AUDIO-VISUAL SPEECH SOURCES SEPARATION

A new approach exploiting the audio-visual coherence of speech stimuli[(1)]

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

Speech Communication Institute (ICP), CNRS UMR 5009, INPG / Université Stendhal, Grenoble France
(*) Image and Signal Processing Laboratory (LIS), CNRS UMR 5083, INPG / UFF, Grenoble France
sodoyer@icp.inpg.fr

ABSTRACT

We present a new approach to the source separation problem in the case of multiple speech signals. The method is based on the use of automatic lipreading: the objective is to extract an acoustic speech signal from other acoustic signals by exploiting its coherence with the speaker's lip movements. We show how, if a statistical model of the joint probability of visual and spectral audio input is learnt to quantify the audio-visual coherence, separation can be achieved by maximising this probability. Then, we present a number of separation results on a corpus of vowel-plosive-vowel sequences uttered by a single speaker, embedded in a mixture of other voices.

1. INTRODUCTION

A number of recent experiments suggest that the audio-visual interaction in speech perception could begin at a very early level, in which the visual input could improve the detection of speech sounds embedded in noise. In a companion paper, we show that this could be part of an "audiovisual scene analysis" module improving speech intelligibility independently of the contribution of lipreading *per se* [1]. In this paper, we study the technological counterpart of this idea: is it possible to enhance the speech sound embedded in various kinds of noise, thanks to a computational process based on the audiovisual coherence? A first work was made on this line, thanks to an original filtering approach [2]. The present work explores a probably more powerful idea: adapt blind source separation (BSS) techniques to audiovisual speech sources.

2. THEORETICAL CONSIDERATIONS

2.1. Architecture

Let us consider the case of a stationary additive mixture of sources, to be separated. The input of an *N*-signals *P*-sensors separation system consists of a set of *P* observations $\boldsymbol{x_j(t)}$, each of them being a mixture of *N* unknown signals $\boldsymbol{s_i(t)}$ to be separated. $\boldsymbol{A}$ is the unknown *(P,N)* mixing matrix, $\boldsymbol{B}$ is the *(N,P)* separation matrix to estimate in order to recover the output signals $\boldsymbol{y_k(t)}$ as close as possible to the sources $\boldsymbol{s_i(t)}$. In our application, these $\boldsymbol{s_i(t)}$ signals are speech acoustic signals, and we assume as many sources as observations, that is *P=N*.

In BSS, the separation coefficients (i.e. the $\boldsymbol{B}$ coefficients) are estimated according to a criterion of maximisation of the independence between the outputs [e.g. 3]. In this study, 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 $\boldsymbol{s_1}$ to be isolated. Typically, $\boldsymbol{V_1}$ contains the trajectory of basic geometric lip shape parameters, which can be automatically extracted by different systems developed in our laboratory [4, 5]. In the present paper, we shall focus on the extraction of one audio-visual source merged in a mixture of two or more acoustic signals (Fig. 1).

2.2. Computational foundations

Most BSS techniques are based on the assumption that the sources are non Gaussian, independent and stationary. In our case, we restrict the independence assumption to a simple *decorrelation*, and add some knowledge on the first source $\boldsymbol{s_1}$, in order to extract it from the mixture. What we know about $\boldsymbol{s_1}$ is the *visual signal* associated with it (the visible speaking face), and it is classical to consider that the visual input is partially linked to the transfer function of the vocal tract. Hence we assume that the additional knowledge about $\boldsymbol{s_1}$ concerns *spectral envelope*. We shall address two possible means to introduce spectral information, through autocorrelation coefficients, or through energy coefficients at the output of a filterbank.

2.2.1. Introducing autocorrelation coefficients in source separation algorithms

To begin with, let us assume that we know something linked to the spectrum, that is a normalised autocorrelation coefficient:

$$\gamma_{\boldsymbol{k}} = \frac{\boldsymbol{R}_{s_1s_1}(\boldsymbol{k})}{\boldsymbol{R}_{s_1s_1}(0)} \qquad (1)$$

where $R_{s1s1}(k)$ is the autocorrelation of the source $\boldsymbol{s_1}$ for a delay *k*, and $R_{s1s1}(0)$ is the same for delay *0*, that is the source power. To simplify further computations, let us introduce the function:

$$\boldsymbol{C_k}(\boldsymbol{y_i y_j}) = \boldsymbol{R}_{y_iy_j}(\boldsymbol{k}) - \gamma_{\boldsymbol{k}} \boldsymbol{R}_{y_iy_j}(0) \qquad (2)$$

At the solution, we expect that one output of the algorithm, say $\boldsymbol{y_1}$, will provide an estimate of $\boldsymbol{s_1}$. In this case, we should obtain:

$$\frac{\boldsymbol{R}_{y_1y_1}(\boldsymbol{k})}{\boldsymbol{R}_{y_1y_1}(0)} = \gamma_{\boldsymbol{k}} \qquad (3)$$

Therefore, we can decide to minimise the following criterion:

$$\boldsymbol{f_{AC}}(\boldsymbol{y}) = (\boldsymbol{R}_{y_1y_1}(\boldsymbol{k}) - \gamma_{\boldsymbol{k}} \boldsymbol{R}_{y_1y_1}(0))^2 = \boldsymbol{C_k}(y_1y_1)^2 \qquad (4)$$

This criterion meets the basic requirement that it is positive or null, and minimum (equal to zero) when the separation is achieved in the restricted sense we consider in the paper, that is when $\mathbf{s_1}$ is separated ($\boldsymbol{y_1}=\boldsymbol{s_1}$). In the case of two sources, it is easy to see that this criterion ensures separation, while for more than two sources, say *N*, we need to know (*N-1*) autocorrelation terms for (*N-1*) delays, which leads to the generalised criterion:

$$\boldsymbol{f_{AC}}(\boldsymbol{y}) = \sum_{\boldsymbol{k}=1}^{\boldsymbol{N}-1} \boldsymbol{C_k}(\boldsymbol{y}_1\boldsymbol{y}_1)^2 \qquad (5)$$

2.2.2. Replacing autocorrelation by spectral coefficients

In the same vein, we can assume that, instead of autocorrelation functions, what we know about $\boldsymbol{s_1}$ is a number of spectral components, defined by a filter bank. Let $\boldsymbol{H_k(f)}$ be the frequency response of a bandpass FIR filter, and $\boldsymbol{h_k(t)}$ 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 $\boldsymbol{h_k * s_1(t)}$. Hence we can assume that we know the normalised energy of $\boldsymbol{s_1}$ in the band corresponding to the filter, that is:

$$\gamma_{\boldsymbol{h_k}} = \frac{\boldsymbol{R}_{(\boldsymbol{h_k s}_1)(\boldsymbol{h_k s}_1)}(0)}{\boldsymbol{R}_{\boldsymbol{s}_1\boldsymbol{s}_1}(0)} \qquad (6)$$

As in the previous case, we can introduce the function:

$$\boldsymbol{C_{h_k}}(\boldsymbol{y_i y_j}) = \boldsymbol{R}_{(\boldsymbol{h_k y_i})(\boldsymbol{h_k y_j})}(0) - \gamma_{\boldsymbol{h_k}} \boldsymbol{R}_{\boldsymbol{y_i y_j}}(0) \qquad (7)$$

and a suitable criterion is provided, similarly to Eq. (5), by:

$$\boldsymbol{f_{SC}}(\boldsymbol{y}) = \sum_{\boldsymbol{k}=1}^{\boldsymbol{N}-1} \boldsymbol{C_{h_k}}(\boldsymbol{y}_1\boldsymbol{y}_1)^2 \qquad (8)$$

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 $\boldsymbol{s_i}$ are different from each other [6].

2.2.3. The audio-visual algorithm

In the case of our application, we do not have at our disposal the exact spectral components of the source $\boldsymbol{s_1}$, but only indirect indications about 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 best be described in statistical terms (see e.g. [7]). Hence, we assume that we can build a statistical model providing the joint probability of a video vector $\boldsymbol{V}$ containing parameters describing the speaking face (e.g., lip characteristics) and of an audio vector $\boldsymbol{S}$ containing spectral characteristics of the sound. Let us call this joint probability $p(\boldsymbol{S},\boldsymbol{V})$. This statistical model is not given for free: it must be designed from a learning corpus. In the present study, we define the probability $p(\boldsymbol{S},\boldsymbol{V})$ as a mixture of Gaussian kernels, and we use the learning corpus to estimate the mean, covariance matrix and weight of each Gaussian kernel, by iterating an Expect. Max. (EM) algorithm.

Then the separation algorithm consists in selecting 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 the following criterion:

$$\text{maximise } \boldsymbol{f_{AV}}(\boldsymbol{y}) = \boldsymbol{p}(\boldsymbol{Y}_1, \boldsymbol{V}_1) \qquad (9)$$

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. Therefore, we introduce the possibility to cumulate the probabilities over time. For this aim, we assume for simplicity that values of audio and visual characteristics at several consecutive time frames are independent from each other, and we define accordingly the cumulated joint audio-visual probability by:

$$\boldsymbol{p}(\boldsymbol{Y}_1(\boldsymbol{t},\ldots,\boldsymbol{t}-\boldsymbol{T}), \boldsymbol{V}_1(\boldsymbol{t},\ldots,\boldsymbol{t}-\boldsymbol{T})) = \boldsymbol{p}(\boldsymbol{Y}_1(\boldsymbol{t}), \boldsymbol{V}_1(\boldsymbol{t}))\ldots\boldsymbol{p}(\boldsymbol{Y}_1(\boldsymbol{t}-\boldsymbol{T}), \boldsymbol{V}_1(\boldsymbol{t}-\boldsymbol{T})) \qquad (10)$$

This product of joint probabilities, for various lengths of temporal integration *(T+1)*, is maximised, instead of criterion (9), to find a better estimation of the separating matrix.

3. EXPERIMENTAL RESULTS

3.1. Data

The audio-visual corpus we attempted to separate from noise consisted of $V_1CV_2CV_1$ sequences uttered by a French speaker. V_1 and V_2 were vowels within [a, i, y, u]. C was within the plosives set [p, t, k, b, d, g, #]. The 112 sequences (4xV_1, 7xC, 4xV_2) were pronounced twice by a single speaker, to generate both a training and a test set. The corrupting signals consisted in continuous meaningful sentences uttered by the same speaker.

The video data consisted in two basic geometric parameters describing the speaker's lip shape, namely internal width (*LW*) and height (*LH*) of the labial contour (see Fig. 2, right box). These parameters were automatically extracted every 20 ms by using the ICP face processing system [4]. Sounds were sampled at 16 kHz. On the same 20-ms windows synchronous with the video analysis, we computed 32 spectral parameters providing power spectral densities (Psd) at the output of a bank of 32 filters linearly spaced between 0 and 5 kHz. Psds were converted in dBs, and a principal component analysis (PCA) was 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 was 14 (12 audio + 2 video).

3.2. Statistical model of the *p(S,V)* probability

The EM Gaussian mixture algorithm was applied to the training data set, containing 2497 audio-visual vectors (112 stimuli, about 24 vectors per stimulus). A preliminary study showed that increasing the number of PCA audio dimensions and of Gaussian laws in the mixture slightly increased performances [8]. In the present work we used 16 Gaussian laws. On Figure 2, we display the projections of the Gaussian covariance matrices on the two video dimensions and on the first two audio dimensions. They can be interpreted in the following manner.

The video space displays a quite classical organisation, with closed lip shapes (bilabials in any context, Gaussian law 1), rounded lip shapes ([y], [u] and dentals and velars in [y]/[u] context, Gaussian laws 2, 3, 4 and 5), spread lip shapes ([i], Gaussian law 8) and opened lip shapes ([a], Gaussian law 16). The other Gaussian laws model the open-to-close and close-to-open gestures of the jaw and lips between these targets. This configuration confirms the basic property of audio-visual speech, that is the complementarity between the two modalities: visually close stimuli are auditorily well separated and vice versa. Thus, different Gaussian kernels of the model whose projection on two specific audio-visual dimensions are confused can be clearly separated when projected on two other dimensions. For example, the four Gaussian kernels 2, 3, 4 and 5 are confused in the (*LW*, *LH*) space around the [y]/[u] round-closed lip shape, while separated in the audio subspace with one kernel around [u] (Gaussian 2), one around [y] (Gaussian 3) and the other two for dentals and velars in rounded context. On the other hand, Gaussian kernels 5 and 13, close in the audio space, are clearly separated in the video space. As we said, this complementarity is essential for the efficiency of our approach.

3.3. Experimental procedure

Most of our study dealt with two-sources mixtures, defined by:

$$x_1 = a_{11}s_1 + a_{12}s_2 \qquad x_2 = a_{21}s_1 + a_{22}s_2 \qquad (11)$$

$\boldsymbol{s_1}$ is the speech source to be separated, $\boldsymbol{s_2}$ is a corrupting speech source to eliminate. Sources were normalised in energy. Hence, the input *SNR*s on each sensor $\boldsymbol{x_i}$ are given by:

$$\boldsymbol{SNR_{input}}_{\_1} = 20\log(\boldsymbol{a}_{11}{}^2 / \boldsymbol{a}_{12}{}^2)$$

$$\boldsymbol{SNR_{input}}_{\_2} = 20\log(\boldsymbol{a}_{21}{}^2 / \boldsymbol{a}_{22}{}^2) \qquad (12)$$

while it is easy to show that the output SNR on $\boldsymbol{y_1}$ is given by:

$$\boldsymbol{SNR_{output}} = 20\log((\boldsymbol{a}_{11} + \boldsymbol{ca}_{21})^2 /(\boldsymbol{a}_{12} + \boldsymbol{ca}_{22}{}^2)) \qquad (13)$$

where *c* is defined as the ratio b_{12}/b_{11}, which controls the $\boldsymbol{y_1}$ output. We tested two mixture matrices. The first one (Mixt. 1 in the following) provided input *SNR* values respectively of –1.2 and –1.6 dB on the two sensors, while the second one (Mixt. 2) respectively provided –14 and –19 dB. The evaluation was made by concatenating all 112 stimuli of the test set (see Section 3.1) into a single file containing 2495 audio-visual frames. For each test frame, and for a given separating matrix ***B***, the procedure consisted in computing *y* = ***B****x*, estimating the spectrum $\boldsymbol{Y_1}$ according to the process described in Section 3.1 (spectral analysis followed by PCA), and computing the probability $p(\boldsymbol{Y_1},\boldsymbol{V_1})$ thanks to the model described in section 3.2, in order to determine the ***B*** matrix maximising this probability. The optimal ***B*** matrix produces an output $\boldsymbol{y_1}$ supposed to provide the best estimation of the source $\boldsymbol{s_1}$. We shall describe in the next section the optimisation procedure.

3.4. Results

Firstly we performed “exhaustive scans”, that is examination of the variations of either the instantaneous version of joint probability $p(\boldsymbol{Y_1(t)},\boldsymbol{V_1(t)})$ (Eq.9) or its temporally integrated version $p(\boldsymbol{Y_1(t\ldots t\text{–}T)},\boldsymbol{V_1(t\ldots t\text{–}T)})$ (Eq.10) when the control parameter *c* was systematically varied. We display on Fig. 3 the variations of the logarithm of inverse probability values for Mixt. 2 and for one frame, for which it appears that the best *c* value (minimum of the curve) is not at the theoretical solution when instantaneous probabilities are considered (Fig. 3, left). This is due to the fact that some visual frames are quite ambiguous in terms of associated spectra (many possible spectra for a lip shape: this is what is called “visemes”, [9]). When temporal integration from 5- to 10-frame length is used, the pattern is much improved (Fig. 3, centre and right).

Then, we implemented an automatic procedure for searching the optimal ***B*** matrix maximising the instantaneous or temporally integrated versions of $p(\boldsymbol{Y_1},\boldsymbol{V_1})$. We used an equivariant algorithm exploiting relative gradient and a serial updating technique with orthogonal contrast [3]. For each mixture, and for three integration lengths 1, 5 and 10 in Eq. 10, we processed the 2495 test frames with this algorithm. For each frame, the algorithm was initialised at the value of the previous frame, and at convergence we computed the output *SNR* value. On Fig. 4 we display the cumulated histograms of output *SNR* values: we observe that for both mixtures, more than 95% of the test frames are separated with an output *SNR* larger than 20 dBs for a 10-frame temporal integration, while input *SNRs* were all lower than 0 dBs. The results are much poorer for smaller temporal integration length *T*, and particularly for *T*=1. Furthermore, the mean number of iterations of the gradient algorithm towards convergence is also much decreased (around 10 times less for *T*=10 than for *T*=1). Therefore, altogether, the convergence time is not increased by temporal integration, while the performances are dramatically improved.

4. CONCLUSION

It appears that 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=10 corresponds to 200 ms, which is rather low), and lead to very good quality of the separated stimuli.

Of course, it must be acknowledged that classical (pure audio) BSS algorithms are able to separate such mixtures with quite the same performances [8]. But this work is promising because we can expect to proceed in the future with much less ideal configurations, and particularly with less sensors than sources, a case in which the visual information should enable to better focus on a particular source and improve its enhancement/separation. Future works should also consider the combination of this approach with standard ICA methods.

Note (1): An enlarged version of this paper has been submitted for publication to EURASIP [6].

REFERENCES

[1] J.L. Schwartz, F. Berthommier & C. Savariaux (2002). Audio-visual scene analysis. ICSLP’2002 (subm.).

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

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

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

[5] 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.

[6] J.L. Schwartz, L. Girin, D. Sodoyer, J. Klinkisch, & C. Jutten (subm.). Separation of audio-visual speech sources. *Eurasip, special issue on AV processing.*

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

[8] D. Sodoyer, L. Girin, C. Jutten, & J.L. Schwartz (subm.). Séparation de sources audio-visuelles. XXIV JEPs, Nancy.

[9] 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.

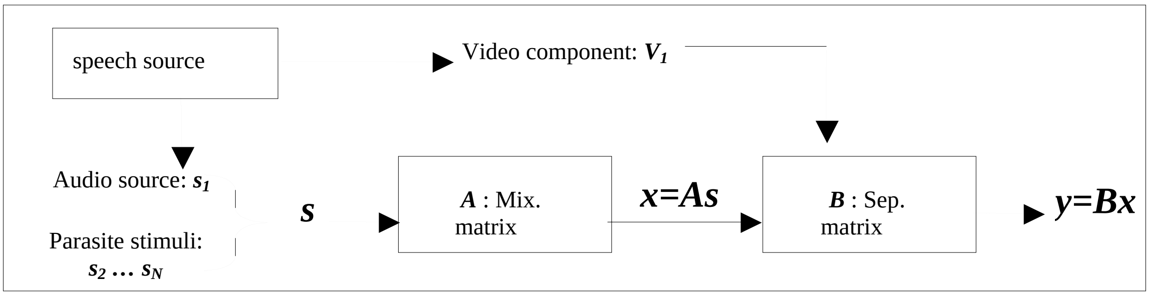

Figure 1 – The audio-visual source separation system

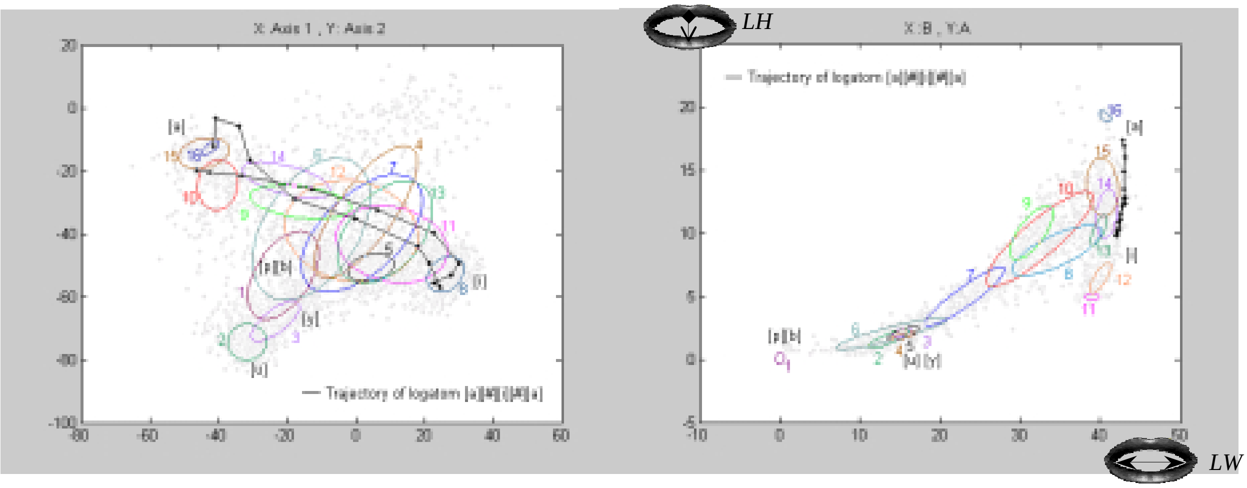

Figure 2 – The audio-visual statistical model (see text)

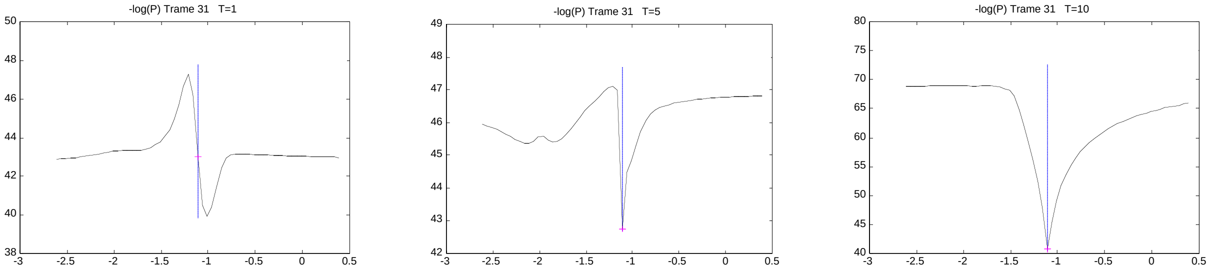

Figure 3 – Variations of the audio-visual probability $p(Y_1,V)$ for a range of c values around the theoretical solution c=–1.11 for mixture 2, and for three temporal integration lengths T: from left to right, T=1, 5 and 10.

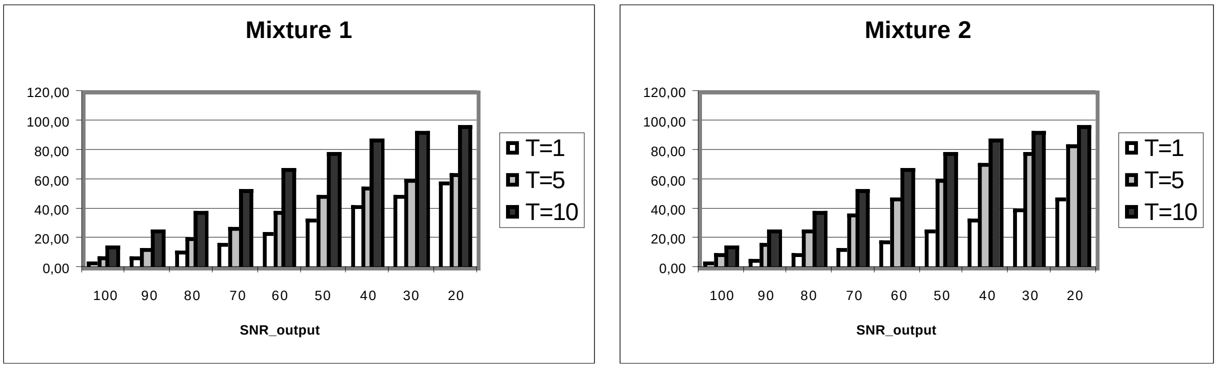

Figure 4 – Separation results: Cumulated histograms (in %) of output *SNR* values on the whole test corpus, for the two mixtures and for compared temporal integration length values *T*.