Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 667–677
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Visual voice activity detection as a help for speech source separation from convolutive mixtures

Bertrand Rivet [a,b,*], Laurent Girin [a], Christian Jutten [b]

[a] Institut de la Communication Parlée (ICP), CNRS UMR 5009, INPG, Université Stendhal, Grenoble, France
[b] Laboratoire des Images et des Signaux (LIS), CNRS UMR 5083, INPG, Université Joseph Fourier, Grenoble, France

Received 31 January 2006; received in revised form 26 January 2007; accepted 11 April 2007

Abstract

Audio–visual speech source separation consists in mixing visual speech processing techniques (e.g., lip parameters tracking) with source separation methods to improve the extraction of a speech source of interest from a mixture of acoustic signals. In this paper, we present a new approach that combines visual information with separation methods based on the sparseness of speech: visual information is used as a voice activity detector (VAD) which is combined with a new geometric method of separation. The proposed audio–visual method is shown to be efficient to extract a real spontaneous speech utterance in the difficult case of convolutive mixtures even if the competing sources are highly non-stationary. Typical gains of 18–20 dB in signal to interference ratios are obtained for a wide range of (2×2) and (3×3) mixtures. Moreover, the overall process is computationally quite simpler than previously proposed audio–visual separation schemes.

Keywords: Speech source separation; Convolutive mixtures; Voice activity detector; Visual speech processing; Speech enhancement; Highly non-stationary environments

1. Introduction

Audio–visual speech source separation (AVSSS) is a growing field of interest to solve the source separation problem when speech signals are involved. It consists of exploiting the bimodal (audio–visual) nature of speech to improve the performance of acoustic speech signal separation (Bernstein and Benoît, 1996; Sumby and Pollack, 1954). For instance, pioneer works by Girin et al. (2001) and then by Sodoyer et al. (2004) have proposed to use a statistical model between the coherence of audio and visual speech features to estimate the separating matrix for additive mixtures. Later, Dansereau (2004) and Rajaram et al. (2004) respectively plugged the visual information in a 2×2 decorrelation system with first-order filters and in the Bayesian framework for a 2×2 linear mixture. Unfortunately, real audio mixtures are generally more complex and better described as convolutive mixtures with quite long filters. Recently, Rivet et al. (2007) have proposed a new approach to exploit visual speech information in such convolutive mixtures. Visual parameters were used to regularize the permutation and the scale factor indeterminacies that arise at each frequency bin in frequency-domain separation methods (Capdevielle et al., 1995; Parra and Spence, 2000; Dapena et al., 2001; Pham et al., 2003). In parallel, the audio–visual (AV) coherence maximization approach was also considered for the estimation of deconvolution filters in Wang et al. (2005).

In this paper, we propose a simpler and more efficient approach for the same problem (extracting one speech source from convolutive mixtures using the visual speech information). First we propose to use visual speech information, for instance lip movements, as a voice activity detector (VAD): the task is to assess the presence or the

* Corresponding author. Address: Institut de la Communication Parlée (ICP), CNRS UMR 5009, INPG, Université Stendhal, Grenoble, France.
E-mail address: rivet@icp.inpg.fr (B. Rivet).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.04.008

absence of a given speaker's speech signal in the mixture, crucial information to be further used in separation processes. Such visual VAD (V-VAD) is characterized by a major advantage as opposed to usual acoustic VADs: it is robust to any acoustic environment, whatever the nature and the number of competing sources (e.g. simultaneous speaker(s), non-stationary noises, convolutive mixture, etc.). Note that previous work on VAD based on visual information can be found in Liu and Wang (2004). The authors proposed to model the distribution of the visual information using two exclusive classes (one for speech non-activity and one for actual speech activity): the decision is then based on likelihood criterion. However, the presented approach is completely different since we exploit the temporal dynamic of lip movements and we do not use an *a priori* statistical model (Section 2).

Secondly, we propose a geometric approach for the extraction process exploiting the sparseness of the speech signals (Section 3). One of the major drawback of the frequency-domain separation methods is the need for regularizing the indeterminacies encountered at each frequency bin (Cardoso, 1998). Indeed, the separation is generally done separately at each frequency bin by statistical considerations, and arbitrary permutations between estimated sources and arbitrary scale factors can occur leading to a wrong reconstruction of the estimated sources. Several solutions to the permutation problem were proposed e.g. exploiting the correlation over frequencies of the reconstructed sources (Parra and Spence, 2000; Dapena et al., 2001), exploiting the smoothness of the separating filters (Pham et al., 2003) or exploiting AV coherence (Rivet et al., 2007). Alternately, other methods try to exploit the sparseness of the sources. For instance, Abrard and Deville (2005) proposed a solution in the case of instantaneous mixture. They exploit the frequency sparseness of the sources: in the time–frequency plane, areas where only one source is present are selected by using an acoustic VAD, allowing the determination of the separating matrix. However, their method has two restrictions: (i) it concerns instantaneous mixtures while real mixtures are often convolutive, (ii) it requires time–frequency areas where only one source is present, which is a very strong assumption (the number of such areas is very small).[1] Recently, Babaie-Zadeh et al. (2004) proposed a geometric approach in the case of instantaneous mixtures of sparse sources. The method is based on the identification of the main directions of the present sources in the mixtures. Our proposed method is also geometric but is quite different from their method, since in our approach (i) only the source to be extracted has to be sparse, (ii) the indexation of the sections where the source to be extracted is absent is done thanks to the proposed V-VAD, (iii) the case of convolutive mixtures is addressed. Also, in addition to intrinsically solve the permutation problem for the reconstructed source, the proposed method is refined by an additional stage to regularize the scale factor ambiguity.

This paper is organized as follows. Section 2 presents the basis of the proposed V-VAD. Section 3 explains the proposed geometrical separation using the V-VAD first in the case of instantaneous mixtures and then in the case of convolutive mixtures. Section 4 presents both the analysis of the V-VAD and the results of the AV separation process before conclusions in Section 5.

2. Visual voice activity detection

In this section, we present our visual voice activity detector (V-VAD) (Sodoyer et al., 2006). For the purpose of developing and assessing this V-VAD, a dedicated audio–visual corpus, denoted C_1, of 45 min of spontaneous speech was recorded. Two male French speakers were set in a spontaneous dialog situation with many speech overlapping and non-speech events. The two speakers were placed and recorded in a different room to collect separately the two audio signals. Each speaker had a microphone and a micro-camera focused on the lip region (Fig. 1a and b).

The visual information consists of the time trajectory of basic lip contour geometric parameters, namely interolabial width $w(k)$ and height $h(k)$, where k represents discrete time index (Fig. 1c). Indeed, several studies have shown

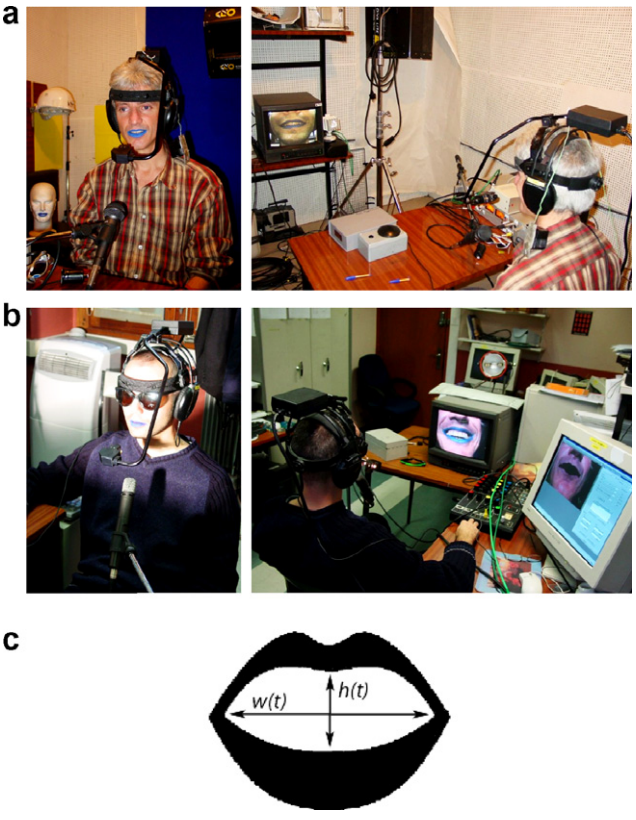

Fig. 1. Experimental conditions used to record the corpus C_1 (a) and (b). (c) presents the video parameters: internal width w and internal height h.

[1] In the case of N sources, if p is the probability of the source absence, the probability that only a given source is present is equal to $p^{N-1}(1-p)$, assuming that the presence of the sources is independent.

that the basic facial lip edge parameters contain most of the visual information according to both intelligibility criterion (Le Goff et al., 1995) and statistical analysis (Elisei et al., 2001). These parameters were automatically extracted by using a device and an algorithm developed at the ICP (Lallouache, 1990). The technique is based on blue make-up, Chroma-Key system and contour tracking algorithms. The parameters are extracted every 20 ms (the video sampling frequency is 50 Hz), synchronously with the acoustic signal which is sampled at 16 kHz. Thus in the following, an audio–visual signal frame is a 20 ms section of acoustic signal associated with a video pair parameters $(w(k), h(k))$.

The aim of a voice activity detector (VAD) is to discriminate speech and non-speech sections of the acoustic signal. However, we prefer to use the distinction between *silence* (defined as vocal inactivity) and *non-silence* sections for a given speaker because non-speech sections are not bound to be silence, since many kinds of non-speech sounds can be produced by the speaker (e.g. laughs, sighs, growls, moans, etc.). Moreover, the separation system of Section 3 is based on the detection of complete non-activity of the speaker to be extracted from the mixture (*i.e.* the detection of time periods where no sound is produced by the speaker is used to extract the speech signal produced during active periods). To provide an objective reference for the detection, we first manually identified and labeled acoustic sections of silence and non-silence. Then, we defined a normalized video vector as $\pi(k) = [w(k)/\mu_w, \kappa h(k)/\mu_h]^{\mathrm{T}}$ where κ is the coefficient of linear regression between $w(k)$ and $h(k)$, μ_w and μ_h the mean values of $w(k)$ and $h(k)$ calculated on the complete corpus for each speaker ($^{\mathrm{T}}$ denotes the transpose operator).

As explained in (Sodoyer et al., 2006), a direct VAD from raw lip parameters cannot lead to satisfactory performances because of the intricate relationship between visual and acoustic speech information. Indeed, Fig. 2 represents the distribution of the first component $\pi_1(k)$ and the second component $\pi_2(k)$ of vector $\pi(k)$ for non-silence frames (Fig. 2a) and silence frames (Fig. 2b). One can see that there is no trivial partition between the two classes (silence vs. non-silence): for instance, closed lip-shapes are present in both distributions and they cannot be systematically associated with a silence frame. The V-VAD (Sodoyer et al., 2006) is based on the fact that silence frames can be better characterized by the lip-shape movements. Indeed, in silence sections, the lip-shape variations are generally small, whereas in speech sections these variations are generally quite stronger. So we proposed the following dynamical video parameter:

$$v(k) = \left|\frac{\partial \pi_1(k)}{\partial k}\right| + \left|\frac{\partial \pi_2(k)}{\partial k}\right|, \tag{1}$$

where the derivations will be implemented as differences. The kth input frame is classified as silence if $v(k)$ is lower than a threshold and it is classified as speech otherwise. However, direct thresholding of $v(k)$ does not provide optimal performance: for instance, the speaker's lips may not move during several frames, while he is actually speaking. Thus, we smooth $v(k)$ by summation over T consecutive frames

$$V(k) = \sum_{i=0}^{T-1} \alpha^i v(k-i), \tag{2}$$

where α is a real coefficient between 0 and 1 and T is chosen large enough so that α^{T-1} is very small compared to α. Finally, the kth frame is classified as silence if $V(k)$ is lower than a threshold $\delta (V(k) < \delta)$, and it is classified as speech otherwise ($V(k) \geqslant \delta$). Fig. 3 shows that the choice of coefficient α must be considered carefully. A too small value of α (or no summation) leads to a detection which is very sensitive to local perturbations (Fig. 3a). On the contrary, a too large α leads to a quite incorrect detection (Fig. 3c). Fig. 3b shows that the choice of α can largely improve the separation of silence and non-silence sections. Since the aim of the V-VAD, as explained in Section 3, is to detect frames where the speaker does *not* produce sounds, we propose an additional stage before the decision in order to decrease the false alarm (silence decision while speech activity). Only sequences of at least L frames of silence are actually considered as silences. The value of L is varied

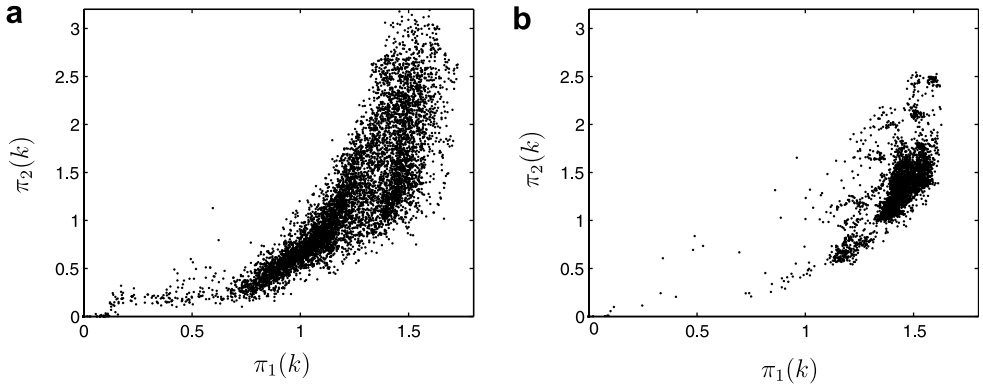

Fig. 2. Distribution of the visual parameter $\pi(t)$ for non-silence frames (a) and silence frames (b). Note that 10% and 36% of the points are at the origin (closed lip-shape) for the (a) and (b) figures respectively. A total (silence and non-silence) of about 13,200 20-ms-frames was used.

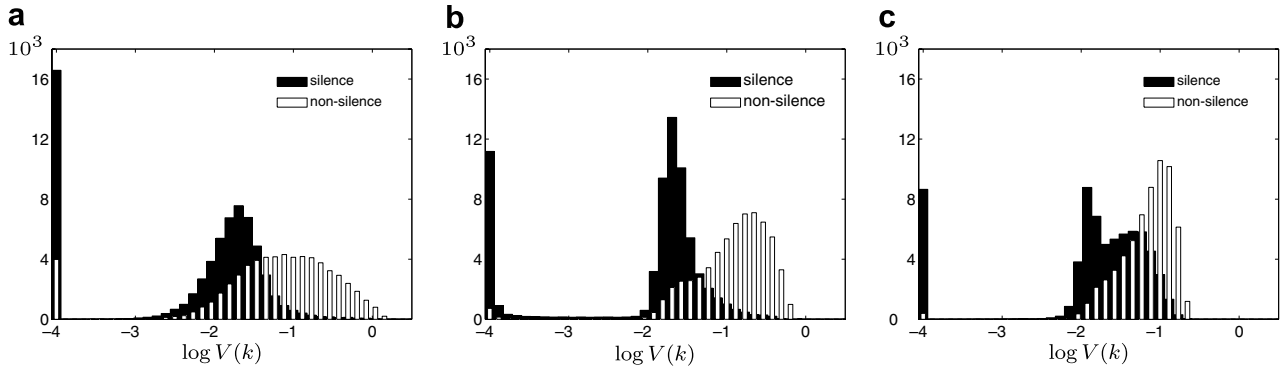

Fig. 3. Histograms of the dynamical visual parameter $V(k)$ (on log-scale) for three values of the pair (α, T). (a) Instantaneous case $\alpha = 0$ and $T = 1$ (*i.e.* $V(k) = v(k)$ since we adopt the classical convention $0^0 = 1$ in (2)). (b) Suitable value of $\alpha = 0.82$, with $T = 50$. (c) Too large value of $\alpha = 0.99$, with $T = 1000$. In each case, the histogram plotted in black represents values $V(k)$ associated with silence sections, and the histogram plotted in white represents values $V(k)$ associated with non-silence sections. A total (silence and non-silence) of about 130,000 20-ms-frames was used for this plot.

in the experiments section and the corresponding results are discussed. Finally, the proposed V-VAD is robust to any acoustic noise and can be exploited even in difficult non-stationary environments. The performance of the proposed V-VAD is given in Section 4.1.

3. Speech source separation using visual voice activity detector

In this section, we present the new geometrical method to extract one source of interest, say $s_1(k)$, from the observations $\mathbf{x}(k)$. The main idea of the proposed method is to exploit both (i) the "sparseness" property of speech signals: in real spontaneous speech situations (e.g. dialog), there exist some periods (denoted "silence") during which each speaker is silent as discussed in Section 2, (ii) the possibility to detect these silent sections by using the V-VAD of Section 2. Note that the method allows the source with detected silence sections to be extracted from the mixtures. If other sources are to be extracted, they should have their own associated silence detector. We first explain the principle of the separation process in the simple case of complex instantaneous mixtures, then we extend it to convolutive mixtures. We use the complex value for purpose of generality because in the case of convolutive mixture, complex spectral values will be considered.

3.1. Case of complex instantaneous mixtures

Let us consider the case of N complex independent centered sources $\mathbf{s}(k) \in \mathbb{C}^N$ and N complex observations $\mathbf{x}(k) \in \mathbb{C}^N$ obtained by a complex mixing matrix $\mathscr{A} \in \mathbb{C}^{\mathscr{N}\times\mathscr{N}}$:

$$\mathbf{x}(k) = \mathscr{A}\mathbf{s}(k), \tag{3}$$

where $\mathbf{s}(k) = [s_1(k), \ldots, s_N(k)]^{\mathrm{T}}$ and $\mathbf{x}(k) = [x_1(k), \ldots, x_N(k)]^{\mathrm{T}}$. We suppose that $\mathscr{A}$ is invertible. Thus to extract the sources $\mathbf{s}(k)$, we have to estimate a separating matrix $\mathscr{B} \in \mathbb{C}^{N\times N}$, which is typically an estimate of the inverse of $\mathscr{A}$. It is a classical property of usual source separation systems that the separation can only be done up to a permutation and a scaling factor (Cardoso, 1998), that is

$$\mathscr{B} \simeq \mathscr{P}\mathscr{D}\mathscr{A}^{-1}, \tag{4}$$

where $\mathscr{P}$ is a permutation matrix and $\mathscr{D}$ is a diagonal matrix. We denote $\mathscr{C} = \mathscr{B}\mathscr{A}$ as the global matrix. Thus, to extract the first source $s_1(k)$, only one row of $\mathscr{B}$ is necessary, we arbitrary choose the first one, denoted $\mathbf{b}_{1,:}$[2]:

$$\hat{s}_1(k) = \mathbf{b}_{1,:}\mathbf{x}(k) = \mathbf{b}_{1,:}\mathscr{A}\mathbf{s}(k) \simeq c_{1,1}s_1(k). \tag{5}$$

In the following, we propose a novel method to estimate $\mathbf{b}_{1,:}$. Moreover, we go one step further by regularizing the scale factor so that the source $s_1(k)$ is estimated up to $a_{1,1}$ instead of $c_{1,1}$. This corresponds to the situation where the estimation of the source $s_1(k)$ is equal to the signal contained in $x_1(k)$ when the other sources vanish (in other words, $s_1(k)$ is estimated up to its mixing matrix ,*i.e.* channel + sensor, coefficient $a_{1,1}$ defined in (3)).

To estimate $\mathbf{b}_{1,:}$, we propose a novel geometric method. The dimension of the spaces $\mathscr{S}_s$ and $\mathscr{S}_x$, spanned by the sources $\mathbf{s}(k)$ and the observations $\mathbf{x}(k)$ respectively, is N (Fig. 4a and b for three uniform distributed real sources). The space spanned by the contribution of source s_1 in $\mathscr{S}_x$ is a straight line denoted $\mathscr{D}_1$ (Fig. 4c). Now, suppose that an oracle[3] gives us $\mathscr{T}$, a set of time indexes when $s_1(k)$ vanishes, then the space $\mathscr{S}'_x$ (respectively $\mathscr{S}'_s$), spanned by $\mathbf{x}(k)$ (respectively $\mathbf{s}(k)$), with $k \in \mathscr{T}$ is a hyper-plane (*i.e.* space of dimension $N-1$) of $\mathscr{S}_x$ (respectively $\mathscr{S}_s$). Moreover, $\mathscr{D}_1$ is a supplementary space of $\mathscr{S}'_x$ in $\mathscr{S}_x$:

$$\mathscr{S}_x = \mathscr{S}'_x \oplus \mathscr{D}_1. \tag{6}$$

Note that $\mathscr{S}'_x$ and $\mathscr{D}_1$ are not necessary orthogonal. Moreover, $\mathscr{S}'_x$ is the space spanned by the contribution of sources $\{s_2(k), \ldots, s_N(k)\}$ in $\mathscr{S}_x$. Thus to extract s_1, we have

[2] In this paper, $\mathbf{b}_{i,:} = [b_{i,1}, \ldots, b_{i,N}]$, and $\mathbf{b}_{:,i} = [b_{1,i}, \ldots, b_{N,i}]^{\mathrm{T}}$, where $b_{i,j}$ is the (i,j)th element of matrix $\mathscr{B}$.

[3] Such an oracle is provided by the V-VAD of Section 2.

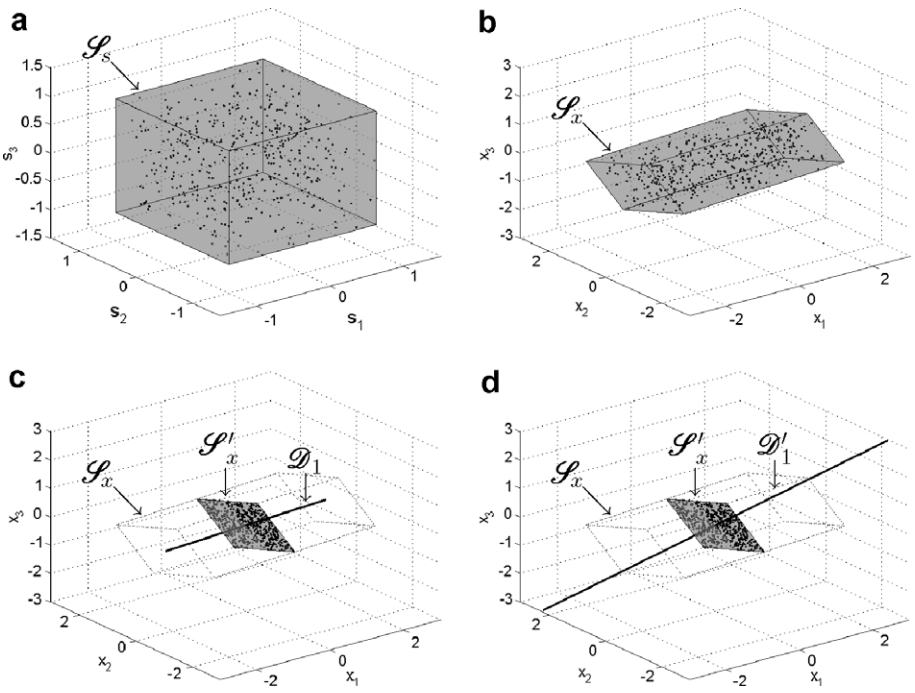

Fig. 4. Illustration of the geometric method in the case of three observations obtained by an instantaneous mixture of three uniform sources. (a) Three independent sources (s_1, s_2, s_3) (dots) with uniform distribution and $\mathscr{S}_s$ the space spanned by them. (b) A 3×3 instantaneous mixture (x_1, x_2, x_3) (dots) of these sources and $\mathscr{S}_x$ the space spanned by the mixture. (c) $\mathscr{S}'_x$ and $\mathscr{D}_1$ (solid line). (d) $\mathscr{S}'_x$ and $\mathscr{D}'_1$ (solid line).

to project the observations $\mathbf{x}(k)$ on a supplementary space of $\mathscr{S}'_x$ (not necessary $\mathscr{D}_1$).

To find this supplementary space, a proposed solution is to use a principal component analysis (PCA). Indeed, performing an eigenvalue decomposition of the covariance matrix $C_{\mathbf{xx}} = E\{\mathbf{x}(k)\mathbf{x}^H(k)\}$ ($\cdot^H$ denotes the complex conjugate transpose) of the observations $\mathbf{x}(k)$ with $k \in \mathscr{T}$ (the set of time indexes when $s_1(k)$ vanishes) provides N orthogonal (since $C_{\mathbf{xx}}$ is Hermitian) eigenvectors associated with N eigenvalues that represent the respective average powers of the N sources during time slots $\mathscr{T}$. Since $s_1(k)$ is absent for $k \in \mathscr{T}$, the smallest eigenvalue within the eigenvalues set is to be associated to this source (this smallest eigenvalue should be close to zero). The straight line $\mathscr{D}'_1$, spanned by the (row) eigenvector, denoted $\mathbf{g} = [1, g_2, \ldots, g_N]$ (g_1 is arbitrary chosen equal to 1), associated with the smallest eigenvalue, defines the orthogonal supplementary space of $\mathscr{S}'_x$ in $\mathscr{S}_x$:

$$\mathscr{S}_x = \mathscr{S}'_x \overset{\perp}{\oplus} \mathscr{D}'_1. \tag{7}$$

Thus, for all time indexes k (now including when source s_1 is active), an estimate of source $s_1(k)$ can be extracted thanks to

$$\hat{s}_1(k) = \mathbf{g}\mathbf{x}(k) \simeq c_{1,1}s_1(k). \tag{8}$$

This way, $\mathbf{b}_{1,:}$ is identified to $\mathbf{g}$ (*i.e.* we set $\mathbf{b}_{1,:} = \mathbf{g}$), and we furthermore have $c_{1,1} = a_{1,1} + b_{1,2}a_{2,1} + \cdots + b_{1,N}a_{N,1}$. Note that scaling factor $c_{1,1}$ can be interpreted as an unchecked distortion since $\mathscr{D}_1$ is, *a priori*, a supplementary space of $\mathscr{S}'_x$ and not necessary the orthogonal supplementary space of $\mathscr{S}'_x$. As explained below (Section 3.2), in the convolutive case this distortion can dramatically alter the estimation of the source.

Now, we address the last issue of fixing the scaling factor to $a_{1,1}$ instead of $c_{1,1}$, *i.e.* we have to find a complex scalar λ such that $s_1^{\dagger}(k) = \lambda\hat{s}_1(k) \simeq a_{1,1}s_1(k)$. Thus λ is given by

$$\lambda = \frac{a_{1,1}}{a_{1,1} + \sum_{i>1} b_{1,i}a_{i,1}} = \frac{1}{1 + \sum_{i>1} b_{1,i}a_{i,1}/a_{1,1}}, \tag{9}$$

where $\forall i, b_{1,i}$ were estimated as explained above (by identifying $\mathbf{g}$ and $\mathbf{b}_{1,:}$) and the set $\{a_{i,1}/a_{1,1}\}_i$ has to be estimated. To estimate these coefficients, we propose a procedure based on the cancellation of the contribution of $\hat{s}_1(k)$ in the different mixtures $x_i(k)$. Thus, let denote $\epsilon_i(\beta_i) = E\{|x_i(k) - \beta_i\hat{s}_1(k)|_2\}$, where $E\{\cdot\}$ denote the statistical expectation operator. Since the sources are independent, we have thanks to (3) and (8)

$$\epsilon_i(\beta_i) = E\{|(a_{i,1} - \beta_i c_{1,1})s_1(k)|^2\} + \sum_{j>1} E\{|a_{i,j}s_j(k)|^2\}. \tag{10}$$

Moreover, $\forall\beta_i$, $\epsilon_i(\beta_i)$ is lower bounded by $\sum_{j>1} E\{|a_{i,j}s_j(k)|^2\}$ and the lower bound is obtained for $\beta_i = a_{i,1}/c_{1,1}$. Let us denote $\hat{\beta}_i$ as the optimal estimation of β_i in the minimum mean square error sense. $\hat{\beta}_i$ is classically given by

$$\hat{\beta}_i = \arg\min_{\beta_i} \epsilon_i(\beta_i) = \frac{E\{x_i^*(k)\hat{s}_1(k)\}}{E\{|\hat{s}_1(k)|^2\}}, \tag{11}$$

where $*$ denotes the complex conjugate. In practice, the expectation is replaced by time averaging and $\hat{\beta}_i$ is given by

$$\hat{\beta}_i = \frac{\sum_{k=1}^{K} x_i^*(k)\hat{s}_1(k)}{\sum_{k=1}^{K}|\hat{s}_1(k)|^2}. \quad (12)$$

So, λ is given by (9) where $a_{i,1}/a_{1,1}$ is replaced by $\hat{\beta}_i/\hat{\beta}_1$. Note that we use the ratio $a_{i,1}/a_{1,1}$ rather than $a_{i,1}$ alone since β_i is equal to $a_{i,1}$ up to the unknown coefficient $c_{1,1}$. Finally, the source $s_1(k)$ is estimated by

$$s_1^{\dagger}(k) = \lambda \mathbf{b}_{1,:}\mathbf{x}(k) \simeq a_{1,1}s_1(k). \quad (13)$$

3.2. Case of convolutive mixtures

Let us now consider the case of convolutive mixtures of N centered sources $\mathbf{s}(k) = [s_1(k), \ldots, s_N(k)]^{\mathrm{T}}$ to be separated from N observations $\mathbf{x}(k) = [x_1(k), \ldots, x_N(k)]^{\mathrm{T}}$:

$$x_m(k) = \sum_{n=1}^{N} h_{m,n}(k) * s_n(k). \quad (14)$$

The filters $h_{m,n}(k)$, which model the impulse response between the nth source and the mth sensor, are the entries of the global mixing filter matrix $\mathscr{H}(k)$.

The aim of the source separation is to recover the sources by using a dual filtering process:

$$s_n^{\dagger}(k) = \sum_{m=1}^{N} g_{n,m}(k) * x_m(k), \quad (15)$$

where $g_{n,m}(k)$ are the entries of the global separating filter matrix $\mathscr{G}(k)$ that must be estimated. The problem is generally considered in the frequency domain (Capdevielle et al., 1995; Parra and Spence, 2000; Dapena et al., 2001; Pham et al., 2003) where the single convolutive problem becomes a set of F (the number of frequency bins) simple linear instantaneous problems with complex entries. For all frequency bins f

$$X_m(k,f) = \sum_{n=1}^{N} H_{m,n}(f)S_n(k,f), \quad (16)$$

$$S_n^{\dagger}(k,f) = \sum_{m=1}^{N} G_{n,m}(f)X_m(k,f), \quad (17)$$

where $S_n(k,f)$, $X_m(k,f)$ and $S_n^{\dagger}(k,f)$ are the short-term Fourier transforms (STFT) of $s_n(k)$, $x_p(k)$ and $s_n^{\dagger}(k)$ respectively. $H_{m,n}(f)$ and $G_{n,m}(f)$ are the frequency responses of the mixing $\mathscr{H}(f)$ and demixing $\mathscr{G}(f)$ filters respectively. Since the mixing process is assumed to be stationary, $\mathscr{H}(f)$ and $\mathscr{G}(f)$ are not time-dependent, although the signals (*i.e.* sources, observations) may be non-stationary. In the frequency domain, the goal of the source separation is to estimate, at each frequency bin f, the separating filter $\mathscr{G}(f)$. This can be done thanks to the geometric method proposed in Section 3.1. Indeed, at each frequency bin f, (16) and (17) can be seen as a case of an instantaneous complex mixture problem. Thus, $\mathbf{b}_{1,:}(f)$ is the eigenvector associated with the smallest eigenvalue of the covariance matrix $C_{\mathbf{xx}}(f) = E\{\mathbf{X}(k,f)\mathbf{X}^H(k,f)\}$ with $k \in \mathscr{T}$. Then, $\beta_i(f)$ is a function of frequency f and is estimated thanks to

$$\hat{\beta}_i(f) = \frac{\sum_{k=1}^{K} X_i^*(k,f)\widehat{S}_1(k,f)}{\sum_{k=1}^{K}|\widehat{S}_1(k,f)|^2}. \quad (18)$$

So $\lambda(f)$ is given by

$$\lambda(f) = \frac{1}{1 + \sum_{i>1} b_{1,i}(f)a_{i,1}(f)/a_{1,1}(f)}, \quad (19)$$

where $a_{i,1}(f)/a_{1,1}(f)$ is replaced by $\hat{\beta}_i(f)/\hat{\beta}_1(f)$. Finally, the source $S_1(k,f)$ is estimated by

$$S_1^{\dagger}(k,f) = \mathbf{G}_{1,:}(f)\mathbf{X}(k,f) \simeq H_{1,1}(f)S_1(k,f), \quad (20)$$

where $\mathbf{G}_{1,:}(f) = \lambda(f)\mathbf{b}_{1,:}(f)$, or

$$s_1^{\dagger}(k) = \mathbf{g}_{1,:}(k) * \mathbf{x}(k) \simeq h_{1,1}(k) * s_1(k). \quad (21)$$

Note that in the convolutive case, if the scale factor regularization $\lambda(f)$ is not ensured, the source $s_1(k)$ is estimated up to an unknown filter which can perceptually alter the estimation of the source. On the contrary, performing the scale factor regularization ensures that the first source is estimated up to the filter $h_{1,1}(k)$ which corresponds to the "channel + sensor" filter of the first observation. The complete method is summarized in the following Algorithm 1.

Algorithm 1 (*Geometric separation in the convolutive case*).

Estimate index silence frames $\mathscr{T}$ using V-VAD (Section 2)
Perform STFT on the audio observations $x_m(k)$ to obtain $X_m(k,f)$
for all frequency bins f **do**
  {Estimation of $\mathbf{b}_{1,:}(f)$}
    Compute $C_{\mathbf{xx}}(f) = E\{\mathbf{X}(k,f)\mathbf{X}^H(k,f)\}$ with $k \in \mathscr{T}$
    Perform eigenvalue decomposition of $C_{\mathbf{xx}}(f)$
    Select $\mathbf{g}(f)$ the eigenvector associated with the smallest eigenvalue
    $\mathbf{b}_{1,:}(f) \Leftarrow \mathbf{g}(f)$
  {Estimation of $\lambda(f)$ to fix the scaling factor}
    Estimate $\beta_i(f)$ with (18)
    $\lambda(f)$ is given by (19) where $a_{i,1}(f)/a_{1,1}(f)$ is replaced by $\hat{\beta}_i(f)/\hat{\beta}_1(f)$
    {Estimation of the demixing filter}
    $\mathbf{G}_{1,:}(f) \Leftarrow \lambda(f)\mathbf{b}_{1,:}(f)$
end for
Perform inverse Fourier transform of $\mathbf{G}_{1,:}(f)$ to obtain $\mathbf{G}_{1,:}(k)$
Estimate source $s_1(k)$ thanks to (15)

4. Numerical experiments

In this section, we first present the results about the V-VAD and next the results of the geometric separation. All these experiments were performed using real speech/acoustic signals. The audio–visual corpus denoted C_1 used for the source to be extracted, say $s_1(k)$, consists of spontaneous male speech recorded in dialog condition (Section 2). Two others corpus, denoted C_2 and C_3 respectively, consist of phonetically well-balanced sentences in French of a

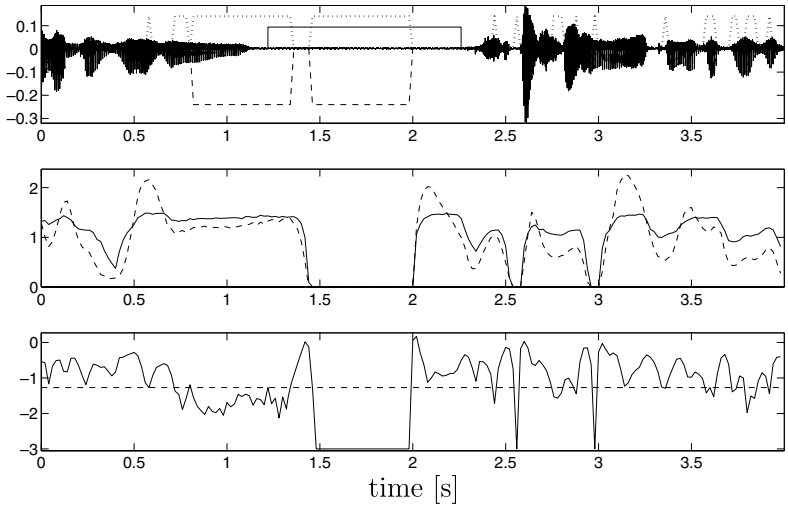

Fig. 5. Silence detection. Top: acoustic speech signal with silence reference (solid line), frames detected as silence (dotted line) and frames eventually retained as silence when $L = 20$ consecutive silence frames (dashed line). Middle: static visual parameters $\pi_1(k)$ (solid line) and $\pi_2(k)$ (dashed line). Bottom: logarithm of the dynamical visual parameter $V(k)$ with $\alpha = 0.82$ (solid line, truncated at -3) and the logarithm of the threshold δ (dashed line).

different male speaker and of acoustic noise recorded in a train, respectively.

4.1. Visual voice activity detector results

We tested the proposed V-VAD on about 13200 20 ms-frames extracted from corpus C_1, representing about 4.4 min of spontaneous speech. First of all, Fig. 5 illustrates the different possible relations between visual and acoustic data: (i) movement of the lips in non-silence (e.g. for time index $k \in [3s, 4s]$), (ii) movement of the lips in silence (e.g. for time index $k \in [2s, 2.3s]$), (iii) non-movement of the lips in silence (e.g. for time index $k \in [1.5s, 2s]$), (iv) non-movement of the lips in non-silence (e.g. for time index $k \in [0.9s, 1.1s]$).

The detection results of the proposed V-VAD are presented as receiver operating characteristics (ROC) (Fig. 6). These curves present the percentage of silence detection (*i.e.* ratio between the number of actual silence frames detected as silence frames and the number of actual silence frames) versus the percentage of false silence detection (*i.e.* ratio between the number of actual non-silence frames detected as silence frames and the number of actual silence frames). Fig. 6a highlights the importance of the summation by a low-pass filter of the video parameter $v(k)$ (1). Indeed, by lessening the influence of short movement of the lips in silence and the influence of the short static lips in speech, the summation (2) improves the performance of the V-VAD: the false silence detection significantly decreases for a given silence detection percentage (e.g. for 80% of correct silence detection, the false silence detection decreases from 20% to 5% with a correct integration). Furthermore, Fig. 6b shows the effect of the post-processing for the unfiltered version of the video parameter $v(k)$. The ROC curves show that a too large duration ($L = 200$ frames corresponding to 4 s) leads to a dramatical decrease in the silence detection ratio. On the contrary, a reasonable duration ($L = 20$ frames corresponding to 400 ms) allows the false silence detection ratio to be reduced without decreasing the silence detection ratio in comparison to the case of no post-processing (*i.e.* $L = 1$ frame). The gain due to post-processing is similar to the gain due to the summation. Eventually, combination of both summation and post-processing leads to a quite robust and reliable V-VAD.

4.2. Separation results

In this subsection, we consider the case of sources mixed by matrices of filters. These filters are finite impulse response (FIR) filters of 320 lags with three significant echoes. They are truncated versions of measured impulse responses recorded in a real room (Pham et al., 2003). All the acoustic signals are sampled at 16 kHz, while the video signal is sampled (synchronously) at 50 Hz. Different configurations of mixing matrices were tested in the case of N sources and N observations denoted ($N \times N$): (2×2) and (3×3). The three corpuses C_1, C_2 and C_3 are used as $s_1(k)$, $s_2(k)$ and $s_3(k)$ respectively.

To compute the STFT, the signals are subdivided into blocks of 320 samples[4] (*i.e.* 20 ms frames). Then, a fast Fourier transform is applied on each block using the zero-padding up to 2048 samples. The length of the separating filters is thus 2048 samples. The blocks are overlapped about 0.85 of the block size.

To evaluate the performance of the proposed geometric method, we use different indexes. Since we are only

[4] Note that the FFT length is here adapted to the size of the mixing filters. Because of potential cyclic convolution effects, longer mixing filters may require a refined processing: e.g. in future work, the FFT-size can be increased according to the length of detected silences.

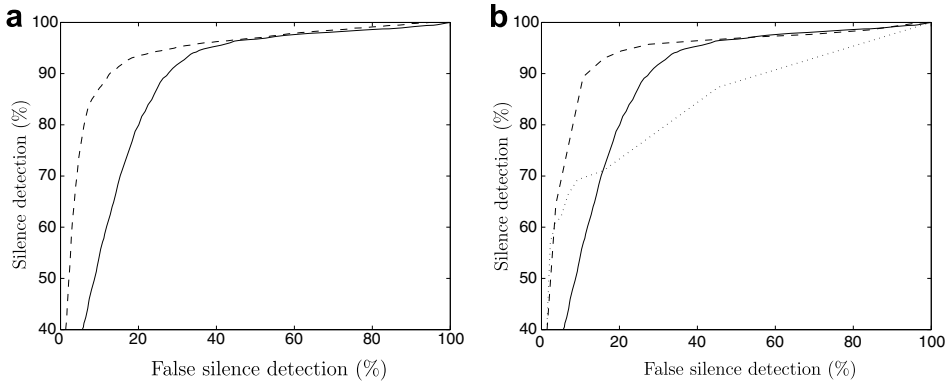

Fig. 6. ROC silence detection curves. (a) ROC curves with two summation coefficients of the visual parameter $V(k)$: instantaneous ($\alpha = 0$, solid line) and suitable summation ($\alpha = 0.82$, dashed line). (b) ROC curves with L consecutive silence frames, in solid line $L = 1$ (*i.e.* instantaneous), in dashed line $L = 20$ frames (400 ms) and in dotted line $L = 200$ frames (4 s).

interested in extracting $s_1(k)$, we define first the performance index $r_1(f)$ for the source $s_1(k)$ as

$$r_1(f) = \sum_{j=1}^{N} \frac{|\mathcal{GH}_{1,j}(f)|}{|\mathcal{GH}_{1,1}(f)|} - 1, \tag{22}$$

where $\mathcal{GH}(f) = \mathcal{G}(f)\mathcal{H}(f)$ is the global system. This index quantifies the quality of the estimated separating matrices $\mathcal{G}(f)$ for the source of interest. For a good separation, this index should be close to zero.

We now define the contribution of a signal $y(k)$ in another signal $z(k)$. In a general way, we can decompose $z(k)$ such that $z(k) = f(y(k)) + n(k)$, where $f(\cdot)$ is a function. In the following $(y\text{—}z)(k)$ denotes the contribution of $y(k)$ in $z(k)$: $(y\text{—}z)(k) = f(y(k))$. Moreover, let denote $\mathcal{P}_y = \frac{1}{K}\sum_{k=1}^{K}|y(k)|^2$ the average power of signal $y(k)$.

Thus, the signal to interference ratio (SIR) for the first source is defined as

$$\mathrm{SIR}_{(s_1|s_1^\dagger)} = \frac{\mathcal{P}_{(s_1|s_1^\dagger)}}{\sum_{s_j \neq s_1} \mathcal{P}_{(s_j|s_1^\dagger)}}. \tag{23}$$

Note that $(s_j|s_1^\dagger)(k) = \sum_{i=1}^{N} g_{1,i}(k) * h_{i,j}(k) * s_j(k)$. This classical index in source separation quantifies the quality of the estimated source $s_1^\dagger(k)$. For a good estimation of the source (*i.e.* $\forall j > 1$, $(s_j|s_1^\dagger)(k) \simeq 0$), this index should be close to infinity. Finally, we define the gain of the first source due to the separation process as

$$G_1 = \frac{\mathrm{SIR}_{(s_1|s_1^\dagger)}}{\max_l \mathrm{SIR}_{(s_1|x_l)}} = \min_l \frac{\mathrm{SIR}_{(s_1|s_1^\dagger)}}{\mathrm{SIR}_{(s_1|x_l)}} \tag{24}$$

with $\mathrm{SIR}_{(\cdot|\cdot)}$ defined by (23) and $(s_j|x_l)(k) = h_{l,j}(k) * s_j(k)$. This gain allows the improvement in SIR before and after the separation process to be quantified. (The reference before separation being taken in the mixture where the contribution of $s_1(k)$ is the strongest.)

Fig. 7 presents a typical result of the separation process in the case of two sources and two sensors (2×2) with an approximate SIR for source $s_1(k)$ equal to 0 dB for both sensors. The two speech sources are plotted in Fig. 7a and b. One can see in Fig. 7c $\ln V_1(k)$, the natural logarithm of video parameter $V(k)$ (continuous line) and $\ln \delta$ the natural logarithm of the threshold $\delta = 0.015$ (dash–dot line) used to estimate the silence frames (Section 2). In this example, and after the results of Section 2, the smoothing coefficient α is set to 0.82 and the minimal number of consecutive silence frames L is chosen equal to 20 (*i.e.* the minimum length of a detected silence is 400 ms). The results of the V-VAD can be seen on Fig. 7a: the frames manually indexed as silence are represented by the dash–dot line and the detected frames as silence (which define the estimation of $\mathcal{T}$) are represented by the dashed line. In this example 154 frames (*i.e.* 3.08 s) were detected as silence representing 63.2% of silence detection while the false detection rate is only 1.3%. The two mixtures obtained from the two sources are plotted in Fig. 7d and f. The result $s_1^\dagger(k)$ of the extraction of source $s_1(k)$ by the proposed method is shown in Fig. 7g. As a reference, the best possible estimation of the first source $(h_{1,1}(k) * s_1(k))$ is plotted in Fig. 7e. In this example, the gain G_1 is equal to 17.4 dB, while $\mathrm{SIR}_{(s_1|s_1^\dagger)} = 18.4$ dB. One can see that the extraction of the first source is quite well performed. This is confirmed by the index performance $r_1(f)$ (Fig. 7h): most values are close to zero.

Beyond this typical example, we processed extensive simulation tests. For each simulation, only 20 s of signals were used. Each configuration of the mixing matrices ($N \times N$) and of the $\mathrm{SIR}_{\mathrm{in}}$ (where $\mathrm{SIR}_{\mathrm{in}}$ is the mean of the SIRs for each mixtures: $\mathrm{SIR}_{\mathrm{in}} = 1/N \times \sum_l \mathrm{SIR}_{(s_1|x_l)}$) was run 50 times and the presented results are given on average. To synthesize different source signals, each speech/acoustic signal is shifted randomly in time. Fig. 8 presents the gain G_1 versus the $\mathrm{SIR}_{\mathrm{in}}$ in both (2×2) and (3×3) cases. One can see that in both cases, the shape of the gains is the same: from low $\mathrm{SIR}_{\mathrm{in}}$ (−20 dB) to high $\mathrm{SIR}_{\mathrm{in}}$ (20 dB), the gains are almost constant at a high value, demonstrating the efficiency of the proposed method: we obtain gains of about 19 dB in the (2×2) case and about 18 dB in the (3×3) case. Then, the gains decrease to 11 dB in the (2×2) case and to 8 dB in the (3×3) case for higher $\mathrm{SIR}_{\mathrm{in}}$.

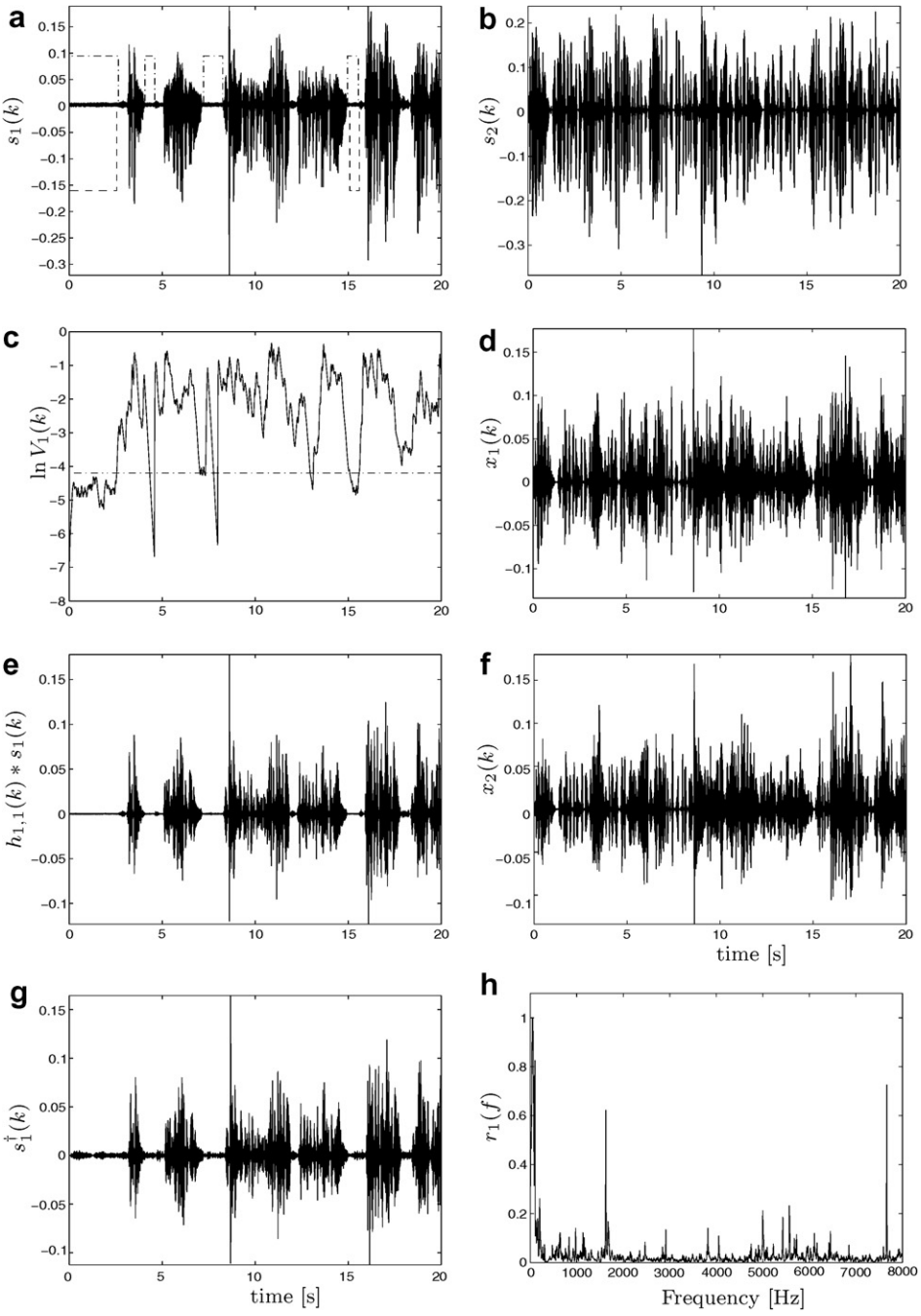

Fig. 7. Example of separation in the (2×2) case. (a) and (b) Sources $s_1(k)$ and $s_2(k)$ respectively. (c) Natural logarithm $\ln V_1(k)$ of the video parameter associated with the first source $s_1(k)$. (d) and (f) Mixtures $x_1(k)$ and $x_2(k)$ respectively. (e) First source $s_1(k)$ up to the filter $h_{1,1}(k)$ (*i.e.* $h_{1,1}(k) * s_1(k)$); tjis signal is used as a reference for the estimation of source $s_1(k)$ by our method. (g) Estimation $s_1^\dagger(k)$ of source $s_1(k)$ given by (21). (h) Performance index $r_1(f)$ (truncated at 1).

It is interesting to note that the gain can happen to be negative for the highest $\mathrm{SIR}_{\mathrm{in}}$ (e.g. (2×2) with $\mathrm{SIR}_{\mathrm{in}} = 20$ dB). However, this is a rare situation since it happens only for isolated high ratio FA/BD (e.g. $FA/BD > 15\%$), *i.e.* when the set $\mathscr{T}$ of detected silence frames contains too many frames for which $s_1(k)$ is active. Indeed in this case, a deeper analysis shows us that, since the average power of $s_1(k)$ is larger than the average power of the other source(s) for high $\mathrm{SIR}_{\mathrm{in}}$, the smallest eigenvalue of the covariance matrix estimated using this set $\mathscr{T}$ is not necessary associated with the first source. On the contrary, even if the ratio FA/BD is high while the $\mathrm{SIR}_{\mathrm{in}}$ is low, the influence of interfering $s_1(k)$ values occurred during false alarms remains poor because these values are small compared to the other source(s). Altogether, the method is efficient and reliable for a large range of $\mathrm{SIR}_{\mathrm{in}}$ values (note that despite the previous remark, the smallest gain obtained at $\mathrm{SIR}_{\mathrm{in}} = 20$ dB for (2×2) mixture is about 11 dB on the average).

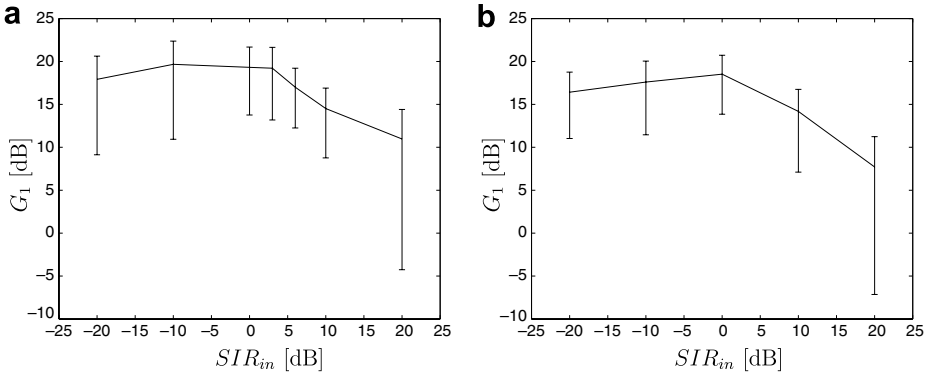

Fig. 8. Expected gain G_1 versus $\mathrm{SIR}_{\mathrm{in}}$ in the (2×2) (a) and (3×3) (b) cases, respectively. The curves show the mean and the standard deviation of G_1 in dB.

5. Conclusion

In this paper, we proposed an efficient method which exploits the complementarity of the bi-modality (audio/visual) of the speech signal. The visual modality is used as a V-VAD, while the audio modality is used to estimate the separation filter matrices exploiting detected silences of the source to be extracted. The proposed V-VAD has a major interest compared to audio only VAD: it is robust in any acoustic noise environment (e.g. in very low signal to noise ratio cases, in highly non-stationary environments with possibly multiple interfering sources, etc.). Moreover, the proposed geometric separation process is based on the sparseness of the speech signal: when the source to be extracted is vanishing (*i.e.* during the silence frames given by the proposed V-VAD), the power of the corresponding estimated source is minimized thanks to the separating filters. The proposed method can be easily extended to extract other/any sparse sources, using associated vanishing oracle. Note that results were presented for (2×2) and (3×3) mixtures but the method can be applied to any arbitrary $(N\times N)$ mixture with $N>3$. Also, compared to other frequency separation methods (Rivet et al., 2007; Capdevielle et al., 1995; Parra and Spence, 2000; Dapena et al., 2001; Pham et al., 2003), the method has the strong advantage to intrinsically regularize the permutation problem: this regularization is an inherent byproduct of the ''smallest eigenvalue'' search. Finally, we can conclude by underlining the low complexity of the method and low associated computation cost, the video parameter extraction being set apart: compared to the methods based on joint diagonalization of several matrices (Rivet et al., 2007; Pham et al., 2003), the proposed method requires the simple diagonalization of one single covariance matrix.

Future works will mainly focus on the extraction of useful visual speech information in more natural conditions (e.g. lips without make-up, moving speaker). Also, we intend to develop an on-line version of the geometric algorithm. These points are expected to allow the implementation of the proposed method for real environment and real-time applications.

References

Abrard, F., Deville, Y., 2005. A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. Signal Processing 85 (7), 1389–1403.

Babaie-Zadeh, M., Mansour, A., Jutten, C., Marvasti, F., 2004. A geometric approach for separating several speech signals. In: Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA), Granada, Spain, pp. 798–806.

Bernstein, L.E., Benoît, C., 1996. For speech perception by humans or machines, three senses are better than one. In: Proc. Int. Conf. Spoken Language Processing (ICSLP), Philadelphia, USA, pp. 1477–1480.

Capdevielle, V., Servière, C., Lacoume, J.-L., 1995. Blind separation of wide-band sources in the frequency domain. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Detroit, USA, 1995, pp. 2080–2083.

Cardoso, J.-F., 1998. Blind signal separation: statistical principles. Proceedings of the IEEE 86 (10), 2009–2025.

Dansereau, R., 2004. Co-channel audiovisual speech separation using spectral matching constraints. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Montréal, Canada, 2004.

Dapena, A., Bugallo, M.F., Castedo, L., 2001. Separation of convolutive mixtures of temporally-white signals: a novel frequency-domain approach. In: Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA), San Diego, USA, pp. 315–320.

Elisei, F., Odisio, M., Bailly, G., Badin, P., 2001. Creating and controlling video-realistic talking heads. In: Proc. Audio-Visual Speech Processing Workshop (AVSP), Aalborg, Denmark, pp. 90–97.

Girin, L., Schwartz, J.-L., Feng, G., 2001. Audio–visual enhancement of speech in noise. J. Acoust. Soc. Am. 109 (6), 3007–3020.

Lallouache, T., 1990. Un poste visage-parole. Acquisition et traitement des contours labiaux, in: Proc. Journées d'Etude sur la Parole (JEP) (French), Montréal.

Le Goff, B., Guiard-Marigny, T., Benoît, C., 1995. Read my lips... and my jaw! How intelligible are the components of a speaker's face? In: Proc. Euro. Conf. on Speech Com. and Tech, Madrid, Spain, pp. 291–294.

Liu, P., Wang, Z., 2004. Voice activity detection using visual information. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Montreal, 2004, pp. 609–612.

Parra, L., Spence, C., 2000. Convolutive blind separation of non stationary sources. IEEE Trans. Speech Audio Process. 8 (3), 320–327.

Pham, D.-T., Servière, C., Boumaraf, H., 2003. Blind separation of convolutive audio mixtures using nonstationarity. In: Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA), Nara, Japan.

Rajaram, S., Nefian, A.V., Huang, T.S., 2004. Bayesian separation of audio–visual speech sources. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Montréal, Canada.

Rivet, B., Girin, L., Jutten, C., 2007. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. IEEE Trans. Audio Speech Lang. Process. 15 (1), 96–108.

Sodoyer, D., Girin, L., Jutten, C., Schwartz, J.-L., 2004. Developing an audio–visual speech source separation algorithm. Speech Comm. 44 (1–4), 113–125.

Sodoyer, D., Rivet, B., Girin, L., Schwartz, J.-L., Jutten, C., 2006. An analysis of visual speech information applied to voice activity detection. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, pp. 601–604.

Sumby, W., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Wang, W., Cosker, D., Hicks, Y., Sanei, S., Chambers, J.A., 2005. Video assisted speech source separation, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, USA, 2005.