Extracting an AV speech source from a mixture of signals

David Sodoyer, Laurent Girin, Christian Jutten(*), Jean-Luc Schwartz

Speech Communication Institute (ICP), CNRS UMR 5009, INPG, Grenoble France
(*) Image and Signal Processing Laboratory (LIS), CNRS UMR 5083, INPG, Grenoble France
sodoyer@icp.inpg.fr girin@icp.inpg.fr schwartz@icp.inpg.fr
Christian.Jutten@lis.inpg.fr

Abstract

We present a new approach to the source separation problem for multiple speech signals. Using the extra visual information of the face speaker, the method aims to extract an acoustic speech signal from other acoustic signals by exploiting its coherence with the speaker's lip movements. We define a statistical model of the joint probability of visual and spectral audio input for quantifying the audio-visual coherence. Then, separation can be achieved by maximising this joint probability. Experiments on additive mixtures of 2, 3 and 5 sources show that the algorithm performs well, and systematically better than the classical BSS algorithm JADE.

1. Introduction

A number of recent experiments suggest that the audio-visual interaction in speech perception could act at a very early level, and that the visual input might improve the *detection* of speech sounds embedded in noise [1]. In this paper, we implement this idea for enhancing speech embedded in various kinds of noise, thanks to a computational process based on the audiovisual input. Enhancement has already been obtained with an original filtering approach [2]. The present work explores another idea: adapt blind source separation (BSS) techniques to audiovisual speech sources.

2. Designing the algorithm

2.1. Background

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

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

where s contains N unknown signals, A is the unknown P x N mixing matrix, x the P observations, and B is the N x P separation matrix to estimate in order to recover the output signals y as close as possible to the sources s. In this application, the signals s are speech signals, and we assume as many sources as observations, that is P=N. Furthermore, we exploit additional observations which consist of a video signal V_1 extracted from speaker 1's face and synchronous with the acoustic signal s_1 that we want to extract. Typically, V_1 contains the trajectory of basic geometric lip shape parameters, which can be automatically extracted by different systems developed in our laboratory [3, 4]. In the present paper, we focus on 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. In the Audio Visual Source Separation (AVSS) approach, we just need decorrelated sources, and we assume that we know the lip motion associated to the source $\boldsymbol{s}_1$ that we want to extract. The lip pattern provides an incomplete information about the vocal tract shape, hence it is classical to consider that the visual input is partially linked to the transfer function of the vocal tract. In the following, we assume that the additional knowledge about $\boldsymbol{s}_1$ concerns the variation of its spectral envelope.

2.2. Principle

First, let us assume that we know a number of spectral components of $\boldsymbol{s}_1$, defined by a filter bank. Let $H_i(f)$ be the frequency response of i-th bandpass FIR filter, and $h_i(t)$ 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 $h_i\{s_1\}(t)=h_i(t)*s_1(t)$. The normalised energy of $\boldsymbol{s}_1$ in the i-th band is:

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

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

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

Moreover, for extracting one source among N, it easy to show that we need $N-1$ spectral coefficients. Therefore, we propose to minimize the following criterion :

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

This criterion, based on a bank of $(N-1)$ band-pass filters, allows the separation of the source $\boldsymbol{s}_1$ provided that the spectra of all sources $\boldsymbol{s}_n$ are different [5].

2.3. The AVSS algorithm

In this work, we don't know the exact spectral components of the source $\boldsymbol{s}_1$, but we can estimate the spectrum through lip characteristics associated to 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. [6]). 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 of

an audio vector $\boldsymbol{S}$ containing spectral characteristics of the sound. Let us denote this joint probability $p_{av}(\boldsymbol{S},\boldsymbol{V})$. This statistical model is designed from a learning corpus, by modelling the probability $p_{av}(\boldsymbol{S},\boldsymbol{V})$ 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, by running an Expectation Maximization (EM) algorithm.
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 maximizing the following Audio-Visual (AV) criterion:

$$J_{av}(\boldsymbol{y}) = p_{av}(\boldsymbol{Y}_1,\boldsymbol{V}_1) \quad (4)$$

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 [7]). 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 the cumulated joint audio-visual probability as:

$$p_{av}\big(\boldsymbol{Y}_1(t,\cdots,t-T-1),\boldsymbol{V}_1(t,\cdots,t-T-1)\big)=\prod_{\tau=0}^{T-1} p_{av}\big(\boldsymbol{Y}_1(t-\tau),\boldsymbol{V}_1(t-\tau)\big) \quad (5)$$

3. Experimental results

3.1. Data

The audio-visual corpus used in the experiments consists of V1-C-V2-C-V1 sequences uttered by a French speaker. V1 and V2 are vowels within [a, i, y, u]. C is a consonant within the plosives set [p, t, k, b, d, g, #] (# means no plosive). The 112 sequences (4xV1, 7xC, 4xV2) are pronounced twice by a single speaker, for generating both a training and a test set. The corrupting signals consist in continuous meaningful sentences uttered by the other speakers. 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 are automatically extracted every 20 ms by using a face processing system [3]. Sounds are sampled at 16 kHz. On the same 20 ms sound windows, synchronous with the video analysis, we compute 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 are converted in dBs, and a principal component analysis (PCA) is applied to reduce the number of spectral components to 12 dimensions (explaining more than 96% of the total variance). Hence the audio-visual space dimension is 14 (12 audio + 2 video). The EM gaussian mixture algorithm is applied to the training data set, containing 2497 audio-visual vectors (112 stimuli, about 24 vectors per stimulus). In the present work, the number of gaussians in each mixture is set to 16.

3.2. Procedure

The AV criterion $J_{av}(\boldsymbol{y})$ (Eq. 4, 5) is optimized by a relative gradient algorithm [8]. We tested several N x N mixtures, with N =2, 3, et 5, where $\mathbf{s}_1$ is the speech source to extract (2495 test frames) and the N-1 other sources are corrupting speech sources. For each N x N mixture, we tested two different mixture matrices $\boldsymbol{A1}$ and $\boldsymbol{A2}$, and we used several temporal integration widths T with T=1, 10 et 20 frames. For each mixture, the N observations are defined by:

$$\boldsymbol{x}_n=\sum_{p=1}^{N} a_{np}\boldsymbol{s}_p \quad (6)$$

which are characterized (the sources being normalized in energy) by input *SNR*s :

$$SNR_{in}(n)=10\log(a_{n1}^2/\sum_{p=2}^{N} a_{np}^2) \quad (7)$$

Table 1: Input SNRs (dB)

	2 sources		3 sources		5 sources	
	A1	*A2*	*A1*	*A2*	*A1*	*A2*
Sensor 1	-1.16	-14	+2.70	+0.73	-6.53	-10.6
Sensor 2	-1.58	-19.1	+3.20	+0.1	-6.47	-3.67
Sensor 3	-	-	+9.60	+2.15	-14.8	-8.43
Sensor 4	-	-	-	-	-3.34	-16.1
Sensor 5	-	-	-	-	-6.02	-10.2

The evaluation was made by concatenating all 112 stimuli of the test set into a single file containing 2495 audio-visual frames. For each test frame, and for a given separating matrix $\boldsymbol{B}$, the procedure consists in computing $\boldsymbol{y}=\boldsymbol{Bx}$, in estimating the spectrum $\boldsymbol{Y_1}$ according to the process described in section 3.1 (spectral analysis followed by PCA), and in computing the probability $p_{av}(\boldsymbol{Y}_1,\boldsymbol{V}_1)$ thanks to the model described in section 2.3. The optimal $\boldsymbol{B}$ matrix, which maximises the probability $p_{av}(\boldsymbol{Y}_1,\boldsymbol{V}_1)$, produces an output $\boldsymbol{y}_1$ which is the best estimation of the source $\boldsymbol{s}_1$. The output *SNR* is given by:

$$SNR_{out}(1)=10\log(g_{11}^2/\sum_{p=2}^{N} g_{1p}^2) \quad (8)$$

where $\boldsymbol{G}$ is the global matrix defined by :

$$\boldsymbol{G=BA}$$

Finally, we tested the same mixtures with the classical BSS algorithm JADE [9]. For taking into account possible indetermination in this algorithm, we systematically selected the best $SNR_{out}(n)$ for the signal $\boldsymbol{s_1}$ among all output sensors:

$$SNR_{out}=\underset{n}{\operatorname{argmax}}\,10\log(g_{n1}^2/\sum_{p=2}^{N} g_{np}^2) \quad (9)$$

The results are presented with histograms of output SNR values on the 2495 test frames (Fig. 1 to 7) from which we summarize some salient values in the tables 3 to 5.

3.3. Results

From Tab. 3-5 and Fig. 1-7, three main features appear :

- *Role of integration width* : It is clear that a large T value improves the performances, both for AVSS and JADE: the output *SNR*s increase from T=1 to 10 and 20 (Fig. 1 to 3, Fig. 4 vs. 5 ; results summarized in Tab. 3 and 4).

The reason for AVSS is that the integration in Eq. (5) allows to improve the estimation of $\boldsymbol{B}=\boldsymbol{A}^{-1}$. For BSS, increasing T improves the estimation of second-order and fourth-order cumulants necessary for the convergence towards $\boldsymbol{A}^{-1}$.

- *Superiority of AVSS* : All along Fig. 1-7, AVSS performs better than JADE, sometimes slightly (as in Fig. 1) sometimes strongly (8 dB mean gain of output *SNRs* in Fig. 6). This shows that the spectral information on $\boldsymbol{s}_1$, even incompletely provided by $\boldsymbol{V}_1$, is a more accurate hint for the extraction of $\boldsymbol{s}_1$, than the only criterion of statistical independence.
- *Equivariance* : The equivariance property implemented in JADE allows a remarkable stability of output *SNR*s from one mixing matrix to the other (compare $\boldsymbol{A1}$ and $\boldsymbol{A2}$ in Tab. 3, 4, 5, for each T value for JADE). Though we implemented a relative gradient descent in AVSS, equivariance is not so well achieved, as it can be seen for example in Tab. 5. This could be due to the non-linear nature of the AV probability function.

Furthermore, two additional criteria should be taken into account in the comparison between AVSS and JADE :

- *Computation cost* : JADE is very fast, thanks to the Jacobi algorithm. Though slower, AVSS is still relatively fast thanks to a gradient descent exploiting the value for a given frame as the initial value for the next frame (less than 15 iterations before convergence with T=20 in the 2-2, 3-3 and 5-5 case).
- *Stability of the selected sensor for extraction* : This is a clear advantage of AVSS, which ensures that $\boldsymbol{s}_1$ is always extracted on $\boldsymbol{y}_1$, since the criterion is focused on $\boldsymbol{V}_1$. On the contrary, the instability of BSS is well-know. Indeed, small fluctuations in the values of second or fourth-order moments result in many permutations of solutions from one frame to the next. In Table 2, we display the number of cases where there was a switch in the selected sensor where $\boldsymbol{s}_1$ appeared, between two consecutive frames. We could envision in the future to apply the AV criterion $J_{av}(\boldsymbol{y})$ at the output of a BSS algorithm, in order to select the good sensor.

Table 2 : Number of JADE permutations for different mixtures and different integration widths.

	2x2			3x3		5x5
	T=1	T=10	T=20	T=10	T=20	T=20
A1	674	218	120	276	362	534
A2	506	49	4	171	148	387

4. Conclusion

The principle of an audio-visual algorithm for speech signals separation is theoretically sound and technically viable. The high separation scores we obtain imply reasonable integration widths (T=20 corresponds to 400 ms, which is rather low), and lead to very good quality of the separated signals. This work is promising because we can expect to proceed in the future with much more realistic (and difficult) configurations. Especially, with less sensors than sources, speech extraction using visual information could be achieved following this idea. Indeed, in this case, AVSS enables to extract the best $\boldsymbol{s}_1$ fit in terms of maximal *SNR*. On the contrary, the BSS cannot find this maximum, since it provides no equivalence in terms of inter-sensor independence. We are presently deriving the formal equations and implementations of this configuration. Finally, we envision in the future a combination of AV coherence cues with classical BSS techniques as they are developed in our group [10].

5. References

[1] J.L. Schwartz, F. Berthommier & C. Savariaux (2002). Audio-visual scene analysis. *Proc. ICSLP'2002*, 1937-1940.

[2] L. Girin, J.L. Schwartz & G. Feng (2001). Audio-visual enhancement of speech in noise. *JASA, 109*, 3007-3020.

[3] M.T. Lallouache (1990). Un poste 'visage-parole'. Acquisition et traitement de contours labiaux. *Proc. XVIII JEPs*, Montréal, 282-286.

[4] L. Revéret & C. Benoît (1998). A new 3D lip model for analysis and synthesis of lip motion in speech production. *Proc. AVSP'98*, 207-212.

[5] D. Sodoyer, J.L. Schwartz, L. Girin, J. Klinkisch, & C. Jutten (2002). Separation of audio-visual speech sources. *Eurasip JASP*, *2002*, 1164-1173.

[6] H. Yehia, P. Rubin, & E. Vatikiotis-Bateson (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*, 23-43.

[7] C. Benoît, T. Lallouache, T. Mohamadi, & C. Abry (1992). A set of visual French visemes for visual speech synthesis. In *Talking Machines*, G. Bailly et al. (eds.). Amsterdam: Elsevier, 485-504.

[8] J.F. Cardoso, & B. Laheld (1996). Equivariant adaptative source separation. *IEEE Trans. SP, 44*, 3017-3030.

[9] J.F. Cardoso, (1993). Blind beamforming for non-gaussian signals. *IEE Proc.-F, 140*, 362-370.

[10] A. Taleb, & C. Jutten. (1997) Entropy optimization - Application to blind separation of sources. *Proc. ICANN 97*, Lausanne, 527-534.

Table 3 : Results for 2 sensors, 2 sources

		A1			A2		
		≥30 dB	≥10 dB	≥0 dB	≥30 dB	≥10 dB	≥0 dB
T=1	AVSS	31%	59%	76%	23%	59%	76%
	JADE	17%	56%	78%	17%	56%	78%
T=10	AVSS	63%	94%	98%	64%	95%	99%
	JADE	39%	91%	98%	39%	91%	98%
T=20	AVSS	81%	98%	99%	80%	99%	100%
	JADE	52%	98%	100%	50%	98%	100%

Table 4 : Results for 3 sensors, 3 sources

		A1			A2		
		≥30 dB	≥10 dB	≥0 dB	≥30 dB	≥10 dB	≥0 dB
T=10	AVSS	35%	88%	98%	38%	90%	98%
	JADE	16%	81%	96%	16%	82%	96%
T=20	AVSS	68%	98%	99%	64%	99%	100%
	JADE	28%	94%	99%	28%	94%	99%

Table 5 : Results for 5 sensors, 5 sources

		A1			A2		
		≥30 dB	≥10 dB	≥0 dB	≥30 dB	≥10 dB	≥0 dB
T=20	AVSS	36%	99%	100%	22%	89%	98%
	JADE	8%	88%	99%	9%	88%	99%

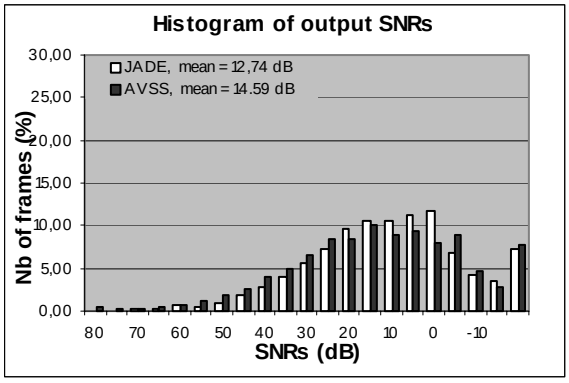

Figure 1 : 2 sensors, 2 sources - Matrix A1 - T=1

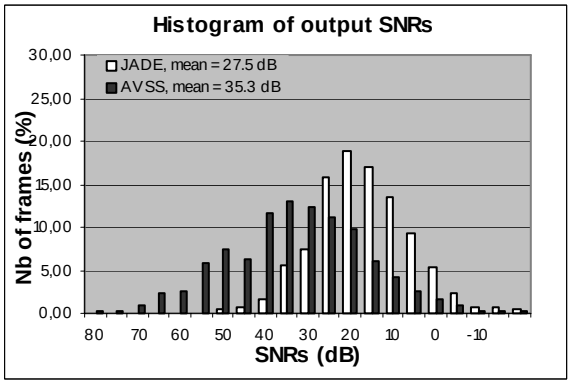

Figure 2 : 2 sensors, 2 sources - Matrix A1 - T=10

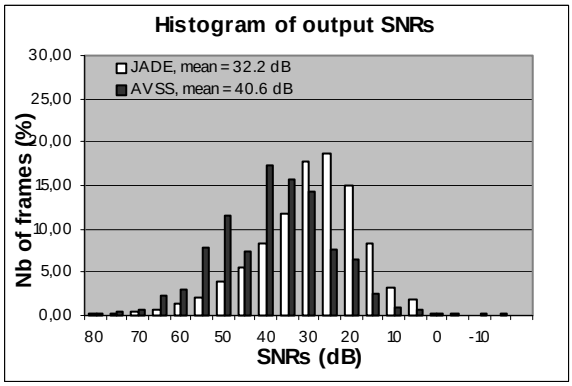

Figure 3 : 2 sensors, 2 sources - Matrix A1 - T=20

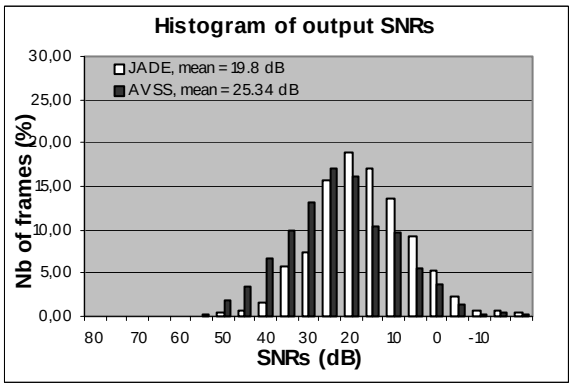

Figure 4 : 3 sensors, 3 sources - Matrix A1 - T=10

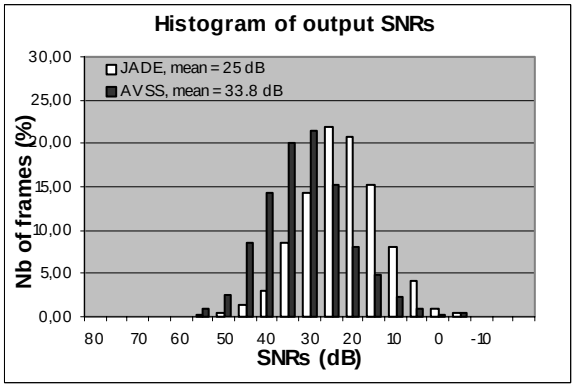

Figure 5 : 3 sensors, 3 sources - Matrix A1 - T=20

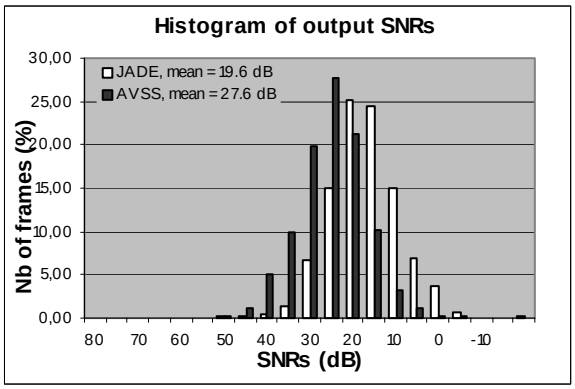

Figure 6 : 5 sensors, 5 sources - Matrix A1 - T=20

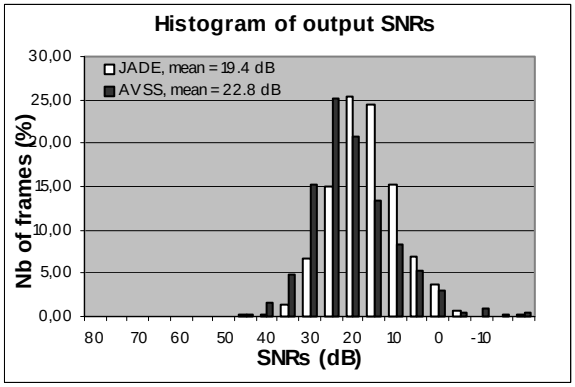

Figure 7 : 5 sensors, 5 sources - Matrix A2 - T=20