Audio-visual enhancement of speech in noise

Laurent Girin,[a] Jean-Luc Schwartz,[b] and Gang Feng[c]
Institut de la Communication Parlée, INPG/Université Stendhal/CNRS UMR 5009,
Université Stendhal, Domaine universitaire, 38040 Grenoble, France

(Received 5 November 1998; revised 26 May 2000; accepted 29 January 2001)

A key problem for telecommunication or human–machine communication systems concerns speech enhancement in noise. In this domain, a certain number of techniques exist, all of them based on an acoustic-only approach—that is, the processing of the audio corrupted signal using audio information (from the corrupted signal only or additive audio information). In this paper, an audio-visual approach to the problem is considered, since it has been demonstrated in several studies that viewing the speaker's face improves message intelligibility, especially in noisy environments. A speech enhancement prototype system that takes advantage of visual inputs is developed. A filtering process approach is proposed that uses enhancement filters estimated with the help of lip shape information. The estimation process is based on linear regression or simple neural networks using a training corpus. A set of experiments assessed by Gaussian classification and perceptual tests demonstrates that it is indeed possible to enhance simple stimuli (vowel-plosive-vowel sequences) embedded in white Gaussian noise. *© 2001 Acoustical Society of America.* [DOI: 10.1121/1.1358887]

PACS numbers: 43.72.Ew, 43.71.Ma [DOS]

I. INTRODUCTION

The bimodal nature of speech is now acknowledged as a basic characteristic, both for understanding speech perception (Summerfield, 1987) and for developing tools for human–human and human–machine communication (Bernstein and Benoît, 1996). One of the most well-known paradigms for the study of audio-visual speech perception is the identification of speech in noise (Sumby and Pollack, 1954; Erber, 1975; MacLeod and Summerfield, 1987; Benoît *et al.*, 1994; Grant and Walden, 1996; Robert-Ribes *et al.*, 1998). In the field of speech technologies, there is an increasing number of works on automatic audio-visual speech recognition systems, evaluated in general through their performance in recognizing speech in adverse conditions (e.g., Stork and Hennecke, 1996).

In all these studies, a common assumption is that the audio and visual sensors process information *independently* for parameter estimation, feature extraction or category estimation (according to the various architectures proposed, see Schwartz *et al.*, 1998) before they are fused by the human brain or the decision recognition algorithm for achieving an audio-visual identification task. The question of the dependence versus independence of processing is seldom addressed (although see Massaro, 1989) but independence is implicit in all recognition systems and cognitive models of audio-visual speech identification.

At the same time, it is obvious that there is some dependence between the *content* of the sensory inputs, that is audio and visual speech, since they are both consequences of one physical cause, the articulatory gestures: e.g., spread lips may be associated in French with the sound of an [i] or a [ti] but not an [y] or a [by], while open lips are compatible with almost nothing but the sound of an [a]. This means that some *predictions about the sound* should be feasible from the image. Indeed, it seems that the visual information coming from a speaker is able to improve the auditory *detection* of speech in noise (Grant and Seitz, 2000). Hence it is likely that audio and video processing are not independent in human perception. This could result in an additional contribution to audio-visual speech perception, in which the visual stream would not only provide an intrinsic benefit through lipreading, but also, at an earlier level, help the extraction of audio cues necessary for identification (Barker *et al.*, 1998). This idea receives some confirmation through data obtained by Driver (1996) which show that seeing a speaker improves the identification of a message produced by *another* unseen speaker, the audio component of which has been mixed with the audio component of the seen speaker's message.

All these facts lead us to suggest that sound enhancement could exploit the information contained in the coherent visible movements of the speaker's face. The objective of this study was to demonstrate the technical feasibility of audio-visual speech enhancement—that is, the enhancement of noisy speech sounds, using the video input—which, to our knowledge, has never before been attempted. We developed a prototype system that generates enhanced speech sounds from noisy speech plus visual information. It is based on a *fusion-and-filtering* algorithm that combines the information provided by the noisy audio and video channels to estimate the parameters of an enhancement filter—in this case, a Wiener filter—and then processes the noisy audio input with this filter.

This study focused on degradations due to additive, stationary, white Gaussian noise. Obviously, such "simple" degradations of the audio input could be efficiently removed using classical pure audio enhancement systems, such as spectral subtraction based on noise estimation in silent periods of speech (Boll, 1979; McAulay and Malpass, 1980;

[a] Electronic mail: girin@icp.inpg.fr
[b] Electronic mail: schwartz@icp.inpg.fr
[c] Electronic mail: feng@icp.inpg.fr

Kang and Fransen, 1989; Le Bouquin-Jeannès and Faucon, 1995), multi-microphone techniques (Widrow *et al.*, 1975; Sambur, 1978; Boll and Pulsipher, 1980; Ferrara and Widrow, 1981; Harrison *et al.*, 1986; Feder *et al.*, 1989), or blind source separation (Comon *et al.*, 1991; Jutten and Hérault, 1991; for an overview of basic acoustic speech enhancement methods, see Lim, 1983). However, our aim was to demonstrate that enhancement may be improved through the use of the video channel. Therefore, we focused on a mono-microphone technique, and special care was given to implement three variants within the fusion-and-filtering process: that is, filter parameters were estimated from only the audio sensor, only the video sensor, or both sensors. The enhancement results were then compared systematically. This enabled us to determine if the system was able to exploit the partial complementarity of the audio and video signals in speech perception: that is, the phonetic contrasts least robust in auditory perception in acoustical noise are the most visible ones, both for consonants (Summerfield, 1987) and vowels (Robert-Ribes *et al.*, 1998). In the future, our project will be to combine the audiovisual technique—if successful—with the previously mentioned pure audio mono- or multi-microphone algorithms to improve speech enhancement: this will be discussed further in Sec. V.

This paper is organized into five sections. In Sec. II, we introduce the basic components of the system: the architecture for estimating filter parameters from audio and video inputs, the filter design, and the nature of these audio and video inputs. In Sec. III, we describe the experimental conditions in more detail and present the main results obtained with an early version of the system. In Sec. IV, we present a second version in which we implemented more powerful (and more successful) processing tools to eliminate difficulties encountered with plosives. Both objective tools such as classification experiments and subjective tools involving perceptual data from identification tests were used to assess the enhancement effect. Finally, in Sec. V, we evaluate the achievements of the system and propose a number of directions for future development.

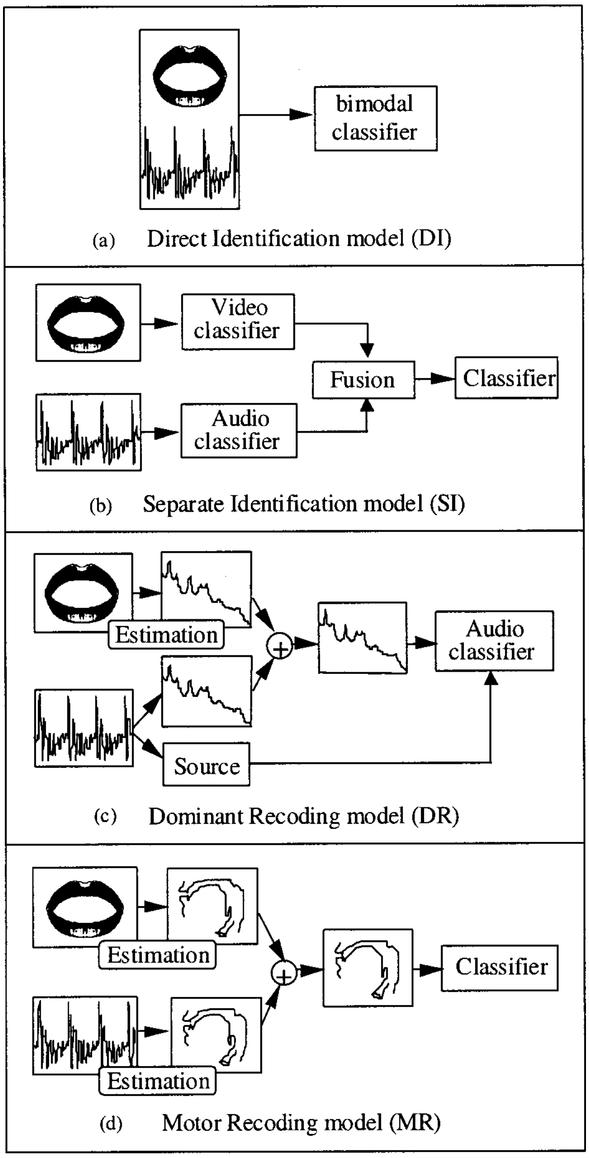
bimodal classifier
(a) Direct Identification model (DI)
Video classifier
Audio classifier
Fusion
Classifier
(b) Separate Identification model (SI)
Estimation
Source
Audio classifier
(c) Dominant Recoding model (DR)
Estimation
Estimation
Classifier
(d) Motor Recoding model (MR)

FIG. 1. Four integration models for audiovisual speech identification.

II. ARCHITECTURE OF THE SPEECH ENHANCEMENT SYSTEM

A. Audio-visual fusion for speech enhancement

Audio-visual speech enhancement is a multi-channel enhancement problem. Most classical multi-sensors noise canceling systems are based on the calculation of correlation functions between the samples of different audio inputs. In our case, the two sensors, that is, the audio and the video input, are quite different in nature. Indeed, the visual input, which is restricted to lip characteristics, has a low sampling frequency compared to the audio one, and it provides only partial information about the vocal tract shape, and no information at all about the source. Hence no usual correlation function can be computed, and we cannot exploit these noise canceling systems as they are. However, the partial information provided by the lips about the vocal tract shape corresponds to some information on the spectral shape of the speech signal. Hence it can be used to define the spectral parameters of an enhancement filter designed to process the audio noisy channel. As additional spectral information should be available from the audio channel, even degraded, the audio and video inputs must be integrated for an accurate estimation of the filter.

Since the problem of mixing an audio and a video speech stream has been addressed extensively in automatic speech recognition, it may be interesting to analyze the proposed solutions. It appears that there exist four basic architectures for fusing sounds and images toward the identification of the speech message (Schwartz *et al.*, 1998). We shall review them briefly, before attempting to study their adaptation to the speech enhancement problem.

In speech recognition, the problem is to estimate a phonetic class from audio and video parameters. In the Direct Identification model (DI), the input stimuli are recognized by a bimodal classifier, which works with vectors of concatenated audio-visual parameters [Fig. 1(a)]. In the Separate Identification model (SI), the audio and video inputs are classified separately before a late-integration fusion process occurs on the separate classification results [Fig. 1(b)]. In contrast, the last two models are based on early-integration (i.e., before classification). In the Dominant Recoding model (DR), audition is supposed to be the dominant modality for speech perception and the visual input is recoded into an auditory representation. Then, classification takes place inside this auditory integration space [Fig. 1(c)]. In the Motor Recoding model (MR), both inputs are projected into an amodal motor representation of articulatory gestures before fusion and classification [Fig. 1(d)].

In this speech enhancement problem, the parameters of an enhancement filter $H(\theta)$ must be estimated from audio and video parameters. The noisy acoustic input is then processed by this filter to provide an enhanced audio signal. By adapting the previous taxonomy to audio-video fusion for filter estimation, four possible architectures can be defined: Direct Estimation (DE) of the filter from the audio+video parameters [Fig. 2(a)]; Separate Estimations (SE) of spectral characteristics of the filter from each input, followed by a fusion process [Fig. 2(b)]; Dominant Recoding (DR) of the video input into spectral characteristics of the filter [Fig. 2(c)]; or implementation of a complete audio-visual inversion process followed by a resynthesis of the filter (Motor Recoding, MR) [Fig. 2(d)].

Within these four architectures, the last one, MR, while quite appealing for theoretical and technical reasons, seems not feasible at present. Indeed, articulatory data are sorely lacking, and neither audiovisual-to-articulatory inversion, nor articulatory-to-acoustic synthesis are sufficient (although see promising advances in, e.g., Bailly *et al.*, 1991; Schroeter and Sondhi, 1994; Yehia *et al.*, 1998). The DR structure, which is the simplest one, has already been tested (Girin *et al.*, 1996). Although preliminary results on vowel enhancement were interesting, they were limited by the fact that the filter was estimated only from the visual input. Therefore, it seems necessary to combine audio and video inputs for filter estimation, which is the case of the two other architectures, SE and DE. Using DE seems to provide a more general framework than SE and requires fewer *a priori* assumptions about audio-visual fusion; this is the one which was selected in this study. In this architecture, the filter estimation is realized from both the video and the noisy audio input. We compared this audio-visual (AV) version with an audio-only (A) condition and with a video-only (V) condition in which only the noisy audio input or only the video input was used for filter estimation. Notice that the V condition is similar to both the DR architecture and the video branch of SE, while the A condition is a special case of the audio branch of SE. This provided a basic homogeneous framework for a comparison with the audio techniques. Such comparisons enabled us to evaluate the true benefit of the video input for speech enhancement, and the role of the audio-visual synergy in this process. In the following, the synergy criterion will be the test of the inequality AV⩾(A or V), that is, performances should be better with two inputs than with one.

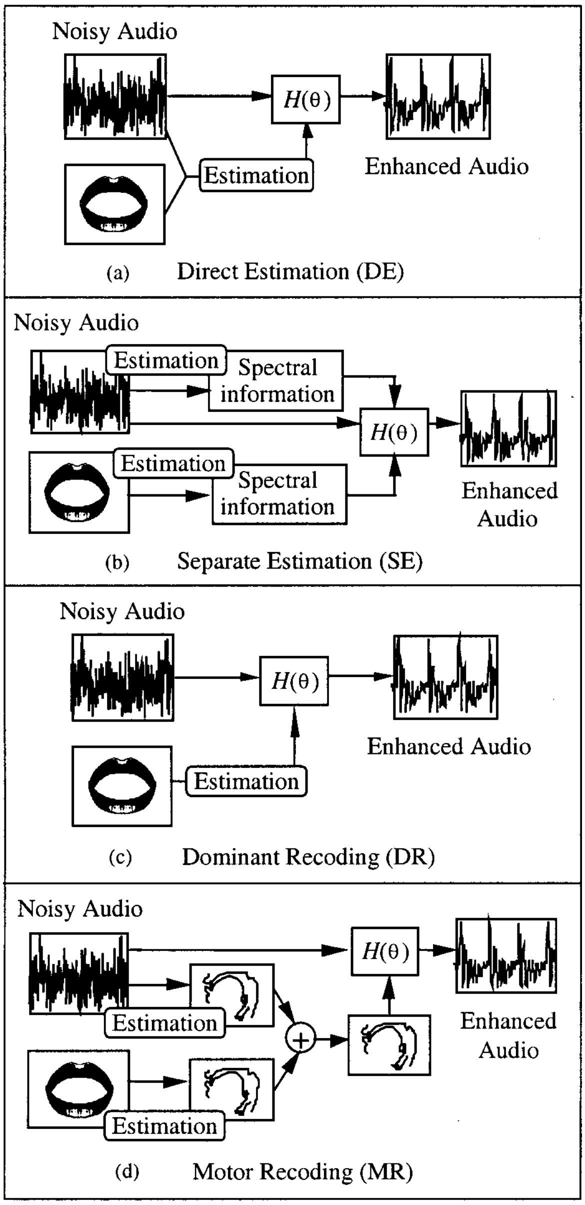

FIG. 2. Four derived integration models for audiovisual speech enhancement.

B. Filter estimation

The sensor fusion process in the DE architecture converges on the estimation of the filter frequency response $H(\theta)$ [Fig. 2(a)]. We shall now specify our choice for this filter and its estimation process. Let us denote, respectively, the speech signal $s(t)$, the noise $v(t)$ [white, Gaussian and assumed to be uncorrelated with $s(t)$], and the observed noisy signal $x(t)=s(t)+v(t)$. The linear, optimal estimator of the signal according to the mean square error criterion is obtained by filtering $x(t)$ with the Wiener filter, of which the expression in the (normalized) frequency domain is (Lim, 1983):

$$H(\theta)=\frac{P_s(\theta)}{P_x(\theta)}, \tag{1}$$

where $P_s(\theta)$ and $P_x(\theta)$ denote, respectively, the power

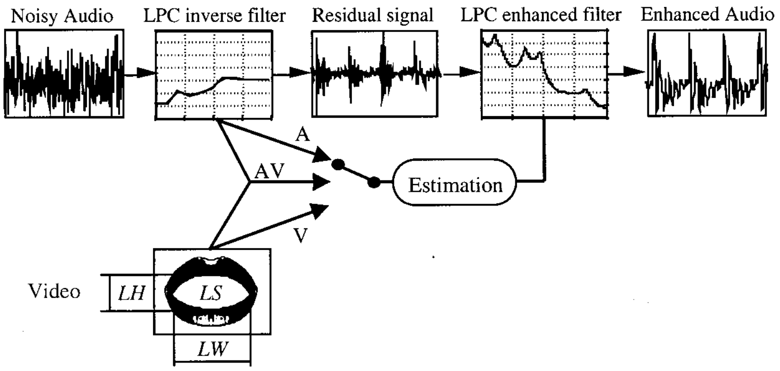

FIG. 3. General schema for the audiovisual speech enhancement system. AV, A, and V correspond, respectively, to the audiovisual, audio-only, and visual-only conditions. The LW, LH, and LS lip contour parameters are defined in Sec. III A 2.

spectral densities (PSD) of the signal and the observed noisy signal.

$P_x(\theta)$ can be estimated directly from the observed noisy signal with classical spectral analysis techniques. The major problem is the estimation of $P_s(\theta)$ since the signal is corrupted by noise. In the present system, $P_s(\theta)$ was estimated using an associative algorithm (henceforth ''associator'') trained on data from a training corpus. This corpus contained clean audio stimuli and their associated video input. The audio stimuli were mixed with noise in a controlled way, so that both clean and noisy spectra (PSD) were available. Then, we used this training set to tune the following three associators:

- an audio-only associator A estimating $P_s(\theta)$ from $P_x(\theta)$ only;
- a video-only associator V estimating $P_s(\theta)$ from the video input only;
- an audio-visual associator AV estimating $P_s(\theta)$ from both $P_x(\theta)$ and the video input.

Concerning the representation of the PSDs, a linear prediction (LP) model was used (Markel and Gray, 1976). There are several justifications for this choice. Since the lip contour provides no information about the source, only the global shape of the short-time spectral amplitude (STSA) of the signal can be estimated from a labial shape, rather than any information on the fine (temporal) structure of the signal.[2] It so happens that the LP model is an efficient method for coding the STSA envelope with a small number of coefficients that can be used by the associators. Furthermore, the LP model has the form of an all-poles transfer function, that is to say a numeric filter, which can be easily used to build the enhancement filter.

Thus the A, V, or AV associators deliver an estimated spectrum of the signal modeled by LP equations, that is,

$$\hat{S}_s(z)=\frac{\hat{G}_s}{1+\Sigma_{i=1}^{p}\hat{a}_{si}z^{-i}}=\frac{\hat{G}_s}{\hat{A}_s(z)}, \tag{2}$$

where $\hat{A}_s(z)=1+\Sigma_{i=1}^{p}\hat{a}_{si}z^{-i}$ is the polynomial of the estimated LP signal model and $\hat{G}_s$ is its gain (the detailed protocol for this estimation will be described in the methodology sub-sections). Then, the PSD $P_s(\theta)$ can be estimated by $|\hat{S}_s(e^{j\theta})|^2$. The observation spectrum is estimated with the LP model calculated directly on the noisy signal, that is,

$$S_x(z)=\frac{G_x}{1+\Sigma_{i=1}^{p}a_{xi}z^{-i}}=\frac{G_x}{A_x(z)}, \tag{3}$$

and $P_x(\theta)$ can be estimated by $|S_x(e^{j\theta})|^2$. Now, the enhancement filter is defined by

$$H(z)=\frac{|\hat{S}_s(z)|^2}{|S_x(z)|^2}=\frac{\hat{G}_s^2}{G_x^2}\,\frac{A_x(z)A_x(z^{-1})}{\hat{A}_s(z)\hat{A}_s(z^{-1})}. \tag{4}$$

This filter was implemented with the two following restrictions. First, since no information about the source energy can be estimated from the lip shape, $\hat{G}_s$ remained unknown. In other words, only the $\hat{a}_{si}$ coefficients could be estimated from the lip shape (i.e., the coefficients that describe the global shape of the spectrum but not its energy level). Thus the filter was defined without its gain, and the output signal was renormalized in energy with the same energy as the input signal. Second, to simplify the model, only the causal factorized part of the filter was considered, that is,

$$H(z)=\frac{A_x(z)}{\hat{A}_s(z)}. \tag{5}$$

Hence, the filtering can be considered as a two-step process, summarized in Fig. 3. In the first step, an LP analysis was performed on the noisy signal and the noisy residual signal was extracted by filtering through the inverse noisy LP filter $A_x(z)$. Then, a new signal was synthesized by filtering the residual through the LP filter $1/\hat{A}_s(z)$, estimated by the associators. Enhancement was expected, provided that the estimated spectrum was close enough to the true spectrum of the signal, or at least, closer than the noisy spectrum $1/A_x(z)$. For continuous speech, the processing was performed frame-by-frame in synchrony with the extraction period of the video parameters. A trapezoidal window was applied to frame junctions to ensure the continuity of the filtered signal.

III. A, V, AND AV ENHANCEMENT OF VOWEL-PLOSIVE-VOWEL SEQUENCES

A. Stimuli

1. Phonetic content and predictions

To evaluate the feasibility of the method and to assess its potential advantages and drawbacks, we worked on a simple corpus consisting of single speaker French vowel–consonant sequences of the form $\mathrm{V_1CV_2CV_1}$, where $\mathrm{V_1}$ and

TABLE I. Phonetic features for the vowel set [a,i,y,u] (a) and the plosive set [p,t,k,b,d,g] (b). Only the height feature is required for the vowel [a]. The [a] stimuli/answers were not used in the computation of transmitted information for the rounding and backness features.

(a) Vowel	Height	Rounding	Backness
a	−	?	?
i	+	−	−
y	+	+	−
u	+	+	+

(b) Consonant	Place	Voicing
p	labial	unvoiced
t	dental	unvoiced
k	velar	unvoiced
b	labial	voiced
d	dental	voiced
g	velar	voiced

V_2 were within the set [a,i,y,u] and C was within [p,t,k, b,d,g]. The 96 $V_1CV_2CV_1$ sequences ($4\times V_1$, $6\times C$, $4\times V_2$) were repeated twice, once for training, the second time for testing. The advantage of this corpus is that it enables us to make a number of predictions based on previous experiments on audio-visual speech perception. Indeed, the vowel set displays the three basic phonetic contrasts for vowels, namely height (e.g., [a] vs [i]), backness (e.g., [y] vs [u]), and rounding (e.g., [i] vs [y]), these last two contrasts being independent in French (Table Ia). It is well-known that rounding and height are visible while backness is not. It has also been shown (Robert-Ribes *et al.*, 1998) that the auditory information of these vowels is distributed in a manner complementary to the visual information: the most visible contrast [i y] is the least robust in acoustical noise. We expected that this complementarity of the audio and video channels should play an important role in the comparison of the AV speech enhancement process with the A-alone or V-alone processes. The same kind of pattern holds for the plosive set, including a visible place contrast (bilabials vs dentals/velars); an almost invisible one, at least when only the lip information is provided (dentals vs velars); and an invisible but quite audible mode contrast between voiced and unvoiced consonants (Table I b). Once more, the audio and video channels are complementary (see Summerfield, 1987), and the AV enhancement process should be expected to increase place intelligibility, if not voicing.

2. Audiovisual characteristics

The video parameters used in this work were geometric parameters describing the lip contour. The choice of such parameters is justified by three main reasons. (i) Lip shape provides the main contribution to visual speech information. By isolating the lip movements in speech perception tests, Summerfield (1979) and Le Goff *et al.* (1996) have shown that lip information represents about two-thirds to three-fourths of the total intelligibility gain obtained when seeing the speaker's face. (ii) It happens that the most relevant information is at the same time the easiest to extract. Different video lip shape tracking systems have been elaborated to focus on this particular region of the face. The lip shape extraction process is usually based on texture contrast with the background skin. Such a system has been developed in our laboratory to record and analyze synchronized sounds and lip movements (Lallouache, 1990). This system, which is used in the present study, can automatically extract basic parameters of the lip shape contour in two steps. First, the lip contour is isolated from the remainder of the image using blue lipstick and a Chroma-Key system. Second, classical pixel-based contour tracking algorithms are applied.[1] (iii) These parameters represent an efficient ''coding'' of the visual input, since information is well concentrated in a small number of coefficients. Several studies have been completed that efficiently characterize lip shape in French (Abry and Boë, 1980, 1986; Benoît *et al.*, 1992). We chose the three parameters that appear to be the most informative, namely interolabial width (LW), height (LH), and area (LS) (see Fig. 3). Notice that LS is highly correlated with the product $LW.LH$, which provided an indirect way to introduce a quadratic term into the linear associator defined later.

Each audio-visual stimulus consisted of an audio segment with a duration around 500 ms, sampled at 16 kHz, paired with a video matrix of $[LW\,LH\,LS]$ vectors extracted every 20 ms, according to the 50 frames-per-second camera sampling.

B. Noise degradation and LP model generation

Stationary white Gaussian noise (generated by computer) was added to each audio stimulus of both the training and test sets with signal-to-noise ratios (SNR) in the set {∞, 18, 12, 6, 0, −6, −12, −18 dB} (SNR=∞ means that no noise was added).

Then, all (clean and noisy, training and test) audio stimuli were cut into 32 ms frames synchronized with the video parameters. This involved an audio window overlap of 12 ms to synchronize with the 20 ms video period. For each audio frame, a 20-order LP spectrum was calculated using the autocorrelation method and the Durbin–Levinson algorithm (Markel and Gray, 1976).

On the one hand, the LP spectra of the training corpus were used for the training of the associators (see Sec. C), and the LP spectra of the test corpus were used for the Gaussian classification test of Sec. E. On the other hand, the noisy stimuli of the test set were processed by the system (in the same configuration as above in what concerns synchronization, overlap, and LP model calculation) and used in the perceptual test of Sec. F. Note that when generating the noisy stimuli, the SNR was defined as the ratio of the signal energy and the noise energy on each entire stimulus so that the noise was stationary over its entirety. But when generating the LP spectra, the SNR was defined as the ratio of the signal energy and the noise energy on each frame. In this latter case, the noise was added frame by frame, so that the complete set of eight SNRs was well controlled in the training process and classification test (see the following sections).

C. Associators for filter parameter estimation

For each noisy frame of the signal to be processed, the A, V, and AV associators were designed to estimate the $1/\hat{A}_s(z)$ spectrum from either $1/A_x(z)$, or the corresponding set of video parameters $[LW\,LH\,LS]$, or both.

The first tool chosen for performing the estimation was linear regression. Its efficiency has already been demonstrated in different works involving estimation of speech spectral parameters from characteristics of the speaking face (Robert-Ribes *et al.*, 1996; Teissier *et al.*, 1999; Yehia *et al.*, 1998). The principle is simply to estimate each output spectral parameter as a linear combination of the input (audio and video) parameters. The regression matrix **M** of linear combination coefficients was obtained by minimizing the mean square error $e=\|\mathbf{M}_I\mathbf{M}-\mathbf{M}_O\|_2$, where $\mathbf{M}_I$ and $\mathbf{M}_O$ were two matrices concatenating the input and corresponding output parameter sets contained in the training corpus.

As far as the input/output LP parameters are concerned, preliminary tests based on spectral distances and classification tests (as described in Sec. E) showed that the best estimation performances were obtained with a spectral amplitude representation, consisting of the logarithmic values of the amplitudes of the LP spectrum taken for 50 values spaced equally on the upper-half unit circle. To obtain the $1/\hat{A}_s(z)$ filter from the 50 output spectral parameters, an inverse FFT was processed on the squared linear-scale coefficients, and a 20-order Levinson procedure was performed on the resulting estimated autocorrelation coefficients.

The A, V, and AV linear associators were trained on the 96 stimuli in the training set, with altogether about 2400 audio-visual vectors (about 25 frames per stimulus). In each condition (A, V, or AV), the associator output consisted of the set of the 50 values of $1/\hat{A}_s(z)$ for each training frame. In the A condition, these values were estimated from $1/A_x(z)$ only. Each input was hence a vector concatenating the 50 values of $1/A_x(z)$ and the value 1 (thereby ensuring that the intercept value of the regression need not be zero). Thus the associator **M** was a 51×50 matrix. In the V condition, $1/\hat{A}_s(z)$ was estimated from the $[LW\,LH\,LS]$ triplet, and **M** was a 4×50 matrix. In the AV condition, $1/\hat{A}_s(z)$ was estimated from both the 50 values of $1/A_x(z)$ and from $[LW\,LH\,LS]$, resulting in a 54×50 matrix for **M**.

In the A and AV conditions, the associator training was realized with audio inputs corrupted at different noise levels for better generalization with respect to SNR. Then the association process combined two associators tuned under two different training/processing conditions. One was dedicated to stimuli with ''large'' SNRs: in the training phase, the stimuli frames were presented at frame SNRs of ∞, 18, 12, 6, and 0 dB. The other one was dedicated to stimuli with ''small'' SNRs: the stimuli frames were presented at frame SNRs of 6, 0, −6, −12, and −18 dB. During the enhancement process, each frame was submitted to a linear discriminant analysis (trained on the same corpus) to decide whether it belonged to the large or small SNR condition, so that the corresponding associator could be applied. Pilot tests carried out on the complete training corpus showed that this linear discriminant analysis could separate frames with SNR lower than 0 dB or higher than 6 dB with less than 1% errors. Between 0 and 6 dB, the two associators provided quite similar outputs.

D. Filtered stimuli

Once the three associators had been trained, the 768 stimuli in the test set (96 sequences, 8 SNRs) were processed in the following way. First, for each 32 ms video-synchronous frame, the LP normalized spectrum $1/A_x(z)$ and the residual signal were computed, and the 50 values of the spectrum log amplitude were extracted. Second, these values and/or the video input $[LW\,LH\,LS]$ were used to estimate the 50 values of $1/\hat{A}_s(z)$ (in the A or AV condition, each frame was first submitted to the linear discriminant analysis to select the large or small SNR associator). Then, an inverse FFT was performed on the linearized 50 estimated values of $1/\hat{A}_s(z)$, providing autocorrelation coefficients; the filter parameters were obtained from these coefficients by a 20-order LP model. Last, the residual signal was processed in this filter; energy was normalized and a trapezoidal windowing was used for frame continuity. This provided us with three sets of filtered stimuli, in the A, V, and AV conditions. The evaluation of these filtered stimuli together with the unprocessed noisy stimuli was made objectively, by a classification test, and subjectively, by a perceptual identification test.

E. Gaussian classification test

1. Methodology

The objective evaluation of the process was made by a classification test performed separately on the vowel and plosive spectra. For each of the 96 sequences in both the training and test corpus, we first manually selected two frames within the vocalic nuclei of each vowel and two frames containing or just preceding the burst of each consonant. Altogether, this provided us with 576 vowel frames in both the training and test corpus (96 stimuli×3 vowels per stimulus×2 frames per vowel)—that is to say 144 per vowel category (four categories)—and 384 consonant frames (96 stimuli×2 plosives per stimulus×2 frames per plosive)—that is to say 64 per plosive category (six categories).

From the frames selected in the training corpus, we found the Gaussian distribution associated with the four vowels and the six plosives. Since the number of data were small compared to the number of input parameters, the number of audio parameters was reduced from 50 to 10 by means of a principal component analysis (PCA). Both the PCA and the Gaussian distribution parameters for the ten classes (means and covariance matrices) were determined with the audio data selected in the training corpus and presented at the three largest SNRs (∞, 18, 12 dB). The ten first components in the PCA represented 97% of the whole variance in this training set.

Then the selected vowel frames in the test set were submitted to Gaussian classification. This means that for each frame, ten PCA spectral components were computed, and *a priori* probabilities of this vector of ten components were calculated for each of the four vowel Gaussian distributions estimated from the training set. The frame was identified as belonging to the category providing the highest *a priori* probability. Similarly, the selected plosive frames in the test set were classified in reference to the six plosive Gaussian distributions. Both unfiltered frames and A, V, and AV filtered frames (that is, the spectra at the output of the A, V, and AV associators) were submitted to the classification process for the eight selected SNRs.

Results are presented in terms of classification scores and transmitted information scores. The classification scores were normalized with respect to chance performance according to the formula

$$\text{Corrected score} = 100\,\frac{\dfrac{\text{correct responses}}{\text{total responses}} - \dfrac{1}{\text{number of categories}}}{1 - \dfrac{1}{\text{number of categories}}}. \qquad (6)$$

The numbers of classification of each phoneme i for presentation of each phoneme j were gathered into confusion matrices. From these matrices, the transmitted information (Miller and Nicely, 1955; Breeuwer and Plomp, 1986; Robert-Ribes *et al.*, 1998) was computed for the three vocalic phonetic features introduced in Sec. A 1, namely height, rounding, and front–back contrast, and the two consonantic phonetic features, namely voicing and place (see Table I). The percentage of transmitted information is defined by:

$$T = 100\,\frac{H(s,r)}{H(s)}, \qquad (7)$$

with $H(s,r)$ the transmitted information from stimuli s to answers r, and $H(s)$ the existing information in the stimuli. These values are defined by:

$$H(s,r) = -\sum_i \sum_j p(s_i,r_j)\log_2\left(\frac{p(s_i)p(r_j)}{p(s_i,r_j)}\right)$$

$$H(s) = -\sum_i p(s_i)\log_2(p(s_i)),$$

with $p(s_i)$ the probability of occurrence of feature s_i in the stimuli, $p(r_j)$ the probability of occurrence of feature r_j in the answers, $p(s_i,r_j)$ the probability of shared occurrence of feature s_i in the stimuli and feature r_j in the answers. If we denote n the total number of stimuli and n_i the number of occurrences of stimulus s_i (both fixed), and n_j the number of occurrences of answer r_j and n_{ij} the number of occurrences of stimulus s_i with answer r_j (both provided by the confusion matrices), then $p(s_i)$ is known as n_i/n; and $p(r_j)$ and $p(s_i,r_j)$ are not known but can be estimated by n_j/n and n_{ij}/n.

2. Results

The correct classification scores are displayed in Fig. 4(a) for the unfiltered and filtered vowel frames. First, we notice that the scores in the unfiltered condition decreased

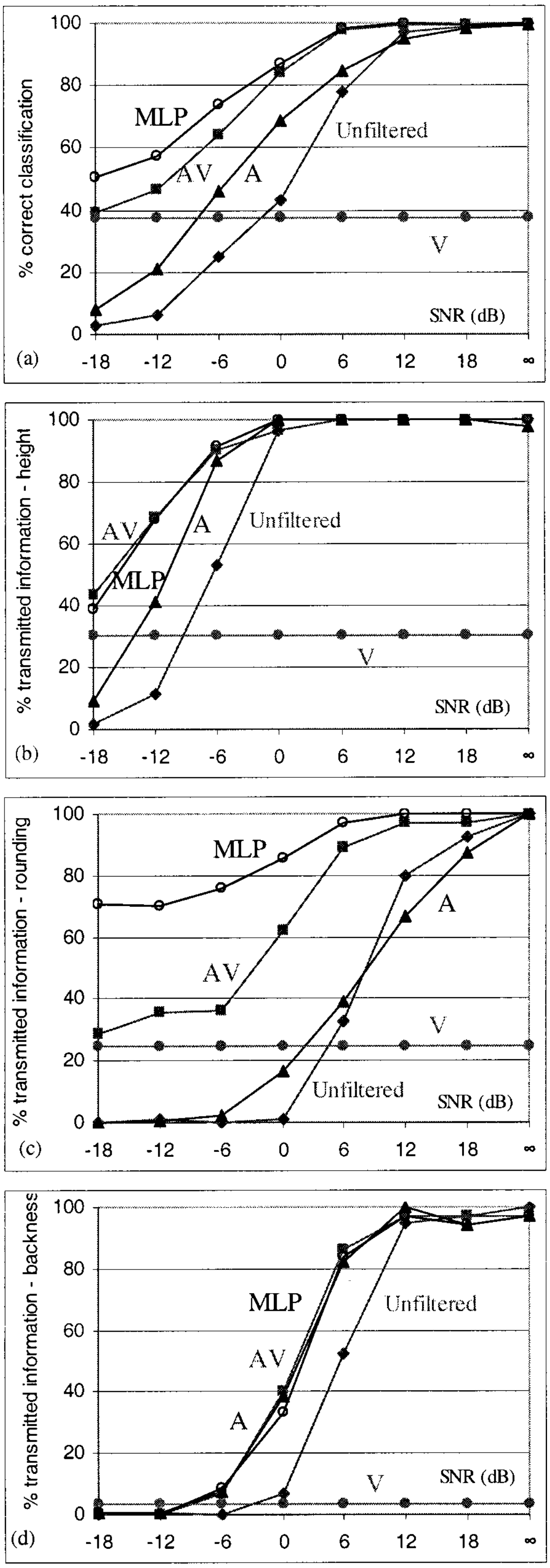

FIG. 4. Gaussian classification scores for the four vowels, in the unfiltered condition and AV, A, and V filtered conditions. MLP stands for the Multi-Layer Perceptron used in Sec. IV. (a) Corrected percentage of correct classification; percentage of transmitted information for the height, (b), rounding, (c), and backness, (d) features.

from 100% with clean stimuli to close to 0% for SNR$=-18$ dB. At such a small SNR value, all spectra were roughly flat, hence identification became impossible. From this baseline, we observe that the A filtered stimuli were better identified, with a gain up to 20% or more around -6 dB to 0 dB [at 0 dB, 43% vs 69%, $\chi^2(1)=77.1$, $p<0.001$]. This shows that the associator did learn some helpful relationship between noisy and clean audio spectra. More interestingly, we notice that the V filtered stimuli were partly recognizable, with a score around 40%. Finally, the AV filtered condition shows the efficiency of the system, with an increase in recognition score around 40% at SNRs lower than 0 dB [at 0 dB, 43% vs 84%, $\chi^2(1)=211.3$, $p<0.001$]. The scores in the AV condition were higher than in the A condition for SNRs below 18 dB [e.g., at 0 dB, 84% for AV vs 69% for A, $\chi^2(1)=38.5$, $p<0.001$] and similar above 18 dB. The AV scores were always higher than the V scores for all SNRs except -18 dB where the AV and V scores were similar. The superiority of AV over V holds even at very low SNRs where the audio information is very poor [e.g., at -12 dB, 46.5% for AV vs 37% for V, $\chi^2(1)=10.1$, $p<0.005$]. To summarize, the inequality AV$\geqslant$(A or V) was verified for the classification scores. This shows the ability of the AV system to ''reshape'' vowel spectra, and to efficiently exploit the complementarity of the A and V sensors.

In terms of individual phonetic features, the transmitted information scores for the vowel features [Figs. 4(b)–(d)] show that each feature was improved by the process in the AV condition. The rounding and height contrasts were well maintained up to the largest amounts of noise in the AV filtered condition compared to the unfiltered (and the A and V filtered) condition(s). The only surprise comes from the low score for the rounding feature in the V condition (and AV condition at SNR$=-18$ dB), since this feature is considered highly visible. We shall come back on this in Sec. IV. For the front–back feature, the AV condition was similar to the A condition, which was expected since the video information is quite poor for this feature, as shown by the V score close to 0%. Altogether, the inequality AV$\geqslant$(A or V) was confirmed for the different features. These results are a first indication of the efficiency of the DE architecture to combine the video and audio information for vowel enhancement.

The results for consonants are more disappointing. Absolute scores were quite poor, but they must be taken cautiously, since it is well known that local information is too restricted. Hence Gaussian classifiers were not powerful enough to achieve fully acceptable performances. However, compared scores are relevant, and they demonstrate that the consonants were poorly improved by the AV filtering process (which was again more efficient than the A or V processes). The scores for large SNRs were even decreased by the filtering procedure while the gains for the small SNRs were less than 15% [max at 0 dB, 5% for unfiltered vs 19% for AV, $\chi^2(1)=36.7$, $p<0.001$] [Fig. 5(a)]. The transmitted information scores [Figs. 5(b)–(c)] show that the voicing feature was severely degraded by both the A and AV filtering procedure at large SNRs. The place feature was degraded by the A filtering over 6 dB SNR and was slightly or not enhanced by the AV filtering for almost all SNRs (some degradation is even induced at 18 dB), in spite of the high visibility of the labial versus nonlabial contrast.

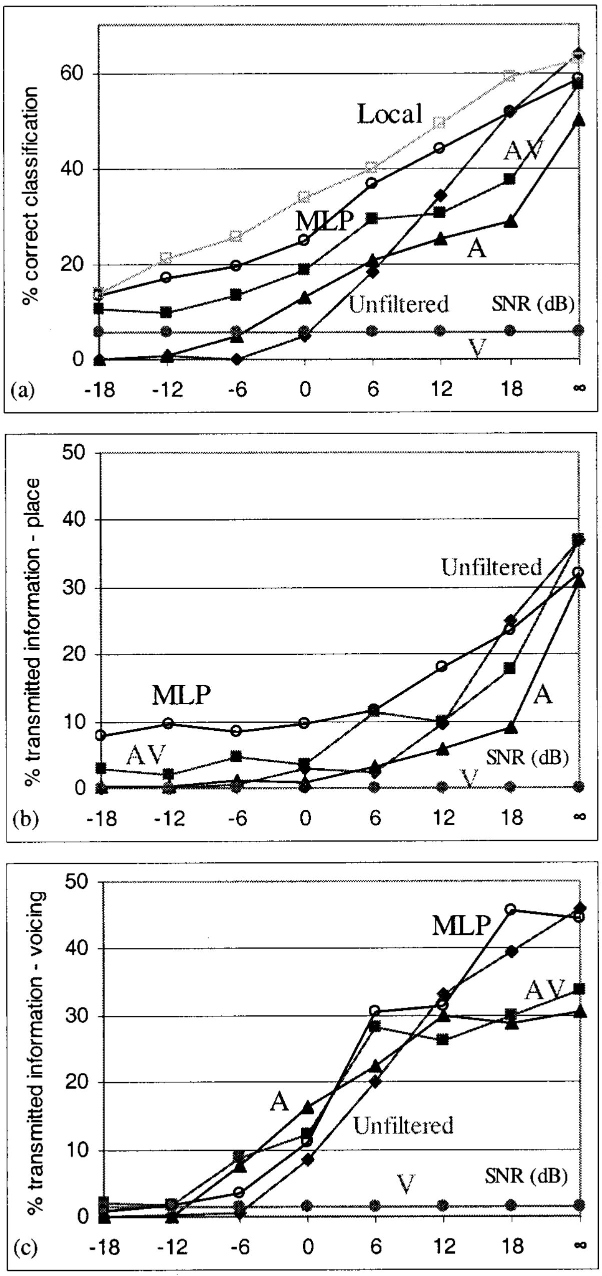

FIG. 5. Gaussian classification scores for the six plosives, in the unfiltered condition and AV, A, and V, filtered conditions. ''Local'' stands for the local consonant associator introduced in Sec. III G, and MLP stands for the Multi-Layer Perceptron used in Sec. IV. (a) Corrected percentage of correct classification; percentage of transmitted information for the place, (b) and voicing (c) features.

F. Perceptual test

1. Methodology

To perform a complete assessment of the system, a perceptual identification experiment was carried out on the 96 stimuli of the test corpus. Sixteen French native subjects, aged from 22 to 31, were tested. They displayed no known problem of hearing and speech production/perception. They were asked to identify both vowels and consonants of the stimuli presented in the unfiltered and A, V, and AV filtered conditions, for the eight SNRs. It should be remembered that in this test, the SNR was defined as the ratio of the signal energy and the noise energy over the complete stimulus (and no longer for each frame).

The $V_1CV_2CV_1$ stimuli were segmented manually into V_1CV_2 and V_2CV_1 items so that the subjects would hear V_1 and C only once for each presentation. The total of 6144 sounds (96 stimuli×2 segments×8 SNRs×4 conditions) was randomized and divided between the 16 subjects. For each stimulus, they were asked to give a response between the four possible vowels for both V_1 and V_2, and between the six possible plosives for C. There were altogether 384 vowel responses and 192 consonant responses for each noise level and each condition (96 stimuli, two VCV segments per stimulus, 2 vowels and 1 plosive per segment), that is to say 96 responses per vowel category (4 categories) and 32 responses per consonant category (6 categories). These responses were analyzed separately for the vowels and the plosives and processed in the same way as for the Gaussian classification test, that is, with corrected global scores, confusion matrices, and transmitted information scores for the individual phonetic features.

2. Results

The results are presented in Fig. 6 for the vowels and Fig. 7 for the plosives. Concerning the vowels, we first notice that the degradation of identification scores for the unfiltered stimuli with low SNRs was similar to classification scores. However, the A filtering was not very efficient: the gain was low at low SNRs [at SNR$=-12$ dB, 20% for unfiltered vs 29% for A, $\chi^2(1)=7.8$, $p<0.005$]. There was even some degradation due to filtering at high SNRs [at SNR$=18$ dB, 96% for unfiltered vs 91% for A, $\chi^2(1)=9.6, p<0.005$]. The V filtering provided some information, with an intelligibility around 20%–30%. Notice that the fluctuations in score with SNR were due to differences in the residual form (see Fig. 3). However, the results in the AV condition were quite good, although less so than in the Gaussian classification test, with an increase in identification scores compared with unfiltered stimuli at all SNRs below 18 dB. The gains reach 5.5% at 12 dB [$\chi^2(1)=9.2, p<0.005$], 9% at 6 dB [$\chi^2(1)=12.6$, $p<0.001$], 17.5% at 0 dB [$\chi^2(1)=24.6$, $p<0.001$], 16% at -6 dB [$\chi^2(1)=20$, $p<0.001$], and about 22% at -12 dB [$\chi^2(1)=41.7$, $p<0.001$], and -18 dB [$\chi^2(1)=77.8, p<0.001$]. Once more, the AV condition was systematically better than the A and V conditions.

The percentages of transmitted information, given in Figs. 6(b)–(d), confirm these results and the ones of the classification test. The scores for the A condition were close to the scores in the unfiltered condition with some improvement for low SNRs and some degradation for high SNRs. The V condition, while providing some information on the height contrast, was surprisingly deceptive for the rounding contrast. However, in the AV condition the system provided much better performances that can be summarized as follows. (i) There was an efficient reinforcement of the round-

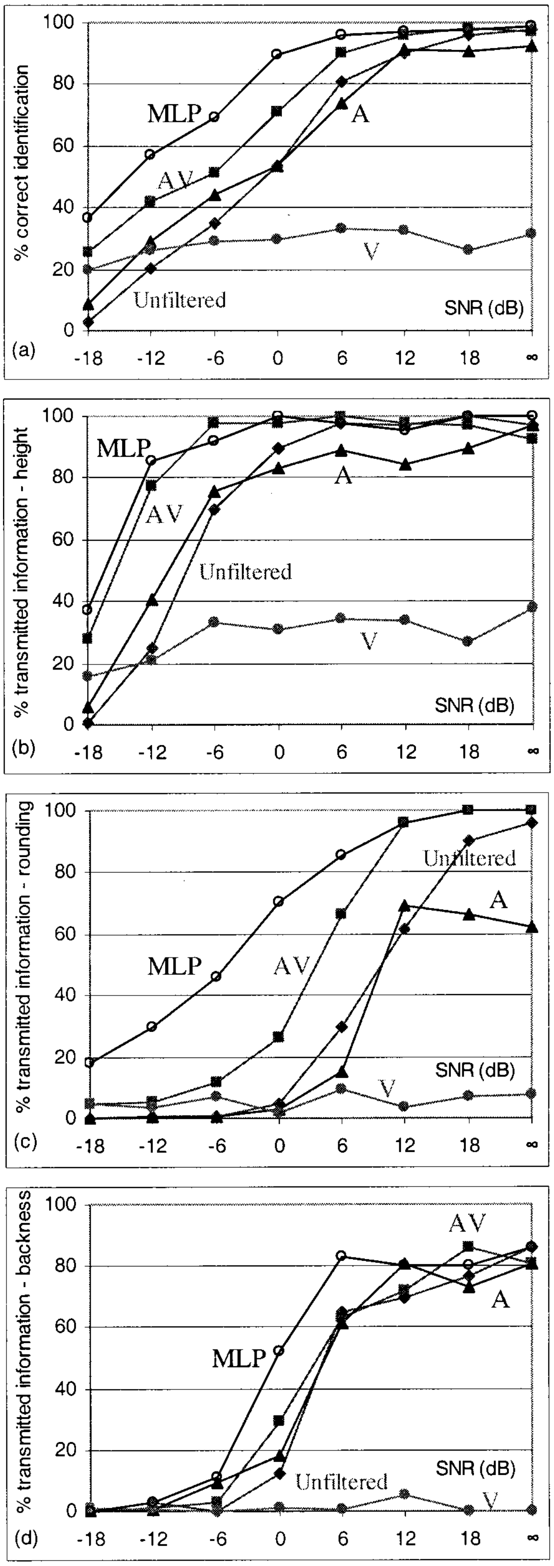

FIG. 6. Perceptual test scores for the vowels, in the unfiltered condition and in the AV, A, and V filtered conditions. MLP stands for the Multi-Layer Perceptron used in Sec. IV. (a) Corrected percentage of correct identification; percentage of transmitted information for the height (b), rounding (c), and backness (d) features.

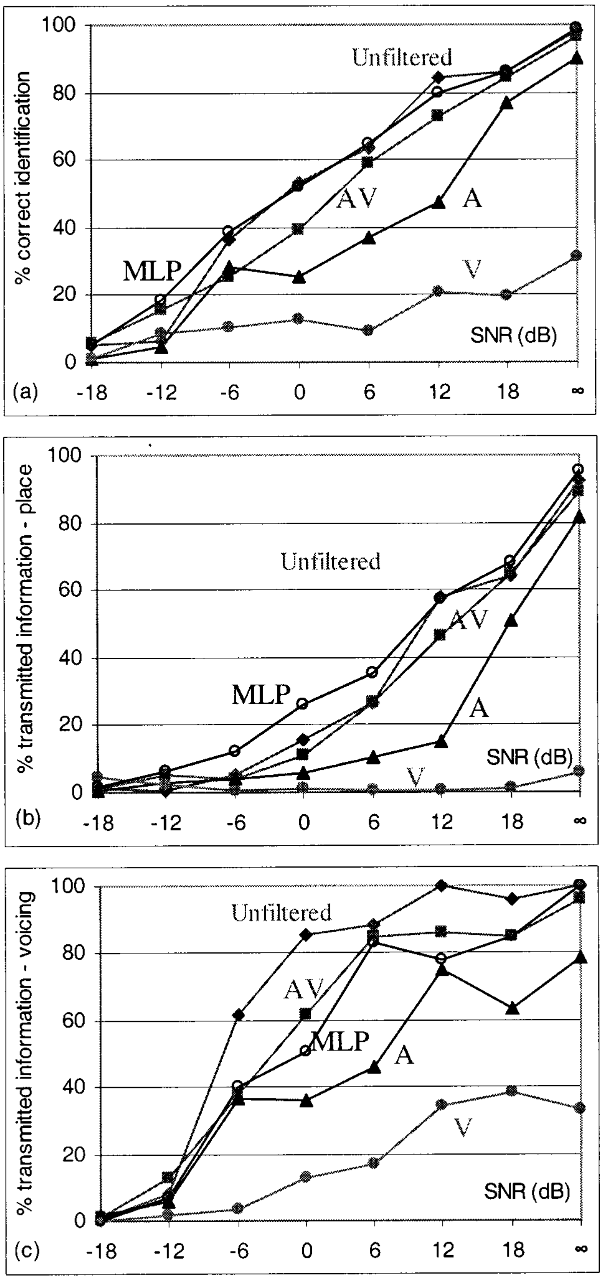

FIG. 7. Perceptual test scores for the consonants, in the unfiltered condition, and in the AV, A, and V filtered conditions. MLP stands for the Multi-Layer Perceptron used in Sec. IV. (a) Corrected percentage of correct identification; percentage of transmitted information for the place (b) and voicing (c) features.

ing feature: the [i,y] contrast, which was rapidly and strongly degraded in noise before enhancement, was well recovered. This case represents a good example of the audio/video complementarity of speech (robust video distinction while it is the weakest audio contrast in noise; see Robert-Ribes *et al.*, 1998). (ii) Contrary to the rounding feature, the height [a] versus [i y u] contrast was robust in the unfiltered condition (until 0 dB SNR). This is due to the good audibility of the first formant region in white noise. Below 0 dB, the AV filtering produced a large improvement of the height feature scores. (iii) At last, the front–back [y,u] contrast, not very robust in the unfiltered condition, was not improved by the AV process.

For consonants, the perceptual results were poor and confirmed the classification test results. While still better than in the A and V conditions, the identification scores [Fig. 7(a)] in the AV condition remained lower than those for unfiltered stimuli [e.g., at SNR=0 dB, 53% for unfiltered vs 39.5% for AV, $\chi^2(1)=7.3$, $p<0.01$], except for the two weakest SNRs [e.g., at SNR=−12 dB, 6% for unfiltered vs 15.5% for AV, $\chi^2(1)=8.7$, $p<0.005$]. This shows that at this point the system produced some degradation of the consonants. As shown by Figs. 7(b)–(c), the voicing feature was degraded by the process, while the place feature was not improved despite the labial information available in the data (the V condition seemed unable to exploit this information as well).

G. Discussion

The Gaussian classification tests on single frames and the perceptual identification of whole sequences provided a consistent pattern of results. AV filtering produced a strong enhancement of vowels, at all SNRs, and always much stronger than both the A and the V filtering. The results for consonants were much more disappointing: A, V, or AV enhancement filters failed to improve on consonant identification or classification except at the poorest SNR conditions, and often resulted in lower scores than those obtained when no filter was applied. Our interpretation is that in the present implementation, the linear associator was more adapted to vowel spectra than to plosives. There are two reasons for this. First, vowellike spectra (consisting of well-defined formant patterns) occurred more frequently in the corpus than plosivelike spectra (including silence, consonantal voicing and bursts). Second, vowel spectral contrasts were larger than plosive ones. Therefore, the intrinsic ''averaging'' process characteristic of linear regression resulted in estimated plosive spectra that looked quite similar to vowels! This was obvious through visual inspection of the filtered stimuli.

To confirm this hypothesis, we carried out a complementary experiment where the training corpus for the AV associator training was restricted to only plosive frames (the two frames per consonant defined in Sec. E 1). In this case, the linear regression algorithm was focused on the available information: it appeared that the plosive test frames filtered by this local AV association process provided much higher scores in the Gaussian classification test. Indeed, these scores, displayed in Fig. 5, were this time systematically better than those of unfiltered stimuli, with a gain close to 30% at 0 dB [5% for unfiltered vs 34% for local associator, $\chi^2(1)=104$, $p<0.001$]. This shows that the poor scores obtained previously were not due to a lack of information in the A and V sensors but to an under-representation of this information in the filtering process. The aim of the next experiment was to exploit an association tool more powerful than linear regression.

IV. NONLINEAR ASSOCIATORS FOR IMPROVING PLOSIVE ENHANCEMENT

A. Improving the association process: From linear regression to perceptrons

The previous section revealed the need for more powerful associators than linear regression (LR in the following), in order to better take into account the plosive parts of the corpus. We decided to look for nonlinear associators in the most efficient condition of the previous experiment, that is, the AV one. Neural Networks have been used extensively for classification tasks in speech recognition, including audio-visual recognition (Stork and Hennecke, 1996). In addition, they are theoretically able to approximate any nonlinear function. Therefore we used classical Multi-Layer Perceptrons (MLP) based on error gradient back-propagation with momentum (Rumelhart *et al.*, 1986), with one hidden layer and sigmoidal neuronal threshold functions.

Different values for the number of neurons in the hidden layer were tested from 20 to 200, using both Gaussian classification and listening tests. It appeared that performances improved slightly above 40 neurons. The following results were obtained with 120 hidden neurons, which was a good compromise between performances and calculation cost. The complete experimentation protocol of Sec. III concerning training and testing was preserved, so that the new results can be compared with the results obtained with linear regression. The training phase involved 200 iterations with the whole training set, which was enough to ensure convergence of the network (i.e., low error and no overtraining).

B. Gaussian classification test

The output spectra obtained with the MLP on the selected frames defined in Sec. III E 1 were presented to the Gaussian classifier of the same section and the results are displayed in Figs. 4 and 5.

For vowels, the classification scores for the MLP were improved compared to the linear regression below 6 dB. The gain reached more than 10% at −18 dB [39.5% vs 51%, $\chi^2(1)=15$, $p<0.001$] and at −12 dB [46.5% vs 57.5%, $\chi^2(1)=14.2$, $p<0.001$]. The transmitted information scores show that this gain was provided by a quite large improvement of the rounding feature, while the height and backness features were not noticeably modified. This provides an interesting correction to the surprisingly low score for the rounding feature noticed in Sec. III E 2.

For consonants, the recognition scores were also increased compared to linear associators, and reached a value almost always higher than the scores for unfiltered stimuli. The gain reached 14% to 20% from −18 dB to 6 dB [at 0 dB, 5% for unfiltered vs 25% for MLP, $\chi^2(1)=62.7$, $p<0.001$], with only a small and not significant loss at SNR=∞ [64% for unfiltered vs 58.5% for MLP, $\chi^2(1)=2.3$, $p>0.1$]. However, the scores of the ''local'' associator described in Sec. III G were not reached. Hence, the available information was not exploited completely. At the phonetic features level, transmitted information scores were almost always higher with the MLP than with linear regression. Consequently, the voicing feature was more or less at the same level for MLP filtered and unfiltered spectra, while the place feature was much improved by the MLP filtering below 18 dB SNR. These encouraging results led us to perform a complementary perceptual test for a final evaluation of the system on this corpus.

C. Perceptual test

The MLP-processed stimuli were presented to the 16 subjects in the same condition as the unfiltered or linear regression filtered stimuli. Global results confirm the important progress from LR to MLP estimation (Figs. 6 and 7). In the MLP condition, the increase of vowel intelligibility with respect to the unfiltered condition was quite large [about 35% for SNRs lower than 0 dB: e.g., at −12 dB, 20% for unfiltered vs 57% for MLP, $\chi^2(1)=110$, $p<0.001$]. These results correspond to a gain in SNR around 9 to 12 dB (Fig. 6). At the phonetic features level, transmitted information scores reveal a large improvement from LR to MLP for vowel rounding below 12 dB SNR, and also some improvement for the vowel backness feature between −6 and +12 dB [Fig. 6(d)]. This demonstrates the ability of the MLP-DE structure to efficiently combine the video and audio information for vowel enhancement.

For consonants, there was also a significant improvement from LR to MLP [13% at SNR=−6 dB: 26% for LR vs 39% for MLP, $\chi^2(1)=7.6$, $p<0.01$, and 12.5% at 0 dB: 39.5% for LR vs 52% for MLP, $\chi^2(1)=6$, $p<0.025$]. As a result, the identification scores of the MLP-processed plosives reached the values for unfiltered stimuli, and the average score across the whole SNR range was almost the same as the one obtained in the unfiltered condition: 55.5% for MLP filtered vs 54.1% for unfiltered stimuli [the difference is not significant, $\chi^2(1)=0.6$, $p>0.4$]. In comparison the LR scores only reached 49.8%, showing an average degradation of about 4% compared with the unfiltered condition [$\chi^2(1)=5.7$, $p<0.025$] [Fig. 7(a)]. The voicing feature was still quite degraded in the MLP filtered condition compared with the unfiltered condition [Fig. 7(c)]. In contrast, the place feature was quite improved by the MLP-based process, especially from −12 to 6 dB SNR [Fig. 7(b)], which explains why the identification scores were comparable with the unfiltered condition. This shows that some visible information (notably on the [p,b] closures) was used efficiently by the filtering process.

V. DISCUSSION

A. Summary of the main achievements

We defined in Sec. III an architecture for audio-visual speech enhancement, expecting some predictability of the audio spectrum from the video input, and based on a fusion-and-filtering procedure in three steps:

(1) separate the sound source from the spectral transfer function characteristics using an LP analysis of the input audio signal;
(2) combine the noisy spectral characteristics and the visual

input parameters to estimate the transfer function of the clean speech sound, using either a linear or a nonlinear associator;

(3) filter the estimated source through the estimated transfer function.

We assessed this system on a vowel-plosive-vowel corpus with both an objective Gaussian classification test and a subjective perceptual identification procedure, and we obtained the following results:

(1) the linear-regression filter estimation provided much better global results when both the audio and the video streams participated to the filter estimation (AV filtered stimuli) than in the audio-only (A filtered stimuli) or video-only (V filtered stimuli) conditions. In more detail, the AV scores were better than the A scores for low SNRs; better than the V scores for high SNRs; better than both the A and V scores for medium SNRs; and in no case lower. This demonstrates the efficiency of the visual contribution and the good exploitation of the audio-visual synergy for speech enhancement by the DE architecture;

(2) in the AV condition, vowels displayed a very large enhancement for both assessment tools, with a significant improvement of the enhancement efficiency from the linear to the nonlinear associator. This was largely due to a better enhancement of the rounding feature. The best filtering algorithm, involving AV nonlinear estimation, provided an increase in recognition and perception scores corresponding to a 9–12 dB gain along the entire SNR range for the vowels;

(3) plosives were not enhanced, and were even degraded by the use of the linear associator. However, the results with the nonlinear associator were less clearcut. Indeed, Gaussian classification displayed a significant enhancement of the stimuli through nonlinear estimation, while perceptual tests showed the same global intelligibility for unfiltered and filtered plosives. At the feature level, voicing was quite degraded for linear and nonlinear estimation, while place was enhanced only for nonlinear estimation.

B. Audio-visual interdependencies

Given this pattern of results, we can return to some issues raised in the introductory sections. The departure point of this study was the assumption that there was some interdependency between the audio and the video streams. The experimental results provide strong support to this assumption. Indeed, it appears that the AV filtering condition produces a gain of about 6 dB in SNR compared with the A condition for both vowels (Fig. 6) and plosives (Fig. 7). Hence the video stream does contain a significant deal of information on the audio spectrum, which is evidence for a statistical dependence. Notice that in this AV versus A comparison, the A condition provides a baseline which is similar to the unfiltered condition for vowels, but unfortunately lower for plosives. In this latter case, the AV associator helped at best to remove the degradation introduced by the A associator.

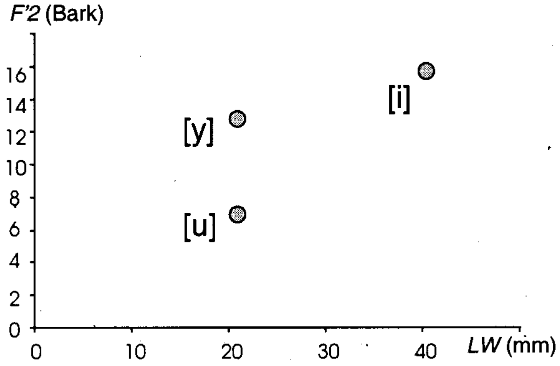

FIG. 8. Typical pattern of distribution of the auditory parameter $F'2$ and the visual parameter LW for the set of high French vowels [i y u].

Audio-visual complementarity is also demonstrated in this work. It enables the poorest audio feature to benefit greatly from audio-visual enhancement, as it was clearly demonstrated for vowels. For example, the rounding feature for unfiltered vowels provided low perceptual intelligibility scores compared to the front–back feature: 5% vs 12% at 0 dB, 30% vs 65% at 6 dB; but the MLP-AV filtered stimuli displayed the inverse pattern: 46% vs 11% at −6 dB, 71% vs 52% at 0 dB, and 100% vs 80% at 18 dB [Figs. 6(b)–(c)].

Finally, it is of interest to notice the importance of introducing nonlinear processes in audio-visual fusion. Indeed, it is commonly considered that linear association between audio and video speech parameters captures a great deal of the information (see, e.g., Yehia *et al.*, 1998; Robert-Ribes *et al.*, 1996). However, it is clear that only gross audio-visual correlations can be captured by linear regression, while audio-visual complementarity is intrinsically associated to a basic nonlinear property of the audio-visual relationship. Consider, for example, the [i,y,u] set of French high vowels. The major auditory parameter characterizing this set would be $F'2$, the ''perceptual second formant,'' while interolabial width LW provides a major correlate of the rounding contrast between [i] and [y] or [u]. In Fig. 8, typical values of $F'2$ and LW for [i y u] for a French male speaker display audio-visual complementarity: in LW, [i] is well separated from [y] and [u] which are almost confounded, while in $F'2$ [i] and [y] are close together and well separated from [u]. Linear regression between LW and $F'2$ would lose this complementary pattern, hence the success of nonlinear association in Sec. IV, particularly for enhancing the rounding feature for vowels.

C. Future directions

Although these results fulfill the initial objective of this study, they are not a final achievement. Indeed, the framework was quite controlled and simple. It will be necessary now to explore new directions. Three main directions can be mentioned.

First, the estimation tools were quite basic. A first effort to switch from linear regression to nonlinear MLPs led to significant progress (see Fig. 6). However, the comparison of LR, MLP, and local associators for plosive enhancement in Fig. 7 shows that all the local information is still not exploited: the MLP went more or less halfway toward what can be considered optimum under the corresponding evaluation tool. More powerful tools, such as multi-expert systems (Jordan and Jacobs, 1994) or neural gas (Fritzke, 1994), could fill this gap. Another important objective will be to explore dynamic associators, able to exploit the regularities in the evolution of audio and video parameters. Hidden Markov Models could provide a natural basis for this, as they do for pure audio enhancement of speech in noise (Ephraim, 1992). This kind of tool is likely to be crucial for plosive enhancement, since it is quite well-known that plosive characterization cannot be achieved correctly without considering spectral dynamics (Kewley-Port *et al.*, 1983; Sussman *et al.*, 1991). More generally, such tools will be also necessary for dealing with more complex conditions, involving extended corpora and noise degradations, and multi-speaker applications. This is part of a global program that will be of increasing interest in the future: i.e., to systematically explore the statistical relationship between sound and image, using such tools as mutual information between groups of parameters in the audio and the video streams.

Second, though it was chosen for this initial demonstration to ignore pure audio enhancement techniques, it will be necessary to re-introduce them in the following of this work. This will be important for dealing with nonstationary noises, and particularly with variations in the spectral patterns of the competing sources, as is the case with ''cocktail-party'' speech. We will study how to extend the so-called DE architecture to a multi-channel framework including various audio and video sensors for performing the filter estimation: we are beginning to explore a generalization of the blind separation approach to multi-modal speech sources (Girin *et al.*, 2000).

Finally, the ''joint processing of the audio and video streams,'' which was applied here to speech enhancement, could be generalized to various problems in the field of human–machine communication and telecommunication. The natural coherence and complementarity of these two data streams are already exploited in speech recognition systems, and in the development of speech synthesis systems. They could also be of benefit to audio-visual compression in videophone technology: in another study, vector quantization algorithms were applied either separately to audio and video data, or to audio-visual vectors (Girin *et al.*, 1998). The latest results showed that it was possible to save 3 bits out of 15 in the second case. This provides a quantitative estimate of the amount of redundancy in the audio and video streams.

Altogether, one can foresee the elaboration of a global platform for audio-visual speech communication, which would involve preprocessing (localization, enhancement, scene analysis...), recognition, coding, transmission and synthesis of audio-visual speech. In any case, the main objective remains to maintain the intrinsic coherence of sound and image at the heart of all the speech processing algorithms.

[1] A new generation of lip contour extraction systems without blue make-up is currently being studied in our laboratory (Revéret and Benoît, 1998). These systems are based on the fitting of a lip model with the speaker's natural lips.

[2] More complete specification of the vocal tract, and even of the source, could be available from facial parameters: see, e.g., Yehia *et al.* (2000).

Abry, C., and Boë, L. J. (**1980**). ''A la recherche de corrélats géométriques discriminants pour l'opposition d'arrondissement vocalique en français,'' in *Labialité et Phonétique*, edited by C. Abry, L. J. Boë, P. Corsi, R. Descout, M. Gentil, and P. Graillot (Publications de l'Université des Langues et Lettres, Grenoble), pp. 217–237.

Abry, C., and Boë, L. J. (**1986**). ''Laws for lips,'' Speech Commun. **5**, 97–104.

Barker, J., Berthommier, F., and Schwartz, J. L. (**1998**). ''Is primitive AV coherence an aid to segment the scene?,'' Proc. AVSP'98, Sydney.

Bailly, G., Laboissière, R., and Schwartz, J. L. (**1991**). ''Formant trajectories as audible gestures: An alternative for speech synthesis,'' J. Phonetics **19**, 9–23.

Benoît, C., Mohamadi, T., and Kandel, S. D. (**1994**). ''Effects of phonetic context on audio-visual intelligibility of French,'' J. Speech Hear. Res. **37**, 1195–1203.

Benoît, C., Lallouache, T., Mohamadi, T., and Abry, C. (**1992**). ''A set of visual French visemes for visual speech synthesis,'' in *Talking Machines: Theories, Models and Designs*, edited by G. Bailly, C. Benoît, and T. R. Sawallis (Elsevier, Amsterdam), pp. 485–504.

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

Boll, S. F. (**1979**). ''Suppression of acoustic noise in speech using spectral subtraction,'' IEEE Trans. Acoust., Speech, Signal Process. **29**, 113–120.

Boll, S. F., and Pulsipher, D. C. (**1980**). ''Suppression of acoustic noise in speech using two microphones adaptive noise cancellation,'' IEEE Trans. Acoust., Speech, Signal Process. **28**, 752–753.

Breeuwer, M., and Plomp, R. (**1986**). ''Speechreading supplemented with auditorily presented speech parameters,'' J. Acoust. Soc. Am. **79**, 481–499.

Comon, P., Jutten, C., and Hérault, J. (**1991**). ''Blind separation of sources, Part II: Problems statement,'' Signal Process. **24**, 11–20.

Driver, J. (**1996**), ''Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading,'' Nature (London) **381**, 66–68.

Ephraim, Y. (**1992**). ''A Bayesian estimation approach for speech enhancement using Hidden Markov Models,'' IEEE Trans. Acoust., Speech, Signal Process. **40**, 725–735.

Erber, N. P. (**1975**). ''Auditory-visual perception of speech,'' J. Speech Hearing Disorders **40**, 481–492.

Feder, M., Oppenheim, A. V., and Weinstein, E. (**1989**). ''Maximum likelihood noise cancellation using the EM algorithm,'' IEEE Trans. Acoust., Speech, Signal Process. **37**, 204–216.

Ferrara, E. R., and Widrow, B. (**1981**). ''Multi-channel adaptive filtering for signal enhancement,'' IEEE Trans. Acoust., Speech, Signal Process. **29**, 766–770.

Fritzke, B. (**1994**). ''Growing cell structures—A self-organizing network for unsupervised and supervised learning,'' Neural Networks **7**, 1441–1460.

Girin, L., Feng, G., and Schwartz, J.-L. (**1996**), ''Débruitage de parole par un filtrage utilisant l'image du locuteur: Une étude de faisabilité,'' Traitement du Signal **13**, 319–334.

Girin, L., Foucher, E., and Feng, G. (**1998**). ''An audio-visual distance for audio-visual vector quantization,'' Proc. IEEE Workshop Multimedia Signal Process., Los Angeles.

Girin, L., Allard, A., Feng, G., and Schwartz, J-L. (**2000**). ''Séparation de sources de parole: Une nouvelle approche utilisant la cohérence audiovisuelle des signaux,'' Proc. XXIII Journées d'Études sur la Parole, Aussois, France, pp. 57–60.

Grant, K. W., and Walden, B. E. (**1996**). ''Evaluating the articulation index for auditory-visual consonant recognition,'' J. Acoust. Soc. Am. **100**, 2415–2424.

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

Harrison, W. A., Lim, J. S., and Singer, E. (**1986**). ''A new application of adaptive noise cancellation,'' IEEE Trans. Acoust., Speech, Signal Process. **34**, 21–27.

Jordan, M. I., and Jacobs, R. A. (**1994**). ''Hierarchical mixtures of experts and the EM algorithm,'' Neural Comput. **6**, 181–214.

Jutten, C., and Hérault, J. (**1991**). ''Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture,'' Signal Process. **24**, 1–10.

Kang, G. S., and Fransen, L. J. (**1989**). ''Quality improvement of LPC-processed noisy speech by using spectral subtraction,'' IEEE Trans. Acoust., Speech, Signal Process. **37**, 939–943.

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (**1983**). ''Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants,'' J. Acoust. Soc. Am. **73**, 1779–1793.

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

Le Bouquin-Jeannès, R., and Faucon, G. (**1995**). ''Study of a voice activity detector and its influence on a noise reduction system,'' Speech Commun. **16**, 245–254.

Le Goff, B., Guiard-Marigny, T., and Benoît, C. (**1996**). ''Analysis-synthesis and intelligibility of a talking face,'' in *Progress in Speech Synthesis*, edited by J. P. H. Van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Springer-Verlag, New York), pp. 235–244.

Lim, J. S. (**1983**). *Speech Enhancement* (Prentice-Hall, Englewood Cliffs, NJ).

MacLeod, A., and Summerfield, Q. (**1987**). ''Quantifying the contribution of vision to speech perception in noise.'' Br. J. Audiol. **21**, 131–141.

McAulay, R. J., and Malpass, M. L. (**1980**). ''Speech enhancement using a soft-decision noise suppression filter,'' IEEE Trans. Acoust., Speech, Signal Process. **28**, 137–145.

Markel, J. D., and Gray, A. H. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, New York).

Massaro, D. W. (**1989**). ''Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry,'' Behav. Brain Sci. **12**, 741–794.

Miller, G. A., and Nicely, P. E. (**1955**). ''An analysis of perceptual confusions among some English consonants,'' J. Acoust. Soc. Am. **27**, 338–358.

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

Robert-Ribes, J., Piquemal, M., Schwartz, J. L., and Escudier, P. (**1996**). ''Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition,'' in *Speechreading by Man and Machine: Models, Systems and Applications*, edited by D. G. Stork and M. Hennecke (Springer-Verlag, Berlin), pp. 193–210.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., and Escudier, P. (**1998**). ''Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise,'' J. Acoust. Soc. Am. **103**, 3677–3689.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (**1986**). ''Learning representations by back-propagating errors,'' Nature (London) **323**, 533–536.

Sambur, M. R. (**1978**). ''Adaptive noise canceling for speech signals,'' IEEE Trans. Acoust., Speech, Signal Process. **26**, 419–423.

Schroeter, J., and Sondhi, M. M. (**1994**). ''Techniques for estimating vocal-tract shapes from the speech signal,'' IEEE Trans. Speech Audio Process. **2**, 133–150.

Schwartz, J. L., Robert-Ribes, J., and Escudier, P. (**1998**). ''Ten years after Summerfield... A taxonomy of models for AV fusion in speech perception,'' in *Hearing by Eye, II. Perspectives and Directions in Research on Audio-visual Aspects of Language Processing*, edited by R. Campbell, B. Dodd, and D. Burnham (Erlbaum/Psychology Press, Hillsdale, NJ), pp. 85–108.

Stork, D. G., and Hennecke, M., Eds. (**1996**). *Speechreading by Man and Machine: Models, Systems and Applications* (Springer-Verlag, Berlin).

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

Summerfield, Q. (**1979**). ''Use of visual information for phonetic perception,'' Phonetica **36**, 314–331.

Summerfield, Q. (**1987**). ''Some preliminaries to a comprehensive account of audio-visual speech perception,'' in *Hearing by Eye: The Psychology of Lipreading*, edited by B. Dodd and R. Campbell (Lawrence Erlbaum, London), pp. 3–51.

Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (**1991**). ''An investigation of locus equations as a source of relational invariance for stop place categorization,'' J. Acoust. Soc. Am. **90**, 1309–1325.

Teissier, P., Robert-Ribes, J., Schwartz, J. L., and Guérin-Dugué, A. (**1999**). ''Comparing models for audio-visual fusion in a noisy-vowel recognition task,'' IEEE Trans. Speech Audio Process. **7**, 629–642.

Widrow, B., Glover, J. R., McCool, J. M., Kaunitz, J., Williams, C. S., Hearn, R. H., Zeidler, J. R., Dong, E., and Goodlin, R. C. (**1975**). ''Adaptive noise cancelling: principles and applications,'' Proc. IEEE **63**, 1692–1716.

Yehia, H., Kuratate, T., and Vatikiotis-Bateson, E. (**2000**). ''Facial animation and head motion driven by speech acoustics,'' Proc. 5th Seminar on Speech Production: Models and Data, Munich, pp. 265–268.

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