AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Further experiments on audio-visual speech source separation

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 Christian.Jutten@lis.inpg.fr
schwartz@icp.inpg.fr

Abstract

Looking at the speaker's face seems useful to better hear 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, we present a set of experiments on a novel algorithm plugging audio-visual coherence estimated by statistical tools, on classical blind source separation algorithms. 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. Audiovisual coherence enables to 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 in reference to a target source.

1. Introduction

For understanding speech, two senses are better than one: to paraphrase the formula used by Lynne Bernstein and Christian Benoît to introduce the AVSP special session in ICSLP'96 [1], we know, since Sumby & Pollack [2] at least, that lipreading improves speech identification in noise, and since Petajan [3], that Audio-Visual Speech Recognition outperforms Audio Speech Recognition in the same conditions. In this framework, the recent discovery by Grant & Seitz [4] (confirmed by Kim & Davis [5]) that vision of the speaker's face intervenes in the audio *detection* of speech in noise, suggests that for *hearing* speech also, two senses are better than one. On this basis, we attempted to show (Schwartz et al. [6], [7]) that vision may *enhance* audio speech in noise and therefore provide what we called a "very early" contribution to 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. [8], [9] 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. A recent contribution by Deligne et al. [10] provides an extension of this work using more powerful nonlinear techniques. The present paper describes a set of new experiments and developments on another approach, hopefully more powerful, exploring the link between two signal processing streams that were almost completely separated (though see an original link between source separation and audio-visual localization in Okuno et al. [11]): sensor fusion in audio-visual (AV) speech processing, and blind source separation (BSS) techniques (see e.g. Jutten & Herault, [12], Taleb & Jutten, [13]). This extends preliminary work providing the basis of the method (Sodoyer et al., [14]).

2. Theoretical 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 $\boldsymbol{s}$ contains N unknown signals, $\boldsymbol{A}$ is the unknown PxN mixing matrix, $\boldsymbol{x}$ are the P observations, and $\boldsymbol{B}$ is the NxP separation matrix to estimate in order to recover the output signals $\boldsymbol{y}$ as close as possible to the sources $\boldsymbol{s}$. In the Audio-Visual 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 just needs decorrelated sources, plus lip motion associated to the source $\boldsymbol{s}_1$ that has to be extracted. The lip pattern provides incomplete information about the vocal tract hence it is classical to consider that the visual input is partially linked to the transfer function. In the following, we assume that the additional knowledge about $\boldsymbol{s}_1$ concerns the variation of its spectral envelope.

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 in our application). Let $H_i(f)$ be the frequency response of the 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 normalized energy of $\boldsymbol{s}_1$ in the i-th band is:

$$\gamma_{h_i} = \sqrt{\frac{r_{\boldsymbol{h}_i\{s_1\}}(0)}{r_{s_1}(0)}} \qquad (1)$$

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

$$\sqrt{\frac{r_{\boldsymbol{h}_i\{y_1\}}(0)}{r_{y_1}(0)}} = \gamma_{h_i} \qquad (2)$$

In an *N*x*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 criterion to minimize:

$$J_{\mathrm{sc}}(\boldsymbol{y}_1) = \sum_{i=1}^{N-1}\left(\sqrt{\frac{r_{\boldsymbol{h}_i\{y_1\}}(0)}{r_{y_1}(0)}} - \gamma_{h_i}\right)^2 \qquad (3)$$

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 other sources $\boldsymbol{s}_n$ are different [14].

2.2. 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. Yehia et al., [15]). 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 $p_{\mathrm{av}}(\boldsymbol{S},\boldsymbol{V})$. This statistical model is 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.

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:

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

It is easy to show that if there is only one Gaussian kernel, it provides a linear regression estimate of the γ_{h_i} terms from $\boldsymbol{V}_1$: hence $J_{\mathrm{av}}(\boldsymbol{y})$ becomes equivalent to $J_{\mathrm{sc}}(\boldsymbol{y})$, replacing γ_{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. [16]). 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:

$$J_{\mathrm{avT}}(\boldsymbol{y}) = \sum_{\varphi=0}^{T-1} J_{\mathrm{av}}(\boldsymbol{y}(\varphi)) \qquad (5)$$

where $\boldsymbol{y}(\varphi)$ is the content of the signal $\boldsymbol{y}$ in the φ^{th} time frame before the current one.

3. Experiment in the $P = N$ case

In this section, we consider as many sources as observations, that is *P=N*. In this case, we know that there is a perfect solution, that is a linear combination of the $\boldsymbol{x}$ sensors providing a perfect estimation of the $\boldsymbol{s}_1$ source. The experimental question concerns the ability of the AVSS algorithm to find this solution, compared with traditional BSS algorithms, which are known to be theoretically able to solve the problem.

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, generating both a training set and a test set. The corrupting signals consist in continuous meaningful sentences uttered by 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[1] (Lallouache, [17]). 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). The number of gaussians in each mixture is set to 16.

3.2. Methodology

The AV criterion $J_{\text{avT}}(\boldsymbol{y})$ (Eq. 5) is optimized by a relative gradient algorithm (Cardoso & Laheld, [18]). We tested several *N*x*N* mixtures, with N =2, 3, et 5, where $\boldsymbol{s}_1$ 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 ***A***1 and ***A***2 (see Table 1), and we used several temporal integration widths T with T=1, 10 and 20 frames. For each mixture, the N observations are defined by:

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

which are characterized (the sources being normalized in energy) by input *SNR*s, in reference to $\boldsymbol{s}_1$, provided by:

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

Table 1: Input SNRs (dB)

	2 sources		3 sources		5 sources	
	A1	*A2*	*A1*	*A2*	*A1*	*A2*
Sens. 1	-1.16	-14	+2.70	+0.73	-6.53	-10.6
Sens. 2	-1.58	-19.1	+3.20	+0.1	-6.47	-3.67
Sens. 3	-	-	+9.60	+2.15	-14.8	-8.43
Sens. 4	-	-	-	-	-3.34	-16.1
Sens. 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 ***B***, the procedure consists in computing ***y=Bx***, in estimating the spectrum $\boldsymbol{Y}_1$ according to the process described in section 3.1 (spectral analysis followed by projection on the selected principal components), and in computing the probability $p_{\text{av}}(\boldsymbol{Y}_1,\boldsymbol{V}_1)$ thanks to the model described in section 2.2. The optimal ***B*** matrix, which minimizes the integrated criterion $J_{\text{avT}}(\boldsymbol{y})$, produces an output $\boldsymbol{y}_1$ which is the best estimation of the source $\boldsymbol{s}_1$. The output *SNR* is given by:

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

where ***G*** is the global matrix defined by : ***G=BA***.

Finally, we tested the same mixtures with the classical BSS algorithm JADE (Cardoso, [19]). An important drawback of BSS algorithms is their indetermination in respect to permutation of sources. We shall come back to this problem in Section 5. In the present assessment of JADE, for taking into account possible indeterminations, we systematically selected the best $SNR_{\text{out}}(n)$ for the signal $\boldsymbol{s}_1$ among all output sensors:

$$SNR_{\text{out}} = \arg\max_{n} 10\log(g_{n1}^2 / \sum_{p=2}^{N} g_{np}^2) \qquad (9)$$

3.3. Results

The results are displayed in Table 2, 3, and 4 with, for each case, the mean output *SNR* averaged over the 2495 test frames. Remember that there is a perfect solution, hence output *SNRs* can be arbitrarily high in all conditions.

Table 2 : Mean output SNRs in the 2x2 case (dB)

		AVSS	JADE
A1	*T=1*	14.6	12.7
	T=10	35.3	27.5
	T=20	40.6	32.2

		AVSS	JADE
A2	*T=1*	14.1	12.9
	T=10	35.6	27.5
	T=20	40.6	33.2

Table 3: Mean output SNRs in the 3x3 case (dB)

		AVSS	JADE
A1	*T=10*	25.3	19.8
	T=20	33.8	25.0

		AVSS	JADE
A2	*T=10*	25.9	19.9
	T=20	33.2	25.0

Table 4: Mean output SNRs in the 5x5 case (dB)

		AVSS	JADE
A1	*T=20*	27.6	19.6

		AVSS	JADE
A2	*T=20*	22.8	19.4

From these data, three main features appear:

- *Role of integration width*: It is clear that increasing T improves the performances, both for AVSS and JADE. The reason for AVSS is that the integration in Eq. (5) allows to smooth the variations of

[1] This face processing system exploits a chroma-key process on lips with blue make-up, and with carefully controlled head position and light.

$J_{\mathrm{avT}}(\mathbf{y})$ and remove spurious local minima. 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*: In all cases, AVSS outperforms JADE by 2 to 8 dBs. 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. However, JADE is still significantly quicker (this is a well-known property of this algorithm).

- *Equivariance*: Equivariance is the property displayed by a source separation algorithm when its performances do not depend on the mixing matrix, but just on the geometry of input sources. It should be realized by implementing a " relative gradient technique" (Cardoso & Laheld, [18]). The algorithm implemented in JADE allows a remarkable stability of output *SNR*s from one mixing matrix to the other (compare $\boldsymbol{A}1$ and $\boldsymbol{A}2$ in all cases). Though we implemented a relative gradient descent in AVSS, equivariance is not so well achieved, probably because of sensitivity to initial conditions.

4. 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. The experimental question now concerns the compared ability of AVSS vs. BSS to find good estimates of $\boldsymbol{s}_1$.

4.1. Maximizing *SNR* through AV fit

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 can't 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:

$$J_{\mathrm{lms}} = \left(\frac{\vec{y}_1}{\|\vec{y}_1\|} - \frac{\vec{s}_1}{\|\vec{s}_1\|} \right)^2 \qquad (10)$$

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

$$\mathrm{E}\left[\left(\frac{y_1(t)}{\|y_1(t)\|} - \frac{s_1(t)}{\|s_1(t)\|} \right)^2 \right] = \int_{-\infty}^{+\infty} \left| \frac{y_1(f)}{\|y_1(t)\|} - \frac{s_1(f)}{\|s_1(t)\|} \right|^2 df \qquad (11)$$

If we assume that the phases of $y_1(f)$ and $s_1(f)$ are equal, we can express J_{lms} as a spectral distance between $\boldsymbol{y}_1$ and $\boldsymbol{s}_1$, or, if we perform a discrete approximation of the Fourier transform by a filter bank:

$$\mathrm{E}\left[\left(\frac{y_1(t)}{\|y_1(t)\|} - \frac{s_1(t)}{\|s_1(t)\|} \right)^2 \right] \approx \sum_{f=1}^{F} \left(\frac{|y_1(f)|}{\sqrt{\sum_{f=1}^{F} |y_1(f)|^2}} - \frac{|s_1(f)|}{\sqrt{\sum_{f=1}^{F} |s_1(f)|^2}} \right)^2 \qquad (12)$$

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 audiovisual approximation of the criterion in Eq. (3), should be able to maximize *SNR* in the estimation of $\boldsymbol{s}_1$. The temporal integration replacing Eq. (4) by 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, it should result in maximizing *SNR* on this integrated window.

4.2. Methodology

We used the same data and paradigm as in the previous section. However, the AV learning procedure was slightly different. Indeed, while the $P=N$ case appeared to be quite robust to a number of variations in the used criterion, the $P<N$ case imposed to precisely define the AV criterion such as to stick as much as possible to the spectral criterion in Eq. (12). Therefore, we had to drop the conversion of psds in dBs, and to normalize spectral parameters over the total temporal width used in Eq. (5). We performed various AV associations with the same technique as in Section 3.1, with one association per integration width. We studied only a 3 sources – 2 sensors case, and estimation was done by exhaustive search of the criterion maximum instead of gradient descent, just to verify in this preliminary set of experiments that there is indeed the good information in the AVSS algorithm to efficiently estimate $\boldsymbol{s}_1$.

4.3. Results

We selected a test case with the matrix:

$$\boldsymbol{A} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

This matrix provides a plan ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$) on which the three sources $\boldsymbol{s}_1$, $\boldsymbol{s}_2$ and $\boldsymbol{s}_3$ project at 120° angles. In consequence, this is the worst case for a BSS algorithm, which has no reason to "prefer" one source over another. In this condition, the input *SNR*s are equal to 0 dB on sensor 1, and $-\infty$ on sensor 2. The optimal output *SNR* value minimizing Eq. (10) is 3 dB (solution: $y = x_1+0.5x_2$).

In Table 5, we display the difference between the optimal *SNR* (3 dB) and the mean output *SNR* for JADE vs. AVSS, at two integration width, that is 20 and 60 frames (400 ms vs. 1.2 s of signal). We also provide the score for the filterbank criterion defined by Eq. (12).

Table 5: Mean decrease in output SNRs (dB) in relation to the optimal 3dB solution in the 3x2 case

	Filterbank	AVSS	JADE
T = 20	0.2	1.65	1.35
T = 60	0.06	0.41	1.45

First, we notice that the phase condition enabling to replace the temporal criterion (10) by a spectral criterion (12) is not too drastic. Indeed, the filterbank solution is very good, and almost perfect for a 60-frames integration (the decrease in output *SNR* compared with the optimal 3 dB value is almost 0). Secondly, we notice that JADE does no more benefit of the temporal integration: indeed, the *SNR* loss compared with the optimal solution is blocked around 1.5 dB (half the 3 dB gain from the 0 dB input value to the optimal 3 dB output value). On the contrary, while AVSS performs slightly less well than JADE at the first integration width because of the lack of precision of the AV correspondence, it benefits of temporal integration to clearly outperform JADE at the larger width. Of course, these results are very preliminary, and should be extended towards a large number of simulations. But they are very encouraging, since they show clearly that the AV coherence provides an efficient way to focus the extraction process towards the target source.

5. A combined JADE – AVSS algorithm

The previous sets of experiments compared a prototypical and already very efficient BSS algorithm, JADE, and an original AVSS algorithm currently under development. However, we are convinced that the properties of these two approaches are very complementary, and that some fusion should be attempted at some stage. As a first try, let us report on a basic simple fusion process. Indeed, JADE is a very rapid and powerful algorithm for source separation, in spite of the limitations displayed in the previous sections. But is suffers from a severe drawback already mentioned: the indetermination in respect to permutation of sources. The result is the impossibility to know where (i.e. on what output sensor) is a given source $\boldsymbol{s}_1$ extracted, the more so since 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 6, we display the number of cases where there was a switch in the selected sensor where $\boldsymbol{s}_1$ appeared, between two consecutive frames in the *N*x*N* cases. It appears that the instability may be really large and lead to severe difficulties in the application of the algorithm.

Table 6: Number of JADE permutations for different NxN mixtures and 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

Therefore, we attempted to apply the AVSS criterion defined by Eq. (5) at the output of JADE, to select the sensor providing the signal most compatible with the source to extract, in terms of AV coherence. In Table 7, we display the results of this selection process in various conditions. We observe that, with 20 frames in the *N*x*N* problem, we select the best sensor (in terms of maximal output *SNR*) in more than 99% of the cases, which leads to a mean output *SNR* almost equal to the value obtained by selecting at each frame the best *SNR* sensor (which is of course impossible to do in a real application). Of course, in the 3x2 problem studied in Section 4, the results are less convincing for a simple reason: it is precisely a case in which the various sensors provide rather close outputs in terms of *SNR*, hence the AVSS selection process is less efficient, unless 60 frames are integrated in the computation of the AV criterion.

Table 7: Performance of the JADE + AV selection algorithm

	% incorrect selection	Mean output SNR loss due to sensor selection from AV probability
2x2, *A1* *T=20*	0.4 %	0.2 dB
2x2, *A2* *T=20*	0.8 %	0.2 dB
3x2, *T=60*	3.8 %	0.35 dB

6. Conclusion

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 now show that the method seems able to deal with the “less sensors than sources” case, thanks to its ability to focus on the source to extract. 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, the route is still long towards a complete demonstration of the efficiency of the technique. It will involve larger multi-speaker corpora, more powerful learning tools for AV association, and it should address the crucial problem of convolutive mixture. But it is already possible 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.

7. References

[1] Bernstein, L.E., Benoît, C. (1996). For speech perception by humans or machines, three senses are better than one. Proc. ICSLP'96, 1477-1480.

[2] Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am., 26, 212-215.

[3] Petajan, E.D. (1984). Automatic lipreading to enhance speech recognition. Doct. Thesis, University of Illinois.

[4] 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.

[5] 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. Proc. AVSP'2001, 127-131.

[6] Schwartz, J.L., Berthommier, F., & Savariaux, C. (2002). Audio-visual scene analysis; Evidence for a "very-early" integration process in audio-visual speech perception. Proc. ICSLP'2002, 1937-1940.

[7] Schwartz, J.L., Berthommier, F., & Savariaux, C. (2003). See to better hear; Evidence for "very early" audiovisual interactions in speech perception". This Workshop.

[8] 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, Proc. AVSP'97, 37-40.

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

[10] Deligne, S., Potamianos, G., & Neti, Chapalathy (2002). Audio-Visual speech enhancement with AVCDCN (AudioVisual Codebook Dependent Cepstral Normalization). Proc. ICSLP'2002, 1449-1452.

[11] Okuno, H.G., Nakadai, K., Lourens, T., and Kitano, H. (2001). Separating Three Simultaneous Speeches with Two Microphones by Integrating Auditory and Visual Processing. Proc Eurospeech 2001, 2643-2646.

[12] Jutten, C., & Herault, J. (1991). Blind separation of sources. Part I: An adaptive algorithm based on a neuromimetic architecture. Signal Processing, 24, 1-10.

[13] Taleb, A., & Jutten, C. (1999). Source separation in postnonlinear mixtures. IEEE Trans SP, 10, 2807-2820.

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

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

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

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

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

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