Comparing Models for Audiovisual Fusion in a Noisy-Vowel Recognition Task

Pascal Teissier, Jordi Robert-Ribes, Jean-Luc Schwartz, and Anne Guérin-Dugué

***Abstract*—Audiovisual speech recognition involves fusion of the audio and video sensors for phonetic identification. There are three basic ways to fuse data streams for taking a decision such as phoneme identification: data-to-decision, decision-to-decision, and data-to-data. This leads to four possible models for audiovisual speech recognition, that is direct identification in the first case, separate identification in the second one, and two variants of the third early integration case, namely dominant recoding or motor recoding. However, no systematic comparison of these models is available in the literature. We propose an implementation of these four models, and submit them to a benchmark test. For this aim, we use a noisy-vowel corpus tested on two recognition paradigms in which the systems are tested at noise levels higher than those used for learning. In one of these paradigms, the signal-to-noise ratio (SNR) value is provided to the recognition systems, in the other it is not. We also introduce a new criterion for evaluating performances, based on transmitted information on individual phonetic features.**

In light of the compared performances of the four models with the two recognition paradigms, we discuss the advantages and drawbacks of these models, leading to proposals for data representation, fusion architecture, and control of the fusion process through sensor reliability.

***Index Terms*— Audiovisual speech, noisy speech recognition, sensor fusion, reliability.**

I. Introduction

SINCE the pioneer work by Petajan [15], [43], a number of specialists of automatic speech recognition have considered audiovisual (AV) speech in order to increase the robustness of their systems, particularly in noise, where the role of visual speech for improving intelligibility has often been demonstrated in humans [5], [22], [56]. However, though there exist at least four different types of architectures for AV fusion, no systematic comparison of their performances has been performed. In this paper, we introduce these four architectures in reference to classical literature on sensor fusion and cognitive psychology, and we provide some *a priori* considerations on their performances (Section II). Section III is devoted to the experimental framework. We present the AV corpus consisting of French oral vowels in acoustical noise, and we introduce a new methodological criterion for comparing AV fusion systems, based on transmitted information at the phonetic level. Section IV describes models implementation. Section V provides the benchmark results, and the discussion in Section VI presents some guides for future research.

Manuscript received March 24, 1997; revised February 1, 1999. This work was supported by the French CNRS-INPG, Fédération de Laboratoires ELESA, and by EEC through Projects HCM-SPHERE and TMR-SPHEAR. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Picone.

P. Teissier is with the Institut de la Communication Parlée, CNRS UPRESA 5009/INPG-U. Stendhal, ICP, INPG, 38031 Grenoble Cedex 1, France, and with the Laboratoire des Images et des Signaux, LIS, INPG, 38031 Grenoble Cedex 1, France (e-mail: teissier@icp.inpg.fr).

J. Robert-Ribes was with the Institut de la Communication Parlée, CNRS UPRESA 5009/INPG-U. Stendhal, ICP, INPG, 38031 Grenoble Cedex 1, France. He is now with CSIRO Division of Information Technology, Macquarie University, North Ryde, NSW 2113, Australia (e-mail: jordi.robert-ribes@syd.dit.CSIRO.AU).

J.-L. Schwartz is with the Institut de la Communication Parlée, CNRS UPRESA 5009/INPG-U. Stendhal, ICP, INPG, 38031 Grenoble Cedex 1 (e-mail: schwartz@icp.inpg.fr).

A. Guérin-Dugué is with Laboratoire des Images et des Signaux, LIS, INPG, 38031 Grenoble Cedex 1, France (e-mail: anne.guerin@inpg.fr).

Publisher Item Identifier S 1063-6676(99)07984-5.

II. Four Basic Architectures for Audiovisual Speech

A. Theoretical Background

A basic ingredient of an automatic AV speech recognition (AAVSR) system is the architecture of the decision fusion system, that is the way the audio (A) and visual (V) inputs are processed in order to reach the phonetic or lexical level. In a recent and very general review of decision fusion systems, Dasarathy [19] introduces three basic architectures for combining several kinds of input data in order to reach the decision level (leaving apart finer-grain distinctions involving an intermediary "feature" level). First, a data-to-decision process directly performs decision from the combined data. Second, a decision-to-decision process is based on separate (partial or preliminary) decisions on individual data streams, followed by a fusion of separate decisions (which can be logical, fuzzy-logical, Bayesian, etc.). Third, a data-to-data process computes an integrated data stream from individual data inputs, and then decision is based on this integrated stream.

We have previously shown that these three basic architectures provide the basis for all AAVSR models existing in the literature [52].

The data-to-decision architecture, that we called *direct identification (DI) model*, had been introduced by Summerfield [57] as an extension of Klatt's *lexical access from spectra* (LAFS) model [29], into a *lexical access from spectra and face parameters*. In this model, the input signals are transmitted directly to the classifier, which is bimodal. This classifier makes decision from the bimodal feature space, in which it has learned bimodal prototypes or bimodal decision rules. There have been various implementations of the DI architecture in speech recognition (see, e.g. [2], [12], [18], [21], [30], [42],

[44], [49], [54], and [59]) and psychophysical modeling (see [11] and 16]).

The decision-to-decision model is based on what cognitive psychologists call "late integration," since integration follows phonetic classification in each separate sensorial pathway [61]. In this model, which we called the *separate identification (SI) model*, there are two parallel recognition processes, one for each modality. Afterwards, the phonemes or phonemic features obtained from each modality are fused. Fusion can be realized on logical values, like in the vision place audition manner (VPAM) model where each modality is in charge for a specific group of phonetic features [39], [57], or on probabilistic (or fuzzy-logical) values (as in the "fuzzy-logical model of perception," FLMP, proposed by Massaro [35]). The fusion is then realized inside a probabilistic framework. The first audiovisual speech recognition system, created by Petajan [43], relied upon a codebook of images to determine the template which best matched the spoken utterance, and used the visual recognition to select one of the two first candidates chosen by the acoustical recognizer, hence it was an implementation of the SI architecture. Later on, most implementations for speech recognition used the SI architecture [2], [14], [17], [21], [24], [27], [28], [30], [34], [38], [40], [41], [42], [49], [53], [55], [59]. In the field of psychophysical models, Massaro's FLMP [35], [36] is the most well-known example of SI architecture.

In the data-to-data architecture, fusion precedes classification, hence it is called "early integration" by cognitive psychologists[1] [61]. The problem is then to decide what can be the common format of sensory representations at the level where they are fused. Cognitive psychology proposes two answers [26], [57]. First, the common representation can be specific to one modality, supposed to be "dominant," and the second route is recoded into the dominant format. In this model, which we called the *dominant recoding (DR) model*, the auditory modality is supposed to be dominant for speech perception and, thus, more adapted to it. The visual input is recoded into a representation of the dominant modality, for instance the transfer function of the vocal tract. This transfer function is estimated independently from the auditory input (e.g., by cepstral analysis) and from the visual input (e.g., by association). These two estimations are then fused. The source characteristics (voiced, nasal, etc.) are estimated only from the auditory information. The whole source-filter set thus estimated is then presented to the phonetic classifier. Implementations of the DR architecture in the field of speech recognition are few [46], [62], [64], [65]—see also an application of the DR architecture for noisy speech enhancement [23]. The second possibility is that the common representation is amodal, and related to the common physical source of both audio and video inputs, that is articulatory gestures and configurations. In this model, that we called *motor recoding (MR) model*, both inputs are projected into an amodal (neither auditory nor visual) common space related to the characteristics of speech gestures (vocal tract configurations or motor programs) and fused in that space. The fusion process can take into account the fact that some dimensions cannot be seen (e.g., the velum) and weights very poorly the visual path for these dimensions. Though an increasing number of works deal with the recovery of articulatory gestures from the acoustical signal (see, e.g. [3], [4], [20], [31], [37], and [51]), we do not know of any implementation of the MR model for audiovisual speech processing, except the one we proposed ourselves [46], [47].

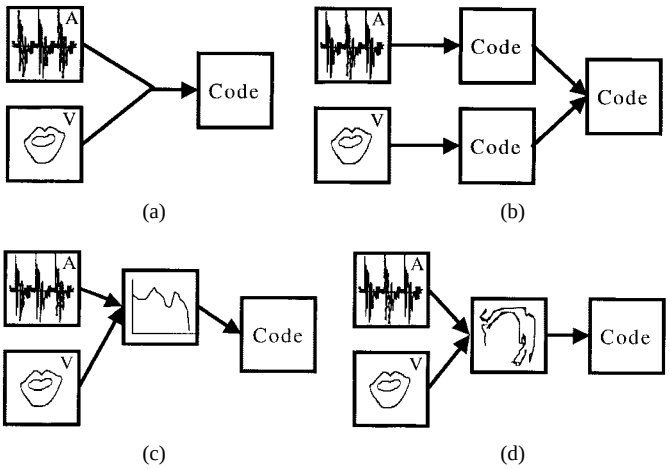


Fig. 1. Four basic models of audiovisual integration: (a) direct integration (DI), (b) separate integration (SI), (c) dominant recoding (DR), and (d) motor recoding (MR).

This provides altogether four architectures (see Fig. 1), which have never been systematically compared on a given recognition task: this is the object of the present work.

B. A Priori Considerations on the Four Architectures

1) DI versus SI: The only comparisons of architectures available in the literature involve the two first models, that is, DI and SI, and no clear-cut conclusion has been drawn from these [2], [42], [49], [59]. Some *a priori* considerations can, however, be attempted. The DI model has a major advantage on the SI model: it is able to exploit the covariations of the A and V inputs. In an extreme configuration where two phonetic classes would have the same means and variances in the A and V space, but different AV covariations, the SI model would be unable to identify them at all, while the DI model could display a significant level of correct discrimination. This is typically the case in the "audiovisual VOT" experiment by Breeuwer and Plomp [13] where the voiced-unvoiced [p] versus [b] contrast is signaled by the temporal covariation of an audio cue for voicing onset and a video cue for bilabial release: subjects perform a [p] versus [b] decision task at random in the A-only or V-only conditions, but their performance is quite good in the AV-condition, which is in line with the DI model and not the SI one.

On the other side, the potential advantage of the SI model is that in case of noisy audio inputs, it processes separately the corrupted input (the sound) and the noncorrupted one (the image). This is exactly the reason of the success of the

[1]The data-to-decision DI model is sometimes considered as a case of early-integration model, since the audiovisual classification process directly operates on input data without intermediary classification process in each sensory pathway. However, early-integration models are generally based on the assumption of a *common metric* for the audio and video inputs, and in various papers only data-to-data models as DR and MR are accepted as early-integration models.

TABLE I
SPECIFICATION OF PHONETIC FEATURES FOR EACH VOWEL CLASS

Vowel	Rounding	Backness	Height
i	-	+	++
e	-	+	+
ɛ	-	+	-
y	+	+	++
ø	+	+	+
œ	+	+	-
u	+	-	++
o	+	-	+
ɔ	+	-	-
a	?	?	--

"multistream approach" in automatic audio speech recognition systems [10], [34].

2) DR versus MR: These two models have been seldom used, and never compared, probably because though they are popular in the cognitive psychology community, they are quite ignored in the AAVSR community. The difference between these two models is the nature of the common representation at the fusion level. The DR model assumes the audio modality to be dominant, which is not bound to be the case. In fact, the natural complementarity between sounds and faces—hard to hear, easy to see (see, e.g., [5] and [57])—is difficult to exploit in the DR model, and we [47] have demonstrated in a simple case that the MR architecture was likely to be more efficient than the DR one for the processing of speech in noise. However, a complete evaluation in a well-controlled benchmark study remains to be done.

III. EXPERIMENTAL BENCHMARK SET-UP

A. Audiovisual Corpus

A benchmark study needs a corpus. We had one at our disposal, simple in its content, but already carefully studied in both acoustical, optical, and perceptual terms [48]. This benchmark corpus consists of 100 repetitions of each of the ten French oral vowels [i, e, a, u, y, o, œ, ϕ, ε, ɔ] pronounced in isolation by a single speaker in a single recording session. The French oral vowel system contrasts three series—front unrounded, front rounded, and back rounded vowels—with four height levels. For further analyses, we define the three basic phonetic contrasts involved in this system by the feature matrix in Table I. Rounding and front-back contrasts have two levels while the height feature has four levels. At the fourth height level, the rounding and front-back contrasts disappear in French, hence we suppose that only the height feature is specified for the vowel [a].

The corpus was recorded with a "video-speech workstation" developed for the study of audiovisual speech [32] which allows to acquire synchronized sounds and images. The audio signal was sampled at 16 kHz. Noisy acoustical signals were obtained by adding various amounts of white Gaussian noise on the temporal stimuli, with eight signal-to-noise ratios (SNR): no noise, 24, 12, 6, 0, −6, −12, and −24 dB. Then, for each acoustical occurrence, a spectrum was computed by the fast Fourier transform (FFT) on the first 64 ms, and each spectrum was characterized for further analysis by the dB values in 20 1-bark wide channels between 0 and 5 kHz. The perceptual frequency bark-scale is defined by the formula proposed by Schroeder *et al.* [50]:

$$z(\text{Bark}) = 7\text{Argsh}\left(\frac{F(\text{Hz})}{650}\right). \quad (1)$$

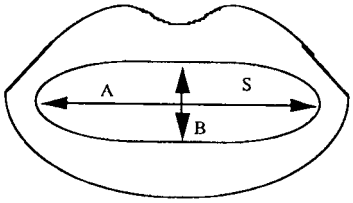


Fig. 2. Definition of the front geometrical parameters of the lips.

The speaker lips were made up in blue, so that a chroma-key process directly connected to the red–green–blue (RGB) output of the camera could turn the lips into black. This procedure allows an easy automatic detection process of the inner and outer contours of the lips for each separate field (20 ms) within a video frame (40 ms). The first video field (20 ms) of each of the 1000 occurrences was processed by the chroma-key system in order to produce images with perfectly black lips, on which our system allowed us to extract three basic front geometrical parameters: inner-lip horizontal width (A), inner-lip vertical height (B), and inner-lip area (S) (see Fig. 2). These features have been shown to be the major discriminant parameters of lip patterns and providing most of the visual intelligibility through resynthesis of facial movements [6].

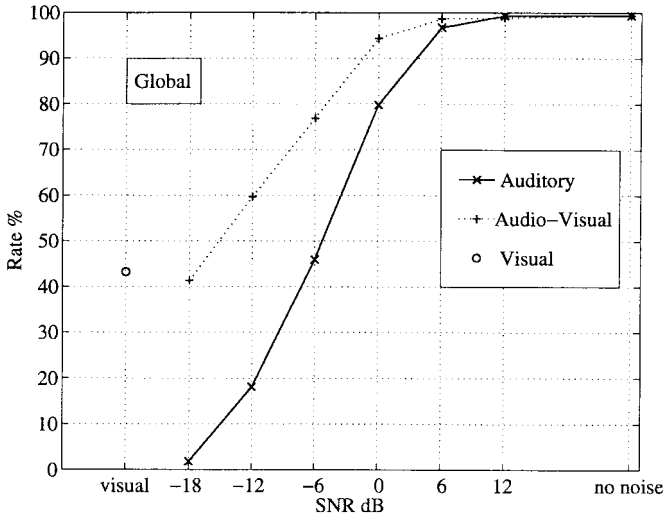
Hence, altogether our stimuli comprised 20 acoustical (spectral) and three optical (geometrical) components.

B. Perceptual Evaluation of This Corpus: A New Criterion for Assessing AAVSR Systems

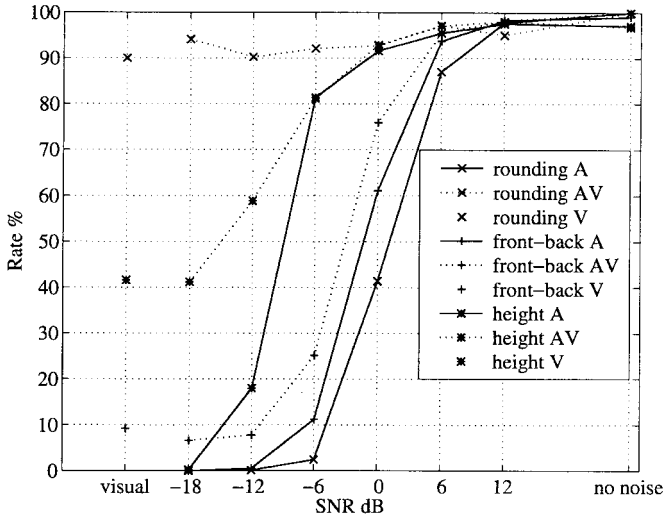
These stimuli were submitted to a series of experiments, investigating the auditory, visual and audiovisual identification at various levels of noise, by 21 French-speaking subjects [48]. These experiments revealed what we called the "double efficacy" of audiovisual speech perception.

First, it appears that there is a certain amount of *complementarity* in the information provided by the audio and video channels (efficacy at the information level): in the audio channel, the height feature is the most robust in noise, followed by the front-back one, the rounding contrast being the poorest, while this contrast is precisely the best identified by the eye. Hence, the least robust feature in the audio channel is the most robust in the video channel.

Second, audiovisual speech perception is also efficient at the information processing level. Indeed, we present in Fig. 3 the auditory-alone, visual-alone, and audiovisual identification scores as a function of SNR, both in global terms [Fig. 3(a)] and for each phonetic feature [Fig. 3(b)]. The results feature by feature are provided by means of transmitted information



(a)



(b)

Fig. 3. Summarized results of the perceptual identification of the auditory, visual and audiovisual stimuli from [48]: (a) correct identification scores as a function of SNR and (b) transmitted information scores for each phonetic feature as a function of SNR. Statistical analyzes confirmed that when audiovisual scores are lower than visual ones (for SNR $= -18$ dB) the difference is not significant.

computation, which will be detailed in Section V. These data show that there is a "*synergy*" between the auditory and visual channels for the bimodal identification of noisy speech. Indeed, the performance is always better with two channels than with only one, be it auditory or visual. Moreover, this synergy, which had already been described in global terms in several studies [5], [22] happens to hold even feature by feature, at least for oral vowels. At the highest levels of noise, the performance is entirely based on vision. Then for increasing SNR audiovisual perception seems to optimally combine the two channels to exhibit performances strictly higher than the individual performances of both of them. At last, without noise perfect identification of all features is realized, mainly thanks to the auditory channel, plus vision for rounding.

This synergy provides a big challenge for the elaboration of audiovisual speech recognition systems robust in adverse conditions. Hence, a basic criterion all along this study will be

$$\text{AV score} \geq \sup(\text{A score, V score})$$

scores being estimated not only in global recognition terms, but also at the level of phonetic feature identification.

C. Recognition Paradigms

Our goal is to assess the compared robustness of the four models introduced in Section II, for the recognition of noisy vowels. For this aim, we have chosen what we call an "extrapolation" paradigm, in which small amounts of noise (or high SNR values) are introduced in the learning phase, and larger noise levels are considered in the test phase.

More precisely, the corpus is divided into two equal parts. One half (50 stimuli per vowel category, 500 stimuli altogether) is used for learning the various parameters for each model (the learning subcorpus) and the other half for testing the models (the test subcorpus). Moreover, for learning we use audio stimuli at four SNR values corresponding to rather small amounts of noise: no noise, 24 dB, 12 dB, and 0 dB ("extrapolation learning corpus" in the following). Hence, the learning corpus contains altogether 500 visual stimuli associated to 2000 (4×500) audio items. Of course, test corpora contain the remaining 500 audiovisual stimuli presented at all noise levels (eight SNR values). All along this work we used ten different partitions of the learning and test corpora, and the results presented in Section V are based on average scores over the test-corpus for these ten partitions.

We also chose to study two different extrapolation paradigms. In the first one, the SNR value is supposed to be known and provided to the recognition system, and it plays the role of a contextual information in the fusion process, which is classical in sensor fusion [7], [19], and has already been used in audiovisual speech recognition [2], [14], [17], [27], [28], [34], [38], [42], [49], [54], [59], [65]. In this paradigm that we call "extrapolation with context," we will have to define for each model how the fusion process can be controlled by such a contextual supplementary input. In the second situation, that we call "extrapolation without context," no SNR value is provided: the interest here will be to know what can be the performances of each model in this more difficult robustness test.

IV. Implementation of the Four Models

A. Game Rules

1) General Principles: Our benchmark study may be considered as a kind of "race" between four systems in competition, and we try to specify here some basic rules of this special race. The four architectures are made up of a given arrangement of three kinds of components, *associators* which enable to quantitatively link two different continuous spaces, *classifiers* which attribute to a stimulus in a given space a set of confidence values for each of the ten vowel classes, and *integrators* which combine two vectors in a given space into some kind of average vector in the same space. Of course,

there is a great diversity of possible implementations for each component, and the choice for an individual component may considerably modify the performance of one model and possibly the final result of our "race." Hence, we attempted to limit as much as possible the role of individual elements in order to emphasize the role of *architectures* in the race. For this aim, we decided to only use very basic processors, simple enough to be easy to tune and, above all, in which the tuning process was analytical and not based on any kind of recursive optimization process sensitive to initial conditions. Moreover, we carefully controlled the number of free parameters to be estimated in the various conditions we studied. Therefore, the results presented in Section V must really be considered as a benchmark study, and not as a set of optimal models directly operational for any more complex speech recognition problem. Let us describe in more detail the rules of our game.

2) Dimensionality of the Data Streams: Our audio input stimuli are 20-dimensional (20-D). This high dimension is somewhat too large in respect to the complexity of our problem and the dimension of our corpora, and we verified in preliminary experiments that the number of dimensions could be lowered down to three by principal component analysis (PCA) without significantly decreasing the performances of audio classification [25]. In comparison, the visual stimuli are three-dimensional (3-D), and typical motor representations for vowels are also defined around three major dimensions, as we shall see later. Hence, we decided to fix the number of all data streams (audio, video, and motor) at three, in order to simplify the control of the number of free parameters in the optimization of the fusion process (see later). This was done on audio stimuli by applying a 3-D PCA computed on the whole corpus (eight SNR values).

3) Associators: The DR and MR models rely on the association between an item in an input space and another one in an output space, where integration occurs. We achieved such associations by means of linear associators, which are simple in their theoretical formulation and in their realization. We obtain an output vector $\boldsymbol{p}$ of dimension m from an input vector $\boldsymbol{q}$ of dimension n by means of the multiplication with a matrix $\boldsymbol{G}$ of dimension $(n+1)^{*}m$:

$$\boldsymbol{p} = [\boldsymbol{q} \quad 1].\boldsymbol{G}. \tag{2}$$

$\boldsymbol{G}$ is the regression matrix between the input (with an extra column of value one) and the output. m and n are equal to three according to the previous point. The extra raw of data of matrix $\boldsymbol{G}$ allows the existence of an additive bias. The tuning of the associator consists in determining the regression matrix $\boldsymbol{G}$ from the input–output pairs in the clean learning corpus (with no added noise).

4) Classifiers: All classifications are realized by means of Gaussian classifiers. In an n-dimensional space, we estimate for each category i its intraclass mean vector $\boldsymbol{m}_i$ and covariance matrix $\boldsymbol{V}_i$. Then we assume that all classes are equiprobable and we estimate the *posterior* probability for an input $\boldsymbol{x}$ to belong to this class by the formula

$$p(i/\boldsymbol{x}) \propto \frac{e^{-0.5(\boldsymbol{x}-\boldsymbol{m}_i)^t \boldsymbol{V}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i)}}{\sqrt{|\boldsymbol{V}_i|}} \tag{3}$$

$|\boldsymbol{V}_i|$ being the determinant of matrix $\boldsymbol{V}_i$. Finally, we compute the probability that $\boldsymbol{x}$ belongs to each of the ten categories (ten vowels) and we classify $\boldsymbol{x}$ in the category for which the probability is highest. The Gaussian classifiers are tuned by estimating for each class the intraclass mean and covariance matrix for the learning corpus.

5) Integrators: The SI, DR, and MR models involve a representation space common to audition and vision. Then a fusion process has to combine two items in the integration space (dimension $m = 3$), one derived from the acoustical input and the other one from the optical input. We used the simplest averaging process, that is an additive process, in models DR and MR [see Section IV-B, (10)–(12)]. Concerning fusion in model SI, we compared various processes [58]: arithmetical or geometrical mean, maximum value, and vote based on Bayesian belief and evidence theory [63]. In this paper, we present only the multiplicative fusion [see Section IV-B, (8) and (9)]. In the first architecture (DI), integration occurs on the covariance matrix of each class [see Section IV-B, (7)].

6) Weighting of the Audio and Video Input Streams in the Fusion Process: A crucial ingredient in any fusion process is the weighting of each sensory pathway, which allows to selectively increase or decrease the role of one input according to its efficiency for decision. In our case, each audio and video component is weighted by a factor $\alpha_{A,k}$ or $\alpha_{V,k}$ (k from one to three, since there are three audio and video dimensions). For example, increasing $\alpha_{A,k}$ results in increasing the contribution of the kth audio component in the global Gaussian distance of each class in model DI, or its contribution in the Gaussian distance of the audio classifier before decision fusion in model SI, or its contribution in the data integration process in models DR or MR.

In the extrapolation-without-context condition where SNR is not provided, the parameters $\alpha_{A,k}$ and $\alpha_{V,k}$ are tuned with a corpus comprising all SNR values including the lowest ones, corresponding to the poorest conditions, in order to be efficient. Whatever the architecture, the parameters are computed by minimizing through gradient descent an error function computed on this learning corpus. The error is defined by summing individual terms over the whole learning set:

$$E = \sum_{i=1}^{10} \sum_{j=1}^{50} \sum_{r=1}^{8} e_{i,j,r}. \tag{4}$$

Individual errors $e_{i,j,r}$ are defined by

$$e_{i,j,r} = (1 - p(i/\boldsymbol{x}_{\mathrm{AV}\,i,j,r}))^2. \tag{5}$$

i is the class index, $\boldsymbol{x}_{\mathrm{AV}\,i,j,r}$ the jth audiovisual sample in the class i in the learning set with r the SNR index, $p(i/\boldsymbol{x}_{\mathrm{AV}\,i,j,r})$ is the *posterior* probability as defined in the last section. Hence $e_{i,j,r}$ should be zero if the sample was perfectly identified.

In the extrapolation-with-context condition, we add the assumption that the reliability of each component depends on SNR. The weight of the kth audio (respectively, video) component at SNR level r is noted $\alpha_{A,k,r}$ (respectively,

$\alpha_{V,k,r}$). The tuning of these parameters is the same as in the extrapolation-without-context case except that it is made for each SNR level. The error to minimize is then

$$E_r = \sum_{i=1}^{10} \sum_{j=1}^{50} e_{i,j,r}. \tag{6}$$

Individual error is defined by (5).

We shall specify in each case how exactly weighting factors are applied and how they control the fusion process.

B. Detailed Implementations

The implementation of the four architectures, using the basic ingredients just described, will now be explained in more detail. For each architecture, we shall describe both how the "audiovisual" recognizer is realized and how it may be tuned with the integration parameters α_A and α_V. We shall also explain how the "monomodal" recognizers using only the auditory or the visual data (respectively, the "auditory" and "visual" recognizer) are implemented. For each model, the input is a six-dimensional (6-D) vector $\boldsymbol{x}_{\mathrm{AV}}$, which is the concatenation of a 3-D acoustical input $\boldsymbol{x}_A$ (resulting from the 3-D PCA of the 20-D dB/Bark spectra), and a 3-D optical input $\boldsymbol{x}_V$ (A, B, S triplet). Intermediary representations in models DR and MR will be, respectively, called $\boldsymbol{y}_{\mathrm{AV}}$, $\boldsymbol{y}_A$, and $\boldsymbol{y}_{\mathrm{V}}$. Then *posterior* probabilities $p(i/\boldsymbol{x})$ are computed for each of the ten classes i, and identification consists in choosing argmax $p(i/\boldsymbol{x})$.

1) Direct Identification (DI) Model [Fig. 4(a)]: In the direct identification model there is no intermediary stage between the physical inputs and the category output. Our implementation of this audiovisual recognition model involves a Gaussian classifier in a 6-D space (three acoustical dimensions and three optical dimensions). Hence, *posterior* audiovisual probabilities are provided by (3), $\boldsymbol{x}$ being in this case $\boldsymbol{x}_{\mathrm{AV}}$. The control of the fusion process through α weights occurs inside the inverse audiovisual covariance matrix $\boldsymbol{V}^{-1}_{\mathrm{AV},i}$ in the following way:

$$\boldsymbol{V}^{-1}_{\mathrm{AV},i} \xrightarrow{\text{weighting}} \boldsymbol{V}^{*-1}_{\mathrm{AV},i} = \boldsymbol{W}_{\mathrm{AV}}.\boldsymbol{V}^{-1}_{\mathrm{AV},i}.\boldsymbol{W}_{\mathrm{AV}}$$

with

$$\boldsymbol{W}_{\mathrm{AV}} = \begin{bmatrix} \alpha_{A,1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha_{A,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha_{A,3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha_{V,1} & 0 & 0 \\ 0 & 0 & 0 & 0 & \alpha_{V,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \alpha_{V,3} \end{bmatrix}. \tag{7}$$

In the extrapolation-with-context case, the parameter integration depends on the SNR values, hence $\alpha_{A,k}$ and $\alpha_{V,k}$ become $\alpha_{A,k,r}$ and $\alpha_{V,k,r}$.

Increasing α_A (respectively, α_V) results in increasing the acoustical (respectively, optical) contribution to the Mahanalobis distance between the input and each class mean. Values $\alpha_A = 1$ and $\alpha_V = 0$ (respectively, $\alpha_A = 0$ and $\alpha_V = 1$) give more or less the same classifier as a Gaussian classifier applied on the three acoustical (respectively, optical) dimensions.[2]

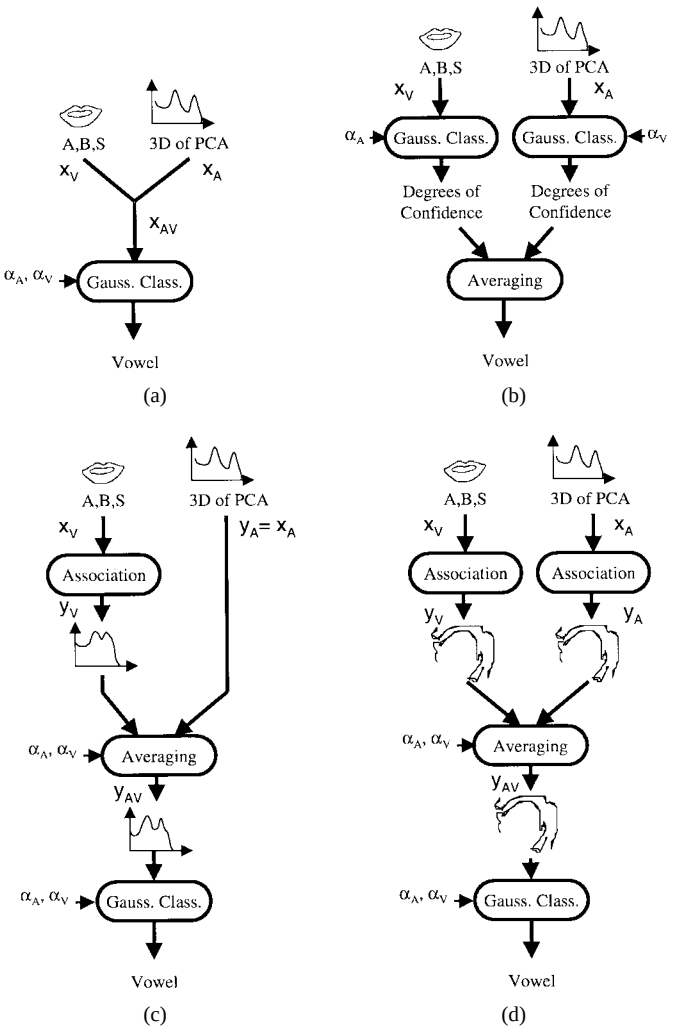


Fig. 4. Benchmark implementation of the four models: (a) DI, (b) SI, (c) DR, and (d) MR.

2) Separate Identification (SI) Model [Fig. 4(b)]: The SI model assumes that the acoustical and optical inputs are separately classified before fusion. To implement this model we first define the "auditory" and "visual" recognizers, which are two Gaussian classifiers, respectively, working on the acoustical and the optical input. Each classifier delivers a probability to belong to each of the ten vowel categories. We have then two sets of values, as follows.

1) $p_A(i) = p(i/\boldsymbol{x}_A)$ giving the *posterior* probability that the acoustical input belongs to category i;
2) $p_V(i) = p(i/\boldsymbol{x}_V)$ giving the *posterior* probability that the optical input belongs to category i.

The control of the fusion process through α weights is similar to what was done in the DI model. We weight the inverse covariance matrix $\boldsymbol{V}^{-1}_{A,i}$ (respectively, $\boldsymbol{V}^{-1}_{V,i}$) to compute $p_A(i)$

[2]A slight deviation from monomodal performance [58] is due to nondiagonal terms linking acoustical and optical terms in the intraclass covariance matrices $\boldsymbol{V}_{\mathrm{AV},i}$, which bias the effective "pure-audio" or "pure-video" scores. However, we verified that the differences are not significant: less than 1% for global scores in comparison with "pure audio" scores, and 0.2% in comparison with "pure-video" scores.

[respectively, $p_V(i)$] from (3) in the following way:

$$\boldsymbol{V}_{Ai}^{-1} \xrightarrow{\text{weighting}} \boldsymbol{V}_{Ai}^{*-1} = \boldsymbol{W}_A.\boldsymbol{V}_{Ai}^{-1}.\boldsymbol{W}_A$$
$$\boldsymbol{V}_{Vi}^{-1} \xrightarrow{\text{weighting}} \boldsymbol{V}_{Vi}^{*-1} = \boldsymbol{W}_V.\boldsymbol{V}_{Vi}^{-1}.\boldsymbol{W}_V$$

with

$$\boldsymbol{W}_A = \begin{bmatrix} \alpha_{A1} & 0 & 0 \\ 0 & \alpha_{A2} & 0 \\ 0 & 0 & \alpha_{A3} \end{bmatrix}$$

and

$$\boldsymbol{W}_V = \begin{bmatrix} \alpha_{V1} & 0 & 0 \\ 0 & \alpha_{V2} & 0 \\ 0 & 0 & \alpha_{V3} \end{bmatrix}. \tag{8}$$

To compute the audiovisual *posterior* probability we use Bayes formula

$$p(i/\boldsymbol{x}_A, \boldsymbol{x}_V) = p_{\text{AV}}(i) = \frac{p_A(i)p_V(i)}{\sum_{j=1}^{10} p_A(j)p_V(j)}. \tag{9}$$

When α_A (respectively, α_V) decreases, the acoustical (respectively, optical) *posterior* probabilities "converge" toward the value 0.5, hence the acoustical (respectively, optical) contribution decreases in the fusion process [see (9)]. Values $\alpha_A = 1$ and $\alpha_V = 0$ (respectively, $\alpha_A = 0$ and $\alpha_V = 1$) provide the same classifier as a Gaussian classifier applied on the three acoustical (respectively, optical) dimensions. Notice that the introduction of the audio and video weights can also be done at the output of the audio and video classifiers, by exponential terms applied to $p_A(i)$ and $p_V(i)$ in (9) [2], [17], [28], [34], [42], [49], [54], [59]. Our procedure provides similar effects, and it can be controlled with the same number of α parameters than the other architectures studied here.

3) Dominant Modality Recoding (DR) Model [Fig. 4(c)]: This model was first implemented for the recognition of English vowels [64]. It assumes that the optical information is recoded into an acoustical space in which integration occurs. The conversion from the visual input to the auditory "equivalent spectrum" is done by means of a linear association which learns the regression between 3-D optical inputs and the 3-D acoustical components of the clean sounds corresponding to the input.

The integration of the "auditory spectrum" with the "visual spectrum" estimated from the visual path is done point by point according to the formula:

$$\boldsymbol{y}_{\text{AV}}(k) = \alpha_{A,k}\boldsymbol{y}_A(k) + \alpha_{V,k}\boldsymbol{y}_V(k) \tag{10}$$

where $\boldsymbol{y}_A$ is equal to the 3-D acoustical input $\boldsymbol{x}_A$, $\boldsymbol{y}_V$ is the 3-D spectrum estimated from the 3-D optical input $\boldsymbol{x}_V$ and $\boldsymbol{y}_{\text{AV}}$ is the integrated audiovisual spectrum. k is the index of one of the three components. In the extrapolation-with-context case, the weighting factors depend on the SNR values, $\alpha_{A,k}$ and $\alpha_{V,k}$ become $\alpha_{A,k,r}$ and $\alpha_{V,k,r}$. Decreasing α_A decreases the weight of the auditory modality in the "audiovisual" recognizer. This is exactly the control process proposed in [65]. Then vowel identification is achieved by a Gaussian classifier applied to $\boldsymbol{y}_{\text{AV}}$. For each class, we estimate $m_{A,i}$ and $\boldsymbol{V}_{A,i}$ (audio mean and covariance matrix for class i), $m_{V,i}$ and $\boldsymbol{V}_{V,i}$ (video mean and covariance matrix for class i). The audiovisual parameters are then computed in the following way:

$$\boldsymbol{m}_{\text{AV},i} = \boldsymbol{W}_A.\boldsymbol{m}_{A,i} + \boldsymbol{W}_V.\boldsymbol{m}_{V,i}$$
$$\boldsymbol{V}_{\text{AV},i} = \boldsymbol{W}_A.\boldsymbol{V}_{A,i}.\boldsymbol{W}_A + \boldsymbol{W}_V.\boldsymbol{V}_{V,i}.\boldsymbol{W}_V + \mathbf{2}.\boldsymbol{W}_A.\mathbf{cov}_i.\boldsymbol{W}_V$$

with

$$\boldsymbol{W}_A = \begin{bmatrix} \alpha_{A1} & 0 & 0 \\ 0 & \alpha_{A2} & 0 \\ 0 & 0 & \alpha_{A3} \end{bmatrix}$$

and

$$\boldsymbol{W}_V = \begin{bmatrix} \alpha_{V1} & 0 & 0 \\ 0 & \alpha_{V2} & 0 \\ 0 & 0 & \alpha_{V3} \end{bmatrix} \tag{11}$$

and with $\boldsymbol{V}_{A,i}$ and $\boldsymbol{V}_{V,i}$, respectively, the audio and video covariance matrix for class i, and $\mathbf{cov}_i$ the audiovisual covariance matrix defined by

$$\boldsymbol{cov}_i = \frac{1}{49}\sum_{j=1}^{50}\left(\boldsymbol{y}_{A,j} - \boldsymbol{m}_{A,i}\right)^t\left(\boldsymbol{y}_{V,j} - \boldsymbol{m}_{V,i}\right). \tag{12}$$

These formula ensure that values $\alpha_A = 1$ and $\alpha_V = 0$ (respectively, $\alpha_A = 0$ and $\alpha_V = 1$) exactly provide the audio-alone (respectively, video-alone) recognizer.

4) Motor Space Recoding (MR) Model [Fig. 4(d)]: Our implementation of model MR is the first implementation of this model in the literature. A crucial choice in the MR model concerns the definition of the "motor space" in which integration should occur [45]. Since we deal with static vowels, we have chosen articulatory representations based on three parameters, X, Y (which are, respectively, the horizontal and vertical coordinates of the highest point of the tongue) and A (the inner-lip width) [see Fig. 4(d)]. X, Y, and A, respectively, provide articulatory correlates of the front-back, open-close and rounding dimensions [1], [8].

The transformation of the inputs into a motor representation is implemented thanks to linear associations. For tuning these associators we must define output articulatory representations for each vowel of the learning set. A values are directly provided by the optical input. However, we do not have at our disposal articulatory estimates of X and Y values. Hence, we used for this aim prototypical values proposed by expert phoneticians, and displayed in Table II. These values are in agreement with a large knowledge on articulatory configurations for French vowels, in connection to both articulatory models and radiographic data [8]. Then, the tuning of the linear associator for the audio and video paths is done by, respectively, calculating the regression matrix from the no-noise learning audio or video subcorpus into these sets of (X, Y, A) triplets. Notice that A is directly transmitted from the visual input, hence only the association between (A, B, S) and (X, Y) has to be learned in the optical path.

$\boldsymbol{y}_A$ and $\boldsymbol{y}_V$ are the 3-D articulatory configurations, respectively, estimated from the audio and the video inputs, and the audiovisual estimate of the (X, Y, A) set $\boldsymbol{y}_{\text{AV}}$ is derived

TABLE II
(X, Y) PROTOTYPICAL VALUES FOR EACH VOWEL CLASS IN THE MR MODEL (SEE TEXT)

Vowel	X	Y
i	0	1
e	0.33	0.66
ɛ	0.66	0.33
y	0	1
ø	0.33	0.66
œ	0.66	0.33
u	1	1
o	1	0.66
ɔ	1	0.33
a	1	0

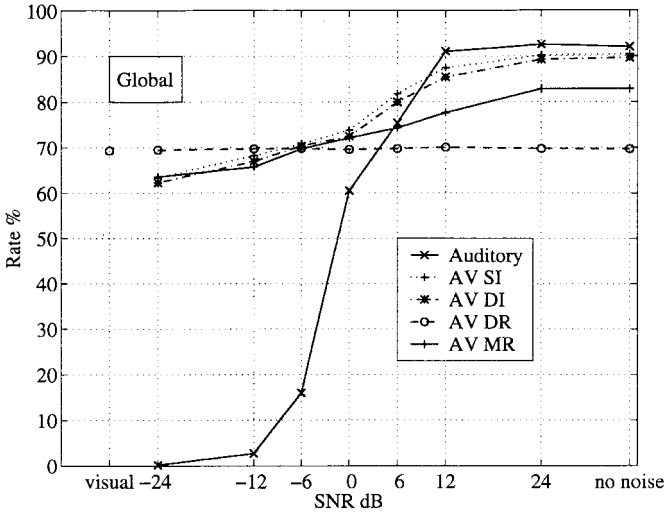


Fig. 5. Auditory, visual, and audiovisual recognition global scores for the four models in the extrapolation-without-context paradigm, in percent ("100" means perfect recognition, "0" means random recognition).

thanks to the same equation as model DR (10). Classification of the final motor representation is achieved by a Gaussian classifier applied to $\boldsymbol{y}_{\mathrm{AV}}$ thanks to (11) and (12).

V. RESULTS

We shall now compare the four models on the two tasks defined in Section III-C. A given recognition experiment depends on four factors: fusion architecture (DI, SI, DR, or MR), recognition paradigm (extrapolation with context or extrapolation without context), acoustical noise level (eight SNR values in the test corpus) and stimuli presentation (auditory, visual, or audiovisual). Each experiment provides a confusion matrix, on which we compute two types of scores.

First, the global identification score in percentage corrected to the random level is computed thanks to the following formula:

$$\text{Corrected score} = 100\,\frac{\left|\dfrac{\text{correct responses}}{\text{total responses}} - \dfrac{1}{\text{number of vowels}}\right|}{1 - \dfrac{1}{\text{number of vowels}}}. \quad (13)$$

Then we compute the "transmitted information" for the three individual phonetic features: rounding, height and front-back contrast. These features have already been defined in Table I. Recall that only the height feature is supposed to be known for vowel [a], hence we use neither the stimuli nor the answers corresponding to [a] in the computation for the rounding and the front-back features. For each feature, the percentage of transmitted information is defined by [13], [60, ch. 18]

$$t = 100\,\frac{h(s, r)}{h(s)} \quad (14)$$

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

$$h(s, r) = -\sum_i \sum_j p(s_i, r_j) \log_2 \left(\frac{p(s_i)p(r_j)}{p(s_i, r_j)}\right) \quad (15)$$

$$h(s) = -\sum_i p(s_i) \log_2 p(s_i) \quad (16)$$

with

1) $p(s_i)$ the probability of occurrence of feature s_i in the stimuli;
2) $p(r_j)$ the probability of occurrence of feature r_j in the answers;
3) $p(s_i, r_j)$ the probability of shared occurrence of feature s_i in the stimuli and feature r_j in the answers.

The probability $p(s_i)$ is known: $p(s_i) = n_{i.}/n$ and the probabilities $p(r_j)$ and $p(s_i, r_j)$ are not known but can be estimated by

$$p(r_j) = n_{.j}/n$$

and

$$p(s_i, r_j) = n_{ij}/n$$

with $n_{i.}$ the number of occurrences of stimulus s_i, $n_{.j}$ the number of occurrences of answer r_j, n_{ij} the number of occurrences of stimulus s_i with answer r_j, and n the total number of stimuli. n and $n_{i.}$ values are fixed, $n_{.j}$ and n_{ij} values are provided by the confusion matrices.

Our basic expectation for a given model is that it optimally *processes* information, which means that the audiovisual scores, both in global terms and in respect to the transmitted information for each individual feature, are always greater than or at least equal to the audio-alone or video-alone scores: we call this the "synergy" criterion. The significance of statistical differences between two different conditions (such as one model compared with another one, or audio scores compared with audiovisual ones) was systematically assessed by χ^2-tests, either on global recognition scores, or restricted to the correct identification of a given feature.

A. Pure Audio and Video Scores

First of all we have tested our four architectures with audio-only or video-only stimuli. The global results are provided as baseline in Fig. 5. It appears that this baseline is the same in the four models. This is due to the way we defined our

implementations in the previous section. Indeed, setting all α_A values to one and α_V values to zero results in (7) to an inverse covariance matrix $\boldsymbol{V}_{\mathrm{AV}}^{-1}$ reduced to $\boldsymbol{V}_{A}^{-1}$, which ensures equal audio scores for DI and SI, and the same is true for video scores (see Note 2). In the case of models DR and MR, the first association step linearly transforms 3-D audio and video vectors into 3-D audio or motor representations, and such a linear 3-D to 3-D transformation does not modify at all the performances of a Gaussian classifier. Then (10)–(12) ensure that audio and video scores for DR and MR are the same as those for SI.

In summary, the four architectures have the same number of dimension inputs, that is, three for both the audio and video streams, and three for the intermediary representation in the case of models DR and MR. They give the same results in "monomodal" conditions ($\alpha_A = 1$ and $\alpha_V = 0$ or $\alpha_A = 0$ and $\alpha_V = 1$). They have the same number of control parameters α, namely six, to be tuned in the optimization of the fusion process. The only difference between the four models is the nature of the fusion process. We shall now see how this intervenes in the AV performances for each experimental paradigm.

B. Extrapolation Without Context

The auditory, visual and audiovisual correct recognition scores for the four models in the extrapolation-without-context paradigm are displayed in Fig. 5. In all cases, there are significant violations of the AV criterion (AV scores should be better than both A and V scores), hence we do not present results on individual phonetic features, which of course also display such violations.

The worst performances are displayed by the DR model, which appears unable to take profit of fusion: the audiovisual scores are similar to the visual scores. The reason is that the weight optimization by gradient descent gives values $\alpha_{V,k}$ near one and $\alpha_{A,k}$ near zero, whatever the partition used, in order to increase the very low audiovisual *posterior* probabilities in the lowest SNR values. This implies, of course, very significant violations of our "AV challenge" between audiovisual and audio scores in medium and large SNR values.

The audiovisual scores obtained with the MR model are better than those of the DR model. Indeed, audiovisual scores vary significantly with SNR, hence fusion is useful, and the global score averaged over all SNR is equal to 73.6%, instead of 69.8% for DR model. However there are significant violations once more: the audiovisual score is lower than the visual score for SNR lower than -12 dB, and lower than the audio score for SNR higher than 12 dB.

The audiovisual scores for the SI and DI model are better than the early integration models scores, and quite similar (though with a slight and significant 1.2% advantage for SI, averaged over all SNR). However, there are violations between the audiovisual and visual scores at -24 dB and -12 dB, and between audiovisual and auditory scores at 12 dB, 24 dB, and no-noise level.

Altogether it appears that in all models, whatever their differences in performance, the optimization process tuning the $\alpha_{A,k}$ and $\alpha_{V,k}$ values leads to a compromise between audio and video performances, which is indeed optimal in the considered paradigm, but does not lead to enough robustness. The logical way to increase robustness is of course to adapt audio and video weights $\alpha_{A,k}$ and $\alpha_{V,k}$ according to the SNR value: this is the case in the extrapolation-with- context paradigm.

C. Extrapolation with Context

In this case the $\alpha_{V,k}$ and $\alpha_{A,k}$ values vary with SNR. The global recognition and transmitted information scores in the extrapolation-with-context paradigm are, respectively, displayed in Fig. 6(a) global recognition, Fig. 6(b) rounding, Fig. 6(c) front-back, and Fig. 6(d) height for the four models. From these figures two conclusions may be drawn.

First, in global terms, the performance of models DI and SI are more or less the same, and higher than those of models DR and MR. The basic difference is the audiovisual score in medium and large SNR: it reaches 9% for SI compared to DR at 0 dB (76.6% versus 85.6%). The performance of model MR is higher than model DR in global (82.2% for DR versus 83.1% for MR in scores averaged over all SNR, difference significant).

Second, in what concerns the global recognition scores, the synergy criterion is more or less acceptably fulfilled in all models. Indeed, the audiovisual score is higher than or equal to both the audio and visual scores at all SNR and for all models, except below -6 dB for models DI and SI, for which the audiovisual scores are slightly lower than the visual score, but the differences are never significant. However there are a number of violations of the "synergy principle" for individual phonetic features, summarized in Table III. For models DI and SI these violations are rather small and not significant, for model MR, these violations are still small (around 3%), though significant. They concern mainly the rounding feature at medium to large SNR values (AV score lower than the V one) and the front-back feature at large SNR (AV lower than A). For the DR model, these violations concern the same kind of cases as MR, but with a wider range of SNR and a larger amplitude: they can reach more than 8% in the case of the AV score compared to the A score for the front-back feature at 24 dB.

VI. Discussion

A. Comparison of the Four Architectures in the Model Race

When one considers both the extrapolation-with-context and the extrapolation-without context paradigms, it is clear that a global hierarchy emerges, which can be summarized by

$$\mathrm{DI} = \mathrm{SI} > \mathrm{MR} > \mathrm{DR}.$$

This hierarchy in two groups, with DI and SI in the leading category and MR and DR in a second group, may be explained by the fact that the final classification process in models MR and DR proceeds on three dimensions only, while the DI model involves a 6-D classification, and the SI model takes profit of two 3-D classification processes. However, in spite of

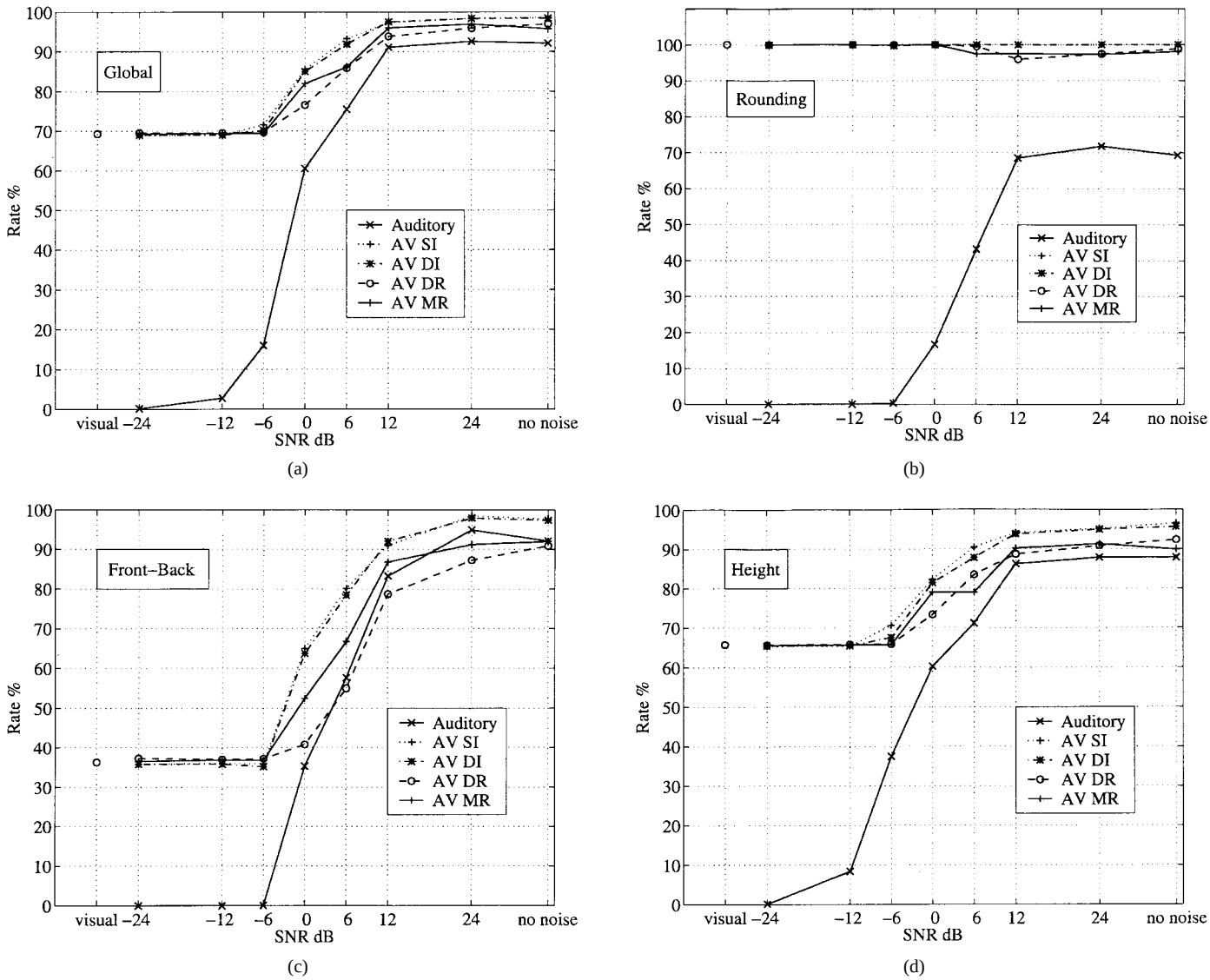


Fig. 6. Auditory, visual, and audiovisual recognition scores for the four models in the extrapolation-with-context paradigm: (a) global scores, (b) rounding scores, (c) front-back scores, and (d) height scores.

this relatively predictable behavior, the hierarchy within each group deserves attention.

1) DI versus SI: In respect to our *a priori* arguments of Section II-B1, the equally good behavior of models DI and SI is quite interesting, since it is, to our knowledge, the first time that a mechanism of control of the fusion process adaptively weighting the audio and video components at the input of the decision process is shown to produce as efficient results for DI as for SI in the recognition of speech in noise. On the other side, the fact that no superiority of DI emerges could appear surprising, since DI should better exploit interclass differences in AV covariations. However, covariations in our corpus are quite regular from class to class. In global terms, it is clear that the video parameter A, which is the major correlate of rounding, is correlated with the second formant $F2$ of the audio spectrum at least for front vowels, while the video parameter B is a good visual cue for height and is correlated with the first formant $F1$. There are some trends for interclass differences in covariations: since B cues height, it is positively correlated with $F2$ for back vowels and negatively for front vowels (it is classical that with an $F1$ increase, $F2$ increases in the first case and decreases in the second one). But these pieces of information are rather marginal—and redundant with audio covariations between $F1$ and $F2$—in this corpus.

2) MR versus DR: In this second group of models, it is clear that MR performs better than DR. This is demonstrated in both extrapolation paradigms, and through all the criteria we use, that is correct recognition scores, correct recognition probabilities and transmitted information on phonetic features. This confirms the *a priori* considerations introduced in Section II-B2, and the basic reason can be understood thanks to Fig. 7. On this figure, we display the two basic correlates of the contrasts between the close vowels [i], [y], and [u], that is for the audio input the second PCA component (which is the best one in terms of auditory contrasts for these three vowels), and for the visual input the parameter A (the pattern would be more or less the same, though with less contrast, with other auditory or visual parameters). From this display, it is easy

TABLE III
CASES OF VIOLATION OF THE AUDIOVISUAL SYNERGY RULE FOR INDIVIDUAL PHONETIC FEATURES IN THE EXTRAPOLATION-WITH-CONTEXT CONDITION. ONLY SIGNIFICANT VIOLATIONS ARE DISPLAYED (NO VIOLATIONS FOR SI AND DI). FOR DEGREE OF SIGNIFICANCE, WE USE THE FOLLOWING CONVENTIONS: $0.01 \leq p < 0.1^{*}$, $0.001 \leq p < 0.01^{**}$, $p \leq 0.001^{***}$

Rounding

SNR	6 dB		12 dB		24 dB		no noise	
Model	AV-V	χ^2(p)	AV-V	χ^2(p)	AV-V	χ^2(p)	AV-V	χ^2(p)
MR	-3.2%	14(***)	-2.8%	12(***)	-3.1%	13(***)	-2.1%	9(**)
DR	not significant		-4.6%	21(***)	-3.1%	14(***)	-1.3%	5(*)

Front-back

SNR	12 dB		24 dB	
Model	AV-A	χ^2(p)	AV-A	χ^2(p)
MR	no violation		-3.6%	4.6(*)
DR	-4.7%	4(*)	-7.9%	28.6(***)

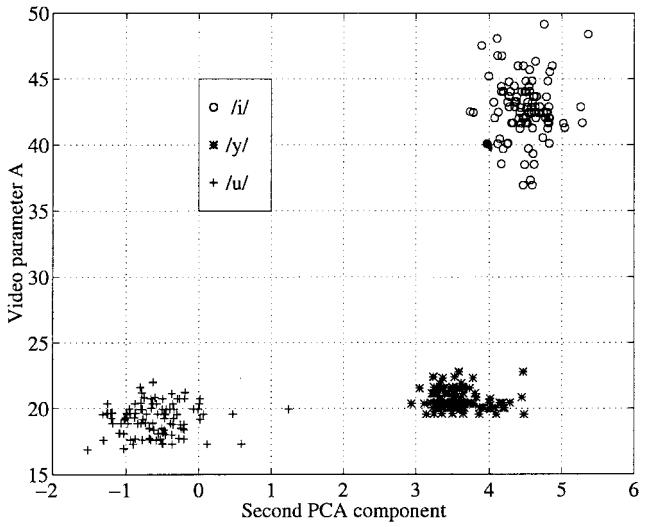


Fig. 7. Audiovisual complementarity in the [i]-[y]-[u] region: a display of the second PCA component versus the video parameter A for all [i], [y], [u] stimuli without noise.

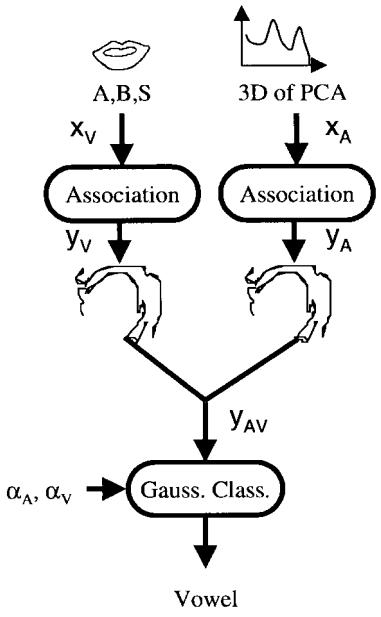


Fig. 8. Hybrid MR-DI model.

to predict a major problem for the DR model: the recoding of vision into audition will lead to project a set of visual patterns as [i], [y], [u], in which [i] is far from [y], which is almost confounded with [u], toward a set of auditory patterns in which [i] is very close to [y], and [u] is quite far apart. This leads to both a theoretical difficulty—how to associate almost identical visual patterns with very different acoustical spectra, e.g. [y] and [u]—and a technical difficulty—how to exploit the good separation between visual [i] and [y] if they are mapped onto close acoustical spectra. This problem emerges because of the nature of the French phonological system for oral vowels, with two independent features in the close series, that is rounding and front-back. It would not exist, e.g., in English, with the single [i] versus [u] contrast with both strong audio and video cues; this explains the limited success by Yuhas *et al.* [65] for AV recognition of static English vowels with DR—an optimal situation for this model. The problem would be, however, much more critical with dynamical stimuli and all the coarticulation problems; hence, the use of the DR model for audiovisual speech recognition seems quite inappropriate.

The interest of the MR model is precisely that it presents the audio and the video data in an almost optimal way in terms of their natural complementarity. Indeed, the articulatory variables X, Y, and A that we have chosen to define the common intermediary representation for fusion, respectively, provide natural articulatory correlates of the front-back, open-close, and rounding dimensions. Therefore the control of their weights α should be easy to optimize, particularly for X, rather robust in the audio modality and almost impossible to estimate from the video one, and A, not robust in the audio spectrum with added noise, and very contrastive in the video channel. Though the performances of the MR model were not very satisfactory, due to the nature of the fusion+classification process, it stroke us that the articulatory representation could be of some interest as a kind of preprocessing in the DI and SI architectures.

Hence, we implemented a kind of "hybrid" model in which the audio and video 3-D input-streams were first transformed by association into two "articulatory" 3-D streams, respectively, (X_A, Y_A, A_A) and (X_V, Y_V, A_V) as in model MR, and then these two streams were concatenated into a 6-D (X_A, Y_A, A_A, X_V, Y_V, A_V) vector feeding a 6-D classifier implementing the DI fusion process (see Fig. 8). It appears that the audiovisual performances with this new model increase a bit, though the difference is not significant. In global

scores averaged over all SNR, we have a 0.2% increase in the extrapolation-with-context and 0.4% increase in the extrapolation-without-context paradigm, and mean probability errors [computed through (4) and (5)] decrease by an amount of about 3% in both conditions.

In conclusion, the DI and SI fusion architectures provide the best performances in our experiments, but some improvement may be expected from an exploration of what could be the "optimal" representations of input data stream improving the efficiency of the control of the fusion process, and articulatory representations in our context seem to provide an efficient first kick in this exploration.

B. Need for Context

Whatever the model chosen, it appears that the only way to reach the AV challenge defined in Section III-B is to introduce context, that is to control the fusion process through the knowledge of SNR. In this case, the best models, that is DI and SI, lead to performances in which there are almost no violations of the challenge even at the fine-grain level of individual phonetic features. On the contrary, when no contextual information is provided, the performances decrease considerably. This is the second strong conclusion of our study: an audiovisual speech recognition system will not behave satisfactorily, whatever its architecture, if the reliability estimation problem is not considered in the implementation of the fusion process. In more general terms, the control of the fusion process is a crucial problem for the elaboration of an audiovisual speech recognition system [41], above all if one wants to provide a model compatible with experimental data on human perception [52].

There are two basic ways to estimate the reliability of the audio channel. First, it can be estimated directly from the input stream, for example through one of the many methods for SNR estimations (e.g., estimation of the noise power spectrum during nonspeech intervals of the voice communication process, [9], [33]; see also a proposal for SNR estimation computed along the vowel trajectory through noise, [59]). Second, a direct reliability estimation can be performed at the output of the classifier (in the case of model SI for example; in model DI, one then needs to add a pure audio classifier to the global audiovisual one). Then, any measure of the coherence between a test input and the statistical properties of the learning set can be used, such as ambiguity [2], [49] or entropy [14], [38]. Similar techniques could be employed to evaluate the reliability both of the visual and audio channel [41].

C. Conclusion and Perspectives

This work raises four major conclusions for future developments in AAVSR.

First, it is crucial to compare various architectures for a given recognition task, and for this aim we have introduced a new criterion: audiovisual scores should be of course as large as possible, but at least higher than both audio and video scores, and this not only in global terms, but also at the level of individual phonetic features. This criterion, inspired by human performances, is never tested, and when confusion matrices are available in published papers, it appears that it is generally not fulfilled for specific features, though global scores are sometimes quite satisfactory.

Second, it seems that data-to-data fusion processes (such as MR or DR) are not as efficient as data-to-decision (DI) or decision-to-decision (SI) models, since they reduce too early the number of dimensions that are necessary to achieve a high level of performances. This could possibly be improved by a specific work on optimal representations at the level of data fusion (see next point). In what concerns the two best models, we do not obtain significant differences, but we show at least that the DI model can be adapted to largely degraded input streams. This could open a route toward future progresses with this model, which has the major interest to be able to take profit of audiovisual subphonetic covariations, which is not the case of model SI. This last model, on the contrary, is probably the easiest to control, since tuning occurs after recognition of each input stream. However, its main problem is that fusion occurs late, at a level where a number of prephonetic characteristics of the sound and the image are already lost.

Third, the representation format of input streams could play a significant part in the quality of audiovisual fusion. Indeed, it is likely that a representation in which the audio and video streams are optimally displayed in terms of audio and video complementarity (with some components easy to see, and others easy to hear, that is robust in noise) should improve recognition: this is what we demonstrated by the "hybrid" model of Fig. 8. Other techniques such as nonlinear projection could be of great interest in this respect [25].

Last, but not least, any AAVSR system should comprise a control of the fusion process driven by estimations of the reliability of each channel at each time for the recognition task. We have presented in this paper a detailed implementation of such control processes for each of the four architectures studied, together with some directions for future research on reliability estimation.

Of course, the complexity of AAVSR tasks in man–machine interfaces is much greater than the task considered here. It involves natural multispeaker speech, various non stationary noises, nonoptimal video inputs. However, the major points raised in this study will remain basically the same. In light of this carefully controlled "laboratory" study, it is our belief that data representation, exploitation of audiovisual covariations, sensor reliability estimation and control of the fusion process will be crucial ingredients for achieving the "challenge of audiovisual speech": improving ASR robustness thanks to the video input without in any case loosing any part of the audio information, to reach audiovisual scores higher than both audio and video ones at all levels of the recognition process.

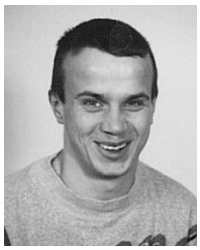
REFERENCES

[1] C. Abry, L. J. Boë, and R. Descout, "Voyelles arrondies et voyelles protruses en français," in *Labialité et Phonétique*, C. Abry *et al.*, Eds. Grenoble, France: Univ. Langues Lettres de Grenoble, 1980, pp. 203–215.

[2] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Man and Machine: Models, Systems and Applications,* D. G. Stork and M. E.

Hennecke, Eds. Berlin, Germany: Springer, NATO ASI Series, 1996, pp. 461–472.
[3] B. S. Atal, J. J. Chang, M. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Amer.*, vol. 63, pp. 1535–1555, 1978.
[4] G. Bailly, "Recovering place of articulation for occlusives in VCV's," in *Proc. XIIIth Int. Cong. Phonetic Sciences*, 1995, vol. 2, pp. 230–233.
[5] C. Benoît, T. Mohamadi, and S. D. Kandel, "Effects of phonetic context on audio-visual intelligibility of French," *J. Speech Hearing Res.*, vol. 37, pp. 1195–1203, 1994.
[6] C. Benoît, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani, "Which components of the face humans and machines best speechread?," in *Speechreading by Man and Machine: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, NATO ASI Series, 1996, pp. 315–328.
[7] I. Bloch, "Information combination operators for data fusion: A comparative review with classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 26, pp. 52–67, Jan. 1996.
[8] L. J. Boë, B. Gabioud, P. Perrier, J. L. Schwartz, and N. Vallée, "Vers une unification des espaces vocaliques," in *Levels in Speech Communication: Relations and Interactions*, C. Sorin *et al.*, Eds. New York: Elsevier B.V., 1995, pp. 63–71.
[9] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, 1979.
[10] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICLSP'96*, pp. 426–429.
[11] L. D. Braida, "Crossmodal integration in the identification of consonant segments," *Quart. J. Exper. Psychol.*, vol. 43A, pp. 647–677, 1991.
[12] L. D. Braida *et al.*, "Use of articulatory signals in automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 80, p. S18, 1986.
[13] M. Breeuwer and R. Plomp, "Speechreading supplemented with auditorily presented speech parameters," *J. Acoust. Soc. Amer.*, vol. 79, pp. 481–499, 1986.
[14] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proc. Int. Joint Conf. Speech and Signal Processing*, Minneapolis, MN, 1993, pp. 557–560.
[15] M. Brooke and E. D. Petajan, "Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics," in *Proc. Int. Conf. Speech Input/Output, Techniques and Applications*, London, U.K., 1986, pp. 104–109.
[16] R. Campbell, "Tracing lip movements: Making speech visible," *Visible Lang.*, vol. 22, pp. 33–57, 1988.
[17] S. Cox, I. Matthews, and A. Bangham, "Combining noise compensation with visual information in speech recognition," in *Proc. ESCA/ESCOP Workshop Audio-Visual and Speech Processing*, Rhodes, Greece, 1997, pp. 53–56.
[18] B. Dalton, R. Kaucic, and A. Blake, "Automatic speechreading using dynamic contours," in *Speechreading by Man and Machine: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, NATO ASI Series, 1996, pp. 373–382.
[19] B. V. Dasarathy, *Decision Fusion.* Los Alamitos, CA: IEEE Comput. Soc. Press, 1994.
[20] L. Deng and D. X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Amer.*, vol. 95, pp. 2702–2719, 1994.
[21] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," in *Proc. Int. Conf. Spoken Language Processing*, Yokohama, Japan, 1994, pp. 547–550.
[22] N. P. Erber, "Auditory-visual perception of speech," *J. Speech Hear. Disord.*, vol. 40, pp. 481–492, 1975.
[23] L. Girin, G. Feng, and J. L. Schwartz, "Speech enhancement with filters estimated from the speaker's image. A feasibility study," *Traite. Signal*, vol. 13, pp. 319–334, 1996.
[24] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Rationale for phoneme-viseme mapping and feature selection in visual speech recognition," in *Speechreading by Man and Machine: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, NATO ASI Series, 1996, pp. 505–518.
[25] A. Guérin-Dugué, P. Teissier, J. L. Schwartz, and J. Hérault, "Nonlinear representation for audio-visual fusion in a noisy-vowel recognition task," in *Proc. NEURAP'97*, Marseille, France, pp. 31–40.
[26] Y. Hatwell, "Transferts intermodaux et intégration intermodale," in *Traité de Psychologie Expérimentale*, M. Richelle, J. Reguin, and M. Robert, Eds. Paris, France: Presses Univ. France, 1993.
[27] T. S. Huang, C. P. Hess, H. Pan, and Z. Liang, "A neuronet approach to information fusion," in *Proc. 1st IEEE Workshop on Multimedia Signal Processing*, Princeton, NJ, 1997, pp. 45–50.
[28] P. Jourlin, "Word-dependent acoustic-labial weights in hmm-based speech recognition," in *Proc. ESCA/ESCOP Workshop Audio-Visual and Speech Processing AVSP'97*, Rhodes, Greece, 1997, pp. 69–72.
[29] D. H. Klatt, "Speech perception: A model of acoustic-phonetic analysis and lexical access," *J. Phonet.*, vol. 7, pp. 279–312, 1979.
[30] G. Krone, B. Talle, A. Wichert, and G. Palm, "Neural architecture for sensor fusion in speech recognition," in *Proc. ESCA/ESCOP Workshop Audio-Visual and Speech Processing*, Rhodes, Greece, 1997, pp. 57–60.
[31] R. Laboissière and A. Galvan, "Inferring the commands of an articulatory model from acoustical specifications of stop-vowel sequences," in *Proc. XIIIth Int. Cong. Phonetic Sciences*, 1995, vol. 1, pp. 358–361.
[32] M. T. Lallouache, "Un poste 'visage-parole'. Acquisition et traitement de contours labiaux," in *Proc. XVIII Journées d'Études sur la Parole*, Montréal, P.Q., Canada, 1990, pp. 282–286.
[33] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech Commun.*, vol. 11, pp. 215–228, 1992.
[34] J. Luettin and S. Dupont, "Continuous audio-visual speech recognition," in *Proc. 5th Eur. Conf. Computer Vision*, 1998.
[35] D. W. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Hillsdale, NJ: Lawrence Erlbaum, 1987.
[36] ———, "Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry," *Behav. Brain Sci.*, vol. 12, pp. 741–794, 1989.
[37] R. McGowan, "Recovering articulator movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary tests," *Speech Commun.*, vol. 14, pp. 19–48, 1994.
[38] U. Meier, W. Hürst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," in *Proc. ICASSP'96*, Atlanta, GA, pp. 833–836.
[39] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
[40] J. R. Movellan and G. Chadderdon, "Channel separability in the audio-visual integration of speech: A Bayesian approach," in *Speechreading by Man and Machine: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, NATO ASI Series, 1996, pp. 473–488.
[41] J. R. Movellan and P. Mineiro, "Modularity and catastrophic fusion: A Bayesian approach with applications to audio-visual speech recognition," Tech. Rep. 97.01, CogSci., Univ. Calif., San Diego, 1996.
[42] S. Nakamura, R. Nagai, and K. Shikano, "Adaptative determination of audio and visual weights for automatic speech recognition," in *Proc. Eurospeech* Rhodes, Greece, 1997, pp. 1623–1626.
[43] E. D. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D. dissertation, Univ. Illinois, Urbana, 1984.
[44] G. Potamianos, E. Cossato, H. P. Graf, and D. B. Roe, "Speaker independent audio-visual data base for bimodal ASR," in *Proc. ESCA/ESCOP Workshop Audio-Visual and Speech Processing* Rhodes, Greece, 1997, pp. 65–68.
[45] J. Robert-Ribes, "Modèles d'intégration audiovisuelle de signaux linguistiques: De la perception humaine à la reconnaissance automatique des voyelles," Ph.D. dissertation, INPG, Signal-Image-Parole, Grenoble, France, 1995.
[46] J. Robert-Ribes, M. Piquemal, J. L. Schwartz, and P. Escudier, "Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition," in *Speechreading by Man and Machine: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, NATO ASI Series, 1996, pp. 193–210.
[47] J. Robert-Ribes, J. L. Schwartz, and P. Escudier, "A comparison of models for fusion of the auditory and visual sensors in speech perception," *Artif. Intell. Rev. J.*, vol. 9, pp. 323–346, 1995.
[48] J. Robert-Ribes, J. L. Schwartz, M. T. Lallouache, and P. Escudier, "The optimality of bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise," *J. Acoust. Soc. Amer.*, vol. 103, pp. 3677–3689, 1998.
[49] A. Rogozan, P. Deléglise, and M. Alissali, "Adaptative determination of audio and visual weigths for automatic speech recognition," in *Proc. ESCA/ESCOP Workshop Audio-Visual and Speech Processing*, Rhodes, 1997, pp. 61–64.
[50] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception," in *Frontiers of Speech Communication Research*, B. Lindblom and S. Ohman, Eds. New York: Academic, 1979, pp.

217–229.
[51] J. Schroeter and M. Sondhi, "Techniques for estimating vocal tract shapes from the speech signal," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 133–150, 1994.
[52] J. L Schwartz, J. Robert-Ribes, and P. Escudier, "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*, R. Campbell, B. Dodd, and D. Burnham, Eds. Psychology Press, 1998.
[53] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 337–351, 1996.
[54] P. L. Silbsee and Q. Su, "Audio-visual sensory integration using hidden Markov models," in *Speechreading by Man and Machine: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, NATO ASI Series, 1996, pp. 489–496.
[55] D. G. Stork, G. Wolff, and E. Levine, "Neural network lipreading system for improved speech recognition," in *Proc. IJCNN'92*, Baltimore, MD, vol. 2, pp. 285–295.
[56] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, pp. 212–215, 1954.
[57] 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. Hillsdale, NJ: Lawrence Erlbaum, 1987, pp. 3–51.
[58] P. Teissier, "Fusion audiovisuelle avec prise en compte de l'environnement," in *DEA Signal-Image-Parole, INPG*, Grenoble, France, 1995.
[59] P. Teissier, J. L. Schwartz, and A. Guérin-Dugué, "Models for audio-visual fusion in a noisy-vowel recognition task," *J. VLSI Signal Process. Syst., Spec. Issue Mulimed.*, vol. 20, pp. 25–44, 1998.
[60] H. Ventsel, *Théorie des probabilités.* Moscow, Russia: Mir, 1973.
[61] J. H. M. Vroomen, "Hearing voices and seeing lips: Investigations in the psychology of the lipreading," Doctoral dissertation, Katholieke Univ., Brabant, The Netherlands, 1992.
[62] T. Watanabe and M. Kohda, "Lip-reading of Japanese vowels using neural networks," in *Proc. Int. Conf. Spoken Language Procesing*, Kobe, Japan, 1990, pp. 1373–1376.
[63] L. Xu, A. Krzyzak, and C.Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, pp. 418–435, 1992.
[64] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Commun. Mag.*, pp. 65–71, Nov. 1989.
[65] B. P. Yuhas, M. H. Goldstein, T. J. Sejnowski, and R. E. Jenkins, "Neural network models of sensory integration for improved vowel recognition," *Proc. IEEE*, vol. 78, pp. 1658–1668, 1990.



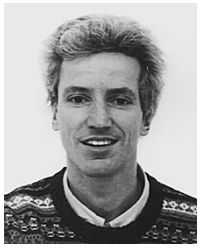
Pascal Teissier received the Electr. Eng. degree in 1995, and the Ph.D. degree in signal, image, and speech processing in 1999 from the Institut National Polytechnique, Grenoble, France.

He is a Researcher at the Institut de la Communication Parlée and the Laboratoire des Images et des Signaux, Grenoble. His main research interests are audiovisual speech recognition, speaker localization, neural networks, and multimodal fusion.

Jordi Robert-Ribes received the Ph.D. degree in signal, image, and speech processing from the Institut National Polytechnique, Grenoble, France in 1995.

He is a Researcher at the CSIRO-Mathematical and Information Sciences, Sydney, Australia. He is also a Visiting Fellow at the Computer Sciences Laboratory, Australian National University, and the editor of the *Australian Speech Science and Technology Newsletter*. His primary research interests are audiovisual speech and lip-reading, audio, and speech processing in general, signal processing for multimedia, automatic speech recognition, and team work effectiveness-communication. He is author and co-author of more than 25 refereed papers in journals, conferences, and book chapters.

Dr. Robert-Ribes co-organized the 1998 International Conference on Auditory-Visual Speech Processing (AVSP'98) and has been a member of many scientific committees of international conferences.



Jean-Luc Schwartz received the degree in physics from the Université d'Orsay 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 modeling and vowel perception in 1987.

Since 1983, he has been with the Centre National de la Recherche Scientifique, ICP. He has been leading the Speech Perception Group at ICP since 1988, and directed ten doctoral theses. He has been involved in various national and European projects on bimodal speech perception and processing (e.g., ESPRIT-BRA ACTS, ESPRIT-BR-Speech Maps, HCM-SPHERE, and TMR-SPHEAR). His main areas of research involve auditory modeling, psychoacoustics, speech perception, auditory front ends for speech recognition, bimodal integration, and perceptuo-motor interactions. He authored or co-authored more than 20 publications in international journals such as *JASA, Journal of Phonetics, Computer Speech and Language, Hearing Research, Artificial Intelligence Review, Speech Communication, Current Psychology of Cognition*, 15 book chapters, and 60 presentations in national and international workshops. He is co-editor of the speech communication journal *Le Bulletin de la Communication Parlée.*

Dr. Schwartz was a member of the Bureau of the French Audition Group from 1987 to 1990.

Anne Guérin-Dugué received the Ph.D. degree in electronics from the Institut National Polytechnique, Grenoble, France, in 1987.

She is currently with the Laboratoire des Images et des Signaux, LIS, INPG, Grenoble. Her main research interests are in computational modeling of human vision, high dimensional nonlinear data representation, and artificial neural networks for vision and, in general, pattern recognition. She is the author or co-author of more than 50 articles in these fields.

Dr. Guérin-Dugué is a member of the Organization and/or Scientific Committee of International Conferences and Journals, such as ESANN, IWANN, EUSIPCO, *Neural Processing Letters*, and *Signal Processing Journal*.