Audio-Visual Speech Modeling for Continuous Speech Recognition

Stéphane Dupont and Juergen Luettin

***Abstract*—This paper describes a speech recognition system that uses both acoustic and visual speech information to improve the recognition performance in noisy environments. The system consists of three components: 1) a visual module; 2) an acoustic module; and 3) a sensor fusion module. The visual module locates and tracks the lip movements of a given speaker and extracts relevant speech features. This task is performed with an appearance-based lip model that is learned from example images. Visual speech features are represented by contour information of the lips and grey-level information of the mouth area. The acoustic module extracts noise-robust features from the audio signal. Finally, the sensor fusion module is responsible for the joint temporal modeling of the acoustic and visual feature streams and is realized using multistream hidden Markov models (HMMs). The multistream method allows the definition of different temporal topologies and levels of stream integration and hence enables the modeling of temporal dependencies more accurately than traditional approaches. We present two different methods to learn the asynchrony between the two modalities and how to incorporate them in the multistream models. The superior performance for the proposed system is demonstrated on a large multispeaker database of continuously spoken digits. On a recognition task at 15 dB acoustic signal-to-noise ratio (SNR), acoustic perceptual linear prediction (PLP) features lead to 56% error rate, noise robust RASTA-PLP (Relative Spectra) acoustic features to 7.2% error rate and combined noise robust acoustic features and visual features to 2.5% error rate.**

***Index Terms*—Joint audio-video sensor integration, multistream hidden Markov models, speech recognition, visual feature extraction.**

I. Introduction

HUMAN speech perception is inherently a multimodal process that involves the analysis of the uttered acoustic signal and includes higher-level knowledge sources such as grammar, semantics, and pragmatics. One information source that is mainly used in the presence of acoustic noise is lipreading or so-called speechreading.[1] Hearing impaired and deaf persons make extensive use of visual speech cues and some few individuals perform lip-reading to such a degree that enables almost perfect speech perception [1]. It is well known that seeing the talker's face in addition to hearing his voice can improve speech intelligibility, particularly in noisy environments [2], [3]. The main advantage of the visual signal is its complementarity to the acoustic signal [4]. Phonemes that are most difficult to perceive in the presence of noise are easier to distinguish visually and vice versa. The visual signal contains that kind of information that is acoustically most sensitive to noise [1]. Studies have also shown that visual information leads to more accurate speech perception even in noise-free environments [5]. The strong influence of visual speech cues on human speech perception is demonstrated by the McGurk effect [6] in which, for example, a person hearing an audio recording of /baba/ and seeing the synchronised video of a person saying /dada/ often resulted in perceiving /gaga/.

Automatic speech recognition (ASR) has been an active research area for several decades, but in spite of the enormous efforts, the performance of current ASR systems is far from the performance achieved by humans: error rates are often one order of magnitude apart [7]. Most state-of-the-art ASR systems make use of the acoustic signal only and ignore visual speech cues. They are therefore susceptible to acoustic noise [8], and essentially all real-world applications are subject to some kind of noise. Much research effort in ASR has therefore been directed toward systems for noisy speech environments and the robustness of speech recognition systems has been identified as one of the biggest challenges in future research [9].

In this paper, we focus on audiovisual feature extraction, modeling, and sensor integration, for noise-robust ASR. Lip-tracking is performed using a model-based image search. Visual features are then extracted from the lip contours and from the mouth region intensity. This is discussed in Section II. Features from the audio signal are obtained using an acoustic front-end based on the perceptual linear prediction (PLP) or on the noise-robust J-RASTA-PLP (relative spectra) methods. In Section III, we tackle the problem of integrating the information obtained from the visual and acoustic front-ends. We are interested in the possibly decoupled dynamics of the two modalities. Both are modeled using hidden Markov models (HMMs) and the joint interaction of visual and acoustic HMMs is realized using multistream topologies. Section IV describes our acoustic, visual, and audio-visual speech recognition systems. Finally, results on a multispeaker digit strings recognition task are reported.

Manuscript received August 31, 1999; revised June 19, 2000. Part of this work was performed in the framework of the ACTS-M2VTS European Project with support from the Swiss Federal Office for Education and Science. The associate editor coordinating the review of this paper and approving it for publication was Dr. Masahiro Iwadare.

S. Dupont was with the TCTS Laboratory, Mons Polytechnical Institute (FPM's), Mons, Belgium. He is now with the International Computer Science Institute, Berkeley, CA 94720 USA (e-mail: dupont@tcts.fpms.ac.be).

J. Luettin was with IDIAP, Martigny, Switzerland. He is now with ASCOM, Maegenwil, Switzerland.

Publisher Item Identifier S 1520-9210(00)08577-1.

[1]Lipreading is the perception of speech purely based on observing the talkers lip movements. Speechreading is the visual perception of speech which also includes observation of facial and manual gestures. Audio-visual speech perception is the perception of speech by combining speechreading with audition.

II. Visual Speech Feature Extraction

Facial feature extraction is a difficult problem due to large appearance differences across persons and due to appearance variability during speech production. Different illumination conditions and different face positions cause further difficulties in image analysis. For a real-world application, whether it is in a car, an office or a factory, the system should be able to deal with these kinds of image variability.

The main approaches for extracting visual speech information from image sequences can be grouped into the following approaches:

1) *image-based;*
2) *visual-motion-based;*
3) *geometric-feature-based; and*
4) *model-based.*

In the *image-based* approach [10]–[13], the grey-level image containing the mouth is either used directly or after some image transform as feature vector whereas the *visual-motion-based* method [14] assumes that visual motion during speech production contains relevant speech information. *Geometric-feature-based* techniques [15], on the other hand, assume that certain measures such as the height or width of the mouth opening are important features. Finally, in the *model-based* approach [16]–[18], a model of the visible speech articulators, usually the lip contours, is built and its configuration is described by a small set of parameters. The advantage of the latter approach is that important features can be represented in a low-dimensional space and can often be made invariant to image transforms like translation, scaling, rotation and lighting. A disadvantage is that the particular model used may not consider all relevant speech information. The main difficulty in the model-based approach is the definition of the model and the development of image search procedures that accurately find the correspondence between the model and the image.

The system presented here falls into the category of model-based feature extraction. We have used an *appearance-based model* of the visual articulators [4], [17]: point distribution models [19] are used to track the lips and to extract relevant speech features from each image. Psychological studies suggest that the inner and outer lip contours are important visual speech features. The shape parameters obtained from the tracking results are therefore used as features for the speech recognition system. Lip shape information provides only part of the visual speech information. Other information is contained in the visibility of teeth and tongue, protrusion, and finer details. We therefore also extract intensity information from the mouth area.

A. Shape Modeling

The lip shape is represented by the coordinates of a point distribution model, outlining the inner and outer lip contours: $x = (x_0, y_0, x_1, y_1, \cdots, x_{Ns-1}, y_{N_s-1})^T$ where (x_j, y_j) are the coordinates of the jth point $(j = 0 \cdots N_s - 1)$. A shape is approximated by a weighted sum of basis shapes which are obtained by a Karhunen-Loéve expansion

$$x = \overline{x} + P_s b_s \tag{1}$$

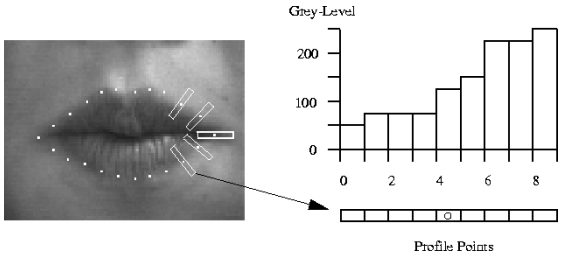

Fig. 1. Grey-level profile extraction. The grey-level vectors are sampled perpendicular to the lip contour and centred at the model points.

where $\overline{x}$ denotes the mean shape vector, $P_s = (p_{s1}, p_{s2}, \cdots, p_{sT_s})$ the matrix of the first T_s $(T_s < N_s)$ column eigenvectors corresponding to the largest eigenvalues and $b_s = (b_{s1}, b_{s2}, \cdots, b_{sT_s})^T$ a vector containing the weights for the eigenvectors, computed for the covariance matrix of a representative set of example images.

B. Intensity Modeling

Intensity modeling serves two purposes: firstly, it is used as a mean for a robust image representation to be used for image search in locating and tracking lips; secondly, it provides visual linguistic features for speech recognition. We therefore need to define dominant image features of the lip contours that we try to match with a certain representation of our model, but which also carry important speech information. Our approach to this problem is as follows. One-dimensional grey-level profiles g_j of length N_p are sampled perpendicular to the contour and centered at point j, as shown in Fig. 1.

The profiles of all model points are concatenated to construct a global profile vector $h = (g_0, g_1, \cdots, g_{N_i-1})^T$ of dimension $N_i = N_s N_p$. Similar to shape modeling, the intensity vector can be approximated by a weighted sum of basis intensities by the K-L expansion using

$$h = \overline{h} + P_i b_i, \tag{2}$$

where $\overline{h}$ denotes the mean intensity vector, $P_i = (p_{i1}, p_{i2}, \cdots, p_{iT_i})$ the $N_i \times T_i$ matrix of the first T_i $(T_i < N_i)$ column eigenvectors corresponding to the largest eigenvalues and b_i a vector containing the weights for each eigenvector. This approach is related to the *local grey-level models* described in [20] and to the *eigen-lips* reported in [11].

C. Image Search

The task of image search is to localize and track the lips in the image and to extract shape and intensity features. We define image search as finding the shape weight vector b_s^* of the model that maximizes the posterior probability (MAP) of the model given the observed image O_i

$$b_s^* = \arg \max_{b_s} P(b_s|O_i) = \arg \max_{b_s} \frac{P(O_i|b_s)P(b_s)}{P(O_i)}. \tag{3}$$

$P(O_i)$ is independent of b_s and can therefore be ignored in the calculation of b_s^*. We assume equal prior shape probabilities $P(b_s)$ within certain limits b_{smax} (e.g., ± 3 standard deviations)

and zero probability otherwise. This reduces the MAP to the likelihood function which is defined as

$$P(O_i|b_s) = (h - \overline{h})^T (h - \overline{h}) - b_i^T b_i \quad (4)$$

where the intensity weight vector b_i can be obtained using

$$b_i = P_i^T (h - \overline{h}) \quad (5)$$

and where h represents the intensity profile of the image corresponding to the model configuration b_s. The intensity weight vector b_i is constrained to stay within certain limits b_{imax} (e.g., ± 3 standard deviations), assuming equal prior intensity probabilities within these limits.

The Downhill Simplex Method [21] is applied to find a minimum of the cost function. We assume that a coarse estimate of the mouth location is given for the first image of a sequence to initialize the search process, for example by a face detection algorithm. Subsequent frames are processed by using the previous search results to initialize the Downhill Simplex Method. The obtained shape weight vector b_s^* and intensity weight vector b_i obtained from image search are used as visual feature vectors.

The accuracy of the lip-tracking algorithm might be estimated by comparing the results with the correct coordinates of the lip countour. These coordinates are however not available and might only be obtained by hand-labeling which is a very labourious and subjective task. Instead of evaluating the tracking performance separately we only evaluate the combined performance of lip-tracking, visual feature extraction, and visual speech modeling using the visual speech recognition performance. Experiments where lip-tracking performance has been evaluated separately can be found in [4].

Much visual speech information is contained in the dynamics of lip movements rather than the actual shape or intensity. Furthermore, dynamic information is likely to be more robust to extra-linguistic variability, i.e., intensity values of the lips and skin will remain fairly constant during speech, while intensity values of the mouth opening will vary during speech. On the other hand, intensity values of the lips and skin will vary between speakers, but temporal intensity changes might be similar for different speakers and robust to illumination. Similar comparisons can be made with shape parameters. Dynamic parameters (Δ parameters) of the shape and intensity vectors were therefore used as additional features.

The feature extraction method described here has been compared with several image-based approaches (low-pass filtering, principal components analysis, optical flow) by Gray *et al.* [22] and was found to outperform all of these methods. It was also found that the performance of image-based approaches can be considerably improved by the use of lip tracking results to normalize the images prior to processing.

III. Audio-Visual Sensor Integration

A. Problem of Audio-Visual Sensor Integration

The strong influence of visual stimuli on human speech perception has notably been demonstrated by the McGurk effect [6]. How humans integrate visual and acoustic information is not well understood. Several models for human integration have been proposed in the literature. They can be divided into early integration (EI) and late integration (LI) models [1]. In the EI model, integration is performed in the feature space to form a composite feature vector of acoustic and visual features. Classification is based on this composite feature vector. The model makes the assumption of conditional dependence between the modes and is therefore more general than the LI model. It can furthermore account for temporal dependencies between the modes, such as the voice-onset-time[2] (VOT), which are important for the discrimination of certain phonemes. In the LI model, each modality is first pre-classified independently of each other. The final classification is based on the fusion of the outputs of both modalities by estimating their joint occurrence. In comparison with the early integration scheme, this method assumes that both data streams are conditionally independent. Furthermore, temporal information between the channels is lost in this approach. Audio-visual speech recognition (AVSR) systems based on EI models have, for example, been described in [11] and [23] and systems based on LI models in [12] and [15]. Although it is still not well known how humans integrate different modalities, it is generally agreed that integration occurs before speech is categorized phonetically [1], [24]. This conclusion is supported by several studies regarding the VOT perception [25], [26] and the McGurk effect. In acoustic speech perception, on the other hand, there is much evidence that humans perform partial recognition across different acoustic frequency bands [27], [28], which assumes conditional independence across bands. The auditory system seems to perform partial recognition which is independent across channels, whereas audio-visual perception seems to be based on early integration, which assumes conditional dependence between both modalities. These two hypotheses are controversial since the audio-visual theory of early integration assumes that no partial categorization is made prior to the integration of both modalities.

The approach described here follows Fletcher's theory of conditional independence [27], [28], but it also allows the modeling of different levels of synchrony/asynchrony between the streams and can therefore account for some temporal dependencies, which otherwise can only be modeled by an EI integration model. Tomlinson *et al.* [23] have addressed the issue of asynchrony between the visual and acoustic streams. Under the independence assumption, composite models were defined from independently trained audio and visual models. Although our work is related with [23], we propose different strategies for modeling (learning) the asynchrony between the two streams.

The bimodal speech signal can be considered as an observation vector consisting of acoustic and visual features. According to Bayesian decision theory, the maximum *a posteriori* probability classifier (MAP) is denoted by

$$\Lambda^* = \arg \max_{\Lambda} P(\Lambda|O^a, O^v) = \frac{P(O^a, O^v|\boldsymbol{\Lambda})P(\boldsymbol{\Lambda})}{P(O^a, O^v)} \quad (6)$$

[2]The time delay between the burst sound, coming from the plosive part of a consonant, and the movement of the vocal folds for the voiced part of a voiced consonant or the subsequent vowel.

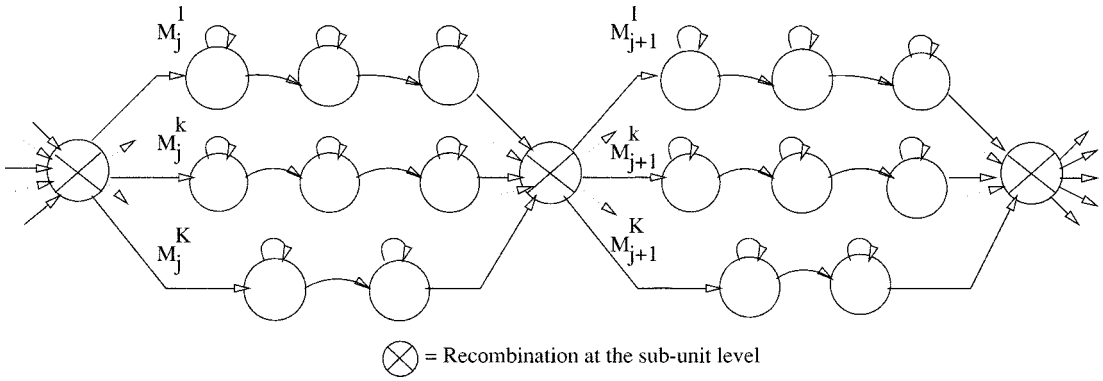

Fig. 2. General form of a K-stream model with anchor-points between speech units, forcing synchrony between the streams.

where Λ represents a particular word string, O^a represents the sequence of acoustic feature vectors, and O^v represents the sequence of visual feature vectors. If the two modalities are independent, the likelihood $P(O^a, O^v|\boldsymbol{\Lambda}_i)$ becomes $P(O^a|\boldsymbol{\Lambda})P(O^v|\boldsymbol{\Lambda})$. In this work, the modalities are assumed to be independent although certain temporal constraints and reliability weights are introduced.

Previous AVSR systems based on conditional independence have essentially addressed the problem of isolated word recognition. Most of these contributions were mainly focused on finding an appropriate automatic weighting scheme so as to guarantee good performance in a wide range of acoustic SNRs. Compared to isolated word recognition, the problem of continuous speech recognition is more tricky. Waiting until the end of the spoken utterance before combining the streams, as in the LI integration model, introduces an undesirable time delay. As the best hypothesis using the acoustic information is not necessarily the same as the best hypothesis using the visual information, it also requires to generate N-best hypothesis lists for the two streams. Identical hypotheses must indeed be matched to combine the scores from the two streams.

B. Multistream Model

The multistream approach, proposed in this work, does not require the use of such an N-best scheme. As we will show, it is an interesting candidate for multimodal continuous speech recognition as it allows for the following:

1) synchronous multimodal continuous speech recognition;
2) asynchrony of the visual and acoustic streams with the possibility to define phonological resynchronization points;
3) specific audio and video word or sub-word models; and
4) asynchrony patterns modeling.

1) Model for Decoupled Dynamics: The multistream approach [29] used in this work is a principled way for merging different sources of information using cooperative HMMs (see [30]). If the streams are supposed to be entirely synchronous and represented by HMMs with the same topologies, they may be accommodated simply. However, it is often the case that the streams are not synchronous, that they do not even have the same frame rate and it might be necessary to define models that do not have the same topology. The multistream approach allows to deal with this. In this framework, the input streams are processed independently of each other (using HMMs) up to certain anchor-points where they have to synchronize and combine their partial segment-based likelihoods. While the phonological level of score combination has to be defined *a priori*, the optimal temporal anchor-points are obtained automatically during recognition.

This structure is meant for processes that evolve independently, i.e., streams that have somewhat decoupled dynamics. With the early integration approach (see Section III-A), several feature vectors are combined into a single feature vector. If the generating processes are only loosely coupled, as it could be assumed for articulatory movements, lip movements and vocal folds movements, this increases the variance of the statistical models, hence reducing the performance of a recognition system. With the LI strategy however, different statistical models are defined for the different feature vectors. Moreover, modality reliability can easily be introduced in this LI approach. Multistream uses the same assumptions but additionally introduces stream synchronization at relevant phonological transitions points, between phonemes, syllable or words for instance.

An observation sequence O, representing the utterance to be recognized, is assumed to be composed of K input streams O_k (possibly with different frame rates). A hypothesized model M associated with O is built by concatenating J sub-unit models $M_j (j = 1, \cdots, J)$ associated with the phonological level at which we want to perform the synchronization of the input streams (e.g., phonemes, syllables, words...). To allow the processing of each of the input streams independently of each other up to the pre-defined sub-unit boundaries, each sub-unit model M_j is composed of parallel HMMs M_j^k (possibly with different topologies). These HMMs are forced to combine their respective segmental scores[3] at the synchronization points. The resulting model is illustrated in Fig. 2.[4] In this model, we note that:

1) the parallel HMMs associated with each of the input streams do not necessarily have the same topology; and
2) the synchronization anchor-point ($\otimes$ in Fig. 2) is not a regular HMM state but combines the scores accumulated over the same temporal segment for all the streams.

[3]From now on, we will simply refer to likelihoods or probabilities as "scores."

[4]Different frameworks for more general networks have also been proposed in [31] and [32].

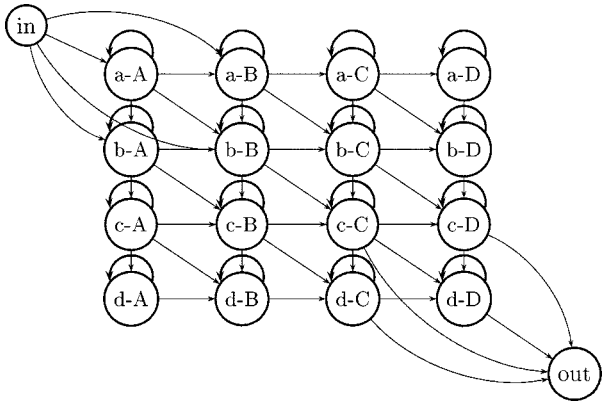

Fig. 3. HMM topology for composite model built from the multistream model presented in Fig. 4.

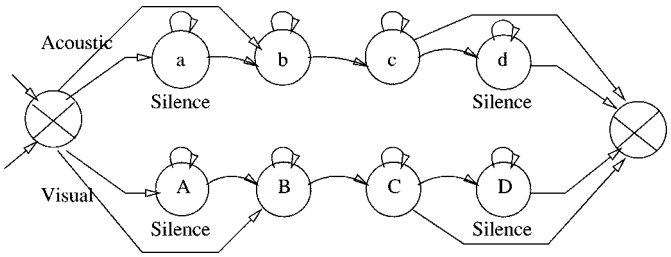

Fig. 4. Multistream model for audio-visual speech recognition with optional begin and end silence states (MODEL 2). These silence states are standard HMM states.

2) Composite Model Formulation: Recognition using the Viterbi approximation [30] appears to be a continuous speech decoding problem where all of the concurrent word segmentations, as well as all of the HMM state segmentations, must be hypothesized. However, as combination of the scores concerns sub-unit paths that must begin at the same time due to the synchrony constraints, and as the best sub-unit state paths are not the same for all of the streams (even if the model topologies are the same), it is necessary to keep track of the dynamic programming paths for all of the sub-unit starting points. Hence, an approach such as the asynchronous two-level dynamic programming [33], or a synchronous formulation of it, is required.

Alternatively, we can define composite HMMs [34] where each state is built by merging a K-tuple of states from the K stream HMMs. The topology of this composite model is defined so as to represent all the possible state paths given the initial HMM topologies. The local scores associated with the composite states are computed as a combination of the local stream scores (see Section III-B4). This model allows to implement independent search within sub-units as well as intra-units synchrony constraints. Fig. 3 shows the composite model obtained from the multistream topology in Fig. 4. This strategy was used in this study.

3) Psychoacoustic Motivations for Multistream: The human integration mechanism seems to be robust to small temporal asynchronies between information streams. In [35], a speech signal is partitioned into 19 quarter-octave frequency bands. These frequency channels are then randomly shifted in time according to a uniform distribution ranging from 0 to a maximum delay $D_{\max}$. Speech intelligibility experiments show that the word accuracy declines progressively as $D_{\max}$ increases. However, it is still above 75% for a strong 140 ms asynchrony condition, although the mean duration of the phonetic segments is 72 ms. It is expected that standard phone-based HMM systems would fail in such conditions.

In the audio-visual field, experiments in [36] introduced systematic asynchronies between the audio and video information sources. These intelligibility experiments, based on /ba/, /da/,/i/ and /u/ stimuli, indicated that the integration process is relatively robust for asynchronies up to 200 ms. Results by Smeele [37] showed that audio-visual intelligibility of CVC stimuli does not degrade for asynchronies of up to 80 ms. The multistream approach proposed here might also provide a robust framework with respect to such asynchronies.

4) Stream Combination: Similar to the LI scheme, the multistream approach requires a formulation to combine the information of the two streams. In our case, this is done at each anchor-point. Combination of the independent likelihoods is done by multiplying the segment likelihoods from the two streams, thus assuming conditional independence of the visual and acoustic streams. This was done according to

$$P(O^a, O^v|\boldsymbol{\Lambda}) \triangleq P(O^a|\boldsymbol{\Lambda}^a)^w P(O^v|\boldsymbol{\Lambda}^v)^{(1-w)}. \quad (7)$$

The weighting factor $w\,(0 \leq w \leq 1)$ represents the reliability of the two modalities. It generally depends on the performance obtained by each modality and on the presence of acoustic or visual noise. Here, we estimate the optimal weighting factor on the development set which is subject to the same noise as the test set. The method used for final experiments however was to automatically estimate the acoustic SNR from the test data and to adjust the weighting factor accordingly. It can be observed empirically that the optimal weight is related almost linearly to the SNR ratio. For our best system (see Section IV), the correlation coefficient between the SNR and the optimal weight is 0.99, and the weight can be estimated using the following empirical linear regression $w = 0.009$ SNR(dB) + 0.512, that is valid from 5 dB to clean speech. With clean speech (SNR $>$ 30 dB), the weighting factor ($w > 0.78$) strongly favors the acoustic information. As the acoustic SNR decreases however, the visual information can become almost as important as the acoustic information (w = 0.56 when SNR = 5 dB).

5) Multistream Training: Given a combination weight, the multistream system parameters can be estimated using traditional (Viterbi-based) maximum likelihood techniques [30] applied to HMM systems, or to HMM systems using artificial neural networks (ANNs) as HMM state probability estimators [38], as was done in this work.

As was shown in [39], maximum-likelihood estimation of the combination weight fails. Indeed, maximizing the likelihood with respect to the weighting factor w yields to the selection of the modality with the highest likelihood (hence $w = 0$ or $w = 1$). The authors also show that additional constraints on the weight can yield to a satisfactory solution. Alternatively, generalized probabilitic descent (GPD) training using a minimum-classification-error criterion can also be used. In this work, we estimate the stream combination weight using a true word-level

classification-error criterion, based on development data. This estimation step is performed at each iteration of the Viterbi-based maximum-likelihood estimation algorithm.

It is important to note that the multistream parameter optimization procedure used here is different than the optimization of two single-stream systems independently, as proposed in [39]. The multistream model topologies, particularly the synchronization anchor-points, introduce additional constraints in the forced alignment of the training data. These constraints can be important to "correct" the alignments of the visual stream. We observed in this study that, without such constraints, the alignments fail for a significant amount of the training: HMM states of particular words were sometimes shown to be aligned on signal portions pertaining to other words. In some cases, single HMM states were also shown to overlap several adjacent words.

The transition probabilities of a multistream model can be different over the two modalities, and this is reflected on the transition probabilities of the composite model used during decoding. These transition probabilities are estimated jointly with the parameters representing the emission probabilities, using the standard maximum likelihood approach. Transition probabilities are often omitted when modeling speech processes, the observation likelihood being dominated by the emission probabilities. The fact is that the standard geometric duration model is not accurate and also that the transition probabilities do not help in the case of matched train/test conditions. Using accurate duration modeling techniques, it has been shown that transition probabilities can improve performance in the case of noisy speech [40]. As explained in the next section, transition probabilities were used here for the modeling of duration and asynchrony patterns.

6) Synchrony/Asynchrony Modeling: Whereas HMMs are mainly used to model a single or several dependent processes, multistream models can be used for processes that evolve independently within predefined anchor-points (transitions between lexical sub-units) where the processes are assumed to resynchronize. However, the definition of these anchor-points is not obvious as the multiple stream dynamics is not known a priori. Moreover, it is very likely that many problems will be characterized by vector streams coming from processes that are neither dependent, nor completely decoupled, leading to tight or loose synchrony, probably depending on the model states. Finally, some processes could also lead to synchrony/asynchrony patterns, one of the streams being in advance, or systematically delayed, with respect to the other streams. Such asynchrony phenomena could be learned to improve the modeling accuracy.

Two approaches have been investigated here. The first one consist of static state pruning. This was done by pruning the multidimensional models by removing the least frequently visited states (based on the prior probabilities of these states).

The second approach for stream synchrony/asynchrony learning consists of modeling the asynchrony patterns resulting from the multistream processes. This was done by explicit modeling of the state durations and transition probabilities of the multidimensional models. Indeed, in the composite HMM, the off-diagonal state (Fig. 3) durations correspond to the stream asynchrony delay and the transition probabilities to these states represent their associated probabilities.

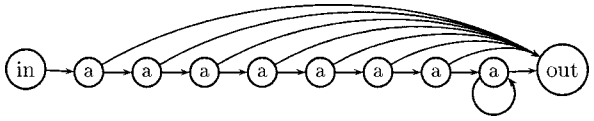

Fig. 5. Example of HMM topology for duration modeling. The emission probability model is common across the different states. The duration is encoded within the transition probabilities.

Explicit modeling of the synchronous and asynchronous state durations was used here to take advantage of the particular structure of state transitions. Viterbi decoding allows to use a particular HMM topology for duration modeling. We implement duration modeling by a chain of identical HMM states and a set of transition probabilities as shown in Fig. 5. The topology is characterized by a maximum length after which the duration model simply becomes exponentially decaying, due to the loop on the last state. Each of the states of the multidimensional model is replaced by a similar topology and the transition probabilities are updated using a counting procedure on a forced Viterbi alignment of the training data. Other duration modeling techniques could also be used [40].

In the presence of noise, the prior information represented by the duration models might be of significant importance. It should be noted however that this model can only be applied at the expense of an important increase of the computational requirements. For a model of duration D, each transition of the one state model is replaced by D transitions toward the following HMM models.

IV. Speech Recognition Experiments

The M2VTS audio-visual database [41] was used for all experiments. It contains 185 recordings of 37 subjects (12 females and 25 males). Each recording contains the acoustic and the video signal of the continuously pronounced French digits from zero to nine. Five recordings have been taken of each speaker, at one week intervals to account for minor face changes like beards. The video sequences consist of 286 $\times$ 360 pixel color images with a 25 Hz frame rate and the audio track was recorded at a 48 kHz sampling frequency and 16 bit PCM coding. The database contains a total of over 27 000 color images which were converted to grey-level images for the experiments reported here.

Although the M2VTS database is one of the largest databases of its type, it is still relatively small compared to reference audio databases used in the field of speech recognition. To increase the significance level of our experiments, we used a jack-knife approach. Five different cuts of the database were used. Each cut consisted of:

1) three pronunciations from the 37 speakers as training set;
2) one pronunciation from the 37 speakers as development set; and
3) one pronunciation from the 37 speakers as test set.

The development set was used to optimize the audio-visual weighting exponent. This procedure allowed to use the whole database as test set (185 utterances) by performing independent experiments for each of the five cuts. The task could be qualified as multispeaker continuous digits speech recognition.

We note here that the digit sequence to be recognized is always the same (digits from "0" to "9"). This somewhat simplifies the task of the speech recognition system which always "sees" the pronounced words in the same context. Moreover, during recognition, only the hypothesis with the correct number of digits (ten digits) were considered. This choice was made to avoid the need to optimize word entrance penalties.

Although highly constrained, this task remains a true continuous speech recognition task. Such constrained problems are often used to evaluate features and acoustic models whereas large vocabulary tasks are mainly used to evaluate language models.

A. Acoustic Speech Recognition

The audio stream was first downsampled to 8 kHz. We used PLP parameters [42] computed every 10 ms on 30 ms sample frames. The complete feature vectors consisted of 25 parameters: 12 PLP coefficients, the first temporal derivatives [43] of these coefficients (12 ΔPLP) and the Δenergy.

We used left-right digit HMM models with between three and nine independent states, depending on the digit mean duration. This yielded a total of 52 states, including a standard HMM state representing the silence. The digit sequences were first segmented into digits using standard Viterbi alignment with a HMM-based recognizer trained on the SWISS-FRENCH POLYPHONE database [44] of 5000 speakers. Each M2VTS digit was then linearly segmented according to the number of states of the corresponding HMM model. This initial segmentation was used to train the HMM-state statistical models using an artificial neural network as HMM state probability estimator [38]. We used a feed-forward Multilayer Perceptron (MLP) trained with speech features at its input to generate HMM state posterior probabilities. Nine adjacent frames of acoustic features were used at the input of the MLP. This allows to model local time correlation and was shown to improve classification performance [38]. Back-propagation was used to adapt the MLP weights using a gradient descent algorithm. A neural network with 150 hidden units was used. Increasing the number of hidden neurons did not yield any performance improvement.

System training and tests were then performed according to the database partitioning described earlier. Results are summarized in Tables I and II for speech corrupted by stationary Gaussian white noise with different SNRs. [5] Comparing Table I with Table II, we can observe that the recognition performance is severely affected by additive noise, even at such moderate noise levels.

B. Visual Speech Recognition

The most dominant 12 shape features and 12 intensity features, described earlier, were used for the recognizer. These features were complemented by 24 temporal derivatives. We used the same HMM topologies and the same initial segmentation as for the previously described acoustic-based recognition system. In this case, the MLP had 70 hidden units.

The mean error rate for the five database cuts defined earlier was 40.3%. Since the visual signal only provides partial information, the error rate for the video-based system was considerably higher than for the audio-based system. This is mainly due to the high visual similarity of certain digits like "quatre," "cinq," "six," and "sept."

[5]In these experiments, the SNR was computed at the sentence level without removing the silence portions of the utterances.

TABLE I
WORD ERROR RATE OF PLP-BASED ACOUSTIC-, VISUAL-, AND ACOUSTIC-VISUAL-BASED (MODEL 1) SPEECH RECOGNITION SYSTEMS ON CLEAN SPEECH. STANDARD DEVIATIONS ACROSS THE FIVE DATABASE CUTS ARE IN BRACKETS

System	Video	Audio	Audio-Visual
Error rate	40.3% (2.8)	1.4% (0.5)	1.2% (0.3)

TABLE II
WORD ERROR RATE FOR DIGIT STRING RECOGNITION WITH NUMBER OF DIGITS KNOWN *A PRIORI*, USING SEVERAL KINDS OF ACOUSTIC-VISUAL-BASED SPEECH RECOGNITION SYSTEMS. THESE RESULTS REPRESENT THE MEAN WORD ERROR RATE ACROSS FIVE NOISE CONDITIONS: CLEAN SIGNAL, 20 dB, 15 dB, 10 dB, AND 5 dB SNR. THE NOISE WAS A STATIONARY GAUSSIAN WHITE NOISE. FOR EACH CONDITION, THE COMBINATION WEIGHT WAS OPTIMIZED ON A DEVELOPMENT SET SUBJECT TO THE SAME NOISE AS THE TEST SET. STANDARD DEVIATIONS ACROSS THE FIVE DATABASE CUTS ARE IN BRACKETS

Error Rate (%)	PLP	J-RASTA-PLP
Acoustic	48.4 (0.6)	12.1 (1.0)
Model0	29.6 (2.9)	6.7 (0.7)
Model1	18.8 (1.9)	**4.3** (0.5)
Model2	17.9 (2.6)	5.1 (0.9)
Model2+Pruning	**16.1** (1.7)	4.8 (0.8)
Acoustic+Duration	46.7 (0.7)	11.1 (1.1)
Model1+Duration	15.2 (1.7)	**4.2** (0.4)
Model2+Duration	15.1 (2.3)	4.6 (0.7)
Model2+Pruning+Duration	**14.9** (2.4)	4.5 (0.8)

C. Audio-Visual Speech Recognition

Audio-visual speech recognition was experimentally investigated. Fig. 6 illustrates the audio-visual system architecture. Three kinds of model topologies were compared. These were based on the HMM word topologies already used in the previous sections. The differences between the models laid notably in the possible asynchrony of the visual stream with respect to the acoustic stream.

The first model (MODEL 0) was based on early integration. Acoustic and visual features were used as input to a single MLP with 150 hidden units.

The second model (MODEL 1) corresponds to a multistream model with combination at the state level and allows to use fusion criteria that can weight differently the two streams according to their respective reliability. However, it did not allow for any asynchrony between the two streams.

The third model (MODEL 2) was a multistream model with combination of the streams at the word level (Fig. 4). This model thus allows the dynamic programming paths to be independent from the beginning up to the end of the words. In this work, the asynchrony was constrained to a difference of one state between the two modalities. In the example of Fig. 3, the following states are not allowed: $a-C$, $a-D$, $b-D$, $c-A$, $d-A$, and $d-B$.

This model also allows the transition from silence to speech and from speech to silence to occur at different time instants for the two streams.[6] Lip movement can occur before and after sound production and conversely. Fig. 7 shows in parallel a

[6]The "visual silence" state is a standard HMM state.

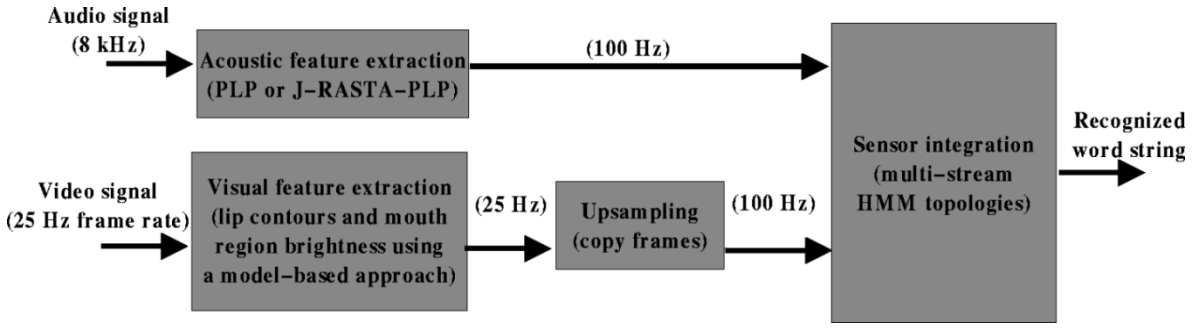

Fig. 6. Audio-visual ASR system architecture.

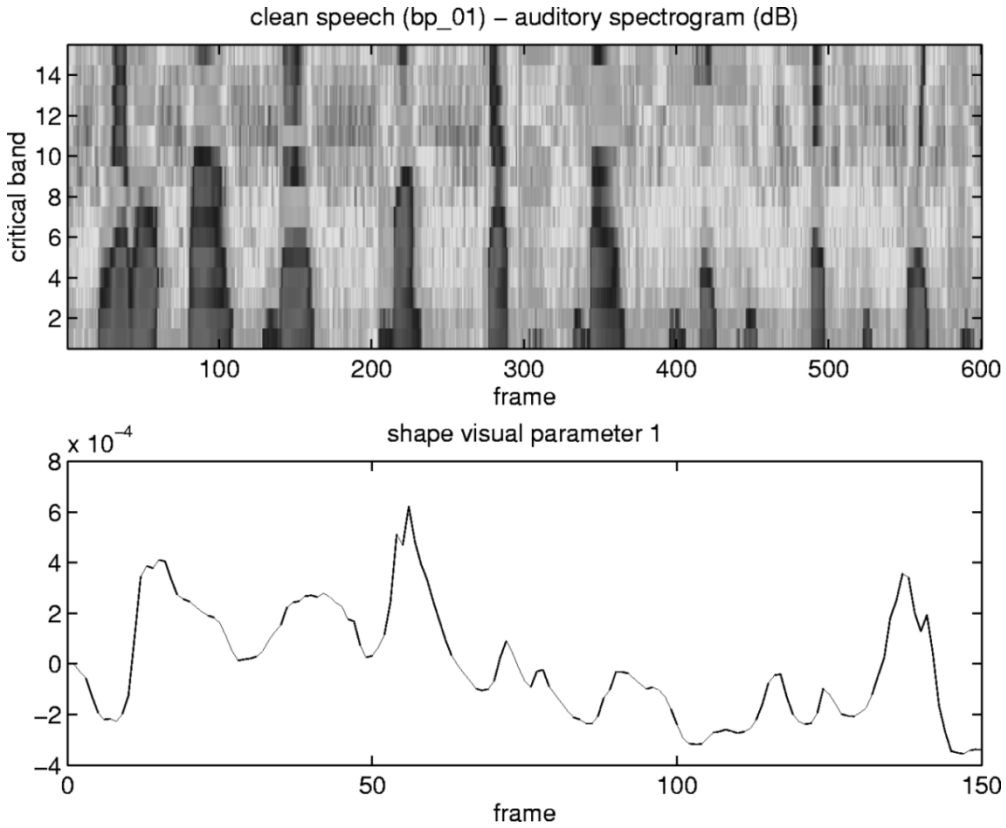

Fig. 7. Auditory spectrogram (evolution of the critical band energies) and evolution of the first visual shape parameter for one portion ("0" to "8") of an M2VTS utterance.

speech spectrogram as well as the evolution of the first visual shape parameter, mainly representing the changes in the position of the lower lip contour [17]. From this figure, as well as from studying the asynchrony of the streams using asynchrony lag histograms (see next section and Fig. 8), it can clearly be seen that the two signals are partially in synchrony and partially asynchronous. Ideally, we would like to have a model which forces the streams to be synchronous where synchrony occurs and asynchronous where the signals are typically in asynchrony. This will be studied in the next section.

We used the same parameterization schemes as in the two previous sections. However, as the visual frame rate (25 Hz) is a quarter of the acoustic frame rate, visual vectors were copied (by copying frames), so that both modalities are synchronously available.

Results are summarized in Tables I and II (PLP column). The optimal weighting factor was estimated on the development set which is subject to the same noise as the test set. In the case of clean speech, using visual information, in addition to the acoustics, does not yield significant performance improvement at $p < .05$. In the case of speech corrupted with noise, significant performance improvement can be obtained by using the visual stream as an additional information source. The slight improvement of MODEL 2 (compared to MODEL 1) is not significant.[7]

Finally, the early integration approach yields inferior results. The gain of combination weight adaptation seems to surpass the possible loss due to the independence assumption.

[7]This is in contradiction with results in [23] which were in favor of a resynchronization level allowing asynchrony although it was limited to the phoneme level (phone models being composed of three different states).

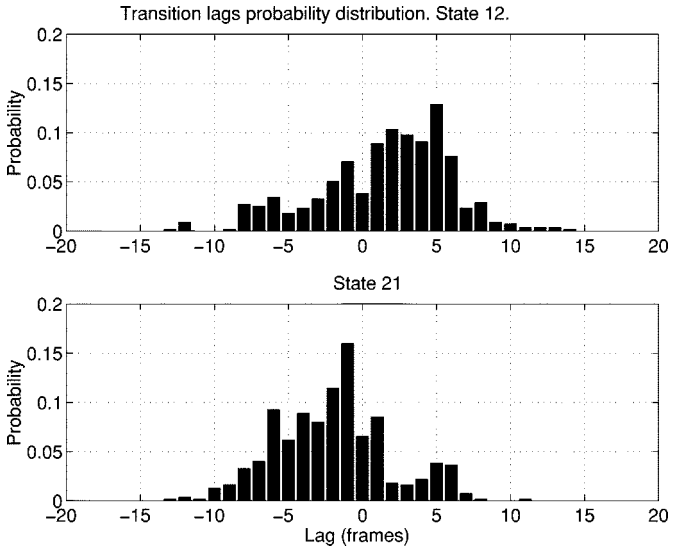

Fig. 8. Transition lags (lips on acoustics) probability distribution. The upper figure is for the first state of the word "trois" and the lower figure is for the third state (out of seven) of "quatre."

D. Learning the Asynchrony Between the Streams

The asynchrony learning approaches proposed in Section III-B6 were then applied. A forced Viterbi alignment of the training data was obtained using MODEL 2. Composite model state priors were computed using this alignment and the states with small prior probability were removed from the model (synchronized states were always kept however). We call this approach *static pruning*. The number of remaining states is optimized to get the best word recognition performance on the cross-validation set. Interestingly, the best models contained 25 off-diagonal states (asynchronized states) in addition to the 52 synchronized states. Speech recognition results on the test set are finally obtained using these simplified composite models. As can be seen in Table II (PLP column), they performed significantly better[8] (16.1% word error rate) than both state-resynchronization models (18.8%) and word-resynchronization model (17.9%).

As proposed in Section III-B6, *duration models* based on the composite models have also been developed. In this case, the states of the multidimensional composite model (Fig. 3) are replaced by particular HMM topologies (Fig. 5) which aim is to model the composite state durations. The length of these HMMs was set to 20 states, hence allowing an accurate duration modeling up to 20 frames. A self-loop on the last state allows longer durations with an exponentially decaying probability.

Transition delays were measured for the *M2VTS* database on a forced Viterbi alignment of the training data. MODEL 2 was used to obtain the alignment. Then, histograms of the transition delays were drawn. Observation of these histograms shows us that some are relatively narrow while other are very wide. Moreover, the transition delay mean is not always close to zero. Further observation even shows that some histograms are significantly shifted toward the positive values, indicating a pattern where the visual transition is generally delayed compared to the acoustic transition, some other are shifted toward the negative values (see Fig. 8). These observations tend to indicate that the transition delays are not only the product of alignment noise (as was hypothesized in [45] for streams based on different frequency bands) but also reflect some structure of the audio-visual asynchrony patterns that could be useful for speech recognition. The analysis of the transition delay distributions, which is however out of the scope of this paper, might also provide some insight into the speech production mechanism.

With this approach, the duration models, which were introduced to model the asynchrony patterns, are also modeling the state durations. Comparing it with a standard HMM model without duration modeling is unfair. The same kind of duration modeling was thus used in additional experiments with the purely synchronized (standard HMM) topologies. It was also applied to the best statically pruned composite models obtained with the previously described procedure. We observed that duration modeling significantly improves the noise robustness of all kinds of models (see Table II, PLP column).

For the systems without duration modeling, MODEL 2 simplified using static pruning performs significantly better than the other models. This suggest that allowing stream asynchronies, with asynchrony patterns learned in the form of multidimensional topologies can yield improved noise robustness. For the systems using duration modeling however, the results are not significantly different across the models.

E. Noise Robust Features

To allow a fair comparison with noise robust acoustic methods, this set of experiments was repeated using noise robust J-RASTA-PLP [46] features for the audio stream. Comparing the results in Table II shows the improved robustness of this kind of features over the PLP parameters. Moreover, using both information sources also results in an important leap forward in terms of robustness. Using decoding schemes allowing for stream asynchronies, even when using asynchrony modeling techniques, did not yield any performance gain (the results are not significantly different across the models). Results concerning asynchrony modeling are thus mitigated. Let us, however, emphasize here a single striking results from these experiments. At 15 dB SNR, PLP features lead to 56.3% error rate, J-RASTA-PLP features lead to 7.2% error rate, and using lip features in addition lead to 2.5% error rate (see Fig. 9).

Finally, speech was corrupted by a highly nonstationary noise from the *Madras* [47] database (moving cars recorded along a motorway). This noise was used because it is more realistic and more difficult to estimate than stationary white noise. The noise level was estimated using the automatic method described in [48]. The technique was shown to yield a 7.6 dB^2 mean square error on this kind of noise. Linear regression was used to dynamically adjust the stream combination weight, according to the estimated SNR and to the optimal weights resulting from the previous experiments on stationary white noise. Results are presented in Table III. Here again, the main conclusion of these results is the important gain resulting from audio-visual integration.

[8]according to a bilateral hypothesis test with $p < .05$ and knowing that the test set contains 5×1850 words because five noise conditions have been used.

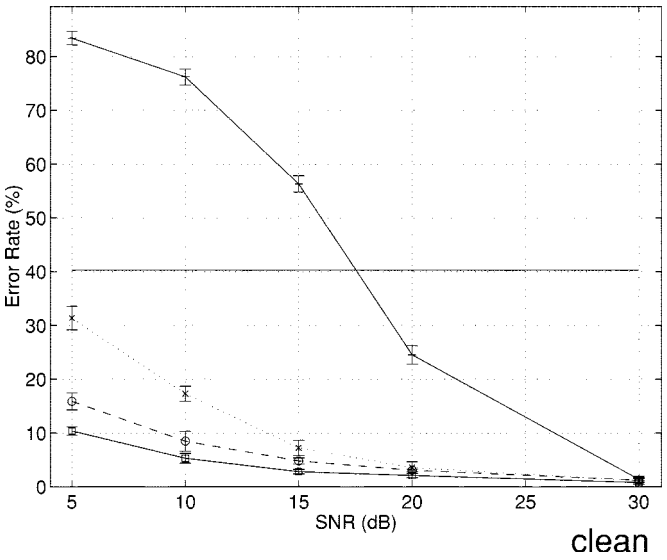

Fig. 9. Word error rate for digit string recognition with number of digits known *a priori*, under various acoustic SNRs. The noise was a stationary Gaussian white noise. The upper most line is for the PLP audio system. The dotted line is for the J-RASTA-PLP audio system. The dashed line is for the early integration system using J-RASTA-PLP and visual features and the last continuous line is for our best audio-visual system using J-RASTA-PLP and visual features. Error bars indicate ± 1 standard deviation across the five database cuts.

TABLE III
WORD ERROR RATE OF ACOUSTIC (A), VISUAL (V), AND SEVERAL KINDS OF ACOUSTIC-VISUAL-BASED SPEECH RECOGNITION SYSTEMS ON SPEECH CORRUPTED BY A HIGHLY NON-STATIONARY CAR NOISE FROM THE *MADRAS* DATABASE (10 dB MEAN SNR). "M0," "M1," AND "M2," RESPECTIVELY, REPRESENT MODEL0, MODEL1, AND MODEL2. J-RASTA-PLP IS USED FOR THE AUDIO STREAM

System	V	A	M0	M1	M2
Error rate (%)	40.3 (2.8)	10 (3.0)	6.9 (2.0)	3.7 (1.4)	4.2 (1.3)

V. CONCLUSIONS

We have described a complete audio-visual speech recognition system that was tested on a multispeaker continuous digit recognition task for different acoustic noise levels and noise sources.

We have described an approach based on appearance-based models for robust lip tracking and feature extraction. This method allows robust lip tracking for a broad range of subjects and without the need of lipstick or other visual aids. Visual speech information is compactly represented in the form of shape and intensity parameters. Visual speech recognition experiments have demonstrated that this technique leads to robust multispeaker continuous speech recognition.

We have presented a framework for the fusion of acoustic and visual information in an audio-visual speech recognition system based on the multistream approach. This provides a way of merging different sources of information using cooperative HMMs. Several significant advances have been achieved using this approach. Firstly, the method enables synchronous audio-visual decoding of continuous speech. Additionally, modality reliability can easily be introduced in the form of adaptive stream weights. It was shown that the gain of weight adaptation for speech recognition in noise is important and surpasses the possible loss due to the independence assumption of our fusion formalism. Finally, the approach allows to model the asynchrony between the two streams. In the case of audio-visual modeling, observations of HMM state alignment for audio and video streams tend to indicate that the transition delays are not only the product of alignment noise but also reflect some structure of the audio-visual asynchrony patterns that could be useful for speech recognition. Experimental results are however mitigated. Stream asynchrony modeling yield significant improvement in terms of noise robustness for ASR system using standard acoustic features whereas no improvement was observed for a system using noise robust features.

Comparisons with acoustic-only recognition systems show that the audio-visual system significantly reduces the error rate in the presence of noise, even in the case where noise robust acoustic features are used. The benefit of asynchrony modeling remains less conclusive and will be subject to further investigations.

REFERENCES

[1] A. Q. Summerfield, "Lipreading and audio-visual speech perception," *Philos. Trans. R. Soc. London B*, vol. 335, pp. 71–78, 1992.
[2] W. H. Sumby and I. Pollak, "Visual contributions to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, pp. 212–215, 1954.
[3] K. W. Grant and L. D. Braida, "Evaluating the articulation index for auditory-visual input," *J. Acoust. Soc. Amer.*, vol. 89, no. 6, pp. 2952–2960, 1991.
[4] J. Luettin, "Visual Speech and Speaker Recognition," Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 1997.
[5] D. Reisberg, J. McLean, and A. Goldfield, "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-Reading*, B. Dodd and R. Campbell, Eds. London, U.K.: Lawrence Erlbaum, 1987, pp. 97–113.
[6] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
[7] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–15, 1997.
[8] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, pp. 261–291, 1995.
[9] R. Cole, L. Hirschmann, and L. Atlas *et al.*, "The challenge of spoken language processing: Research directions for the nineties," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 1–20, 1995.
[10] 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, Oct. 1990.
[11] C. Bregler and S. M. Omohundro, "Nonlinear manifold learning for visual speech recognition," in *IEEE Int. Conf. Computer Vision*, Piscataway, NJ, 1995, pp. 494–499.
[12] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 337–351, 1996.
[13] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. IEEE Int. Conf. Image Processing*, 1998, pp. 173–177.
[14] K. Mase and A. Pentland, "Automatic lipreading by optical flow analysis," *Syst. Comput. Jpn.*, vol. 22, no. 6, 1991.
[15] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1985, pp. 40–47.
[16] T. Coianiz, L. Torresani, and B. Capril, "2D deformable models for visual speech analysis," in *Speechreading by Humans and Machines: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, 1996, vol. 150 of NATO ASI Series, Series F: Computer and Systems Sciences, pp. 391–398.
[17] J. Luettin and N. A. Thacker, "Speechreading using probabilistic models," *Comput. Vis. Image Understand.*, vol. 65, no. 2, pp. 163–178, Feb. 1997.
[18] S. Basu, N. Oliver, and A. Pentland, "3D modeling and tracking of human lip motion," in *Proc. IEEE Int. Conf. Computer Vision*, 1998.

[19] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," *Comput. Vis. Image Understand.*, vol. 61, pp. 38–59, Jan. 1995.
[20] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 743–756, July 1997.
[21] J. A. Nelder and R. Mead, "A simplex method for function optimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, 1965.
[22] M. S. Gray, J. R. Movellan, and T. J. Sejnowski, "Dynamic features for visual speechreading: A systematic comparison," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, vol. 9.
[23] M. J. Tomlinson, M. J. Russel, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing*, 1996, pp. 821–824.
[24] L. Braida, "Crossmodal integration in the identification of consonants," *Q. J. Exp. Psych.*, vol. 43A, no. 3, pp. 647–677, 1991.
[25] N. P. Erber and C. L. De Filippo, "Voice-mouth synthesis of tactual/visual perception of /pa, ba, ma/," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1015–1019, 1978.
[26] K. P. Green and J. L. Miller, "On the role of visual rate information in phonetic perception," *Percept. Psychophys.*, vol. 38, no. 3, pp. 269–276, 1985.
[27] H. Fletcher, *Speech and Hearing in Communication*. New York: Krieger, 1953.
[28] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.
[29] H. Bourlard and S. Dupont, "Sub-band-based speech recognition," in *Proc. IEEE Int.. Conf. Acoustic Speech and Signal Processing*, Apr. 1997, pp. 1251–1254.
[30] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall Signal Processing Series, 1993.
[31] M. I. Jordan, Z. Ghahramani, and L. K. Saul, "Hidden Markov decision trees," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, vol. 9.
[32] G. Gravier, M. Sigelle, and G. Chollet, "Toward Markov random field modeling of speech," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, Dec. 1998.
[33] H. Sakoe, "Two level DP matching—a dynamic time warping based pattern matching algorithm for continuous speech recognition," *IEEE Trans. IECE Jpn.*, vol. 3, 1979.
[34] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoustic Speech and Signal Processing*, 1990, pp. 845–848.
[35] T. Arai and S. Greenberg, "Speech intelligibility in the presence of cross-channel spectral asynchrony," in *Proc. ICASSP*, 1998, pp. 933–936.
[36] D. W. Massaro and M. M. Cohen, "Perceiving asynchronous bimodal speech in consonant vowel and vowel syllables," *Speech Commun.*, vol. 13, pp. 127–134, 1993.
[37] P. M. Smeele *et al.*, "Intelligibility of audio-visually desynchronized speech: Asymmetrical effect of phoneme position," in *Proc. Int. Conf. Spoken Language Processing*, Alberta, Canada, 1992, pp. 65–68.
[38] H. Bourlard and N. Morgan, *Connectionist Speech Recognition—A Hybrid Approach*. Norwell, MA: Kluwer, 1994.
[39] G. Potamianos and H. P. Graf, "Discriminative training of hmm stream exponents for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustic Speech and Signal Processing*, Seattle, WA, 1998, pp. 3733–3736.
[40] N. B. Yoma, F. R. McInnes, and M. A. Jack, "Weighted viterbi algorithm and state duration modeling for speech recognition in noise," in *Proc. IEEE Int. Conf. Acoustic Speech and Signal Processing*, Seattle, WA, 1998, pp. 709–712.
[41] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (release 1.00)," in *Proc. of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Crans-Montana, Switzerland, 1997.
[42] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.
[43] S. Furui, "Speaker independant isolated word recognizer using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.
[44] G. Chollet, J. L. Cochard, C. Jaboulet, A. Constantinescu, and P. Langlais, Swiss French polyphone and polyvar: Telephone speech databases to model inter and intra-speaker variability, , IDIAP, Martigny, Switzerland, 1996.
[45] N. N. Mirghafori, "A Multi-Band Approach to Automatic Speech Recognition," Ph.D. dissertation, Int. Comput. Sci. Inst., Berkeley, CA, Jan. 1999.
[46] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.
[47] ULg, "ULg—acoustics laboratory—The MADRAS project (1998, Aug.). [Online] Available http://www.montefiore.ulg.ac.be/services/acous/homelab.html
[48] S. Dupont and C. Ris, "Assessing local noise level estimation methods," in *Workshop on Robust Methods for Speech Recognition in Adverse Conditions (Nokia, COST249, IEEE)* Tampere, Finland, May 1999, pp. 115–118.

Stéphane Dupont received the Ph.D. degree in electrical engineering from the Mons Polytechnical Institute, Mons, Belgium, in 2000.

He has been a Visiting Researcher at IDIAP, Martigny, Switzerland, in 1996. He is currently a Research Assistant at the International Computer Science Institute, Berkeley, CA. His research interests include speech processing, neural networks, pattern recognition, and computer music. He has authored/coauthored over 20 papers on these topics.

Juergen Luettin received the Ph.D. degree in electronic and electrical engineering from the University of Sheffield, U.K.

He joined IDIAP, Martigny, Switzerland, in 1996 as a Research Assistant and became head of the computer vision group in 1997. He was a Visiting Researcher at the Center for Language and Speech Processing at Johns Hopkins University, Baltimore, MD, in 1997 and 2000. His research areas include computer vision, speech recognition, biometrics, and multimodal recognition.