Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 44 (2004) 113–125
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Developing an audio-visual speech source separation algorithm

David Sodoyer [a,b,*], Laurent Girin [a], Christian Jutten [b], Jean-Luc Schwartz [a]

[a] *Institut de la Communication Parlée, (ICP: Speech Communication Institute), INPG/Université Stendhal/CNRS UMR 5009, ICP, INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex 1, France*

[b] *Laboratoire des Images et des Signaux, (LIS: Laboratory of Images and Signals), INPG/Université Joseph Fourier/CNRS UMR 5083*

Received 2 March 2004; received in revised form 12 October 2004; accepted 13 October 2004

Abstract

Looking at the speaker's face is useful to hear better a speech signal and extract it from competing sources before identification. This might result in elaborating new speech enhancement or extraction techniques exploiting the audio-visual coherence of speech stimuli. In this paper, a novel algorithm plugging audio-visual coherence estimated by statistical tools on classical blind source separation algorithms is presented, and its assessment is described. We show, in the case of additive mixtures, that this algorithm performs better than classical blind tools both when there are as many sensors as sources, and when there are less sensors than sources. Audio-visual coherence enables a focus on the speech source to extract. It may also be used at the output of a classical source separation algorithm, to select the "best" sensor with reference to a target source.

Keywords: Blind source separation; Audio-visual coherence; Speech enhancement; Audio-visual joint probability; Spectral information

1. Introduction

For understanding speech, two senses are better than one: to paraphrase the formula used by Bernstein and Besnoît (1996) to introduce the AVSP special session in ICSLP'96, we know, since Sumby and Pollack (1954) at least, that lipreading improves speech identification in noise, and since Petajan (1984), that Audio-Visual Speech Recognition outperforms Audio Speech Recognition in the same conditions. Recently, Grant and Seitz (2000) discovered that vision of the speaker's face also intervenes in the audio *detection* of speech in noise. This result (confirmed by Kim and Davis, 2001, 2004; Bernstein et al., 2004) lead us show (Schwartz et al., 2002, 2004) that vision may *enhance* audio speech in noise and therefore provide what we called a "very early" contribution to

* Corresponding author. Tel.: +33 0476574850; fax: +33 0476574710.
E-mail address: sodoyer@icp.inpg.fr (D. Sodoyer).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.10.002

speech intelligibility, different and complementary to the classical lipreading effect. In parallel, we exploited, since the middle of the 90s, a technological counterpart of this idea. Girin et al. (1997, 2001) developed a first system for enhancing audio speech embedded in white noise, thanks to a filtering approach, with filter parameters estimated from the video input (see recent developments by Deligne et al., 2002 and Goecke et al., 2002; and also Berthommier, 2003, 2004). The present paper describes a set of new experiments and developments on another approach, exploring the link between two signal processing streams that were almost completely separated: sensor fusion in audio-visual (AV) speech processing, and blind source separation (BSS) techniques (see e.g. Jutten and Herault, 1991; Taleb and Jutten, 1999). This extends preliminary work providing the basis of the method (Sodoyer et al., 2002, 2003) (see also an original link of a different kind between source separation and audio-visual localization in Okuno et al., 2001; Nakadai et al., 2004).

In this paper, the theoretical foundations are presented (Section 2). The evaluation corpora are described in Section 3, together with an analysis of audio-visual coherence in the corresponding material. Comparison methodology and results are provided in Section 4 for the case with as many sensors as sources. The more difficult case involving less sensors than sources is addressed in Section 5, before a general discussion in Section 6 and a final conclusion.

2. Theory

Let us consider the case of a stationary additive mixture of sources, to be separated:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$$

where $\boldsymbol{s}$ contains N unknown signals, $\boldsymbol{A}$ is the unknown $P \times N$ mixing matrix, $\boldsymbol{x}$ are the P observations. Separation consists in estimating output signals $\boldsymbol{y}$ as close as possible to the sources $\boldsymbol{s}$. Output signals are computed from the observations $\boldsymbol{x}$ by applying an $N \times P$ matrix $\boldsymbol{B}$ which is called the separation matrix:

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$$

If the number of sources N equals the number of sensors P, $\boldsymbol{A}$ is a square matrix. If it is regular, perfect separation is possible by taking $\boldsymbol{B} = \boldsymbol{A}^{-1}$. If P is lower than N, then $\boldsymbol{A}$ is no more invertible and there is no exact solution, which makes the problem much more complex.

In the Audio-Visual Speech source Separation (AVSS) approach, we suppose that one source, say $\boldsymbol{s}_1$, is a speech signal, and we exploit additional observations which consist of a video signal $\boldsymbol{V}_1$ extracted from speaker 1's face and synchronous with the acoustic signal $\boldsymbol{s}_1$ that we want to extract. Typically, $\boldsymbol{V}_1$ contains the trajectory of basic geometric lip shape parameters, supposing that they can be automatically estimated by any kind of lip-tracking system. The goal is hence the *extraction* of one audio-visual source merged in a mixture of two or more acoustic signals.

Classical BSS algorithms consider statistically independent sources, and basically involve higher (than 2) order statistics. The AVSS algorithm considers decorrelated sources, and in addition lip motion associated to the source $\boldsymbol{s}_1$ that has to be extracted. The lip pattern provides no information on the glottal source, and incomplete information about the vocal tract. Hence it is classical to consider that the visual input is partially linked to the transfer function in a source-filter model of the speech signal.

2.1. Exploiting spectral information

First, let us assume that we know a number of spectral components of $\boldsymbol{s}_1$, defined by a filter bank on a given time frame (typically, 20-ms windows separated by 20-ms intervals in our application). Let $\boldsymbol{H}_i(\boldsymbol{f})$ be the frequency response of the ith bandpass FIR filter, and $\boldsymbol{h}_i(t)$ be its temporal impulse response. The energy of the source $\boldsymbol{s}_1$ at the output of the filtering process is provided by the autocorrelation with zero delay of the filtered signal $\boldsymbol{h}_i\{\boldsymbol{s}_1\}(t) = \boldsymbol{h}_i(t)*\boldsymbol{s}_1(t)$. The normalized energy of $\boldsymbol{s}_1$ in the ith band is:

$$\gamma_{\boldsymbol{h}_i} = \sqrt{\frac{r_{\boldsymbol{h}_i\{\boldsymbol{s}_1\}}(0)}{r_{\boldsymbol{s}_1}(0)}} \tag{1}$$

where $r_{\boldsymbol{u}}(t)$ denotes the autocorrelation function of signal $\boldsymbol{u}$. If one output, say $\boldsymbol{y}_1$, provides an estimate of $\boldsymbol{s}_1$, we should obtain:

$$\sqrt{\frac{r_{\boldsymbol{h}_i\{\boldsymbol{y}_1\}}(0)}{r_{\boldsymbol{y}_1}(0)}} = \gamma_{\boldsymbol{h}_i} \tag{2}$$

In an $N \times N$ mixture, the direction of $\boldsymbol{s}_1$ is defined by $N-1$ parameters, hence it is easy to show that $N-1$ spectral coefficients are necessary and sufficient to extract $\boldsymbol{s}_1$ (keeping a gain indeterminacy). Therefore, we introduce the following "spectral coefficient" criterion J_{sc}, based on a bank of $(N-1)$ band-pass filters:

$$J_{\mathrm{sc}}(\boldsymbol{y}_1) = \sum_{i=1}^{N-1} \left(\sqrt{\frac{r_{\boldsymbol{h}_i\{\boldsymbol{y}_1\}}(0)}{r_{\boldsymbol{y}_1}(0)}} - \gamma_{\boldsymbol{h}_i} \right)^2 \tag{3}$$

The minimization of this criterion allows the separation of the source $\boldsymbol{s}_1$, provided that the $N \times N$ matrix of the $\boldsymbol{s}_n$ spectral coefficients is regular (Sodoyer et al., 2002).

2.2. The AVSS algorithm

In the real case, we do not know the exact spectral components of the source $\boldsymbol{s}_1$, but we can estimate the spectrum through lip characteristics associated with the sound $\boldsymbol{s}_1$. It is classical to consider that the visual parameters of the speaking face and the spectral characteristics of the acoustic transfer function of the vocal tract are related by a complex relationship which can be described in statistical terms (see e.g. Yehia et al., 1998). Hence, we assume that we can build a statistical model providing the joint probability of a video vector $\boldsymbol{V}$ containing parameters describing the speaker's face (e.g., lip characteristics) and an audio vector $\boldsymbol{S}$ containing spectral characteristics of the sound (i.e. $\gamma_{\boldsymbol{h}_i}$ terms at the output of a filter bank). Let us denote this joint probability $p_{\mathrm{av}}(\boldsymbol{S}, \boldsymbol{V})$.

This statistical model can be designed from a learning corpus, by modeling the probability $p_{\mathrm{av}}(\boldsymbol{S}, \boldsymbol{V})$ as a mixture of Gaussian kernels. The learning corpus is used for estimating the mean, the covariance matrix and the weight of each Gaussian kernel, through an Expectation Maximization (EM) algorithm (Dempster et al., 1977).

Then the separation algorithm consists in estimating a separation matrix $\boldsymbol{B}$ for which the first output $\boldsymbol{y}_1$ produces a spectral vector $\boldsymbol{Y}_1$ as coherent as possible with the video input $\boldsymbol{V}_1$. This results in minimizing the following Audio-Visual (AV) criterion:

$$J_{\mathrm{av}}(\boldsymbol{y}) = -\log(p_{\mathrm{av}}(\boldsymbol{Y}_1, \boldsymbol{V}_1)) \tag{4}$$

It is easy to show that, if there is only one Gaussian kernel, this AV criterion provides a linear regression estimate of the $\gamma_{\boldsymbol{h}_i}$ terms from $\boldsymbol{V}_1$: hence $J_{\mathrm{av}}(\boldsymbol{y})$ becomes equivalent to $J_{\mathrm{sc}}(\boldsymbol{y})$, replacing $\gamma_{\boldsymbol{h}_i}$ by their visual estimate. However, it may happen that the video input $\boldsymbol{V}_1$, at some instants, is associated to a large series of possible spectra, and hence produces very poor separation (the "viseme" problem, see Benoît et al., 1992). For solving this problem, we introduce the possibility to cumulate the probabilities over time. For this purpose, we assume that the values of audio and visual characteristics at several consecutive time frames are independent from each other, and we define an integrated audio-visual criterion by:

$$J_{\mathrm{av}T}(\boldsymbol{y}) = \sum_{k=0}^{T-1} J_{\mathrm{av}}(\boldsymbol{y}(k)) \tag{5}$$

where $\boldsymbol{y}(k)$ is the content of the signal $\boldsymbol{y}$ in the kth time frame before the current one.

2.3. Definition of a reference BSS algorithm: from JADE to $JADE_{track}$

It is necessary to compare the efficiency of our AVSS algorithm with a classical BSS algorithm, in order to be able to assess the interest of the audio-visual approach. A number of reference BSS algorithms exploiting statistical independence between the sources are available in the literature (e.g. Jutten and Herault, 1991; Cardoso and Souloumiac, 1993; Hyvärinen, 1999). They are able to perfectly solve the source separation problem in simple cases (such as linear additive models with as many sensors as sources). The reference BSS algorithm we will use is JADE (Cardoso and Souloumiac, 1993) well known for its simplicity and speed.

BSS algorithms suffer however from two indeterminacy problems. Firstly, the lack of knowledge about the energy of the sources leads to a "gain indeterminacy" (that is, the separating matrix can be multiplied by a diagonal matrix without modifying the value of the criterion to minimize). Secondly, and more seriously, an important drawback of this family of algorithms is their indeterminacy with respect to the permutation of sources. The consequence is the impossibility to know where (i.e. on which sensor) a given source is extracted. For a non-stationary signal, like speech, the energy varies in each frame. Experimentally, it leads to permutation which can vary from one frame to the next one. Of course, this leads to severe difficulties in the application of the algorithm.

A simple way to deal with the permutation problem is to search for the permutation that should be applied at each time frame, in order to maximize the continuity of the output of each sensor, in reference to the previous frame. For this aim, considering that the BSS algorithm has found a given separating matrix $\boldsymbol{B}(n-1)$ for frame $(n-1)$ and another separating matrix $\boldsymbol{B}(n)$ for frame $(\boldsymbol{n})$, we chose to apply to $\boldsymbol{B}(n)$ all possible permutations, and we defined a corrected separating matrix $\boldsymbol{B}^{*}(n)$ by the algorithm:

$$\boldsymbol{B}^{*}(n) = \arg\min_{\text{perm}} |\boldsymbol{B}_{\text{perm}}(n) - \boldsymbol{B}(n-1)| \qquad (6)$$

where $|\boldsymbol{M}| = \text{trace}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})$ defines the square of the Frobenius norm of a given matrix $\boldsymbol{M}$.

This results in minimizing the distance between the coefficients characterizing each sensor after separation, between time $n-1$ and time n. In this equation, as throughout this work, the gain indeterminacy is solved by normalizing each line of $\boldsymbol{B}$ matrices so that the $\boldsymbol{B}$ diagonal values are all "1". The algorithm resulting from the application of this procedure on JADE is called $\text{JADE}_{\text{track}}$. Both JADE and $\text{JADE}_{\text{track}}$ will be used as reference algorithms in the following.

It is important to mention at this point that the AVSS algorithm does not suffer from the permutation indeterminacy problem, since maximizing the coherence of the video component of the target source $\boldsymbol{s}_1$ and the spectral characteristics of the sensor output $\boldsymbol{y}_1$ naturally imposes that $\boldsymbol{s}_1$ is estimated by $\boldsymbol{y}_1$. The gain indeterminacy is solved, as for JADE, by normalizing each line of $\boldsymbol{B}$ matrices so that the $\boldsymbol{B}$ diagonal values are all "1".

2.4. A combined BSS–AVSS algorithm

For relaxing the permutation indeterminacy of BSS algorithms, we purpose to combine the properties of BSS and AVSS criteria, by first estimating a separating matrix by a BSS technique, and then applying an AVSS criterion for selecting the sensor providing the best $\boldsymbol{s}_1$ estimation at the output of BSS. By applying this principle to JADE, we take profit of all its qualities of performance and speed for estimating quickly and efficiently the separation matrix. Then, the selected sensor is the one which maximizes the AVSS criterion in Eq. (5). The resulting algorithm is called hereafter the $\text{JADE}_{\text{AVSS}}$ algorithm.

3. Audio-visual material

3.1. Corpora

We used two types of audio-visual corpora for assessing the separation algorithms.

The first corpus is a corpus of French logatoms, that is non-sense V1–C–V2–C–V1 utterances, where V1 and V2 are same or different vowels within [a, i, y, u] and C is a consonant within the plosives set [p, t, k, b, d, g, #] (# means no plosive). The 112 sequences ($4\times$V1, $7\times$C, $4\times$V2) were pronounced twice by a single male speaker, which resulted in a training set (first repetition) and a test set (second repetition). This logatom corpus presents the interest that it groups in a restricted set all the basic problems to be addressed by audio-visual studies. Indeed, it contains stimuli with similar lips and different sounds (such as [y] vs. [u] or [p] vs. [b]). It also contains pairs of sounds difficult to distinguish, particularly in noise, while their lips are quite distinctive (e.g. [i] vs. [y], or [b] vs. [d]). It is the corpus on which all preliminary studies have been realized.

The second corpus consists in 107 meaningful continuous sentences uttered by the same French speaker, of which we used the first 54 sentences for the training set and the remainder 53 for the

test set. This corpus represents a large jump in difficulty compared with the previous one, for two reasons. Firstly, the complexity of the audio-visual material is much larger, and secondly the test set is quite different from the training set, which is not the case in the logatom corpus.

3.2. Audio and visual parameter extraction

Both corpora were completely analyzed in order to extract sequences of audio and visual parameters aligned in time. These parameters then provided the input to the AVSS separation algorithm. The video data consist of two basic geometric parameters describing the speaker's lip shape, namely width (LW) and height (LH) of the labial internal contour. These parameters were automatically extracted every 20 ms by using a face processing system (Lallouache, 1990). This system exploits a chroma-key process on lips with blue make-up, and with carefully controlled head position and light.

Sounds were sampled at 16 kHz. On 20-ms windows synchronous with the video frames, we computed 32 spectral parameters providing power spectral densities (PSD) at the output of a bank of 32 filters equally spaced between 0 and 5 kHz. PSDs were converted into dBs, and a principal component analysis (PCA) was applied to reduce the number of spectral components to $N_a = 1$, 5 or 8 dimensions. Hence the total dimension of the audio-visual space is $(N_a + 2)$.

3.3. Modeling and assessing audio-visual coherence

For each corpus, the Gaussian mixture model of the $p_{av}(\boldsymbol{S}_1, \boldsymbol{V}_1)$ probability was tuned by an EM algorithm applied to the training data set, containing 2495 audio-visual vectors (112 stimuli, about 24 vectors per stimulus) in the logatom case and 5297 vectors in the sentences case. The simplicity of the logatom corpus allows a careful study of the repartition of Gaussian kernels, in relation with the visual and auditory properties of the corresponding vowels and plosives in the corpus (see Sodoyer et al., 2002). For the sentence corpus, the study is more complex. Therefore, for assessing the validity of the audio-visual modeling in this case, we compared the value of the integrated AV criterion $J_{avT}(\boldsymbol{y})$ applied to two different audio stimuli: (*i*) the true audio source $\boldsymbol{s}_1$ coherent with the video input at each time frame and (*ii*) an audio input $\boldsymbol{s}_2$ coming from another similar sentence corpus, uttered by another male speaker. Comparison of $J_{avT}(\boldsymbol{s}_1)$ and $J_{avT}(\boldsymbol{s}_2)$ was systematically done for the 2606 frames of the sentence test corpus, and for each frame, we selected the signal associated to the smaller $J_{avT}(\boldsymbol{s}_i)$. This was done for various values of N_a (i.e. 1, 5 and 8), various numbers of Gaussian kernels in the audio-visual probability modeling ($N_G = 12$, 18 and 24) and various temporal window integration widths $T = 10$, 20, 40.

On Fig. 1, we display the recognition rate, that is the percentage of cases where $\boldsymbol{s}_1$ is selected rather than $\boldsymbol{s}_2$ ($J_{avT}(\boldsymbol{s}_1) < J_{avT}(\boldsymbol{s}_2)$). It appears that the performance increases largely with T and N_a, and marginally with N_G. It reaches a very high level (about 99%) for $T = 40$, for 8 (or possibly 5) audio dimensions, and for 18 or 24 Gaussian kernels (see Fig. 1). This shows that the audio-visual probability modeling captures the audio-visual natural coherence quite well. Incidentally, the fact that the performance is already very high for $T = 20$ frames (more than 95%) is quite interesting. A duration of 20 frames corresponds to 400 ms, that is roughly two syllables. This seems enough to clearly distinguish audio-visual coherence from incoherence, and the data on the detection of audio-visual asynchronies (see Grant et al., 2004), suggest that human subjects would not perform much better (200 ms instead of 400 ms). Notice that even very poor spectral information may be quite efficient: one spectral parameter suffices to produce 85% correct recognition with 40 frames in Fig. 1. This is also coherent with psychophysical data (Grant et al., 2004).

4. Experiments in the $N \times N$ case

4.1. Methodology

In this first series of experiments, there are as many sensors as sources. In this case, the mixing matrix $\boldsymbol{A}$ is an $N \times N$ matrix supposed to be

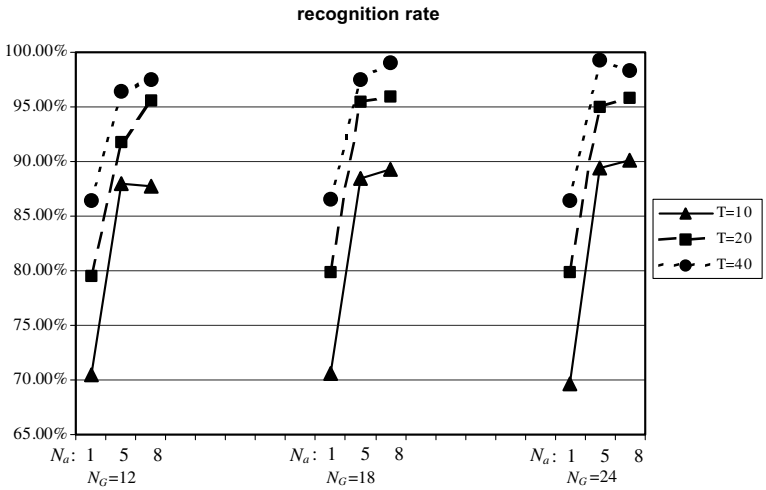

Fig. 1. Correct recognition scores (see text) for various numbers of spectral components, $N_a = 1$, 5, and 8; various numbers of Gaussian kernels in the audio-visual probability modeling, $N_G = 12$, 18 and 24; and various temporal window integration widths $T = 10$, 20, 40.

non-singular. Hence if $\boldsymbol{B}$ is a good estimation of $\boldsymbol{A}^{-1}$, it is trivial that $\boldsymbol{y}$ is a good estimation of the original sources $\boldsymbol{s}$. We tested three values of N, that is $N = 2$, 3 and 5.

Tests were performed on both corpora. $\boldsymbol{s}_1$ is the speech source to extract (2606 test frames for the logatom and the sentence corpus) and the $N-1$ other sources are corrupting speech sources borrowed from another sentence corpus uttered by other speakers. A property sought for a BSS algorithm is equivariance, which is characterized by the fact that performance (independently of permutation and gain indeterminacy problems) does not depend on the mixing matrix, but just on properties of input sources. Equivariance can be achieved by implementing a "relative gradient technique" (Cardoso and Laheld, 1996) or the closely related "natural gradient technique" (Amari, 1998). Therefore, throughout this work, the optimisation of the AVSS criterion $J_{\mathrm{av}T}(\boldsymbol{y})$ (Eq. 5) was realized by a relative gradient in order to make the AVSS algorithm equivariant. To check that this property was ensured, for each value of N we tested two different mixture matrices $\boldsymbol{A}1$ and $\boldsymbol{A}2$. We compared several temporal integration widths T with $T = 1$, 10, 20 and 40 frames. For each mixture, the N observations are defined by:

$$\boldsymbol{x}_n = \sum_{p=1}^{N} a_{np}\boldsymbol{s}_p \tag{7}$$

which are characterized by input SNRs, in reference to $\boldsymbol{s}_1$, provided by:

$$\mathrm{SNR}_{\mathrm{in}}(n) = 10\log\left(a_{n1}^2 E_{s_1} \Big/ \sum_{p=2}^{N} a_{np}^2 E_{s_p}\right) \tag{8}$$

with E_{s_i} the energy of source i in the current frame. Input SNR values displayed in Table 1 are mean values averaged over the test frames.

For each test frame, and for a given separating matrix $\boldsymbol{B}$, the AVSS procedure consists in computing $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$, in estimating the spectrum $\boldsymbol{Y}_1$ according to the process described in Section 3.2 (spectral analysis followed by projection on the selected principal components), and in computing the probability $p_{\mathrm{av}}(\boldsymbol{Y}_1, \boldsymbol{V}_1)$ thanks to the model described in Section 3.3. The number of Gaussian kernels is $N_G = 18$, and the audio dimension is $N_a = 8$. The optimal $\boldsymbol{B}$ matrix, which minimizes the integrated criterion $J_{\mathrm{av}T}(\boldsymbol{y})$, produces an output $\boldsymbol{y}_1$ which is the estimation of the source $\boldsymbol{s}_1$. The output SNR is given by:

$$\mathrm{SNR}_{\mathrm{out}} = 10\log\left(g_{11}^2 E_{s_1} \Big/ \sum_{p=2}^{N} g_{1p}^2 E_{s_p}\right) \tag{9}$$

Table 1
Mean input SNRs (dB) relative to the energy of s_1 for the logatom (panel A) and sentence (panel B) corpus, for various $N \times N$ configurations

	Two sources		Three sources		Five sources	
	$A1$	$A2$	$A1$	$A2$	$A1$	$A2$
Panel A						
Sens. 1	2.6	−10.8	−3.6	−16.2	−12.7	−17.6
Sens. 2	1.6	−15.9	−3.8	−21.8	−14.6	−11.7
Sens. 3	–	–	−1.1	−9.4	−22.3	−14.6
Sens. 4	–	–	–	–	−9.3	−20.3
Sens. 5	–	–	–	–	−13.7	−17.1
Panel B						
Sens. 1	−1.1	−13.9	−6.4	−18.9	−15.7	−20.6
Sens. 2	−1.6	−19.1	−6.6	−24.6	−17.5	−14.6
Sens. 3	–	–	−3.75	−12.1	−26.3	−17.5
Sens. 4	–	–	–	–	−12.3	−23.3
Sens. 5	–	–	–	–	−16.4	−20.1

where $\boldsymbol{G}$ is the global matrix defined by: $\boldsymbol{G} = \boldsymbol{B}A$. We also estimated $\boldsymbol{B}$ separation matrices with JADE, $\text{JADE}_{\text{track}}$ and $\text{JADE}_{\text{AVSS}}$. Output SNR for the four algorithms were then systematically computed and compared for all experimental conditions.

4.2. Results

The results are displayed in Tables 2–4 with, for each case, the mean output SNR averaged over the test frames. Since JADE and $\text{JADE}_{\text{track}}$ do not guarantee that the source s_1 is estimated by the sensor y_1, we systematically selected the best sensor in terms of mean output SNR, for further algorithmic comparison. Let us recall that there is a perfect solution in the $N \times N$ case, hence output SNRs can be arbitrarily high in all conditions. From the results displayed in Tables 2–4, three main features appear:

Role of integration width: It is clear that increasing T improves the performances, for each algorithm. The reason for AVSS is that the integration in Eq. (5) allows the smoothing of the variations of $J_{\text{av}T}(\boldsymbol{y})$, which removes spurious local minima. For JADE and $\text{JADE}_{\text{track}}$, increasing T improves the estimation of second-order and fourth-order cumulants necessary for the convergence towards $\boldsymbol{A}^{-1}$. Furthermore, increasing T also decreases fluctuations of these cumulants, which decreases the number of permutation

Table 2
Mean output SNRs (dB) for the 2×2 configuration (panel A: logatoms, panel B: sentences)

Two sensors, two sources		JADE	$\text{JADE}_{\text{track}}$	$\text{JADE}_{\text{AVSS}}$	AVSS
Panel A: logatom corpus					
$A1$	$T = 1$	0.5	1.0	12.3	16.7
	$T = 10$	4.0	18.2	29.9	36.1
	$T = 20$	9.0	29.3	33.9	41.8
	$T = 40$	15.8	37.6	37.6	45.2
$A2$	$T = 1$	9.0	9.4	12.3	17.0
	$T = 10$	27.7	25.7	29.9	36.3
	$T = 20$	34.1	30.4	33.9	41.4
	$T = 40$	37.6	30.9	37.6	45.2
Panel B: sentence corpus					
$A1$	$T = 1$	0.6	1.9	7.2	7.5
	$T = 10$	7.0	16.8	23.4	20.5
	$T = 20$	12.7	21.8	31.2	31.0
	$T = 40$	16.6	37.1	36.5	36.9
$A2$	$T = 1$	10.3	7.9	7.2	6.9
	$T = 10$	27.7	25.3	23.4	20.2
	$T = 20$	32.8	32.0	31.2	30.7
	$T = 40$	37.1	36.9	36.5	38.0

Table 3
Mean output SNRs (dB) for the 3×3 configuration (panel A: logatoms, panel B: sentences)

Three sensors, three sources		JADE	$\text{JADE}_{\text{track}}$	$\text{JADE}_{\text{AVSS}}$	AVSS
Panel A: logatom corpus					
$A1$	$T = 10$	3.7	−5.4	18.3	24.8
	$T = 20$	9.2	7.7	23.1	29.9
	$T = 40$	14.3	27.0	27.0	33.5
$A2$	$T = 10$	3.5	13.4	18.3	25.2
	$T = 20$	15.7	15.0	23.1	29.8
	$T = 40$	24.7	18.1	27.0	33.8
Panel B: sentence corpus					
$A1$	$T = 10$	−6.5	−8.2	11.1	8.9
	$T = 20$	−3.1	−3.0	19.2	18.4
	$T = 40$	7.1	18.5	26.2	25.8
$A2$	$T = 10$	5.9	14.4	11.1	9.6
	$T = 20$	15.5	22.2	19.0	18.7
	$T = 40$	22.7	26.3	26.2	25.6

Table 4
Mean output SNRs (dB) for the 5×5 configuration (panel A: logatoms, panel B: sentences)

Five sensors, five sources		JADE	$\text{JADE}_{\text{track}}$	$\text{JADE}_{\text{AVSS}}$	AVSS
Panel A: logatom corpus					
$A1$	$T = 20$	−22.5	−20.2	15.1	22.5
	$T = 40$	−22.1	0.7	19.1	26.5
$A2$	$T = 20$	−17.7	−8.7	15.1	22.3
	$T = 40$	−15.6	11.4	19.1	26.7
Panel B: sentence corpus					
$A1$	$T = 20$	−18.0	−9.2	9.7	10.7
	$T = 40$	−15.3	16.7	16.7	16.9
$A2$	$T = 20$	−12.3	−18.1	9.7	11.9
	$T = 40$	−10.0	−7.9	16.7	16.7

switches from one frame to the next. Of course, $\text{JADE}_{\text{AVSS}}$ enjoys the same property as both JADE and AVSS. In the remainder of the study, we shall concentrate on the values $T = 20$ and $T = 40$.

Separation performance: In all cases, the JADE algorithm provides poor results because of permutation problems. $\text{JADE}_{\text{track}}$ corrects this problem rather well, and provides a good baseline for further assessment of AVSS and $\text{JADE}_{\text{AVSS}}$. On the logatom corpus, both techniques largely outperform $\text{JADE}_{\text{track}}$. Furthermore, AVSS itself largely outperforms $\text{JADE}_{\text{AVSS}}$ for both integration widths. The result is much less contrasted for sentences. AVSS and $\text{JADE}_{\text{AVSS}}$ outperform $\text{JADE}_{\text{track}}$ only slightly, and depending on configurations, while performances of AVSS and $\text{JADE}_{\text{AVSS}}$ are quite similar in all conditions.

Equivariance: The JADE algorithm is shown to be equivariant (see Cardoso and Souloumiac, 1993) but the permutation problems do not allow to verify this property. However, solving this problems thanks to the audio-visual selector ensures that $\text{JADE}_{\text{AVSS}}$ displays a remarkable stability of output SNRs from one mixing matrix to the other (compare $A1$ and $A2$ in all cases). Though we implemented a relative gradient descent in AVSS, equivariance is slightly less well achieved, probably because of the sensitivity of the gradient descent to initial conditions.

5. Experiments in the $P < N$ case

In this section, we consider mixtures with less observations than sources, that is $P < N$. In this case, it is known that there is no perfect solution, since $\boldsymbol{s}_1$ does not in general belong to the hyperplane defined by the $\boldsymbol{x}$ sensors. In other words, the inverse matrix of $\boldsymbol{A}$ does not exist, and the identification of $\boldsymbol{A}$ is not sufficient for perfectly recovering the sources. The experimental question now concerns the compared ability of our different algorithms to find good estimates of $\boldsymbol{s}_1$.

5.1. Maximizing SNR through audio-visual coherence

The $P < N$ case is likely to provide a very good test bed for our algorithm. Indeed, in this case, BSS algorithms suffer from an intrinsic limitation. They must find a solution minimizing various kinds of independence criteria (e.g. higher-order statistical moments) but they cannot focus on one or the other source. On the contrary, the AVSS criterion is directed towards the source to extract.

In the hyperplane defined by the set of sensor observations ($\boldsymbol{x}$), the best estimate of $\boldsymbol{s}_1$ maximizing the signal-to-noise ratio SNR should minimize a criterion of least mean square error, J_{lms}:

$$J_{\text{lms}} = E\left[\left(\frac{\boldsymbol{y}_1(t)}{\|\boldsymbol{y}_1(t)\|} - \frac{\boldsymbol{s}_1(t)}{\|\boldsymbol{s}_1(t)\|}\right)^2\right] \tag{10}$$

With the Bessel–Parseval formula, we can transform the cumulated distance in time into a cumulated distance in frequency:

$$E\left[\left(\frac{\boldsymbol{y}_1(t)}{\|\boldsymbol{y}_1(t)\|} - \frac{\boldsymbol{s}_1(t)}{\|\boldsymbol{s}_1(t)\|}\right)^2\right] = \int_{-\infty}^{+\infty} \left|\frac{\boldsymbol{y}_1(f)}{\|\boldsymbol{y}_1(f)\|} - \frac{\boldsymbol{s}_1(f)}{\|\boldsymbol{s}_1(f)\|}\right|^2 \mathrm{d}f \tag{11}$$

where $\boldsymbol{y}_1(f)$ and $\boldsymbol{s}_1(f)$ are the Fourier transforms of $\boldsymbol{y}_1(t)$ and $\boldsymbol{s}_1(t)$.

If we assume that the phases of $\boldsymbol{y}_1(f)$ and $\boldsymbol{s}_1(f)$ are equal, we can express J_{lms} as a spectral distance between $\boldsymbol{y}_1$ and $\boldsymbol{s}_1$:

$$E\left[\left(\frac{\boldsymbol{y}_1(t)}{\|\boldsymbol{y}_1(t)\|}-\frac{\boldsymbol{s}_1(t)}{\|\boldsymbol{s}_1(t)\|}\right)^2\right] = \int_{-\infty}^{+\infty}\left(\frac{|\,\boldsymbol{y}_(f)\,|}{\|\boldsymbol{y}_1(f)\|}-\frac{|\,\boldsymbol{s}_1(f)\,|}{\|\boldsymbol{s}_1(f)\|}\right)^2 \mathrm{d}f$$

If we perform a discrete approximation of the Fourier transform by a filter bank, we have:

$$E\left[\left(\frac{\boldsymbol{y}_1(t)}{\|\boldsymbol{y}_1(t)\|}-\frac{\boldsymbol{s}_1(t)}{\|\boldsymbol{s}_1(t)\|}\right)^2\right] \approx \sum_{f=1}^{F}\left(\frac{|\,\boldsymbol{y}_1(f)\,|}{\sqrt{\sum_{f=1}^{f}|\,\boldsymbol{y}_1(f)|^2}}-\frac{|\,\boldsymbol{s}_1(f)\,|}{\sqrt{\sum_{f=1}^{F}|\,\boldsymbol{s}_1(f)|^2}}\right)^2 \quad (12)$$

which is quite close to the criterion defined by Eq. (3). Hence, it appears that the AV criterion defined in Eq. (4), which provides an audio-visual approximation of the criterion in Eq. (3), should lead to an estimation of $\boldsymbol{s}_1$ with a close to maximal SNR. The temporal integration in Eq. (5) is the AV approximation of an integrated spectral criterion cumulating spectral distances between $\boldsymbol{s}_1$ and $\boldsymbol{y}_1$ on T consecutive temporal windows. Therefore, minimizing $J_{\text{av}T}$ should be close to maximizing the SNR on this integrated window. In fact, the logarithmic transform of PSDs in dBs, induces a discrepancy between the theoretical J_{lms} criterion and the practical $J_{\text{av}T}$ one.

5.2. Methodology

We tested a simple $P < N$ configuration, with two sensors and three sources. In this case, the equivariance property cannot be satisfied since there is no exact solution. The solution found by AVSS depends on both the mixing matrix $\boldsymbol{A}$ and the energy of the sources. We tested four different mixing matrices (Table 5). Notice that the non-stationarity of the speech sources results in variations of the geometry of the problem from frame to frame. Hence, each mixing matrix corresponds in fact to testing many different configurations. The study was done on both corpora, with the same methodology as in Section 4.

Table 5
Mean Input SNRs (dB) in reference to $\boldsymbol{s}_1$ for the logatom (panel A) and sentence (panel B) corpus, for various $P \times N$ configurations

	Two sensors, three sources			
	$\boldsymbol{A}1$	$\boldsymbol{A}2$	$\boldsymbol{A}3$	$\boldsymbol{A}4$
Panel A				
Sens. 1	0.9	−4.4	−15.2	−17.9
Sens. 2	−0.3	−11.3	−5.3	−1.8
Panel B				
Sens. 1	−1.9	−7.1	−17.7	−20.7
Sens. 2	−2.8	−14.2	−8.1	−4.7

5.3. Results

The results are shown in Table 6 for the logatom and sentence corpora. Mean output SNRs are not very large for all algorithms, which is

Table 6
Mean output SNRs (dB for the 2×3 configuration (panel A: logatoms, panel B: sentences)

Two sensors, three sources		JADE	$\text{JADE}_{\text{track}}$	$\text{JADE}_{\text{AVSS}}$	AVSS
Panel A: logatom corpus					
$\boldsymbol{A}1$	$T = 20$	1.1	2.4	4.2	4.4
	$T = 40$	1.9	4.3	4.5	5.1
$\boldsymbol{A}2$	$T = 20$	−7.8	−6.0	−1.0	−0.5
	$T = 40$	−7.9	−6.7	−0.6	0.1
$\boldsymbol{A}3$	$T = 20$	−2.7	−1.2	−1.5	−1.3
	$T = 40$	−0.8	−0.4	−0.6	−0.4
$\boldsymbol{A}4$	$T = 20$	2.6	7.8	11.7	12.7
	$T = 40$	5.6	7.2	11.9	12.2
Panel B: sentence corpus					
$\boldsymbol{A}1$	$T = 20$	−2.6	0.5	0.1	−0.4
	$T = 40$	−1.2	1.7	1.4	0.7
$\boldsymbol{A}2$	$T = 20$	−9.0	−10.2	−4.4	−5.3
	$T = 40$	−10.9	−11.0	−4.4	−4.8
$\boldsymbol{A}3$	$T = 20$	−5.0	−3.5	−5.4	−6.3
	$T = 40$	−3.6	−3.2	−4.7	−4.6
$\boldsymbol{A}4$	$T = 20$	0.8	9.5	7.3	6.7
	$T = 40$	2.4	9.2	8.2	7.2

logical since perfect recovery of the sources is impossible in undetermined mixtures ($P < N$, see here above). They are generally larger for AVSS and $\text{JADE}_{\text{AVSS}}$ than for $\text{JADE}_{\text{track}}$ (and of course JADE), both for logatoms and sentences, though the difference is not very large, and depends on the mixing matrix. Once more, the performances are similar for AVSS and $\text{JADE}_{\text{AVSS}}$. Notice that temporal integration here does not produce much increase in separation. The reason is probably that non-stationarity in the $P < N$ case leads to fluctuations of the separation matrix, which blurs the efficiency of integration for all methods.

6. Discussion

The series of experiments presented in this paper confirm interest in the AVSS technique. Firstly, the theoretical foundations introduced in Sodoyer et al. (2002) for the $N \times N$ case, and developed here for the $P < N$ case, seem sound. Secondly, the extension to a sentence corpus demonstrates that the method can indeed be applied to realistic data, without suffering too much of the increasing complexity of the phonetic material. The fact that 400 ms are enough to distinguish a coherent audio-visual stimulus from an incoher-

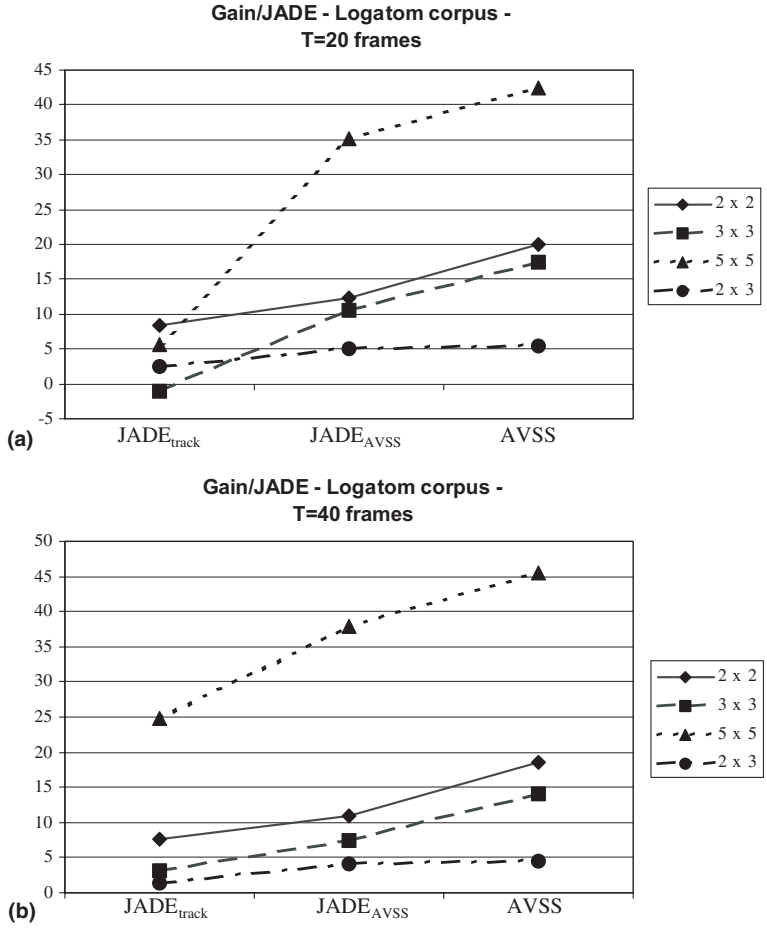

Fig. 2. Mean output SNR gain in reference to the JADE algorithm for the logatom corpus, with a temporal window integration width $T = 20$ (a) and $T = 40$ (b), for various $P \times N$ settings (averaging Results for all tested matrices) and for the three algorithms $\text{JADE}_{\text{track}}$, $\text{JADE}_{\text{AVSS}}$ and AVSS.

ent one in the probability distribution model (Fig. 1) is quite encouraging in this respect.

However, we must admit that the robustness in increasing phonetic complexity is limited. In the summary picture displayed in Figs. 2 and 3 it is clear that for logatoms (Fig. 2) there is a strong hierarchy $\text{AVSS} \gg \text{JADE}_{\text{AVSS}} \gg \text{JADE}_{\text{track}} \gg \text{JADE}$, while for sentences (Fig. 3), the pattern is severely reduced, with something like: $\text{AVSS} = \text{JADE}_{\text{AVSS}} > \text{JADE}_{\text{track}} \gg \text{JADE}$ (denoting by $\gg$ and $>$ the assertions "has a largely/slightly better performance than"). It is nevertheless interesting that algorithms incorporating the AVSS criterion remain better than $\text{JADE}_{\text{track}}$ for sentences, particularly in the "less sensors than sources" condition.

The comparison between $\text{JADE}_{\text{AVSS}}$ and AVSS is a bit disappointing, considering that the large advantage displayed by AVSS for logatoms disappears for sentences. This clearly illustrates the need for more powerful models for estimating the $p_{\text{av}}(\boldsymbol{S}, \boldsymbol{V})$ probability, able to deal with large continuous speech corpora, particularly when we shall deal with multi-speaker applications. However the good performance of the $\text{JADE}_{\text{AVSS}}$ algorithm is interesting. It shows that AVSS does introduce a significant gain at the output of JADE, by its

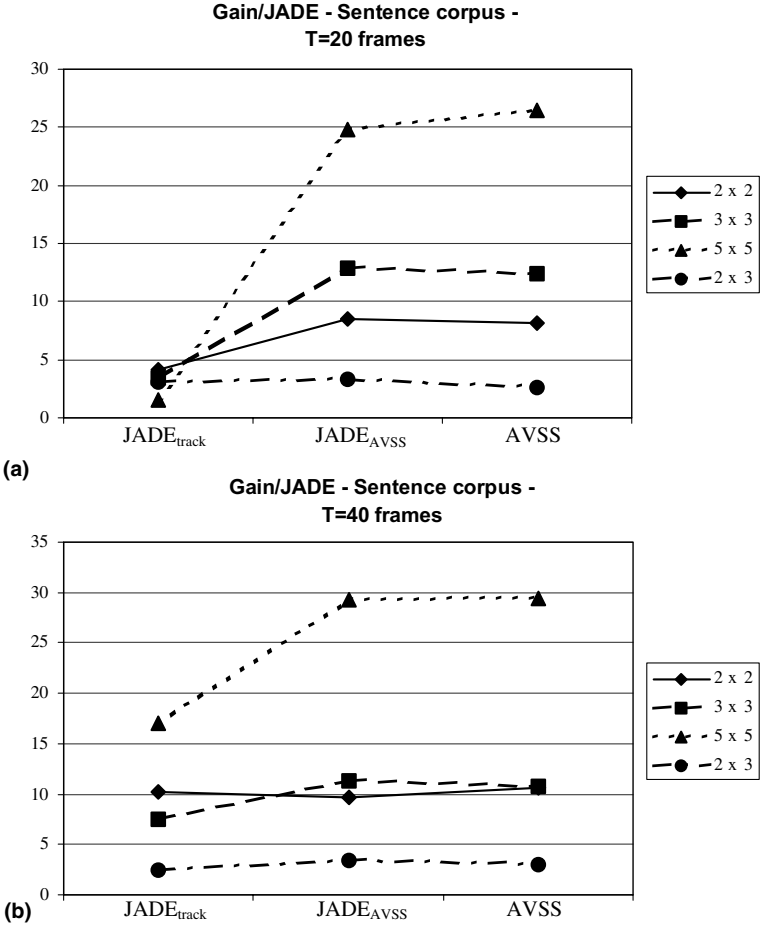

Fig. 3. Mean output SNR gain in reference to the JADE algorithm for the sentence corpus, with a temporal window integration width $T = 20$ (a) and $T = 40$ (b), for various $P \times N$ settings (averaging Results for all tested matrices) and for the three algorithms $\text{JADE}_{\text{track}}$, $\text{JADE}_{\text{AVSS}}$ and AVSS.

systematic selection of a coherent output, as close as possible to the source to extract. The crucial point is that the $JADE_{AVSS}$ algorithm keeps the nice properties of JADE (speed and equivariance), which could be important for future applications. We are presently considering other ways to combine BSS and AVSS techniques, e.g. for more complex mixtures of sources.

7. Conclusion

Altogether, the technological counterpart of the "very early" visual enhancement of audio speech looks quite promising. The method is very efficient in the case of additive mixtures of sources with as many sensors as sources. In this paper, we show that the method seems able to deal with less sensors than sources, thanks to its ability to focus on the target source. This might also lead to efficient BSS/AVSS combined algorithms exploiting both independence criteria, and AV coherence criteria to select a given source in a mixture. Of course, further developments are still necessary for a complete demonstration of the efficiency of the technique. They will involve larger multi-speaker corpora, more powerful learning tools for AV association, and they should address more complex mixtures (including convolutive ones). It is already possible however to assert that the connection of BSS techniques with the field of AV speech processing is an exciting new challenge for future research in both communities.

References

Amari, S.-L., 1998. Natural gradient works efficiently in learning. Neural Comput. 10, 251–276.

Benoît, C., Lallouache, M.T., Mohamadi, T., Abry, C., 1992. A set of visual French visemes for visual speech synthesis. In: Bailly, G. et al. (Eds.), Talking Machines. Elsevier, Amsterdam, pp. 485–504.

Bernstein, L.E., Benoît, C., 1996. For speech perception by humans or machines, three senses are better than one. In: Proc. ICSLP'96, pp. 1477–1480.

Bernstein, L.E., Takayanagi, S., Auer E.T., Jr., 2004. Enhanced auditory detection with AV speech: Perceptual evidence for speech and non-speech mechanisms. This volume.

Berthommier, F., 2003. Audiovisual speech enhancement based on the association between speech envelope and video feature. In: Proc. Eurospeech'03, Geneva, pp. 1045–1048.

Berthommier, F., 2004. A phonetically neutral model of the low-level audiovisual interaction. This volume.

Cardoso, J.F., Souloumiac, A., 1993. Blind beamforming for non-Gaussian signals. IEE Proc.—F 140, 362–370.

Cardoso, J.F., Laheld, B., 1996. Equivariant adaptative source separation. IEEE Trans. SP 44, 3017–3030.

Deligne, S., Potamianos, G., Neti, C., 2002. Audio-visual speech enhancement with AVCDCN (AudioVisual Codebook Dependent Cepstral Normalization). In: Proc. ICSLP'2002, pp. 1449–1452.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. 39, 1–38.

Girin, L., Feng, G., Schwartz, J.-L., 1997. Can the visual input make the audio signal pop out in noise? A first study of the enhancement of noisy VCV acoustic sequences by audiovisual fusion. In: Proc. AVSP'97, pp. 37–40.

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

Goecke, R., Potamianos, G., Neti, C., 2002. Noisy audio feature enhancement using audio-visual speech data. In: Proc. Internat. Conf. Acoust., Speech, Signal Process., pp. 2025–2028.

Grant, K.W., Seitz, P., 2000. The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. 108, 1197–1208.

Grant, K.W., van Wassenhove, V., Poeppel, D., 2004. Detection of auditory and auditory–visual synchrony. This volume.

Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for Independent Component Analysis. IEEE Trans. Neural Networks 10, 626–634.

Jutten, C., Herault, J., 1991. Blind separation of sources. Part I: An adaptive algorithm based on a neuromimetic architecture. Signal Process. 24, 1–10.

Kim, J., Davis, C., 2001. Visible speech cues and auditory detection of spoken sentences: an effect of degree of correlation between acoustic and visual properties. In: Proc. AVSP'2001, pp. 127–131.

Kim, J., Davis, C., 2004. Testing the cuing hypothesis for the AV speech detection advantage. This volume.

Lallouache, M.T., 1990. Un poste 'visage-parole'. Acquisition et traitement de contours labiaux. In: Proc. XVIII JEPs, Montréal, pp. 282–286.

Nakadai, K., Matsuura, D., Okuno, H.G., Tsujino, H., 2004. Improvement of three simultaneous speech recognition by using AV integration and scattering theory for humanoid. This volume.

Okuno, H.G., Nakadai, K., Lourens, T., Kitano, H., 2001. Separating three simultaneous speeches with two microphones by integrating auditory and visual processing. In: Proc. Eurospeech 2001, pp. 2643–2646.

Petajan, E.D., 1984. Automatic Lipreading to Enhance Speech Recognition. Doct. Thesis, University of Illinois.

Schwartz, J.L., Berthommier, F., Savariaux, C., 2002. Audio-visual scene analysis: Evidence for a "very-early" integration process in audio-visual speech perception. In: Proc. ICSLP'2002, pp. 1937–1940.

Schwartz, J.L., Berthommier, F., Savariaux, C., 2004. Seeing to hear better: Evidence for early audio-visual interactions in speech identification. Cognition 93, B69–B78.

Sodoyer, D., Schwartz, J.L., Girin, L., Klinkisch, J., Jutten, C., 2002. Separation of audio-visual speech sources. Eurasip JASP 2002, 1164–1173.

Sodoyer, D., Girin, L., Jutten, C., Schwartz, J.L., 2003. Extracting an AV speech source from a mixture of signals. In: Proc. Eurospeech 2003, pp. 1393–1396.

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

Taleb, A., Jutten, C., 1999. Source separation in postnonlinear mixtures. IEEE Trans. SP 10, 2807–2820.

Yehia, H., Rubin, P., Vatikiotis-Bateson, E., 1998. Quantitative association of vocal-tract and facial behavior. Speech Communication 26, 23–43.