When silence is gold

Bertrand Rivet, Christian Jutten

WHEN SILENCE IS GOLD

Bertrand Rivet and Christian Jutten

GIPSA-lab, CNRS UMR-5216, Grenoble University
email: firstname.lastname@gipsa-lab.grenoble-inp.fr

ABSTRACT

Audiovisual source separation is a fascinating approach to source extraction. Several algorithms have already been proposed for extracting speech sources from audio mixtures by exploiting audiovisual coherence. One of the main property of speech signals is that they are highly non-stationary: there are periods during which speakers do not produce sounds. In this work, the audiovisual coherence is used to estimate such silent periods which are then useful to extract corresponding speech signals.

1. INTRODUCTION

Have you ever wondered why it could be so difficult to hear the driver of a car from a back seat, particularly in a noisy environment? It is well known that the human brain is able to discriminate sounds: it can enhance the different audio sources so that the interest source(s) can be more easily extracted. However, if the environment noise becomes too loud, this faculty could be deteriorated. As a consequence, how can it be explained that in the same noisy conditions, you can understand what your neighbor says? The answer comes from the cognitive sciences: your brain is able to merge the heard signals and what you see, especially the motion of speech articulators as lips or tongue. Indeed, speech is multimodal [10, 32, 33]. The bimodal nature of speech is now a basic feature, both for the understanding of speech perception [33] and for the development of tools for human-machine communication [31, 22]. It is thus well known that the vision of a speaker's face influences what is perceived. One of the most famous example is the McGurk's effect [19] which is an audiovisual illusion: superimposing an audio stimulus [ba] with a video stimulus [ga] leads to perceive [da]. Indeed, most of people are able to lip read (e.g., [3, 10, 32]) even if they are not conscious of doing so: the recognition rate with audiovisual speech increases compared to audio speech only [10]. In addition, the vision of the speaker's face not only favors understanding but also enhances speech detection in a noisy environment [4, 13, 16]. All of these studies finally exploit the redundancies and complementarities between visual and audio modalities of speech. It seems actually intuitive that there are strong links between the motion of the speaker's lips and the speech signal.

On the other hand, blind source separation (BSS) is an exciting field of signal processing. It aims at recovering unknown sources from mixtures of them [7]. The lack of prior knowledge about the sources and the mixing process is generally overcame by strong assumptions. For instance, the mutual independence between the sources like in independent component analysis (ICA) e.g., [6, 5] or the sparsity of the signals (sparse component analysis, e.g., [14]). Another additional prior information, close to the sparsity, is the inactivity of some sources in some time windows [24, 1, 9]. These inactivity periods can be silences in the case of speech signals.

In the last decade, these two fields of signal processing (BSS and audiovisual speech processing) have merged into the audiovisual source separation problem which aims at extracting speech signal(s) from audio mixtures when visual information of the speaker(s) are available. Several approaches (e.g., [30, 28, 8, 23, 34, 27, 25, 26, 17, 18]) have been developped to deal with this problem from the very basic filtering process using enhancement filters estimated with the help of lip shape information [11] to the definitively more complex case of moving sources [21]. By exploiting visual information to this problem, one want to provide an easier extraction of speaker(s) than purely audio methods. This paper deals with the extraction of speaker(s) in a multimodal way by exploiting the silence periods in the speech signal. Compared to previous work [26], the proposed study provides a better visual detection of silent periods and a more robust estimation of the extraction vector related to the target speaker.

This paper is organized as follows. Section 2 briefly describes an audiovisual voice activity detection (V-VAD) exploited by the proposed method (Section 3) to extract speech signals. Numerical experiments and results are given in Section 4 before conclusions and perspective in Section 5.

2. AUDIOVISUAL COHERENCE: VISUAL VOICE ACTIVITY DETECTION

Speech is at least a bimodal signal, both acoustic and visual, leading thus to exploit these two modalities in a complementary fashion to improve methods generally derived from audio information only. The possibility of using visual information to detect speech and silence periods in a given audio channel is considered in this section.

Visual voice activity detection (V-VAD) has an advantage over audio only based voice activity detection in that it is not susceptible to the problems associated with the acoustic environment (e.g., noise, simultaneous speakers, reverberations, etc.). The V-VAD used in this study is based on a retinal filter [15, 2] in order to detect the movement of the speaker's lips. This method requires no prior information: it is based on the assumption that silence frames can be mostly characterized by the lip-shape movements. Indeed, in silence

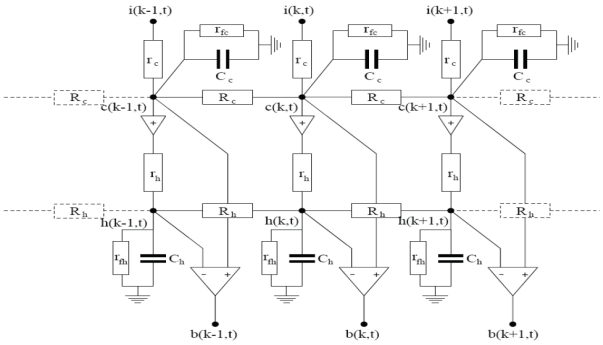

Figure 1: Mono-dimensional electrical scheme to model human retina from [15].

Figure 2: Illustration of the output (right) of the retinal filter [15] applied on the image flow (left).

sections, the lip-shape variations are often small. On the contrary, during speech sections these variations are generally quite stronger. Given these observations, the silence sections can be identified with one or several visual parameters of dynamical nature. The used V-VAD applies a retinal filter to the images flow and calculates the change in energy to classify voice activity. The used spatio-temporal filter (Fig. 1) mimics the output of the parvo cells in a human's retina which enhances the moving contours in the image flow [15] as illustrated in Fig. 2 in which white pixels correspond to moving objects in the image. Let denote by $R_{i,j}(t)$ the output of the retinal filter at time t in the (i,j)-th pixel. The change in energy is then simply computed by

$$\eta(t) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} R_{i,j}(t), \tag{1}$$

where I and J are the number of rows and columns pixels respectively. The probability that at time t the speaker is silent given $\eta(t)$ is finally provided by

$$P_V(t) = .98\exp\left(-\frac{\eta(t)}{45}\right). \tag{2}$$

The parameters of this expression are estimated from the analysis of the database and they are speaker dependent (see Section 4.1).

3. AUDIOVISUAL SOURCE EXTRACTION

In this section, the proposed method to extract audio signal related to one (or several) speaker(s) who is (are) filmed is presented (Subsection 3.1) before a brief recall of the used framework (Subsection 3.2).

3.1 Extraction of intermittent sources

Let consider the linear instantaneous mixing model

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

with as many sources $\mathbf{s}(t) = [s_1(t), \cdots, s_N(t)]^T$ as mixtures $\mathbf{x}(t) = [x_1(t), \cdots, x_N(t)]^T$. Let us denote the covariance matrix of mixtures $\mathbf{x}(t)$ at sample t estimated from a T-length window by

$$R_{T,\mathbf{x}}(t) \stackrel{\triangle}{=} E_T[\mathbf{x}(t)\mathbf{x}(t)^T] = \sum_{i=1}^{N}\sigma^2_{T,i}(t)\mathbf{a}_i\mathbf{a}_i^T \tag{4}$$

where $\sigma^2_{T,i}(t)$ is the average power of the i-th windowed source at sample t and $\mathbf{a}_i$ is the i-th column of the mixing matrix $A = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$. Let now suppose that

- N_1 sources[1] are simultaneously inactive at time τ while they are active at time t, i.e. $1 \leq i \leq N_1$, $\sigma^2_{T,i}(\tau) = 0$ and $\sigma^2_{T,i}(t) \neq 0$,
- the $N - N_1$ other sources have the same average power at times τ and t, i.e. $N_1 + 1 \leq i \leq N$, $\sigma^2_{T,i}(\tau) = \sigma^2_{T,i}(t) \neq 0$.

It is easy to show that the generalized eigenvalue decomposition [12] of pair $(R_{T,\mathbf{x}}(\tau), R_{T,\mathbf{x}}(t))$ admits only two distinct generalized eigenvalues: 0 degenerated N_1 times whose eigensubspace $\mathcal{E}_0(\tau)$ is orthogonal to the space spanned by $\{\mathbf{a}_{N_1+1}, \cdots, \mathbf{a}_N\}$ and 1 degenerated $N - N_1$ times. As a consequence, the projection of observations $\mathbf{x}(t)$ onto $\mathcal{E}_0(\tau)$ cancels the contribution of the sources $s_i(t)$, $N_1 + 1 \leq i \leq N$. This means that the separation vectors $\mathbf{b}_i$, $1 \leq i \leq N_1$, lie in $\mathcal{E}_0(\tau)$, where $\mathbf{b}_i$ is the i-th column of the separation matrix B.

This method allows us to detect how many sources are vanishing by testing the generalized eigenvalues and then to extract the space spanned by the corresponding sources by projecting the observations onto the generalized eigenvectors associated with the eigenvalues equal to zero. Nevertheless, it is worth noting that this method does not allow to identify which sources are inactive. However, even if this difficulty can be overcome by a clustering stage [24], the bimodality of speech is used in this study to extract particular speakers.

3.2 Application to speech source extraction

In this section, the proposed method mixing audio and video processes to extract specific speaker(s) from mixtures is presented. It is a two stages method which first estimates the periods of the speaker's silence and then estimates the corresponding separation vector.

To estimate the periods of silence related to the speaker of interest, the audio and video informations are merged into the probability that this specific speaker is silent at time τ by

$$P_{AV}(\tau) = P_A(\tau)P_V(\tau), \tag{5}$$

where $P_V(\tau)$ is the visual probability that the speaker of interest is silent (2) and $P_A(\tau)$ is the audio probability

[1] Without loss of generality, the inactive sources are arbitrary the N_1 first sources.

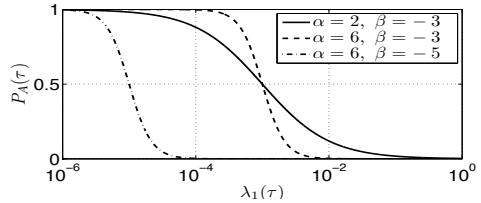

Figure 3: Probability $P_A(\tau)$ of at least one inactive source (6) for several values of parameters α and β.

that at least one source is inactive. This latter probability is obtained by mapping the smallest eigenvalue $\lambda_1(\tau)$ of pair $(R_{T,\mathbf{x}}(\tau), R_{T,\mathbf{x}}(t))$ into $[0,1]$ thanks to the following sigmoid function

$$P_A(\tau) = 1 - \frac{1}{1+\exp\big(-\alpha(\log(\lambda_1(\tau)) - \beta)\big)} \qquad (6)$$

where α and β are parameters that allow to customize the detection of inactive sources (Fig. 3): if $\log(\lambda_1(\tau)) - \beta$ is large (resp. small) compared to $1/\alpha$ then $P_A(\tau)$ is about 0 (resp. 1). The periods of silence are then estimated as the time indices τ such that $P_{AV}(\tau) > .5$.

In the second step, the separating vector $\mathbf{b}_1$ of the target speaker[2] is estimated from the inactivity periods estimated in the first step of the method. While only one generalized eigenvector is theoretically needed to estimate the separating vector, to obtain a better estimation of this latter vector several eigenvectors are used. If only the target speaker is silent ($s_1(t) = 0$), then the corresponding eigenvector is align with the separating vector $\mathbf{b}_1$. However, if several sources are simultaneously inactive with the target speaker, then the corresponding eigenvectors lying in $\mathcal{E}_0(\tau)$ are not necessary align with the related separating vectors $\mathbf{b}_i$. It is worth noting that if $\dim(\mathcal{E}_0(\tau) \cap \mathcal{E}_O(\tau')) = 1$, the support vector $\mathbf{u}_{\tau,\tau'}$ lying in this intersection is then necessary aligned with the separating vector $\mathbf{b}_1$ related to the target speaker. As a consequence, one can estimate the separating vector $\mathbf{b}_1$ as a weighted average of the eigenvectors when only one source is silent and of the support vectors $\mathbf{u}_{\tau,\tau'}$. But as already mentioned, to have a better estimation of $\mathbf{b}_1$ a weighted kernel-PCA is proposed [20, 24] with kernel

$$\psi(\mathbf{u}_{\tau_1,\tau_2}, \mathbf{u}_{\tau_3,\tau_4}) = \frac{|\mathbf{u}_{\tau_1,\tau_2}^T \mathbf{u}_{\tau_3,\tau_4}| - \cos(\theta_0)}{1 - \cos(\theta_0)},$$

if $|\mathbf{u}_{\tau_1,\tau_2}^T \mathbf{u}_{\tau_3,\tau_4}| \geq \cos(\theta_0)$ and $\psi(\mathbf{u}_{\tau_1,\tau_2}, \mathbf{u}_{\tau_3,\tau_4}) = 0$ else, with the convention that $\mathbf{u}_{\tau,\tau}$ is a generalized eigenvector when only one source is inactive. θ_0 is an *a priori* chosen angle which defines the maximum angle of the beam formed by several estimations of $\mathbf{b}_1$. It corresponds to the angular resolution. The probability of the inactivity of the target speaker is included in this model by weighting the kernel $\psi_w(\mathbf{u}_{\tau_1,\tau_2}, \mathbf{u}_{\tau_3,\tau_4}) = w_{\tau_1,\tau_2} w_{\tau_3,\tau_4} \psi(\mathbf{u}_{\tau_1,\tau_2}, \mathbf{u}_{\tau_3,\tau_4})$ with $w_{\tau_i,\tau_j} = \sqrt{P_{AV}(\tau_i) P_{AV}(\tau_j)}$. Finally, the separating vector $\mathbf{b}_1$ is estimated by

$$\mathbf{b}_1 = U \Psi_w \boldsymbol{\phi}_1, \qquad (7)$$

[2]Without loose of generality the target speaker is arbitrary chosen to be $s_1(t)$.

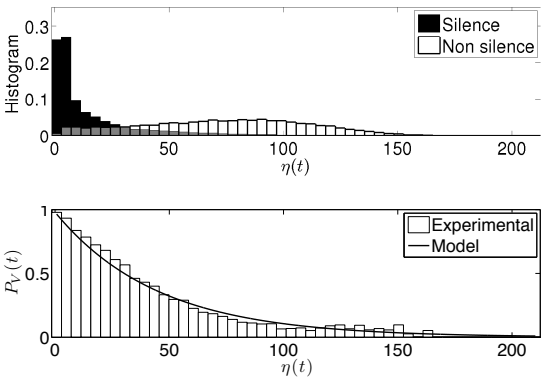

Figure 4: Analysis of change in energy (1) (top plot) and visual probability of silence (2) (bottom plot).

where Ψ_w is the symmetric kernel matrix whose entries are $\psi_w(\mathbf{u}_{\tau_1,\tau_2}, \mathbf{u}_{\tau_3,\tau_4})$ and $\boldsymbol{\phi}_1$ is the eigenvector of Ψ_w associated with the largest eigenvalue. The target speaker is then estimated by

$$\hat{s}_1(t) = \mathbf{b}_1^T \mathbf{x} \qquad (8)$$

for all samples t including those when the target speaker is speaking.

4. RESULTS

In this study, two distinct databases are used to the target speaker (database D1) and to the concurrent audio sources (database D2). The database D1 [29] has been recorded by two French speakers in a spontaneous dialog condition. Both of them have been simultaneously recorded by a microphone (sampling rate of 16kHz) and a camera (sampling rate of 25 frames per second) focussed on the lips region. The second database D2 is composed of speech signals and of music songs. All these signals have a sampling rate of 16kHz.

4.1 Visual voice activity detection

In this section, the analysis of the change in energy (1) is presented (Fig. 4). As one can see, the values of η for the speech and silence classes are not perfectly separated. Besides, even during silence, the speaker's mouth may sometimes slightly move: however the less the lips are moving, the higher the probability of silence is (bottom plot). Indeed, a direct VAD from raw lip shape cannot lead to satisfactory performances because of the intricate relationship between visual and acoustic speech information. Furthermore, the proposed model of the visual probability of silence (2) given the change in energy (1) provides a good fitting of the experimental values. As shown on Fig. 5 which compares the time-moving power of audio signal (when there is no additive noise) and the change in energy $\eta(t)$, these two values are highly correlated as well as the visual probability of silence (2) with the ground truth indexation of silence. This model finally provides a smooth visual decision between silence and non-silence frames. It is worth noting that in real conditions (i.e. with several competitive sources) $\eta(t)$ remains unchanged while the profile of the audio power is altered by the competitive sources.

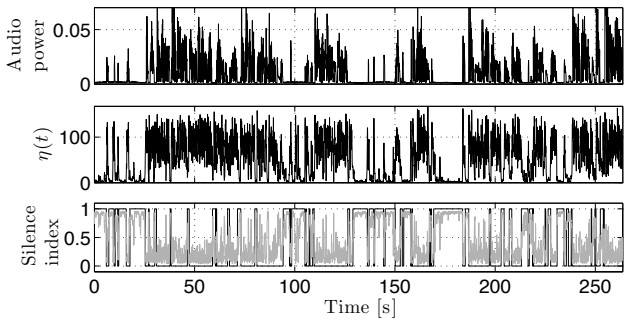

Figure 5: Illustration of visual voice activity detector. In the bottom plot, the black line is the ground truth of silence (1 means silence while 0 means speech) and the grey line is the visual probability of silence $p_V(t)$ (2), with $\alpha = 2$ and $\beta = -1$.

4.2 Audiovisual source extraction

In this experiment, one target speaker from D1 is linearly mixed with two other sources from D2 (Fig. 6). In this example, the two concurrent sources are speech (s_2) and music (s_3), respectively. The audio signals have been windowed synchronously with the video signal resulting a 40ms frames without overlapping. To compute the two covariance matrices $R_{T_1,\mathbf{x}}(t)$ and $R_{T,\mathbf{x}}(\tau)$, T_1 is equal to 5 seconds to have a reference of the average powers of the non-stationary sources and T is equal to 40ms to follow the variations of power of the sources. Fig. 6(c) presents the probabilities of silence obtained from audio only $P_A(t)$, video only $P_V(t)$ and audiovisual $P_{AV}(t)$, respectively. As one can see, the highest values of the audio probability of silence (6) correspond to the silent frames of all sources. In this example the two speech sources (s_1 and s_2) present inactive periods. However, as already mentioned the visual probability of silence is highly correlated with the actual periods of silence of the target speaker (s_1). As a consequence, the proposed audiovisual probability of silence $P_{AV}(t)$ has two main advantages. First, it allows to select among all the periods of silence of at least one source those which mainly correspond to the target speaker. Secondly, it allows to weight the related eigenvectors which are used by the kernel-PCA to estimate the extraction vector. Finally, the target speaker is correctly extracted from the mixtures by the proposed method.

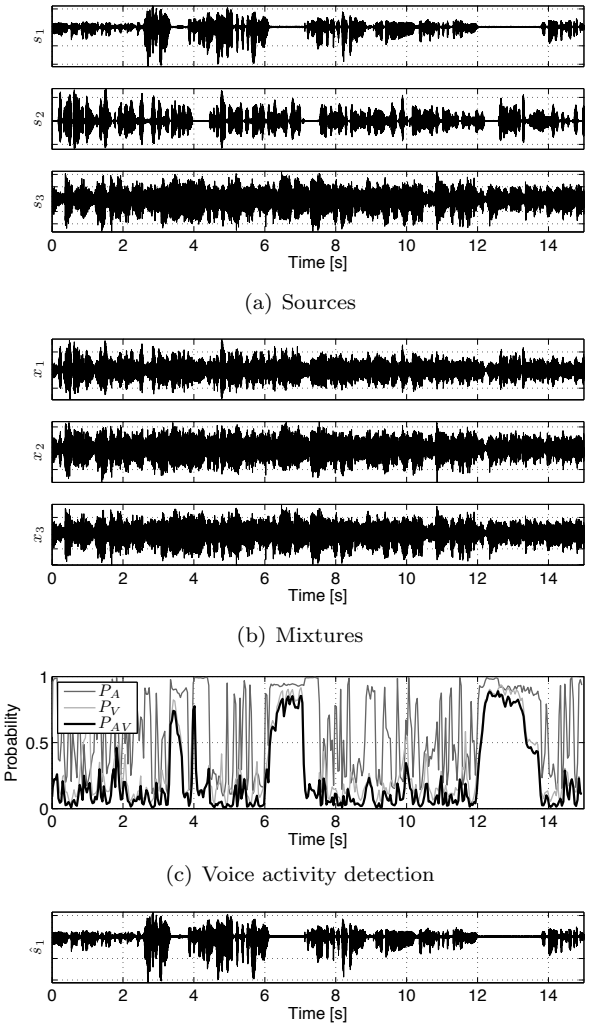

Figure 6: Illustration of the audiovisual speaker extraction.

5. CONCLUSIONS AND PERSPECTIVES

In this study, nothing more besides the periods of silence of a specific speaker are exploited in a peculiar way to extract this source when it is active. The proposed visual voice activity detection based on a smooth criterion as well as the weighted kernel-PCA provide a soft algorithm to take into account the confidence level of the related silent periods. Moreover, an advantage of this algorithm compared to other methods is that it allows to extract and to identify only the target sources thanks to the visual information. It is worth noting that even if the proposed algorithm is applied on a linear instantaneous mixtures, it can easily be extended to the convolutive case by applying it in the frequency domain.

While almost all the classical audiovisual methods to extract speech sources are generally focussed on the periods of activity, conversely the main original assumption of this method is to highlight the inactivity periods of the sources. The latter periods, which are too often ignored, are in fact of a great help to estimate the extraction vector as shown in this study. Future works will focus to propose audiovisual methods to extract moving sources, i.e. with a non-stationary mixing process.

REFERENCES

[1] S. Araki, H. Sawada, and S. Makino. Blind speech separation in a meeting situation with maximum snr beamformers. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–41–I–44, 2007.

[2] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten. Two novel visual voice activity detectors based on appearance models and retinal filltering. In *Proc. European Signal Processing Conference (EU-*

SIPCO), pages 2409–2413, Poznan, Poland, September 2007.

[3] C. Benoît, T. Mohamadi, and S. Kandel. Effects of phonetic context on audio-visual intelligibility of French. *J. Speech and Hearing Research*, 37:1195–1293, 1994.

[4] L. E. Bernstein, E. T. J. Auer, and S. Takayanagi. Auditory speech detection in noise enhanced by lipreading. *Speech Comm.*, 44(1–4):5–18, 2004.

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

[6] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994.

[7] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation Independent Component Analysis and Applications*. Academic Press, 2010.

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

[9] Y. Deville and M. Puigt. Temporal and time-frequency correlation-based blind source separation methods. part I: Determined and underdetermined linear instantaneous mixtures. *Signal Processing*, 87(3):374–407, Mar. 2007.

[10] N. P. Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *J. Speech and Hearing Research*, 12:423–425, 1969.

[11] L. Girin, J.-L. Schwartz, and G. Feng. Audio-visual enhancement of speech in noise. *J. Acoust. Soc. Am.*, 109(6):3007–3020, June 2001.

[12] G. H. Golub and C. F. Van Loan. *Matrix Computation*. Johns Hopkins University Press, third edition, 1996.

[13] K. W. Grant and P.-F. Seitz. The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.*, 108:1197–1208, 2000.

[14] R. Gribonval and S. Lesage. A survey of Sparse Component Analysis for Blind Source Separation: principles, perspectives, and new challenges. In *Proc. European Symposium on Artificial Neural Networks, Advances in Computational Intelligence and Learning (ESANN)*, pages 323–330, Bruges, April 2006.

[15] J. Hérault and W. Beaudot. Motion processing in the retina: About a velocity matched filter. In *Proc. European Symposium on Artificial Neural Networks, Advances in Computational Intelligence and Learning (ESANN)*, Brussels, April 1993.

[16] J. Kim and D. Chris. Investigating the audio–visual speech detection advantage. *Speech Comm.*, 44(1–4):19–30, 2004.

[17] Q. Liu, W. Wang, and P. Jackson. Use of bimodal coherence to resolve spectral indeterminacy in convolutive bss. In *Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, pages 131–139, Saint-Malo, France, September 2010.

[18] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval. Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 12(5):358–371, 2010.

[19] H. McGurk and J. McDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[20] K.-R. Müller, S. Mika, G. Ratsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.*, 12(2):181–201, March 2001.

[21] S. Naqvi, M. Yu, and J. Chambers. A multimodal approach to blind source separation of moving sources. *IEEE Journal of Selected Topics in Signal Processing*, 4(5):895–910, 2010.

[22] G. Potamianos, C. Neti, and S. Deligne. Joint Audio-Visual Speech Processing for Recognition and Enhancement. In *Proc. AVSP'03*, 2003.

[23] 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.

[24] B. Rivet, T. L. Duarte, and C. Jutten. Blind extraction of intermittent sources. In *Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, pages 402–409, Saint-Malo, France, September 2010.

[25] 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 Process.*, 15(1):96–108, January 2007.

[26] B. Rivet, L. Girin, and C. Jutten. Visual voice activity detection as a help for speech source separation from convolutive mixtures. *Speech Comm.*, 49(7-8):667–677, 2007.

[27] C. Sigg, B. Fischer, B. Ommer, V. Roth, and J. Buhmann. Nonnegative CCA for audiovisual source separation. In *Proc. IEEE Int. Workshop on Machine Learning and Signal Processing (MLSP)*, pages 253 – 258, 2007.

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

[29] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten. A study of lip movements during spontaneous dialog and its application to voice activity detection. *J. Acoust. Soc. Am.*, 125(2):1184–1196, February 2009.

[30] 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 Journal on Applied Signal Processing*, 2002(11):1165–1173, 2002.

[31] D. G. Stork and M. E. Hennecke. *Speechreadings by Humans and Machines*. Berlin, Germany : Springer-Verlag, 1996.

[32] W. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.*, 26:212–215, 1954.

[33] Q. Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell, editors, *Hearing by Eye: The Psychology of Lipreading*, pages 3–51. Lawrence Erlbaum Associates, 1987.

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