationale Supérieure d'Electronique et de Radioélectricité, Grenoble, from 1982 to 1989. He was a Visiting Professor with the Swiss Federal Polytechnic Institute, Lausanne, Switzerland, in 1989, before becoming a Full Professor with the Université Joseph Fourier, Grenoble, more precisely in Polytech'Grenoble Institute. He is currently an Associate Director of the Images and Signals Laboratory (100 people). For 25 years, his research interests have been in blind source separation, independent component analysis, and learning in neural networks, including theoretical aspects (separability, source separation in nonlinear mixtures), applications in signal processing (biomedical, seismic, speech), and data analysis. He is author or coauthor of more than 40 papers in international journals, 16 invited paper, and 100 communications in international conferences.

Prof. Jutten was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS (1994–19 95) and coorganizer the First International Conference on Blind Signal Separation and Independent Component Analysis (Aussois, France, January 1999). He is a Reviewer of main international journals (IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE SIGNAL PROCESSING LETTERS, IEEE TRANSACTIONS ON NEURAL NETWORKS, *Signal Processing*, *Neural Computation*, *Neurocomputing*, etc.) and conferences in signal processing and neural networks (ICASSP, ISCASS, EUSIPCO, IJCNN, ICA, ESANN, IWANN, etc.).