Audiovisual speech source separation: a regularization method based on visual voice activity detection

Bertrand Rivet[1,2], *Laurent Girin*[1], *Christine Servière*[2], *Dinh-Tuan Pham*[3], *Christian Jutten*[2]

[1,2]Grenoble Image Parole Signal Automatique (GIPSA - [1]ICP/[2]LIS)
CNRS UMR 5216 , Grenoble Institute of Technology (INPG), Grenoble, France
emails: {rivet,girin}@icp.inpg.fr, {rivet,serviere,jutten}@lis.inpg.fr
[3] Laboratoire Jean Kuntzmann)
CNRS UMR 5524, Grenoble Institute of Technology (INPG), Université Joseph Fourier, Grenoble, France
email: Dinh-Tuan.Pham@imag.fr

Abstract

Audio-visual speech source separation consists in mixing visual speech processing techniques (e.g. lip parameters tracking) with source separation methods to improve and/or simplify the extraction of a speech signal from a mixture of acoustic signals. In this paper, we present a new approach to this problem: visual information is used here as a voice activity detector (VAD). Results show that, in the difficult case of realistic convolutive mixtures, the classic problem of the permutation of the output frequency channels can be solved using the visual information with a simpler processing than when using only audio information.

Index Terms: blind source separation, convolutive mixtures, visual voice activity detection, audiovisual speech

1. Introduction

Blind source separation (BSS) consists in retrieving source signals from mixtures of them, without any knowledge on the mixing nature, or on the sources themselves. As far as speech signals are concerned, the separation is no more completely blind since speech signals have specific properties that can be exploited in the separation process. For instance, non-stationarity of speech has been exploited in [1, 2]. However, accurate separation is still a difficult task, notably in the case where less sensors than sources are available, and also because of the permutation and scale factor indeterminacies: output signals can only be reconstructed up to a gain and a permutation on the output channels [3].

Audiovisual (AV) speech source separation is an attractive field to solve the source separation problem when speech signals are involved (e.g. [4, 5, 6]). It consists in exploiting the (audio-visual) bi-modality of speech, especially the speaker's lip movements, to improve and/or simplify the performance of acoustic speech source separation. For instance, Sodoyer *et al.* [4], and then Wang *et al.* [5] and Rivet *et al.* [6] have proposed to use a statistical model of the coherence of audio and visual speech features to extract a speech source in the case of instantaneous and convolutive mixtures respectively.

In this paper, we propose a new different and simpler but even so efficient approach for the permutation problem. We propose to use the visual speech information of a speaker as a voice activity detector (VAD): the task is to assess the presence or the absence of the speaker in the mixture. Such information allows the extraction of the particular (filmed) speaker from the mixture thanks to a very simple proposed method.

This paper is organized as follows. Section 2 presents the basis of the proposed visual VAD (V-VAD). Section 3 recalls the principle of source separation in the frequency domain for convolutive mixtures and explains how the V-VAD corresponding to a particular speaker can be useful to solve the permutation ambiguity for this speaker. Section 4 presents numerical experiments.

2. Visual voice activity detection

The visual voice activity detector (V-VAD) that we combine in this study with source separation, has been described in details in [7]. We thus give here a succinct description. The main idea of this V-VAD is that during speech, lips are generally moving whereas they are not moving (so much) during silences. So we use the video parameter

$$v(m) = \left|\frac{\partial A(m)}{\partial m}\right| + \left|\frac{\partial B(m)}{\partial m}\right| \quad (1)$$

where $A(m)$ (resp. $B(m)$) is the speaker's lip contour internal width (resp. height). Such parameters are automatically extracted every 20ms (a speech *frame* length) synchronously with the audio signal (sampled at 16kHz) by using the "face processing system" of the GIPSA/ICP laboratory [8]. To improve the silence detection, we smooth $v(m)$ over T consecutive frames

$$V(m) = \sum_{l=0}^{T-1} a^l \, v(m-l), \quad (2)$$

where $a = 0.82$. The m-th input frame is then classified as silence if $V(m)$ is lower than a threshold δ and it is classified as speech otherwise. As explained in Section 3, the aim of the V-VAD is to actually detect *silences*, *i.e.* frames where the speaker do *not* produce sounds. Therefore, to decrease the false alarm (silence decision while speech activity) rate, only sequences of at least $L = 20$ frames (*i.e.* 400ms) of silence are actually considered as silences [7]. This leads to 80% of good detection for only 15% of false alarms. Finally, the proposed V-VAD is robust to any acoustic noise, even in highly non-stationary environment, whatever the nature and the number of competing sources.

3. BSS with visual VAD

In this section, we first briefly present the general framework of BSS for convolutive mixtures and then we explain how the V-VAD can solve the permutation problem.

3.1. BSS of convolutive mixtures

Let us consider N sources $\mathbf{s}(m) = [s_1(m), \cdots, s_N(m)]^T$ (^{T} denoting the transpose) to be separated from P observations $\mathbf{x}(m) = [x_1(m), \cdots, x_P(m)]^T$ defined by $x_p(m) = \sum_{n=1}^{N} h_{p,n}(m) * s_n(m)$. The filters $h_{p,n}(m)$ that model the impulse response between $s_n(m)$ and the p-th sensor are entries of the mixing filter matrix $H(m)$. The goal of the BSS is to recover the sources by using a dual filtering process: $\hat{s}_n(m) = \sum_{p=1}^{P} g_{n,p}(m) * x_p(m)$ where $g_{n,p}(m)$ are entries of the demixing filter matrix $G(m)$ which are estimated such that the components of the output vectors (the estimated sources) $\hat{\mathbf{s}}(m) = [\hat{s}_1(m), \cdots, \hat{s}_N(m)]^T$ are as mutually independent as possible. This problem is generally considered in the frequency domain (e.g. [1, 2]) where we have

$$X_p(m,f) = \sum_{n=1}^{N} H_{p,n}(f) S_n(m,f) \quad (3)$$

$$\hat{S}_n(m,f) = \sum_{p=1}^{P} G_{n,p}(f) X_p(m,f) \quad (4)$$

where $S_n(m,f)$, $X_p(m,f)$ and $\hat{S}_n(m,f)$ are the Short-Term Fourier Transforms (STFT) of $s_n(m)$, $x_p(m)$ and $\hat{s}_n(m)$ respectively. $H_{p,n}(f)$ and $G_{n,p}(f)$ are the frequency responses of the mixing and demixing filters respectively. From (3) and (4), basic algebra manipulation leads to

$$\Gamma_x(m,f) = H(f) \Gamma_s(m,f) H^H(f) \quad (5)$$

$$\Gamma_{\hat{s}}(m,f) = G(f) \Gamma_x(m,f) G^H(f) \quad (6)$$

where $\Gamma_y(m,f)$ denotes the time-varying power spectrum density (PSD) matrices of a signal vector $\mathbf{y}(m)$. $H(f)$ and $G(f)$ are the frequency response matrices of the mixing and demixing filter matrices (^{H} denotes the conjugate transpose).

If the sources are assumed to be mutually independent (or at least decorrelated), $\Gamma_s(m,f)$ is diagonal and an efficient separation must lead to a diagonal matrix $\Gamma_{\hat{s}}(m,f)$. A basic criterion for BSS [2] is to calculate $\Gamma_x(m,f)$ from the observations and adjust the matrix $G(f)$ so that $\Gamma_{\hat{s}}(m,f)$ is as diagonal as possible. Since this condition must be verified for any time index m, this can be done by a joint diagonalization method (*i.e.* best approximate simultaneous diagonalization of several matrices), and in the following we use the algorithm of [9].

3.2. Canceling the permutation indeterminacy

The well-known crucial limitation of the BSS problem is that for each frequency bin, $G(f)$ can only be provided up to a scale factor and a permutation between the sources:

$$G(f) = P(f)\, D(f)\, \hat{H}^{-1}(f), \quad (7)$$

where $P(f)$ and $D(f)$ are arbitrary permutation and diagonal matrices. Several audio approaches to the permutation indeterminacy were proposed (e.g. [1, 2, 10]). In [6], we proposed to use a statistical model of the coherence of visual and acoustic speech features to cancel the permutation and scale factor indeterminacies of audio separation. Although effective, the method had the drawbacks to require an off-line training and to be computationally expensive.

In this new study, we simplify this approach by directly exploiting the V-VAD focusing on the lips of a specific speaker. The audiovisual model of [6] is replaced by the (purely visual) V-VAD of Section 2 and the detection of the *absence* of a source allows to solve the permutation problem for that peculiar source when this source is *present* in the mixtures. Indeed, at each frequency bin f, the separation process (Subsection 3.1) provides a separating matrix $G(f)$ which leads to a diagonal PSD matrix $\Gamma_{\hat{s}}(m,f)$ of the estimated sources. The k-th diagonal element of $\Gamma_{\hat{s}}(m,f)$ is the spectral energy of the k-th estimated source at frequency bin f and time m. The logarithm of $\Gamma_{\hat{s}}(m,f)$ is called here a *profile* and is denoted $E(f,m;k)$:

$$E(f,m;k) = \log\left(\Gamma_{\hat{s}}(m,f)\right)_{k,k}, \quad (8)$$

where $\left(\Gamma_{\hat{s}}(m,f)\right)_{k,k}$ is the k-th diagonal element of $\Gamma_{\hat{s}}(m,f)$. Let denote $\mathcal{T}$ the set of all time indexes. The V-VAD associated with a particular source, say $s_1(m)$, provides the set of time indexes $\mathcal{T}_1$ when this source vanishes ($\mathcal{T}_1 \subset \mathcal{T}$). Then the profile $E(f,m;\cdot)$, with $m \in \mathcal{T}_1$, corresponding to the estimation of $s_1(m)$ must be close to $-\infty$. Therefore, at the output of the joint diagonalization algorithm, we compute centered profiles $E_{\mathcal{T}_1}(f;k)$ calculated during $s_1(m)$ absence detection $m \in \mathcal{T}_1$:

$$E_{\mathcal{T}_1}(f;k) = \frac{1}{|\mathcal{T}_1|} \sum_{m \in \mathcal{T}_1} E(f,m;k) - \frac{1}{|\mathcal{T}|} \sum_{m \in \mathcal{T}} E(f,m;k) \quad (9)$$

where $|\mathcal{T}_1|$ is the cardinal number of the set $\mathcal{T}_1$. Note that since each source can only be estimated up to a gain factor, the profiles are defined up to an additive constant. Hence by centering all profiles (by subtracting their time average) this additive constant is eliminated. Then, based on the fact that the centered profile $E_{\mathcal{T}_1}(f;\cdot)$ corresponding to $s_1(m)$ must tend toward $-\infty$, for all frequencies f, we search for the smallest centered profile. Finally, we set $P(f)$ so that this smallest centered profile corresponds to $E_{\mathcal{T}_1}(f;1)$. Applying this set of permutation matrices $P(f)$ to the demixing matrices $G(f)$ for all time indexes $\mathcal{T}$ (*i.e.* including the ones where $s_1(m)$ is present) allows to reconstruct $s_1(m)$ without frequency permutations when it is present in the mixtures.

Note that, the proposed scheme enables to solve frequency permutations for a given source if it has an associated V-VAD for absence detection, but frequency permutations can remain on the other sources without consequences for the extraction of $s_1(m)$. To extract more than one source, it is necessary to have additional corresponding detectors and to apply the same method.

4. Numerical experiments

In this section, we consider two sources mixed by 2×2 matrices of FIR filters of 512 lags with three significant echoes, which are truncated impulse responses measured in a real 3.5m×7m×3m conference room[1]. The source to be extracted, say $s_1(m)$, consists of spontaneous male speech recorded in dialog condition. The second source consists of continuous speech produced by another male speaker. In each experiments, ten seconds of signals, randomly chosen from the two databases, were mixed and then used to estimate separating filters of 4096 lags (thus it is the size of all STFTs).

[1] They can be found at http://sound.medi.mit.edu/ica-bench.

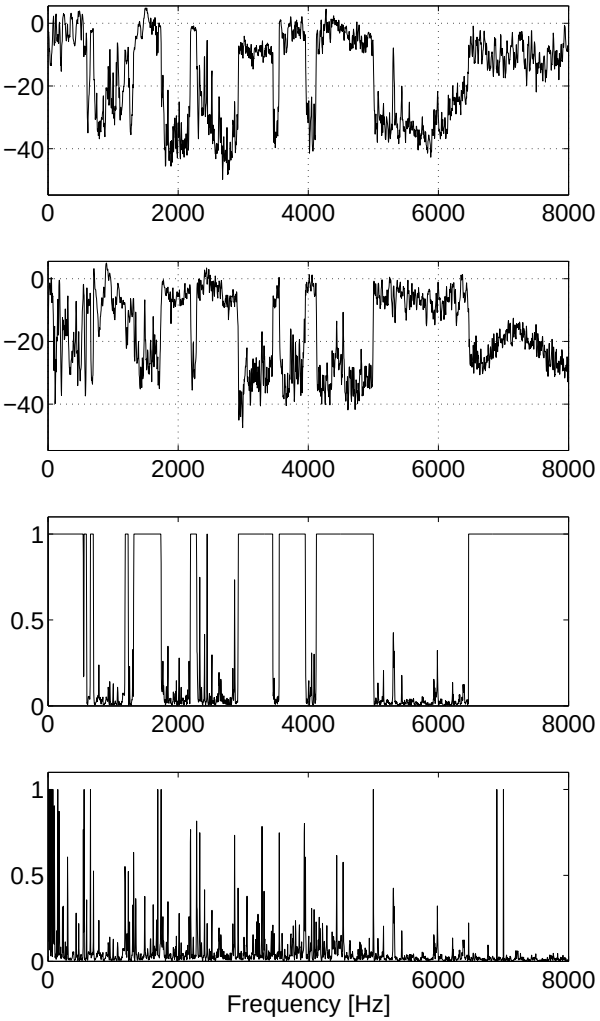


Figure 1: Permutation estimation. From top to bottom: centered profiles $E_{\mathcal{T}_1}(f;1)$ and $E_{\mathcal{T}_1}(f;2)$ before permutation cancellation; performance index $r_1(f)$ (truncated at 1) before and after permutation cancellation respectively.

Since we are only interested in extracting $s_1(m)$ we define a performance index as

$$r_1(f) = |GH_{12}(f)/GH_{11}(f)|\,, \tag{10}$$

where $GH_{i,j}(f)$ is the (i,j)-th element of the global system

$$GH(f) = G(f)\,H(f). \tag{11}$$

For a good separation, this index should be close to 0, or close to infinity if a permutation has occurred: the performance index is thus also an efficient flag to detect if a permutation has occurred.

First, we present performance of the proposed permutation cancellation method (Fig. 2 and Fig. 1). In a real life application context the mixing filters are unknown, so it is impossible to compute the performance index $r_1(f)$. However, one can see (Fig. 1) that the proposed centered profiles (9) are very correlated with the performance index $r_1(f)$, leading to a simple and efficient estimation of $r_1(f)$. Finally, let denote $(P_1/P_2)_{\mathcal{T}_1}$ the ratio of the averaged powers P_1 and P_2 of the two sources s_1 and s_2 respectively during time indexes $\mathcal{T}_1$ (the silence of s_1).

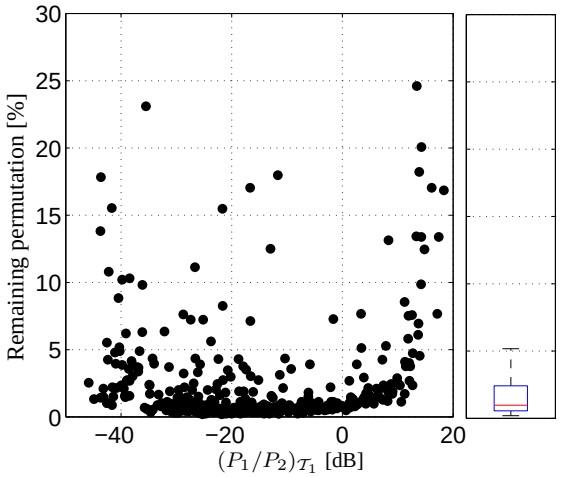


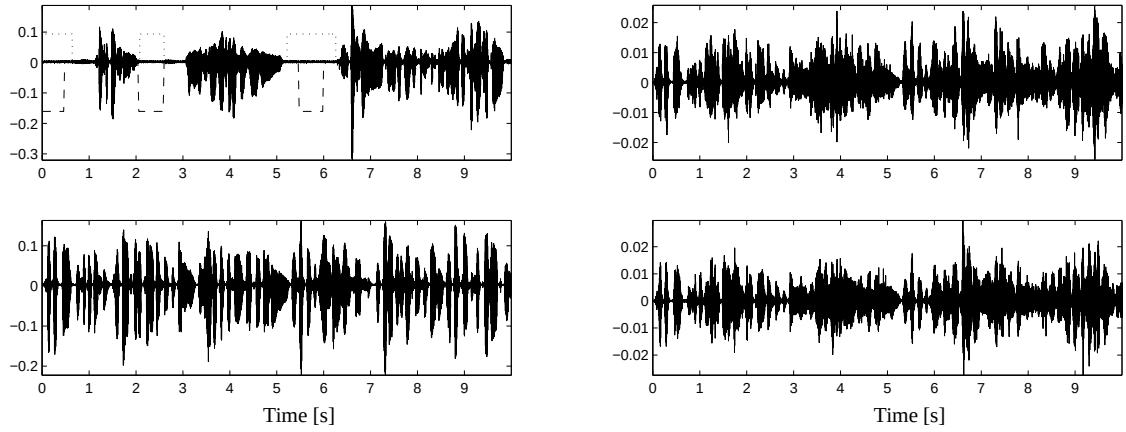
Figure 2: Percentage of remaining permutation versus ratio $(P_1/P_2)_{\mathcal{T}_1}$ (On the right: repartition of the 400 results)
.

The proposed permutation cancellation method performs quite well as shown in Fig. 2 which plot the percentage of remaining permutations versus the ratio $(P_1/P_2)_{\mathcal{T}_1}$. Indeed, 75% of the 400 tested situations leads to less than 2.4% of remaining permutations (2.4% is the median value) and the good detection rate increased to 89% for only 5% of remaining permutations. However, one can see that the residual permutations correspond to isolated permutations (Fig. 1 bottom) which are shown to have minor influence on the separation quality: they are generally assumed to correspond to spectral bins with both sources of low energy.

Our system was compared to the baseline frequency domain ICA without permutation cancellation as well as to an audio-based permutation cancellation system [2]. In this example, the two sources (resp. the two mixtures) are plotted in Fig. 3(a) (resp. in Fig. 3(b)). In this example, the dotted line represents a manual indexation of silence and the dashed line represents the automatic detection obtained by the V-VAD, which is quite good (see more detailed results in [7]). In the first experiment (Fig. 3(c)), the source s_1 is estimated by the baseline frequency domain ICA without permutation cancellation. One can see on the global filter (Fig. 3(c)-right) the consequences of unsolved permutations: $(G*H)_{1,1}(n)$ is not significantly larger than $(G*H)_{1,2}(n)$, so the estimation of s_1 is quite poor (Fig. 3(c)-left). In the second experiment (Fig. 3(d)), the source s_1 is estimated by the baseline frequency domain ICA with an audio-based permutation cancellation system [2] followed by a manual selection of $\hat{s}_1$ among the two estimated sources. In the last experiment (Fig. 3(e)), the source s_1 is estimated by the baseline frequency domain ICA with the proposed audiovisual permutation cancellation system. In these two experiments, one can see that the sources are well estimated ($(G*H)_{1,1}(n)$ is much larger than $(G*H)_{1,2}(n)$) and very close source estimations are obtained.

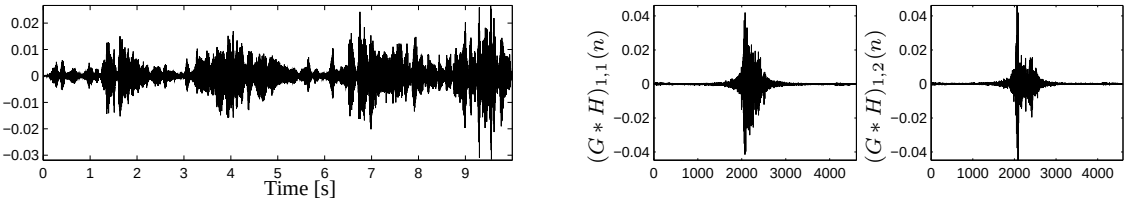
5. Conclusion

The proposed combined audiovisual method provides a very simple scheme to solve the permutations of a baseline frequency domain ICA. Indeed, given the time indexes of absence of a pe-

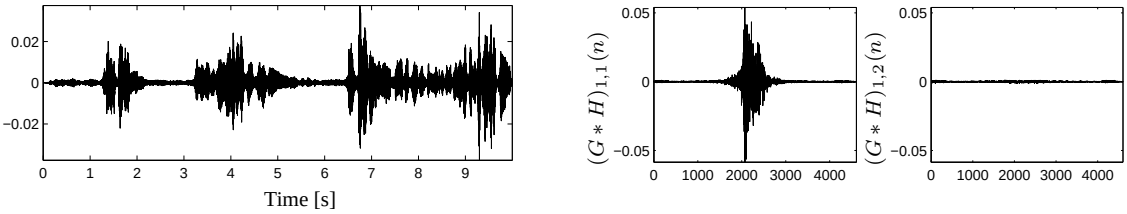


(a) Original sources

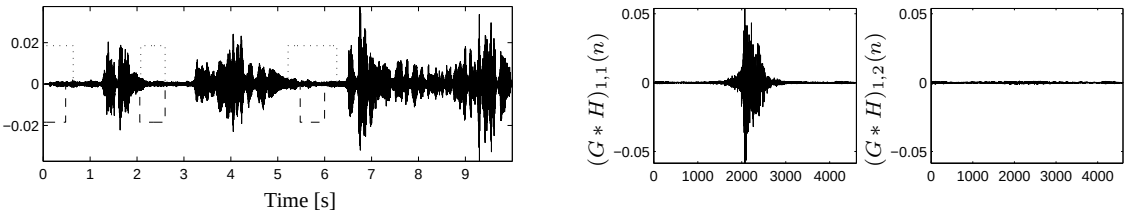
(b) Mixtures



(c) Estimation of s_1 by the baseline frequency domain ICA without permutation cancellation



(d) Estimation of s_1 by the baseline frequency domain ICA with an audio-based permutation cancellation system [2]



(e) Estimation of s_1 by the baseline frequency domain ICA with the proposed audiovisual permutation cancellation system

Figure 3: Illustration of the extraction of s_1 from mixtures using different systems.

culiar source provided by the visual voice activity detection, it is simple to solve the permutation corresponding of this source thanks to the proposed centered profiles.

Beyond the presented example, the proposed combined audiovisual method was tested on several experimental mixture conditions (e.g. nature of competing sources, length of the mixing filters, *etc*.) and yields very good source extraction. This method has three major advantages compared to a purely audio approach (e.g. [2]): (i) it is computationally much simpler (given that the video information is available), especially when more than two sources are involved; (ii) the visual proposed method implicitly extracts the estimated sour-ce corresponding to a filmed speaker, while purely audio regularization provides the estimated sources in an arbitrary order (*i.e.* up to a *global* unknown permutation of the regularized sources across speakers); (iii) more generally the visual approach to voice activity detection [7] is robust to any acoustic environment (unlike a purely audio voice activity detection).

In this work, all processes were made off-line, that is to say on a large section of signals (about 10 seconds). Future

works concern a pseudo real-time version where the processes are updated on-line. Also, the use of visual parameters extracted from natural face processing in natural environment is currently being explored. All this will contribute to build a system usable in real life conditions.

6. References

[1] L. Parra and C. Spence, "Convolutive blind separation of non stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[2] C. Servière and D.-T. Pham, "A novel method for permutation correction in frequency-domain in blind separation of speech mixtures," in *Proc. ICA*, Granada, Spain, 2004, pp. 807–815.

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

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

[5] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video assisted speech source separation," in *Proc. ICASSP*, Philadelphia, USA, March 2005.

[6] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 1, pp. 96–108, January 2007.

[7] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 601–604.

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

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

[10] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency Domain Blind Source Separation for Many Speech Signals," in *Proc. ICA*, Granada, Spain, 2004, pp. 461–469.