EURASIP Journal on Applied Signal Processing 2002:11, 1165–1173

Separation of Audio-Visual Speech Sources: A New Approach Exploiting the Audio-Visual Coherence of Speech Stimuli

David Sodoyer

Institut de la Communication Parlée, Institut National Polytechnique de Grenoble, Université Stendhal, CNRS UMR 5009, ICP, INPG, 46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Email: sodoyer@icp.inpg.fr

Jean-Luc Schwartz

Institut de la Communication Parlée, Institut National Polytechnique de Grenoble, Université Stendhal, CNRS UMR 5009, ICP, INPG, 46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Email: schwartz@icp.inpg.fr

Laurent Girin

Institut de la Communication Parlée, Institut National Polytechnique de Grenoble, Université Stendhal, CNRS UMR 5009, ICP, INPG, 46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Email: girin@icp.inpg.fr

Jacob Klinkisch

Institut de la Communication Parlée, Institut National Polytechnique de Grenoble, Université Stendhal, CNRS UMR 5009, ICP, INPG, 46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Email: jacob.klinkisch@gmx.de

Christian Jutten

Laboratoire des Images et des Signaux, Institut National Polytechnique de Grenoble, Université Joseph Fourier, CNRS UMR 5083, LIS, INPG, 46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Email: christian.jutten@inpg.fr

Received 19 October 2001 and in revised form 7 May 2002

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 consider the case of an additive stationary mixture of decorrelated sources, with no further assumptions on independence or non-Gaussian character. Firstly, we present a theoretical framework showing that it is indeed possible to separate a source when some of its spectral characteristics are provided to the system. Then we address the case of audio-visual sources. 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 maximizing this probability. Finally, 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. We show that separation can be quite good for mixtures of 2, 3, and 5 sources. These results, while very preliminary, are encouraging, and are discussed in respect to their potential complementarity with traditional pure audio separation or enhancement techniques.

Keywords and phrases: blind source separation, lipreading, audio-visual speech processing.

1. INTRODUCTION

There exists an intrinsic coherence and even a complementarity between audition and vision for speech perception [1]. Indeed, the phonetic contrasts least robust in auditory perception in acoustic noise are the most visible ones, both for consonants and vowels [2]. Thus, visual cues can compensate to a certain extent the deficiency of the auditory ones. This explains that the fusion of auditory and visual information meets a great success in several speech applications, mainly in speech recognition in noisy environments [3].

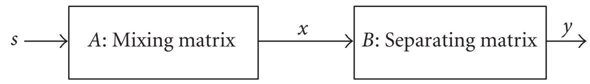


Figure 1: The source separation problem.

In a previous work [4], we tested a slightly different idea, we presented a prototype system which was able to exploit the visual input to *enhance* the audio signal corrupted by acoustic additive white noise. The principle was to estimate enhancing filters from both lip shape and noisy acoustic information. This paper is an extension of this work to the more complex case of a mixture of speech signals (the cocktail-party effect). The goal is to separate such signals, that is to recover the individual signals from the mixture. This problem, generalised to any kind of signals under the label *source separation*, has recently met a great success in the signal processing community. Many methods have been proposed, most of them being based on independence criteria between the mixed signals, leading to the concept of independent components analysis (ICA) [5]. In this paper, we propose a new approach in the case of speech signal separation. This approach, presented in Section 2, is based on the use of the bimodality of speech and on the intrinsic coherence between audio and video speech. Results are provided in Section 3. Firstly, we present our audio-visual speech corpus together with the statistical model used to characterise the audio-visual speech coherence on this corpus. Then we show that the audio-visual separation technique is indeed very promising, and compares well with classical ICA techniques. In Section 4, we conclude by presenting some perspectives of improvement and further development of this new technique.

2. AN ARCHITECTURE FOR SEPARATING AUDIO-VISUAL SPEECH SOURCES

2.1. *The problem*

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 $x_j(t)$, each of them being a mixture of N unknown signals $s_i(t)$ to be separated. A is the unknown (P, N) mixing matrix, B is the (N, P) separation matrix to estimate in order to recover the output signals $y_k(t)$ as close as possible to the sources $s_i(t)$ (Figure 1). In our application, these $s_i(t)$ signals are speech acoustic signals, and we assume as many sources as observations, that is $P = N$.

In ICA, the separation coefficients (i.e., the B coefficients) are estimated according to a criterion of maximization of the independence between the outputs (e.g., [6]). In this study, 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 to be isolated. Typically, V_1 contains the trajectory of basic geometric lip shape parameters, which can be automatically extracted by different systems developed in our laboratory [7, 8]. In the present paper, we will focus on the extraction of one audio-visual source merged in a mixture of two or more acoustic signals (Figure 2).

2.2. *Theoretical considerations*

Most ICA techniques are based on the assumption that the sources are non Gaussian, independent and stationary. In our case, we will attempt to restrict the independence assumption to a simple *decorrelation*, and add some knowledge on the first source s_1, in order to extract it from the mixture. What we know about 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 will assume that the additional knowledge about s_1 concerns its *spectral envelope*. We will address here 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*

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

$$\gamma_k = \frac{R_{s_1 s_1}(k)}{R_{s_1 s_1}(0)}, \tag{1}$$

where $R_{s_1 s_1}(k)$ is the autocorrelation of the acoustic source s_1 for a delay k, and $R_{s_1 s_1}(0)$ is the same for delay 0, that is the source power. To simplify further computations, we introduce the function

$$C_k(y_i y_j) = R_{y_i y_j}(k) - \gamma_k R_{y_i y_j}(0). \tag{2}$$

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

$$\frac{R_{y_1 y_1}(k)}{R_{y_1 y_1}(0)} = \gamma_k. \tag{3}$$

Therefore, we can decide to minimize the following criterion:

$$f_{AC}(y) = \left(R_{y_1 y_1}(k) - \gamma_k R_{y_1 y_1}(0)\right)^2 = C_k(y_1 y_1)^2. \tag{4}$$

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 which we consider in the paper, that is when s_1 is separated ($y_1 = s_1$). However, we must determine if the criterion ensures separation, or on the contrary if it may be zero even in nonseparated configurations. To study this point, we introduce the whole mixing-separating matrix $G = BA$, with the following notation:

$$y_p = \sum_n g_{pn} s_n. \tag{5}$$

Our separation criterion means that we expect all g_{1n} entries to be zero except the first one g_{11}. We determine what happens when the criterion $f_{AC}(y)$ is minimum, that is equal to

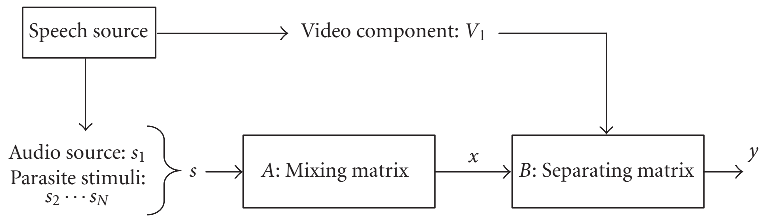


Figure 2: The audio-visual source separation system.

zero. Incorporating (5) into the definition of autocorrelation leads to

$$R_{y_1y_1}(k) = \sum_{m,n} g_{1m}g_{1n}R_{s_ms_n}(k). \tag{6}$$

Assuming that the sources are not correlated, we obtain

$$R_{y_1y_1}(k) = \sum_{n} {g_{1n}}^2 R_{s_ns_n}(k). \tag{7}$$

Introducing (7) into (4) shows that cancelling $f_{AC}(y)$ leads to

$$C_k(y_1y_1) = \sum_{n} {g_{1n}}^2 C_k(s_ns_n) = 0. \tag{8}$$

By construction, we know that $C_k(s_1s_1)$ equals 0, hence (8) becomes

$$C_k(y_1y_1) = \sum_{n\geq 2} {g_{1n}}^2 C_k(s_ns_n) = 0. \tag{9}$$

In the case of a mixture of two sources, if we assume that $C_k(s_2s_2)$ is not zero, (9) ensures the cancellation of g_{12}, which leads to the correct separation of source s_1 ($y_1 = g_{11}s_1$ such that there remains a gain unspecification). Notice that the hypothesis of a nonzero value for $C_k(s_2s_2)$ just means that the second source is assumed not to have the same autocorrelation property than the first one, which is of course necessary for separating them.

In the case of a mixture of more than two sources, the situation is not the same. Indeed, (9) is not sufficient to cancel all g_{1n} values for n from 2 to N, and we must add other constraints. For this aim, we must introduce more knowledge about source s_1, in terms of autocorrelations for other delays. More precisely, we need at least $(N-1)$ different values of delay k, for which we assume that we know the value of γ_k defined according to (1). Then, we may modify the criterion $f_{AC}(y)$ by changing (4) into

$$f_{AC}(y) = \sum_{k=1}^{N-1} C_k(y_1y_1)^2. \tag{10}$$

Here, $f_{AC}(y)$ is zero if and only if

$$C_k(y_1y_1) = 0 \quad \forall k \in \{1, \ldots, (N-1)\}. \tag{11}$$

Introducing (9) into (11) leads to

$$\begin{bmatrix} C_1(s_2s_2) & C_1(s_ns_n) & C_1(s_Ns_N) \\ C_k(s_2s_2) & C_k(s_ns_n) & C_k(s_Ns_N) \\ C_{N-1}(s_2s_2) & C_{N-1}(s_ns_n) & C_{N-1}(s_Ns_N) \end{bmatrix} \begin{bmatrix} {g_{12}}^2 \\ {g_{1n}}^2 \\ {g_{1N}}^2 \end{bmatrix} = 0. \tag{12}$$

If the matrix of $C_k(s_ns_n)$ values in the previous equation is not singular, this leads to cancel the vector of g_{1n} values for n from 2 to N, which is exactly what we want for separating s_1. The nonsingular assumption is a generalisation of the nonzero assumption for $C_k(s_2s_2)$. It means that the sources s_i must have different shapes of correlation sets $R(k)$, that is, different spectra.

2.2.2 Introducing spectral energy coefficients in source separation

In the same vein, we can assume that, instead of autocorrelation functions, what we know about s_1 is a number of spectral components, defined by a filter bank. Let $H_k(f)$ be the frequency response of a bandpass FIR filter, and $h_k(t)$ be its temporal impulse response (Figure 3). The energy of the source s_1 at the output of the filtering process is provided by the autocorrelation with zero delay of the filtered signal $h_k * s_1(t)$. Hence we can assume that we know the normalised energy of s_1 in the band corresponding to the filter, that is,

$$\gamma_{h_k} = \frac{R_{(h_ks_1)(h_ks_1)}(0)}{R_{s_1s_1}(0)}. \tag{13}$$

As in the previous case, we can introduce the function

$$C_{h_k}(y_iy_j) = R_{(h_ky_i)(h_ky_j)}(0) - \gamma_{h_k}R_{(y_iy_j)}(0), \tag{14}$$

and a suitable criterion is provided, similarly to (10), by

$$f_{SC}(y) = \sum_{k=1}^{N-1} C_{h_k}(y_1y_1)^2. \tag{15}$$

This criterion, based on a bank of $(N-1)$ band-pass filters, leads to the same kind of equation as (12), and it allows separation of the source s_1 provided that the spectra of all sources s_i are different from each other (to ensure that the matrix of $C_{hk}(s_ns_n)$ values is not singular).

In summary, this theoretical analysis shows that the knowledge of $(N-1)$ spectral components of a given source s_1 (e.g., autocorrelation or narrowband energy coefficients)

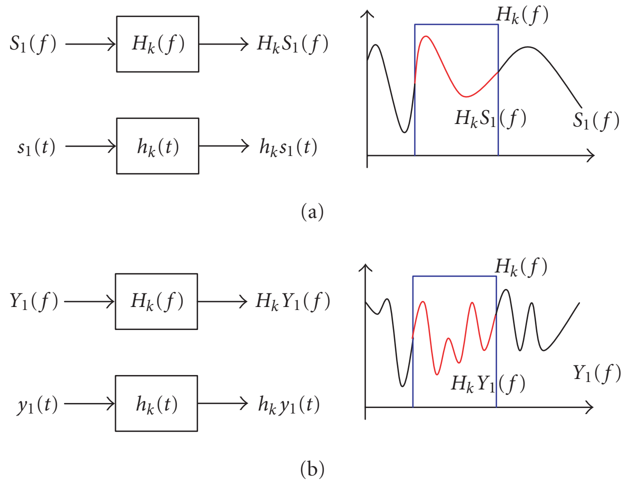


FIGURE 3: Filtering the source s_1, (a, top) or the output y_1 (b, bottom).

enables to separate the source from $(N-1)$ other decorrelated sources in an unknown additive stationary mixture, by minimizing the criterion defined in (10) or (15). We are currently studying the implementation of gradient techniques in an "equivariance" scheme [9] for the minimization of this kind of criterion [10]. Notice that all along this theoretical discussion, we discussed the evaluation of g_{1n} values, but said nothing about other g_{in} values, hence nothing about the extraction of the other sources, since it was not our objective in this study. Separating all sources would involve either spectral information on the other sources, or other ingredients based on independence cues.

In the case of our application, we do not have at our disposal the exact spectral components of the source s_1, but only indirect indications about the spectrum through lip characteristics associated to the sound s_1. In Section 2.3, we present a practical algorithm able to deal with this situation.

2.3. *The audio-visual algorithm*

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., [11]). Hence, we assume that we can build a statistical model providing the joint probability of a video vector V containing parameters describing the speaking face (e.g., lip characteristics) and of an audio vector S containing spectral characteristics of the sound. We call this joint probability $p(S,V)$. This statistical model is not given for free, it must be designed from a learning corpus. In the present study, we define the probability $p(S,V)$ 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 Expectation-Maximization (EM) algorithm.

Then the separation algorithm consists in selecting a separation matrix B for which the first output y_1 produces a spectral vector Y_1 as coherent as possible with the video input V_1. This results in the following criterion:

$$\text{maximize } f_{AV}(y) = p(Y_1, V_1). \tag{16}$$

It consists in maximizing the a posteriori estimate of y knowing V_1, from a trained probability model. Notice that once more the criterion is focused on y_1, hence it does not guarantee the separation of the other sources. However, the present method displays a very important property, it ensures that the first source is extracted on the first output channel, while classical blind source separation techniques do not let know which source corresponds to which output.

Though (15) and (16) seem to provide very different criteria, the link is in fact very direct. Indeed, consider what happens in the case of a probability function $p(Y_1, V_1)$ containing only one Gaussian, that is,

$$p(Y_1, V_1) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\left[-\frac{1}{2}[Y_1 V_1]^t C^{-1} [Y_1 V_1]\right], \tag{17}$$

where Y_1 is a spectral column vector defined, as in Section 2.1, by a number $\dim_S$ of energy values at the output of $\dim_S$ bandpass filters, V_1 is a column vector of $\dim_V$ facial characteristics, $[Y_1 V_1]$ is the concatenation of vectors Y_1 and V_1, that is, a column vector of dimension d equal to the sum of $\dim_S$ and $\dim_V$, and C is the covariance matrix of the Gaussian model, estimated from a learning corpus (we take the mean of the Gaussian law to be zero, for sake of simplicity). Maximizing $p(Y_1, V_1)$ results in minimizing the matrix product $[Y_1 V_1]^t C^{-1} [Y_1 V_1]$. If we decompose the symmetric matrix C^{-1} in the following way:

$$C^{-1} = \begin{bmatrix} D & E \\ E^t & F \end{bmatrix}, \tag{18}$$

we can introduce another criterion that should be *minimized* for separation

$$f_{AV2}(y) = Y_1^t D Y_1 + 2V_1^t E^t Y_1 + V_1^t F V_1. \tag{19}$$

By factorising this quadratic function of Y_1, we can obtain another equivalent criterion, differing from the previous one by a constant term

$$f_{AV3}(y) = (Y_1 - HV_1)^t D (Y_1 - HV_1) \tag{20}$$

with

$$H = -D^{-1}E, \tag{21}$$

where H is the regression matrix from V to S, that is the optimal linear estimator of S given V in the least mean square error sense. Criterion (20) is quite similar to criterion (15), the knowledge of normalized spectral terms (γ_{h_k}) being replaced by the knowledge of a spectrum HV_1 estimated from the visual input V_1. The D term in (20) replaces the simple summation in (15) by a slightly more complex summation process involving rotation and weights.

To better understand how the algorithm works in the case of a true mixture of several Gaussian laws, we consider the case of a two-source mixture. In this case, the B matrix contains 4 terms. Focusing on y_1, we may impose $b_{11} = 1$ since there is a gain underspecification, and we change b_{12} into b to simplify notations. The two-source audio-visual separation problem may hence be reduced to

$$y_1 = x_1 + bx_2 \quad \text{with } b = \arg\max\left(p(Y_1, V_1)\right). \tag{22}$$

When y_1 is defined by the first part of (22), the spectrum Y_1 describes a curve in the spectral space, and the second part of (22) specifies the optimal b value. For an audio-visual probability function with a two-Gaussians mixture, we display on Figure 4 how the algorithm works. We can observe on this figure that the audio-visual complementarity plays an important role here. Indeed, even though the visual input V_1 may underspecify the spectrum and provide two possible kinds of spectral configurations, the noisy spectral information contained in the mixture constraints the path of possible spectral configurations, and the optimal b value can be chosen with a good accuracy. However, it may happen that the video input 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

$$\begin{aligned} &p\left(Y_1(t, \ldots, t-T), V_1(t, \ldots, t-T)\right) \\ &\quad = p\left(Y_1(t), V_1(t)\right) \cdots p\left(Y_1(t-T), V_1(t-T)\right). \end{aligned} \tag{23}$$

This product of joint probabilities, for various lengths of temporal integration $(T+1)$, is maximized, instead of criterion (16), to find a better estimation of the separating matrix.

3. EXPERIMENTAL RESULTS

3.1. Data

For this preliminary study, the audio-visual data 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 96 sequences ($4\times V_1$, $6\times C$, $4\times V_2$) 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 other French speakers.

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 Figure 5a). These parameters were automatically extracted every 20 ms by using the ICP face processing system [7]. Sounds were sampled at 16 kHz. The audio spectra envelopes were estimated by 20-order linear prediction (LP) models, which were calculated synchronously with the video parameters (every 20 ms), on 32 ms frames (involving a 12 ms overlap), by using

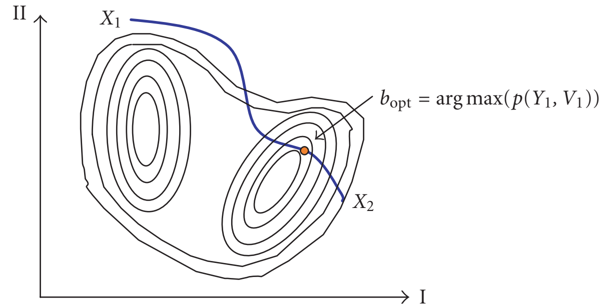


Figure 4: Variations of the audio spectrum Y_1 with b and selection of the optimal b value. Audio dimensions I and II are arbitrary in the figure. The curve between X_1 and X_2 displays the possible variations of Y_1 according to (22). The evolution of the two-Gaussian $p(Y_1, V_1)$ law is displayed by the contour lines.

the auto-correlation method. The log spectrum amplitudes were sampled at 32 frequency values equally spaced from 0 to 5 kHz. Then a principal component analysis (PCA) was applied to reduce the number of spectral components. We used either 5 or 8 dimensions (explaining, respectively 85.5% or 92.5% of the total variance). Therefore, the dimension of the audio-visual space was 7 (5 audio + 2 video) or 10 (8 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 about 2300 audio-visual vectors (96 stimuli, about 24 vectors per stimulus). We tested various numbers of Gaussian laws, from 6 to 12, to model the training data set, that is to correctly represent the mapping of the audio-visual vectors. On Figure 5, we present the results for a mixture of 8 Gaussian laws applied to vectors with 8 PCA audio components, we display the projections of the Gaussian covariance matrices on the two video dimensions (a) and on the first two audio dimensions (b).

The video space displays a quite classical organization, 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, and 8), spread lip shapes ([i], Gaussian law 7) and opened lip shapes ([a], Gaussian law 5). Gaussian laws 4 and 6 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 three Gaussian kernels 2, 3, and 8 are confused in the (LW, LH) space around the [y]/[u] round-closed lip shape, while separated in the audio subspaces according to the [y], [u] and [ty/tu/dy/du/ky/ku/gy/gu] distinction. On the other hand, Gaussian kernels 1, 4, and 8, close and overlapping in the audio space, are clearly separated in the video space. As

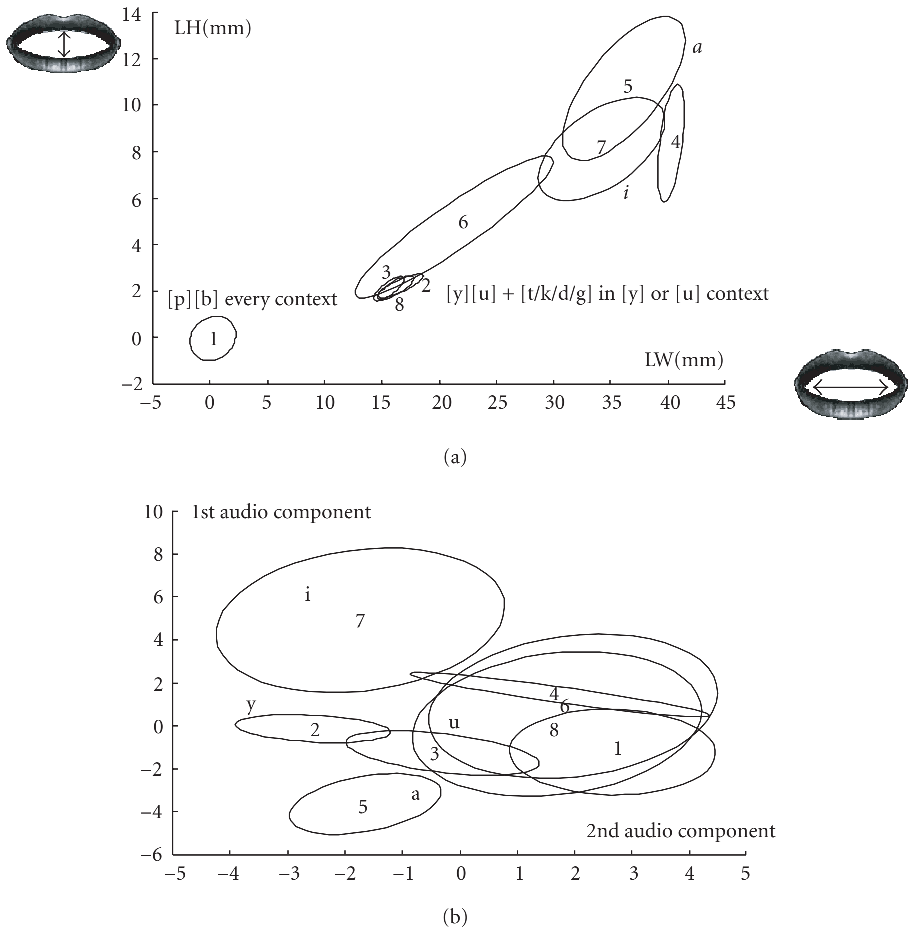


Figure 5: Projection of the standard deviation ellipses of the 8 Gaussian kernels in the video subspace (LW, LH) (a, top) and in the first two principle audio planes (b, bottom). Typical locations of the 4 vowels [i], [a], [u], and [y] are displayed.

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

$$x_1 = a_{11}s_1 + a_{12}s_2; \qquad x_2 = a_{21}s_1 + a_{22}s_2, \tag{24}$$

with the solution defined by (22). The source s_1 is the speech source to be separated, s_2 is a corrupting speech source to eliminate. Sources were normalised in energy. Hence, the input SNRs on each signal x_i are given by

$$\begin{aligned} \mathrm{SNR}_{\mathrm{input1}} &= 20\log\left(a_{11}{}^2/a_{12}{}^2\right), \\ \mathrm{SNR}_{\mathrm{input2}} &= 20\log\left(a_{21}{}^2/a_{22}{}^2\right), \end{aligned} \tag{25}$$

while it is easy to show that the output SNR on y_1 is given by

$$\mathrm{SNR}_{\mathrm{output}} = 20\log\left(\left(a_{11} + ca_{21}\right)^2/\left(a_{12} + ca_{22}{}^2\right)\right). \tag{26}$$

In the following, we will present results obtained with the following mixture matrix

$$a_{11} = 2 \quad a_{12} = 1 \quad a_{21} = 3 \quad a_{22} = 5 \tag{27}$$

corresponding to a theoretical solution $b = -0.2$, and to input SNR values, respectively of 6 and -4.4 dBs. The evaluation was made by concatenating all 96 stimuli of the test set (see Section 3.1) into a single file containing about 2300 audio-visual frames. For each test frame, and for a given b value, the procedure consisted in computing y_1 through (22), estimating the spectrum Y_1 according to the process described in Section 3.1 (LP model followed by PCA analysis), and computing the probability $p(Y_1, V_1)$, thanks to the model described in Section 3.2, in order to determine the best b value maximizing this probability. The optimal b value produces an output y_1 supposed to provide the best estimation of the source s_1.

3.4. Results

Firstly, we studied the variations of either the instantaneous version of joint probability $p(Y_1(t), V_1(t))$, or the temporally integrated version $p(Y_1(t, \ldots, t-T), V_1(t, \ldots, t-T))$, when

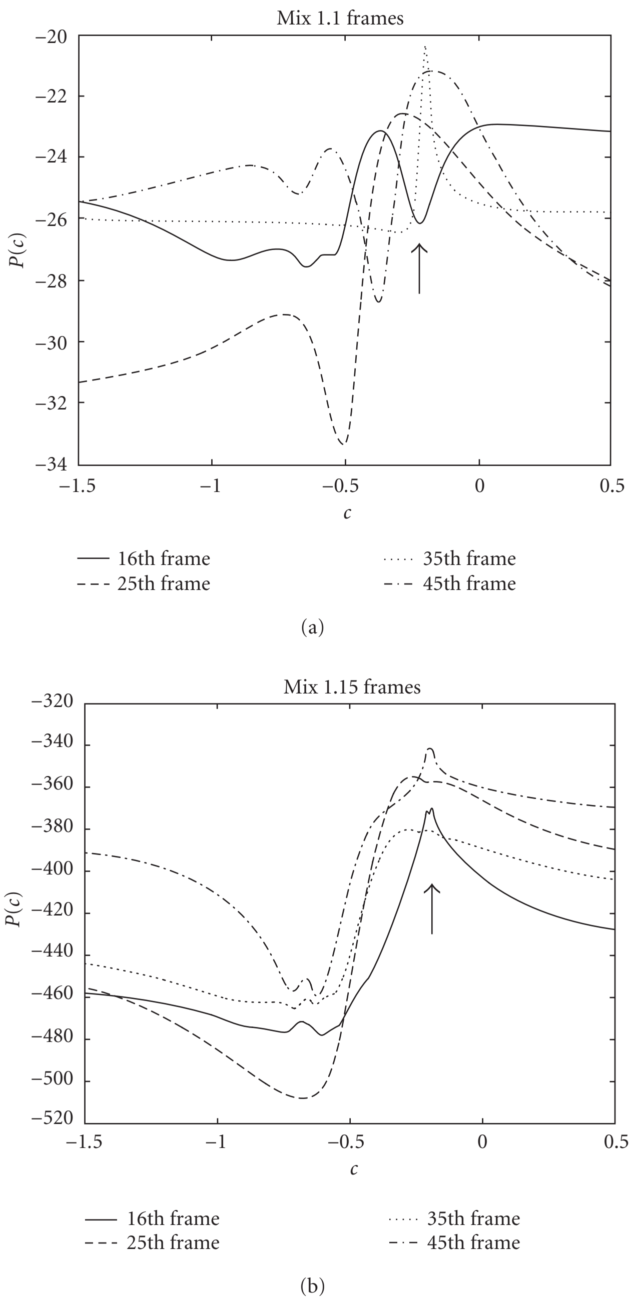


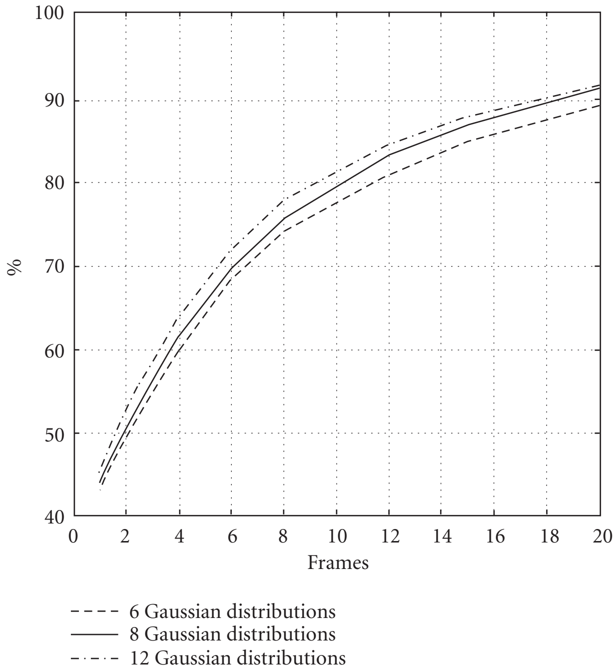
FIGURE 6: Variations of $\log(p(y_1, V_1))$ with c, for instantaneous probabilites (a) or temporal integration over 15 frames (b). The arrow points on the theoretical solution $b = -0.2$.

b was systematically varied around the theoretical solution -0.2. We display on Figure 6 the variations of the logarithm of probability values, for four frames inside the 2300 test frames, and for 2 temporal integration values, namely $T = 0$ (instantaneous probability) and $T = 14$ (temporal integration over 15 consecutive frames). It is obvious on this figure that these variations are quite noisy in the first case (no integration, Figure 6a), while they are extremely coherent in the second one (integration over 15 frames, Figure 6b). In this second case, the probability values display a clear maximum around the theoretical solution $b = -0.2$, and also a minimum around the "antisolution" $b = -0.67$, corresponding to the extraction of s_2 instead of s_1. Then, we implemented an automatic procedure for searching the b value maximizing $p(Y_1, V_1)$ in various conditions. Optimisation was based on the Nelder-Mead "simplex" algorithm [12]. Conditions involved varying the number of PCA audio components from 5 to 8, varying the number of Gaussian kernels in the tuning of the function $p(S_1, V_1)$ from 6 to 12, and varying the integration length for computing $p(Y_1, V_1)$ from one frame to 20 frames. For each condition, we scanned each of the 2300 frames in the test corpus, and counted the percentage of cases where the b value determined by the automatic optimisation procedure was between -0.15 and -0.25. This corresponds to output SNRs higher than 14 dBs (while input SNRs are, respectively 6 dB and -4.4 dB). Results are displayed on Figure 7. It appears that the temporal integration is indeed crucial. Integrating the joint audio-visual probability over 20 frames (i.e., 400 ms, which stays reasonably short) increases success scores very significantly. On this basic pattern, increasing both the number of Gaussian kernels (Figure 7a) and the number of PCA components slightly improves the performances. Selecting the best configuration, that is using 8 PCA components, 12 Gaussian kernels and integrating over 20 frames, enables us to produce a score of 96% of frames displaying an output SNR greater than 14 dB. Listening tests show that the enhancement of source s_1 is indeed quite important. The results are also good for three-source and five-source mixtures [13]; and they compare well with the performances of the JADE algorithm [6] which is a reference in the domain of Blind Source Separation techniques.

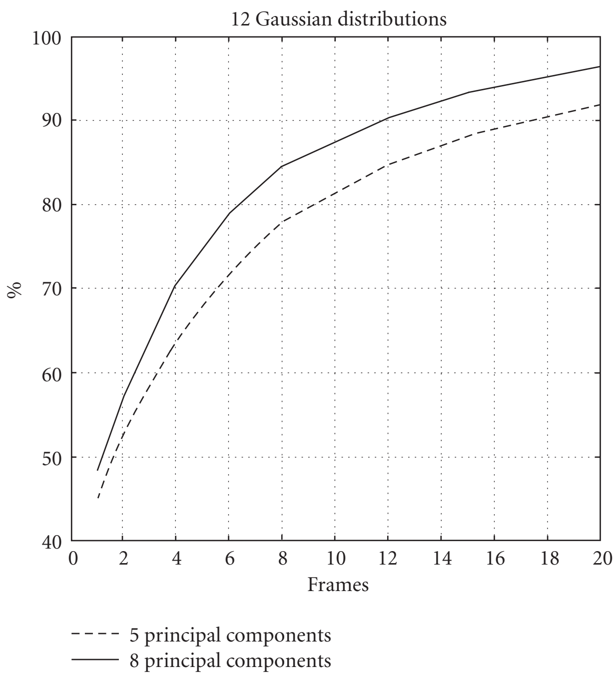
4. CONCLUSION

It appears that the principle of an audio-visual algorithm for speech signals separation is theoretically sound and technically viable. Of course, we are far from the end, and a number of problems are still to be solved. We mention three main ones. Firstly, the optimisation procedure we used here is very rough, and we presently study more powerful gradient-based techniques that should be very important to speed up the algorithm, which is presently rather slow. Secondly, the statistical models of the joint audio-visual probability could be based on more sophisticated functions, and particularly the assumption of temporal independence between consecutive frames could be replaced by more general assumptions, possibly involving hidden Markov models. Last but not least, the speech source we used here is very simple, with only plosives and vowels uttered by one speaker, and the passage to more complex stimuli will considerably increase the difficulty.

Moreover, it must be acknowledged that classical (pure audio) separation algorithms are able to almost perfectly separate such mixtures. However, while these techniques are known to be very sensitive to additive noise, we may hope that audio-visual separation should be much more robust



(a)



(b)

Figure 7: Separation scores for the two-source mixture. Percentage of correct b estimates over the 2300 frames of the test corpus, depending on the length of temporal integration. (a): 5 PCA dimensions, 6 vs. 8 vs. 12 Gaussian kernels; (b): 12 Gaussian kernels, 5 vs. 8 PCA dimensions.

due to the video information, which is generally not corrupted by the noise. This will be studied in the near future.

More generally, our method provides some kind of extension to a number of proposals based on the introduction of spectral knowledge in blind separation algorithms (e.g., [14, 15, 16, 17]). Furthermore, this work is promising because we can expect to proceed in the future with much less ideal configuration: for example fewer sensors than sources and more complex mixtures (e.g., convolutive). In such cases, where the audio information may not be sufficient, the visual information could enable to better focus on a particular source and improve its enhancement/separation. Furthermore, it solves continuity problems that can be quite complex in general ICA techniques, for which there can be a permutation of sources between sensors for each computation frame. Thus, future works should consider the combination of this approach with standard ICA methods.

REFERENCES

[1] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds., pp. 3–51, Lawrence Erlbaum Associates, London, 1987.

[2] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, and P. Escudier, "Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise," *Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3677–3689, 1998.

[3] D. G. Stork and M. E. Hennecke, Eds., *Speechreading by Humans and Machines: Models, Systems and Applications*, Springer-Verlag, Berlin, 1996.

[4] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[5] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994, Special issue on Higher-Order Statistics.

[6] J. F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.

[7] M. T. Lallouache, "Un poste 'visage-parole'. Acquisition et traitement de contours labiaux," in *Proc. XVIII Journées d'Études sur la Parole*, pp. 282–286, Montréal, Canada, 1990.

[8] L. Revéret and C. Benoît, "A new 3D lip model for analysis and synthesis of lip motion in speech production," in *Proc. the International Conference on Auditory-Visual Speech Processing*, pp. 207–212, Sydney, Australia, 1998.

[9] J. F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.

[10] D. Sodoyer, "Séparation de sources audiovisuelles; de la formalisation à l'expérimentation," M.S. thesis, National Polytechnical Institute of Grenoble, France, 2001, Signal-Image-Parole.

[11] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[12] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder-Mead simplex method in low dimensions," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[13] J. Klinkisch, "Séparation de sources audiovisuelles pour la reconnaissance de chiffres dans le bruit," M.S. thesis, National Polytechnical Institute of Grenoble, France, 2000.

[14] L. Tong, V. C. Soon, Y. F. Huang, and R. Liu, "AMUSE: A new blind identification algorithm," in *Proc. IEEE Int. Symp. Circuits and Systems*, pp. 1784–1786, New Orleans, La, USA, 1990.
[15] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.
[16] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. Signal Processing*, vol. 45, no. 7, pp. 1712–1725, 1997.
[17] S. Hosseini, C. Jutten, and D. T. Pham, "Blind separation of temporally correlated sources using a quasi maximum likelihood approach," in *Proc. ICA '2001*, pp. 586–590, San Diego, Calif, USA, 2001.

David Sodoyer was born in Soissons, France, in 1976. He received the Bachelor degree in electronics, electrotechnics and automatics in 2000 and received in 2001 a Diplôme d'Etudes Approfondies in signal, image, speech and telecommunications at the National Polytechnic Institute of Grenoble. He is preparing the Ph.D. thesis on blind source separation techniques applied to mixtures of audio-visual speech sources, under the codirection of Jean-Luc Schwartz, Christian Jutten, and Laurent Girin.

Jean-Luc Schwartz received the M.S. degree in physics from the Université de Paris-Sud in 1979, and the Ph.D. degree in psychoacoustics from the Institut de la Communication Parlée (ICP), Grenoble, France, in 1981. He obtained the State Thesis in the field of auditory modelling and vowel perception in 1987. Since 1983, he is employed by the Centre National de la Recherche Scientifique, and leads the Speech Perception Group at ICP. His main areas of research involve auditory modelling, psychoacoustics, speech perception, auditory front-ends for speech recognition, bimodal integration in speech perception and source separation, perceptuo-motor interactions and speech robotics. He has been involved in various national and European projects, and authored or coauthored more than 25 publications in international journals such as IEEE SAP, JASA, Journal of Phonetics, Computer Speech and Language, Artificial Intelligence Review, Speech Communication, Behavioural and Brain Sciences, Hearing Research, etc., about 20 book chapters, and 80 presentations in national and international workshops.

Laurent Girin was born in Moutiers, France, in 1969. He received the M.S. and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble, France, in 1994 and 1997, respectively. In 1997, he joined the École Nationale d'Electronique et de Radioélectricité de Grenoble, where he is currently an Associate Teacher in electrical engineering and signal processing. His current research interests are in audiovisual speech processing with application to speech source separation/enhancement and speech coding.

Jacob Klinkisch was born in Orange (France) in 1975. After finishing his Abitur on the Albert-Schweitzer-Gymnasium in Hamburg (Germany) he joined the Karlsruhe University and the ENSERG/INP Grenoble (France) for studying electrical engineering, taking part on a French-German double degree program, and he got a German and French Master degree in 2000. He obtained his practical experience at DASA (Munich/Germany) and Libertel (Maastricht/Holland) during studies. For his final work of 6 months he joined the ICP to work on audiovisual source separation. Then he got further practical experience on digital signal processing at Cochlear Ltd. in Sydney (Australia) before starting to work in 2001 at MobilCom (Germany).

Christian Jutten received the Ph.D. degree in 1981 and the Docteur ès Sciences degree in 1987 from the Institut National Polytechnique of Grenoble (France). He taught as associate professor in École Nationale Supérieure d'Electronique et de Radioélectricité of Grenoble from 1982 to 1989. He was a visiting professor in Swiss Federal Polytechnic Institute in Lausanne in 1989, he became full professor in the Sciences and Techniques Institute, Université Joseph Fourier of Grenoble. For 20 years, his research interests are source separation and independent component analysis and learning in neural networks. He has been associate editor of IEEE Trans. on Circuits and Systems (1994–95), and co-organizer with Dr. J.-F. Cardoso and Prof. Ph. Loubaton of the 1st International Conference on Blind Signal Separation and Independent Component Analysis (Aussois, France, January 1999). He is currently member of a technical committee of IEEE Circuits and Systems society on blind signal processing.

EURASIP JOURNAL ON APPLIED SIGNAL PROCESSING

Special Issue on

Advanced Signal Processing and Computational Intelligence Techniques for Power Line Communications

Call for Papers

In recent years, increased demand for fast Internet access and new multimedia services, the development of new and feasible signal processing techniques associated with faster and low-cost digital signal processors, as well as the deregulation of the telecommunications market have placed major emphasis on the value of investigating hostile media, such as powerline (PL) channels for high-rate data transmissions.

Nowadays, some companies are offering powerline communications (PLC) modems with mean and peak bit-rates around 100 Mbps and 200 Mbps, respectively. However, advanced broadband powerline communications (BPLC) modems will surpass this performance. For accomplishing it, some special schemes or solutions for coping with the following issues should be addressed: (i) considerable differences between powerline network topologies; (ii) hostile properties of PL channels, such as attenuation proportional to high frequencies and long distances, high-power impulse noise occurrences, time-varying behavior, and strong inter-symbol interference (ISI) effects; (iv) electromagnetic compatibility with other well-established communication systems working in the same spectrum, (v) climatic conditions in different parts of the world; (vii) reliability and QoS guarantee for video and voice transmissions; and (vi) different demands and needs from developed, developing, and poor countries.

These issues can lead to exciting research frontiers with very promising results if signal processing, digital communication, and computational intelligence techniques are effectively and efficiently combined.

The goal of this special issue is to introduce signal processing, digital communication, and computational intelligence tools either individually or in combined form for advancing reliable and powerful future generations of powerline communication solutions that can be suited with for applications in developed, developing, and poor countries.

Topics of interest include (but are not limited to)

- Multicarrier, spread spectrum, and single carrier techniques
- Channel modeling
- Channel coding and equalization techniques
- Multiuser detection and multiple access techniques
- Synchronization techniques
- Impulse noise cancellation techniques
- FPGA, ASIC, and DSP implementation issues of PLC modems
- Error resilience, error concealment, and Joint source-channel design methods for video transmission through PL channels

Authors should follow the EURASIP JASP manuscript format described at the journal site http://asp.hindawi.com/. Prospective authors should submit an electronic copy of their complete manuscripts through the EURASIP JASP manuscript tracking system at http://www.hindawi.com/mts/, according to the following timetable:

Manuscript Due	October 1, 2006
Acceptance Notification	January 1, 2007
Final Manuscript Due	April 1, 2007
Publication Date	3rd Quarter, 2007

GUEST EDITORS:

Moisés Vidal Ribeiro, Federal University of Juiz de Fora, Brazil; mribeiro@ieee.org

Lutz Lampe, University of British Columbia, Canada; lampe@ece.ubc.ca

Sanjit K. Mitra, University of California, Santa Barbara, USA; mitra@ece.ucsb.edu

Klaus Dostert, University of Karlsruhe, Germany; klaus.dostert@etec.uni-karlsruhe.de

Halid Hrasnica, Dresden University of Technology, Germany hrasnica@ifn.et.tu-dresden.de

Special Issue on

Numerical Linear Algebra in Signal Processing Applications

Call for Papers

The cross-fertilization between numerical linear algebra and digital signal processing has been very fruitful in the last decades. The interaction between them has been growing, leading to many new algorithms.

Numerical linear algebra tools, such as eigenvalue and singular value decomposition and their higher-extension, least squares, total least squares, recursive least squares, regularization, orthogonality, and projections, are the kernels of powerful and numerically robust algorithms.

The goal of this special issue is to present new efficient and reliable numerical linear algebra tools for signal processing applications. Areas and topics of interest for this special issue include (but are not limited to):

- Singular value and eigenvalue decompositions, including applications.
- Fourier, Toeplitz, Cauchy, Vandermonde and semi-separable matrices, including special algorithms and architectures.
- Recursive least squares in digital signal processing.
- Updating and downdating techniques in linear algebra and signal processing.
- Stability and sensitivity analysis of special recursive least-squares problems.
- Numerical linear algebra in:
  - Biomedical signal processing applications.
  - Adaptive filters.
  - Remote sensing.
  - Acoustic echo cancellation.
  - Blind signal separation and multiuser detection.
  - Multidimensional harmonic retrieval and direction-of-arrival estimation.
  - Applications in wireless communications.
  - Applications in pattern analysis and statistical modeling.
  - Sensor array processing.

Authors should follow the EURASIP JASP manuscript format described at http://www.hindawi.com/journals/asp/. Prospective authors should submit an electronic copy of their complete manuscript through the EURASIP JASP manuscript tracking system at http://www.hindawi.com/mts/, according to the following timetable:

Manuscript Due	October 1, 2006
Acceptance Notification	February 1, 2007
Final Manuscript Due	May 1, 2007
Publication Date	3rd Quarter, 2007

GUEST EDITORS:

Shivkumar Chandrasekaran, Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA; shiv@ece.ucsb.edu

Gene H. Golub, Department of Computer Science, Stanford University, USA; golub@sccm.stanford.edu

Nicola Mastronardi, Istituto per le Applicazioni del Calcolo "Mauro Picone," Consiglio Nazionale delle Ricerche, Bari, Italy; n.mastronardi@ba.iac.cnr.it

Marc Moonen, Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium; marc.moonen@esat.kuleuven.be

Paul Van Dooren, Department of Mathematical Engineering, Catholic University of Louvain, Belgium; vdooren@csam.ucl.ac.be

Sabine Van Huffel, Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium; sabine.vanhuffel@esat.kuleuven.be

Special Issue on Human-Activity Analysis in Multimedia Data

Call for Papers

Many important applications of multimedia revolve around the detection of humans and the interpretation of human behavior, for example, surveillance and intrusion detection, automatic analysis of sports videos, broadcasts, movies, ambient assisted living applications, video conferencing applications, and so forth. Success in this task requires the integration of various data modalities including video, audio, and associated text, and a host of methods from the field of machine learning. Additionally, the computational efficiency of the resulting algorithms is critical since the amount of data to be processed in videos is typically large and real-time systems are required for practical implementations.

Recently, there have been several special issues on the human detection and human-activity analysis in video. The emphasis has been on the use of video data only. This special issue is concerned with contributions that rely on the use of multimedia information, that is, audio, video, and, if available, the associated text information.

Papers on the following and related topics are solicited:

- Video characterization, classification, and semantic annotation using both audio and video, and text (if available).
- Video indexing and retrieval using multimedia information.
- Segmentation of broadcast and sport videos based on audio and video.
- Detection of speaker turns and speaker clustering in broadcast video.
- Separation of speech and music/jingles in broadcast videos by taking advantage of multimedia information.
- Video conferencing applications taking advantage of both audio and video.
- Human mood detection, and classification of interactivity in duplexed multimedia signals as in conversations.
- Human computer interaction, ubiquitous computing using multimedia.
- Intelligent audio-video surveillance and other security-related applications.

Authors should follow the EURASIP JASP manuscript format described at the journal site bellow http://www.hindawi.com/GetJournal.aspx?journal=ASP. Prospective authors should submit an electronic copy of their complete manuscript through the EURASIP JASP manuscript tracking system at the following site http://www.hindawi.com/mts/, according to the following timetable:

Manuscript Due	February 1, 2007
Acceptance Notification	June 1, 2007
Final Manuscript Due	October 1, 2007
Publication Date	1st Quarter, 2008

GUEST EDITORS:

A. Enis Cetin, Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey; cetin@ee.bilkent.edu.tr

Eric Pauwels, Signals and Images Research Group, Centre for Mathematics and Computer Science (CWI), 1098 SJ Amsterdam, The Netherlands; eric.pauwels@cwi.nl

Ovidio Salvetti, Institute of Information Science and Technologies (ISTI), Italian National Research Council (CNR), 56124 Pisa, Italy; ovidio.salvetti@isti.cnr.it

EURASIP JOURNAL ON APPLIED SIGNAL PROCESSING

Special Issue on

Advanced Signal Processing and Pattern Recognition Methods for Biometrics

Call for Papers

Biometric identification has established itself as a very important research area primarily due to the pronounced need for more reliable and secure authentication architectures in several civilian and commercial applications. The recent integration of biometrics in large-scale authentication systems such as border control operations has further underscored the importance of conducting systematic research in biometrics. Despite the tremendous progress made over the past few years, biometric systems still have to reckon with a number of problems, which illustrate the importance of developing new biometric processing algorithms as well as the consideration of novel data acquisition techniques. Undoubtedly, the simultaneous use of several biometrics would improve the accuracy of an identification system. For example the use of palmprints can boost the performance of hand geometry systems. Therefore, the development of biometric fusion schemes is an important area of study. Topics related to the correlation between biometric traits, diversity measures for comparing multiple algorithms, incorporation of multiple quality measures, and so forth need to be studied in more detail in the context of multibiometrics systems. Issues related to the individuality of traits and the scalability of biometric systems also require further research. The possibility of using biometric information to generate cryptographic keys is also an emerging area of study. Thus, there is a definite need for advanced signal processing, computer vision, and pattern recognition techniques to bring the current biometric systems to maturity and allow for their large-scale deployment.

This special issue aims to focus on emerging biometric technologies and comprehensively cover their system, processing, and application aspects. Submitted articles must not have been previously published and must not be currently submitted for publication elsewhere. Topics of interest include, but are not limited to, the following:

- Fusion of biometrics
- Analysis of facial/iris/palm/fingerprint/hand images
- Unobtrusive capturing and extraction of biometric information from images/video
- Biometric identification systems based on face/iris/palm/fingerprint/voice/gait/signature
- Emerging biometrics: ear, teeth, ground reaction force, ECG, retina, skin, DNA
- Biometric systems based on 3D information
- User-specific parameterization
- Biometric individuality
- Biometric cryptosystems
- Quality measure of biometrics data
- Sensor interoperability
- Performance evaluation and statistical analysis

Authors should follow the EURASIP JASP manuscript format described at http://www.hindawi.com/journals/asp/. Prospective authors should submit an electronic copy of their complete manuscript through the EURASIP JASP manuscript tracking system at http://www.hindawi.com/mts/, according to the following timetable:

Manuscript Due	May 1, 2007
Acceptance Notification	September 1, 2007
Final Manuscript Due	December 1, 2007
Publication Date	1st Quarter, 2008

GUEST EDITORS:

Nikolaos V. Boulgouris, Department of Electronic Engineering, Division of Engineering, King's College London, London WC2R 2LS, UK; nikolaos.boulgouris@kcl.ac.uk

Juwei Lu, EPSON Edge, EPSON Canada Ltd., Toronto, Ontario M1W 3Z5, Canada; juwei@ieee.org

Konstantinos N. Plataniotis, The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada, M5S 3G4; kostas@dsp.utoronto.ca

Arun Ross, Lane Department of Computer Science & Electrical Engineering, West Virginia University, Morgantown WV, 26506, USA; arun.ross@mail.wvu.edu

EURASIP JOURNAL ON BIOINFORMATICS AND SYSTEMS BIOLOGY

Special Issue on Information Theoretic Methods for Bioinformatics

Call for Papers

Information theoretic methods for modeling are at the center of the current efforts to interpret bioinformatics data. The high pace at which new technologies are developed for collecting genomic and proteomic data requires a sustained effort to provide powerful methods for modeling the data acquired. Recent advances in universal modeling and minimum description length techniques have been shown to be well suited for modeling and analyzing such data. This special issue calls for contributions to modeling of data arising in bioinformatics and systems biology by information theoretic means. Submissions should address theoretical developments, computational aspects, or specific applications. Suitable topics for this special issue include but are not limited to:

- Normalized maximum-likelihood (NML) universal models
- Minimum description length (MDL) techniques
- Microarray data modeling
- Denoising of genomic data
- Pattern recognition
- Data compression-based modeling

Authors should follow the EURASIP JBSB manuscript format described at http://www.hindawi.com/journals/bsb/. Prospective authors should submit an electronic copy of their complete manuscript through the EURASIP JBSB's manuscript tracking system at http://www.hindawi.com/mts/, according to the following timetable.

Manuscript Due	February 1, 2007
Acceptance Notification	May 1, 2007
Final Manuscript Due	July 1, 2007
Publication Date	3rd Quarter, 2007

GUEST EDITORS:

Jorma Rissanen, Computer Learning Research Center, University of London, Royal Holloway, TW20 0EX, UK; jorma.rissanen@mdl-research.org

Peter Grünwald, Centrum voor Wiskunde en Informatica (CWI), National Research Institute for Mathematics and Computer Science, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands; pdg@cwi.nl

Jukka Heikkonen, Laboratory of Computational Engineering, Helsinki University of Technology, P.O. Box 9203, 02015 HUT, Finland; jukka.heikkonen@tkk.fi

Petri Myllymäki, Department of Computer Science, University of Helsinki, P.O. Box 68 (Gustaf Hällströmin katu 2b), 00014, Finland; petri.myllymaki@cs.helsinki.fi

Teemu Roos, Complex Systems Computation Group, Helsinki Institute for Information Technology, University of Helsinki, P.O.Box 68, 00014, Finland; teemu.roos@hiit.fi

Juho Rousu, Department of Computer Science, University of Helsinki, P.O. Box 68 (Gustaf Hällströmin katu 2b), 00014, Finland; juho.rousu@cs.helsinki.fi