IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Continuous Audio-Visual Speech Recognition

Juergen Luettin [1]   Stéphane Dupont [2,1]

IDIAP–RR 98-02

PUBLISHED IN
5th European Conference on Computer Vision, Freiburg, 1998

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

1 IDIAP, email: luettin@idiap.ch
2 Faculté Polytechnique de Mons — TCTS 31, Bld. Dolez, B-7000 Mons, Belgium, email: dupont@tcts.fpms.ac.be

IDIAP Research Report 98-02

Continuous Audio-Visual Speech Recognition

Juergen Luettin Stéphane Dupont

PUBLISHED IN
5th European Conference on Computer Vision, Freiburg, 1998

Abstract. We address the problem of robust lip tracking, visual speech feature extraction, and sensor integration for audio-visual speech recognition applications. An appearance based model of the articulators, which represents linguistically important features, is learned from example images and is used to locate, track, and recover visual speech information. We tackle the problem of joint temporal modelling of the acoustic and visual speech signals by applying Multi-Stream hidden Markov models. This approach allows the use of different temporal topologies and levels of stream integration and hence enables to model temporal dependencies more accurately. The system has been evaluated for a continuously spoken digit recognition task of 37 subjects.

1 Introduction

Human speech perception is inherently a multi-modal process, which involves the analysis of the uttered acoustic signal and which includes higher level knowledge sources such as grammar, semantics, and pragmatics. One information source which is mainly used in the presence of acoustic noise is lipreading or so-called speechreading. It is well known that seeing the talker's face in addition to audition can improve speech intelligibility, particularly in noisy environments.

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. Most state-of-the-art ASR systems make use of the acoustic signal only and ignore the visual speech cues. They are therefore susceptible to acoustic noise [15], and essentially all real-world applications are subject to some kind of noise. Much research effort in ASR has therefore been directed towards 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].

2 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 has to 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 *image based*, *geometric feature based*, *visual motion based*, and *model based* approaches. In the *image based* approach [34, 6, 30], the grey-level image containing the mouth is either used directly or after some pre-processing as feature vector. The advantage of this method is that no data is disregarded. The disadvantage is that it is left to the classifier to learn the nontrivial task of finding the generalisation for image variability (translation, scaling, 3D rotation, illumination) and linguistic variability (inter/intra speaker variability). The *visual motion based* approach [25] assumes that visual motion during speech production contains relevant speech information. This information is likely to be robust to different speakers and to different skin reflectance, however, the algorithms usually do not calculate the actual flow field but the visual flow field. A further difficulty consists in the extraction of relevant and robust features from the flow field. The *geometric feature based* approach [27] assumes that certain measures such as the height or width of the mouth opening are important features. Their automatic extraction is however not trivial and most of these systems have used semi-automatic methods or have painted the lips of the talker to facilitate feature extraction. In the *model based* approach 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 is that important features can be represented in a low dimensional space and can often be made invariant to translation, scaling, rotation, and lighting. A disadvantage is that the particular model used may not consider all relevant speech information. Some of the most successful *model based* approaches have been based on colour information [8, 29, 2], although it was found that individual chromaticity models are necessary for each subject if the method is being used for several persons [29]. In comparison, our system [23] is based on grey-level information only. A technique which enables lip tracking from different head poses and the recovery of 3D lip shape from the 2D view has been described in [2].

The system presented here falls into the category of *model based* feature extraction. An important issue is to choose an appropriate description of the visible articulators. We are modelling a physical process, so we could describe this process in terms of physical movements and positions of the articulators that determine the vocal tract. Specifically, for visual analysis, we could attempt to estimate muscle action from the image such as in [25, 13]. However, the musculature of the face is complex, 3D information is not present, muscle motion is not directly observable, and there are at least thirteen groups of muscles involved in the lip movements alone [18]. We have chosen to use an *appearance based* *model* of the visual articulators [23] based on point distribution models [11].

2.1 Shape Modelling

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

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

where $\mathbf{P_s} = (\mathbf{p}_{s1}, \mathbf{p}_{s2}, \ldots, \mathbf{p}_{sT_s})$ is the matrix of the first T_s $(T_s < N_s)$ column eigenvectors corresponding to the largest eigenvalues and $\mathbf{b_s} = (b_{s1}, b_{s2}, \ldots, b_{sT_s})$ a vector containing the weights for the eigenvectors.

The approach assumes that the principal modes are linearly independent, although there might be non-linear dependencies present. For objects with non-linear behaviour, linear models reduce the specificity of the model and can generate implausible shapes, which lead to less robust image search. They also require more modes of variation than the true number of degrees of freedom of the object. The specificity of a model can however be improved by a nonlinear process, e.g. by nonlinear PCA.

2.2 Intensity Modelling

Intensity modelling 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 which we try to match with a certain representation of our model, but which also carry important speech information. Our solution to this problem is as follows. One dimensional grey-level profiles $\mathbf{g}_{ij}$ of length N_p are sampled perpendicular to the contour and centred at point j, as described in [10]. The profiles of all model points are concatenated to construct a global profile vector $\mathbf{h}_i = (\mathbf{g}_{i0}, \mathbf{g}_{i1}, \ldots, \mathbf{g}_{iN_s-1})^T$ of dimension $N_i = N_s N_p$. Similar to shape modelling, the model intensity can be approximated by a weighted sum of basis intensities using a K-L expansion

$$\mathbf{h} = \overline{\mathbf{h}} + \mathbf{P}_i \mathbf{b}_i, \tag{2}$$

where $\mathbf{P}_i = (\mathbf{p}_{i1}, \mathbf{p}_{i2}, \ldots, \mathbf{p}_{iT_i})$ is the $N_i \times T_i$ matrix of the first T_i $(T_i < N_i)$ column eigenvectors corresponding to the largest eigenvalues and $\mathbf{b}_i$ a vector containing the weights for each eigenvector. This approach is related to the *local grey-level models* described in [22] and to the *eigen-lips* reported in [6].

We define image search as finding the shape of the model which maximises the posterior probability (MAP) of the model given the observed image O_i:

$$\mathbf{b}_s^* = \arg\max_{\mathbf{b}_s} P(\mathbf{b}_s|O_i) = \arg\max_{\mathbf{b}_s} \frac{P(O_i|\mathbf{b}_s)P(\mathbf{b}_s)}{P(O_i)} \tag{3}$$

$P(O_i)$ is independent of $\mathbf{b}_s$ and can therefore be ignored in the calculation of $\mathbf{b}_s^*$. We assume equal prior shape probabilities $P(\mathbf{b}_s)$ within certain limits $\mathbf{b}_{smax}$ (e.g. $\pm$ 3 *s.d.*) and zero probability otherwise. This reduces the MAP to the likelihood function which is defined as

$$P(O_i|\mathbf{b}_s) = E_i^2 = (\mathbf{h} - \overline{\mathbf{h}})^T(\mathbf{h} - \overline{\mathbf{h}}) - \mathbf{b}_i^T \mathbf{b}_i \tag{4}$$

where $\mathbf{b}_i$ can be obtained using

$$\mathbf{b}_i = \mathbf{P_i}^T(\mathbf{h} - \overline{\mathbf{h}}) \tag{5}$$

and where $\mathbf{h}$ represents the intensity profile of the image corresponding to the model configuration $\mathbf{b}_s$.

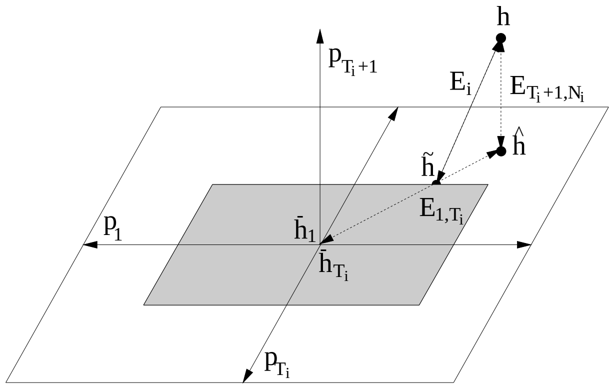

Figure 1: A simple model for $T_i = 2$ and $N_i = 3$ with mean $\overline{\mathbf{h}} = (\overline{h}_1, \overline{h}_T)^T$ and two eigenvectors $\mathbf{p}_1, \mathbf{p}_{T_i}$. The *best fit* $\hat{\mathbf{h}}$ is the projection of $\mathbf{h}$ onto the surface spanned by $\mathbf{p}_1$ and $\mathbf{p}_{T_i}$ which results in the residual error E_{T_i+1,N_i}. Constraining all parameters b_i to stay within a certain limit $\mathbf{b}_{imax}$ (shaded area) results in the *limited fit* $\tilde{\mathbf{h}}$ and the residual error E_i.

Cootes et al. [10] have used the following measure to estimate how well the model fits the profile:

$$E_c^2 = \sum_{j=1}^{T_i} \frac{b_{ij}^2}{\lambda_j} + \frac{E_i^2}{0.5\lambda_{T_i}} \tag{6}$$

where λ_i is the eigenvalue corresponding to the ith eigenvector with $\lambda_i \leq \lambda_{i+1}$. This measure considers both the distances between the considered modes from the mean $E_{1,T}$ (first term) and the distance not explained by the considered modes $E_{T+1,N}$ (second term). The notation $E_{n,m}$ refers to the residual error for the modes n to m. The relative weighting of both terms assumes that the sum of squares of residuals are Gaussian distributed and have a variance of $0.5\lambda_{T_i}$. It has been shown [26] that the optimal value for $0.5\lambda_{T_i}$ is the arithmetic mean of the eigenvalues $(\lambda_{T_i+1}, \ldots, \lambda_{N_i})$. Since this measure penalises values far from the mean, it is unlikely to be appropriate for the application of lip localisation and tracking, where the intensities vary considerably for different subjects and mouth opening. In this case it is more desirable to assign equal prior probabilities to instances within a certain limit and to constrain the model parameters to stay within these limits. This strategy was implemented here, by using the sum of residual square errors E_i^2 as distance measure but forcing all intensity modes to stay within certain limits $\mathbf{b}_{imax}$.

Figure 1 illustrates the different error measures for a simple model with two modes of variation. Along the directions $\mathbf{p}_i$ for which $i \leq T_i$, the weights b_i are not considered in the error function E_i, but they are constrained to lie within the limits $\mathbf{b}_{imax}$ (e.g. $\pm$ 3 *s.d.*). For the point $\mathbf{h}$ the *best fit* $\hat{\mathbf{h}}$ is the projection of $\mathbf{h}$ onto the surface spanned by $\mathbf{p}_1$ and $\mathbf{p}_{T_i}$, resulting in the residual error E_{T_i+1,N_i}. The *limited fit* $\tilde{h}$ is obtained by limiting the weight vectors which results in the residual error E_i. Examples of tracking results are shown in Fig. 2.

2.3 Feature Extraction

Psychological studies suggest that the inner and outer lip contour are important visual speech features. The shape parameters $\mathbf{b}_s$ obtained from the tracking results are therefore used as features for the speech recognition system. Translation, rotation and scale parameters are disregarded since they are unlikely to provide speech information. The shape features are invariant to translation, rotation (2D), scale, and illumination.

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. The intensity parameters

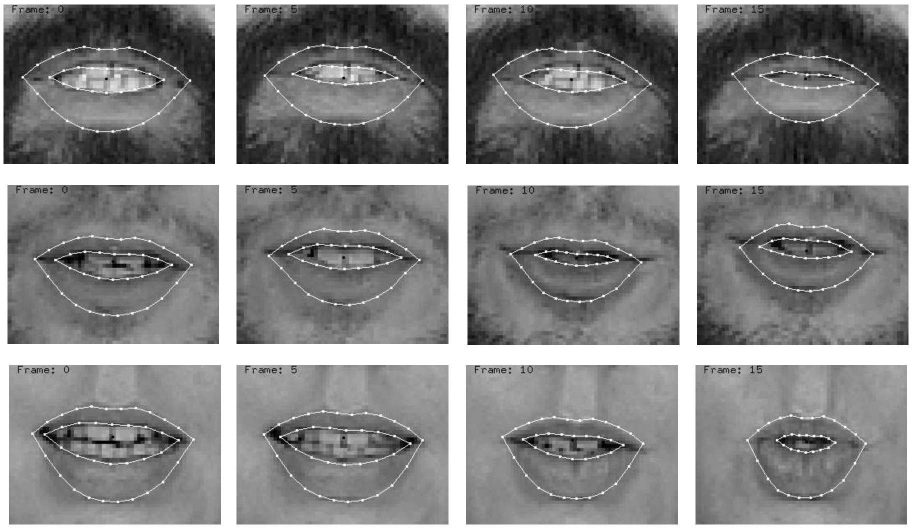

Figure 2: Examples of lip tracking results. The examples demonstrate the robustness of the algorithm for appearance variability across subjects (e.g. beards) and appearance variability during speech (e.g. visibility of teeth).

$\mathbf{b}_i$ of the lip model are therefore used as features to provide information complementary to the shape features. The intensity of a 2D image reflects its actual 3D shape and provides information not covered by the shape model. In general, the intensity image depends on the shape of the object, its reflectance properties, the light source, and the viewing angle. For a Lambertian surface the image radiance or brightness $I_r(x, y)$ at a particular point in the image is proportional to the irradiance or illumination $I_i(x, y)$ at that point. The radiance depends on the irradiance of the illumination source $I_i(x, y)$ and the angle θ between the surface normal and the direction toward the illumination source:

$$I_r(x, y) = \frac{1}{\pi} I_i(x, y) \cos\theta \qquad \text{for } \theta \geq 0. \tag{7}$$

This equation directly relates the 3D shape to intensity and is fundamental to the methods for recovering shape from shading [20]. The recovery of 3D shape from shading is possible under certain constraints, i.e. it is normally assumed that the angle of the illumination source is known and that the surface is smooth. For our application of the face image, the shape from shading problem becomes very difficult. The illumination source is generally not known, the surface is disrupted at the oral opening, the oral opening itself is not smooth, and the reflectance properties of facial parts are not homogeneous and generally not known. Here, the motivation behind the use of intensity information is therefore not to reconstruct the 3D shape but to use it to implicitly represent 3D information and to represent information about the position and configuration of facial parts based on their individual brightness.

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 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 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, PCA, optical flow) by Gray et al. [16] and was found to outperform all of these methods. It was also found that the performance of image-based approaches can be considerably

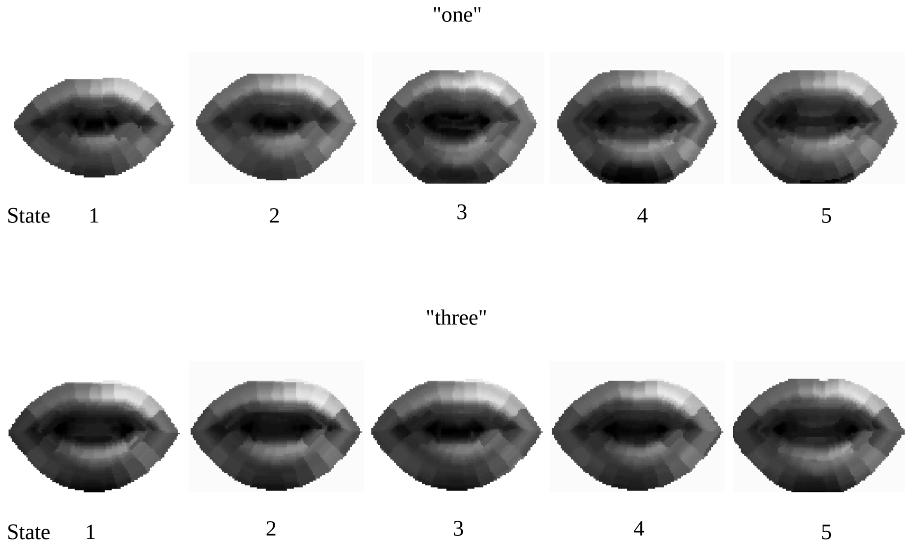

Figure 3: Learned sequence of quasi-stationary states for the words “one” and “two” learned by the HMM. The images represent the mean shape and intensities learned from training data of 11 subjects. The images represent the lips, the mouth opening, and the skin region around the lips.

improved by the use of the lip tracking results to normalise the images prior to processing.

Figure 3 displays learned hidden Markov models (HMMs) for the words “one” and “two” using extracted visual shape and intensity features from example utterances of 11 subjects. The images represent the mean images synthesised by the shape features and intensity features of HMMs with one Gaussian distribution per state.

3 Audio-Visual Sensor Integration

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 [31]. 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[1] (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 EI scheme, this method assumes that both data streams are conditionally independent. Furthermore, temporal information between the channels is lost in this approach. AVSR systems based on EI models have for example been described in [6, 32] and systems based on LI models in [27, 30]. Although it is still not well known how humans integrate different modalities, it is generally agreed that integration occurs before speech is categorised phonetically [5, 31]. Furthermore, several studies have shown that consonants which differ by the VOT such as “bi” and “pi”, are distinguished based on the evidence of both modalities [12, 17]. It was concluded that integration, therefore, must take place before phonetic categorisation. In acoustic speech perception, on the other hand, there is much evidence that humans perform partial recognition across different acoustic frequency bands [14, 1] which assumes conditional independence across bands. The auditory

[1] The time delay between the burst sound and the movement of the vocal folds

system seems to perform partial recognition which is independent across channels, whereas audio-visual perception seems to be based on some kind of 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 categorisation is made prior to the integration of both modalities.

The approach described here follows Fletcher's theory of conditional independence [14, 1], but it also allows the modelling of different levels of synchrony/asynchrony between the streams and can therefore account for speech features like the VOT, which otherwise can only be modelled by an EI integration model. Tomlinson et al. [32] have already addressed the issue of asynchrony between the visual and acoustic streams by the use of HMM decomposition. Under the independence assumption, composite models were defined from independently trained audio and visual models. Although our work is strongly related with [32], it allows to consider different recombination formalisms and enables the decoding of continuous speech. Moreover, the scope of asynchrony between the two streams was here extended from the phone level to the word level.

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

$$\Lambda^* = \arg\max_{\Lambda} P(\Lambda|\mathbf{O}^a, \mathbf{O}^v) = \frac{P(\mathbf{O}^a, \mathbf{O}^v|\Lambda)P(\Lambda)}{P(\mathbf{O}^a, \mathbf{O}^v)} \tag{8}$$

where Λ represents a particular word string, $\mathbf{O}^a$ represents the sequence of acoustic feature vectors $\mathbf{O}^a = \mathbf{o}^a(1), \mathbf{o}^a(2), \ldots, \mathbf{o}^a(T)$ and $\mathbf{O}^v$ the sequence of visual feature vectors $\mathbf{O}^v = \mathbf{o}^v(1), \mathbf{o}^v(2), \ldots, \mathbf{o}^v(T)$. If the two modalities are independent, the likelihood $P(\mathbf{O}^a, \mathbf{O}^v|\Lambda_i)$ becomes

$$P(\mathbf{O}^a, \mathbf{O}^v|\Lambda) = P(\mathbf{O}^a|\Lambda)P(\mathbf{O}^v|\Lambda). \tag{9}$$

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 signal-to-noise ratios. Compared to isolated word recognition, the problem of continuous speech recognition is more tricky as we do not want to wait until the end of the spoken utterance before recombining the streams. This introduces a time delay and it also requires to generate N-best hypothesis lists for the two streams. Indeed, one can only recombine the scores from identical hypothesis. As the best hypothesis for the acoustic stream is not necessarily the same as the best hypothesis for the visual stream, techniques such as N-best lists are required. Identical hypothesis must then be matched to recombine the scores from the two streams. An alternative approach would be to generate an N-best list for one of the two streams, to compute the score of these best hypothesis for the other stream, and finally to recombine the scores.

The Multi-Stream approach, proposed in this work, does not require to use such an N-best scheme. As we will show, it is an interesting candidate for multimodal continuous speech recognition as it allows for: (1) synchronous multimodal continuous speech recognition, (2) asynchrony of the visual and acoustic streams with the possibility to define phonological resynchronisation points, (3) specific audio and video word or sub-word HMM topologies.

3.1 Multi-Stream Statistical Model

The Multi-Stream approach [4, 3] used in this work is a principled way for merging different sources of information using cooperative HMMs[2]. If the streams are supposed to be entirely synchronous, 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 Multi-Stream approach allows to deal with this. In this framework,

[2] A different framework for more general networks has also been proposed in [21].

the input streams are processed independently of each other up to certain anchor points where they have to synchronise and recombine their partial segment-based likelihoods. While the phonological level of recombination has to be defined a priori, the optimal temporal anchor points are obtained automatically during recognition.

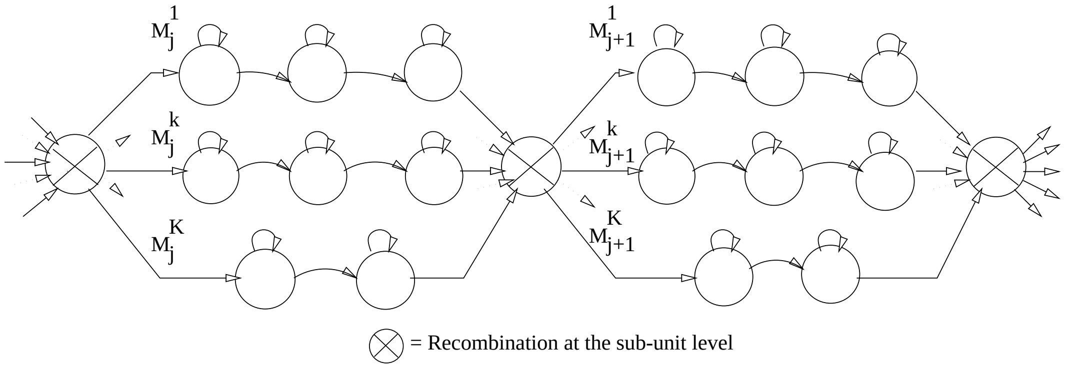

Figure 4: General form of a K-stream model with anchor-points between speech units, forcing synchrony between the streams.

An observation sequence **O**, representing the utterance to be recognised, is assumed to be composed of K input streams X_k (possibly with different frame rates). A hypothesised model M associated with **O** is built by concatenating J sub-unit models M_j $(j = 1, \ldots, J)$ associated with the phonological level at which we want to perform the recombination of the input streams (e.g., syllables). 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 models M_j^k (possibly with different topologies). These models are forced to recombine their respective segmental scores at some temporal anchor points. The resulting model is illustrated in Fig. 4. In this model we note that:

- The parallel HMMs, associated with each of the input streams, do not necessarily have the same topology.
- The recombination state ($\otimes$ in Figure 4) is not a regular HMM state since it will be responsible for recombining probabilities (or likelihoods) accumulated over the same temporal segment for all the streams.

The recombination has to be done for all possible segmentation points. The problem appears to be similar to the continuous speech recognition problem where all of the concurrent word segmentations, as well as all of the phone segmentations, must be hypothesised. However, as recombination concerns sub-unit paths that must begin at the same time, and as the best state path is not the same for all of the sub-stream models, (even if the 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, or a synchronous formulation of it, is required.

Alternatively, composite models can be used in the same spirit as HMM decomposition [33]. The HMM decomposition algorithm is a time-synchronous Viterbi search which allows the decomposition of a single stream (speech signal) into two independent components (typically speech and noise), each component being modelled by its own set of HMMs. Composite states are defined for each of the combined model states of the different components. This allows to use a classical Viterbi decoding as far as observation probabilities for the combined states can be computed. This idea was exploited in this work to replace the multi-dimensional search (required for decoding using the model in Figure 4) by a one-dimensional search. Composite sub-unit models are built up from corresponding sub-unit models from each stream. This allows to implement independent search within sub-units as well as inter-units synchrony constraints.

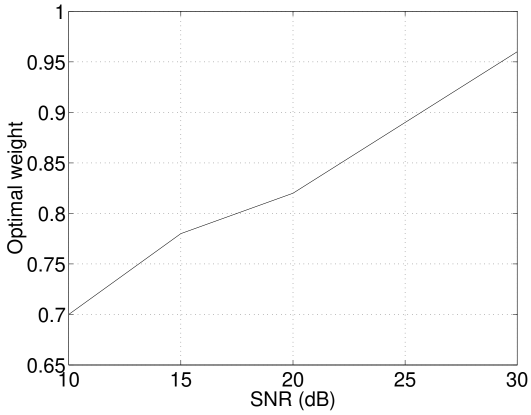

Figure 5: Mapping between the optimal recombination weight w and the acoustic SNR.

As discussed in [3], the training and recognition problems (including automatic segmentation and recombination) can be coined into different statistical formalisms based on likelihoods or posterior probabilities and using linear or nonlinear recombination schemes. During recognition, we will have to find the best sentence model according to (8).

In this work, recombination of the independent likelihoods is done linearly, by multiplying segment likelihoods from the two streams, thus assuming conditional independence of the visual and acoustic streams. This was done according to:

$$P(\mathbf{O}^a, \mathbf{O}^v|\Lambda) = P(\mathbf{O}^a|\Lambda^a)^w P(\mathbf{O}^v|\Lambda^v)^{(1-w)}, \tag{10}$$

The weighting factor w represents the reliability of the modalities which generally depends 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. Another possibility is to estimate the sound-noise ratio (SNR) from the test data and adjust the weighting factor accordingly. Figure 5 displays the mapping between the SNR and the weighting factor found in our experiments. It can be seen that the optimal weight is related almost linearly to the SNR ratio and can easily be estimated from it.

4 Speech Recognition Experiments

The M2VTS audio-visual database [28] was used for all experiments. This database is publicly available and hence allows the comparison of algorithms by other researchers. 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. For each person, the shot with the largest imperfection was labelled as shot 5. This shot differs from the others in face variation (head tilted, unshaved beards), voice variation (poor voice SNR) or shot imperfections (poor focus, different zoom factor). Additional imperfections apart from those of shot 5 are due to some people who were smiling while speaking. The video sequences consist of 286*360 pixel colour 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 colour 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:

- 3 pronunciations from the 37 speakers as training set.
- 1 pronunciation from the 37 speakers as development set. It was used to optimise the weighting coefficients between audio and video streams.
- 1 pronunciation from the 37 speakers as test set.

This procedure allowed to use the whole database as test set (185 utterances) by developing five independent speech recognition systems for each of the compared approaches. These systems could be qualified as multi-speaker (but speaker dependent) continuous digit recognition systems. We note here that the digit sequence to be recognised is always the same (digits from '0' to '9'). This somewhat simplify the task of the speech recognition system which always "see" the pronounced words in the same context.

4.1 Acoustic Speech Recognition

The audio stream was first down sampled to 8 kHz. We used perceptual linear prediction (PLP) parameters [19] computed every 10 ms on 30 ms sample frames. The complete feature vectors consisted of 25 parameters: 12 PLP coefficients, 12 ΔPLP coefficients and the Δenergy.

We used left-right digit HMMs with between 3 and 9 independent states, depending on the digit mean duration. This yielded a total of 52 states. The digit sequences were first segmented into digits using standard Viterbi alignment with a HMM based recogniser trained on the Swiss-French Polyphone database [7] of 5000 speakers. Each M2VTS digit was then linearly segmented according to the number of states of the corresponding HMM model. This segmentation was used to train the HMM states which were represented by a mixture of two multidimensional Gaussian distributions with diagonal covariance matrices, yielding to 5200 parameters.

System training and tests were then performed according to the database partitioning described earlier using the Viterbi algorithm. Results are summarised in Figure 8 for clean speech as well as for speech corrupted by additive white noise with different signal-to-noise ratios. As can be observed, recognition performance is severely affected by additive noise, even at such moderate noise levels.

4.2 Visual Speech Recognition

The most dominant 12 shape features and 12 intensity features, described earlier, were used for the recogniser. These features were complemented by 24 temporal difference parameters (delta parameters). We used the same HMM topologies and the same initial segmentation as for the previously described acoustic-based recognition system. In this case, the HMM-states were represented by a single multidimensional Gaussian distribution with diagonal covariance matrix.

The mean error rate for the five database cuts defined earlier was 44.0%. Since the visual signal only provides partial information, the error rate for the video-based system was considerably lower than for the audio-based system. This is mainly due to the high visual similarity of certain digits like "quatre", "cinq", "six", and "sept", which accounted for about half of the errors. Most of the other errors were deletion errors (i.g. fewer words than the actual number of words were recognised) which are also likely to be due to the high similarity of visually confusable digit models. A more detailed analysis can be found in [24].

4.3 Audio-Visual Speech Recognition

Audio-Visual speech recognition was experimentally investigated and 2 kinds of model topologies were compared. These were based on the HMM word topologies used in the two previous sections. The differences between the models lay in the possible asynchrony of the visual stream with respect to the acoustic stream.

The first model (MODEL 1) did not allow for any asynchrony between the two streams. It corresponds to a Multi-Stream model with recombination at the state level and allows to use fusion criteria that can weight differently the two streams according to their respective reliability.

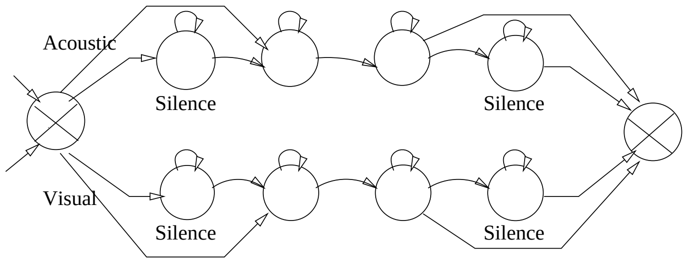

Figure 6: Multi-Stream model for Audio-Visual speech recognition with optional silence states.

The second model (MODEL 2) was a Multi-Stream model with recombination of the streams at the word level. This model thus allows the dynamic programming paths to be independent from the beginning up to the end of the words. This relaxes the assumption of piecewise stationarity by allowing the stationarity of the two streams to occur on different time regions, while still forcing the modalities to resynchronise at word boundaries. This also accounts for the possible asynchrony of the streams inherent to the production mechanism. Indeed, lip movements and changes in the vocal tract shape are only synchronous up to a certain point.

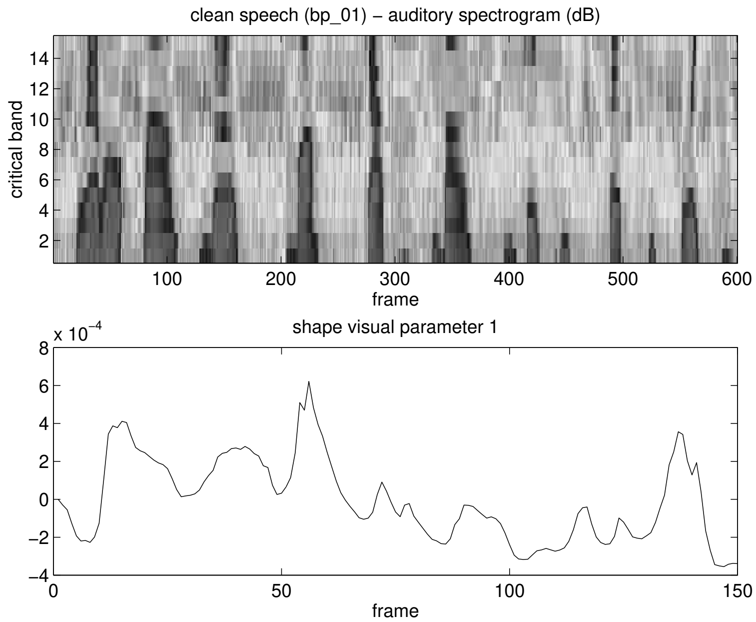

Figure 7: Acoustic spectrogram (evolution of the critical band energies) and evolution of the first visual shape parameter for a portion ('0' to '8') of an M2VTS utterance.

MODEL 2 also allows the transition from silence to speech and from speech to silence to occur at different time instants for the two streams[3]. Indeed, lip movement can occur before and after sound production and conversely. Figure 7 shows in parallel a speech spectrogram as well as the evolution of the first visual shape parameter, mainly representing the changes in the position of the lower

[3] 'Visual silence' could be defined as a portion of the visual signal that doesn't carry any relevant linguistic information.

Table 1: Word error rate of acoustic-, visual- and acoustic-visual-based (Model 2) speech recognition systems on clean speech.

System	Video	Audio	Audio-Visual
Error rate	43.9%	3.4%	2.6%

lip contour [23]. 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. Model 2 is presented in Figure 6.

We used the same parameterisation 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 added at the frame level (by copying frames), so that both signals were synchronously available.

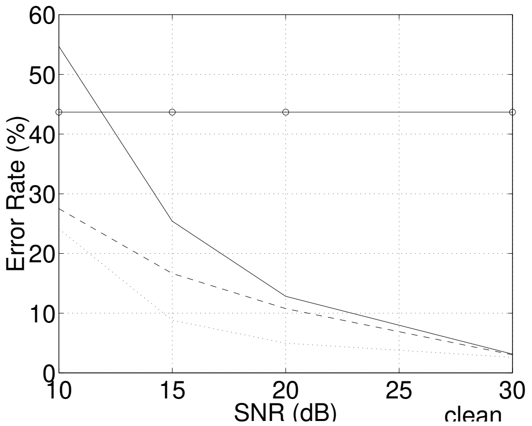

Figure 8: This graph presents the results obtained after embedded training for the visual models, the acoustic models, and the two audio-visual models. All models were trained on clean speech only. The solid line represents the acoustic system, the dashed line Model1 and the dotted line Model2. The horizontal line represents the performance of the visual-only system.

Results are summarised in Figure 8 for different levels of noise degradation. In the case of clean speech, using visual information, in addition to the acoustic signal, does not yield significant performance improvements (see Table 1). The confidence level of the hypothesis test was 0.95. In the case of speech corrupted with additive stationary Gaussian white noise, significant performance improvement can be obtained by using the visual stream as an additional information source. The results also clearly show that we can get a significant performance improvement with Model2 compared to Model1 by allowing the acoustic and visual decoding paths to be asynchronous and by the inclusion of "silence models".

5 Conclusions

We have described an approach based on appearance based models for robust lip tracking and feature extraction. This method allows robust lip tracking in grey-level images 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 multi-speaker continuous speech recognition.

We have presented a framework for the fusion of acoustic and visual information for speech recognition based on the Multi-Stream approach. Several significant advances have been achieved by this approach. Firstly, the method enables synchronous audio-visual decoding of continuous speech and we have presented one of the first continuous audio-visual speech recognition experiments. Secondly, it allows for asynchronous modelling of the two streams, which is inherent in the acoustic and visual speech signal and which has been shown to lead to more accurate modelling and to improved performance. Thirdly, the approach allows to design specific audio-visual word or sub-word topologies, including "silence models", which leads to more accurate audio-visual models.

Acknowledgements

We would like to thank Hervé Bourlard for his support and for many useful discussions.

References

[1] J. B. Allen. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, 1994.

[2] S. Basu, N. Oliver, and A. Pentland. 3D Modeling and Tracking of Human Lip Motion. In *IEEE International Conference on Computer Vision*, 1998.

[3] H. Bourlard, S. Dupont, and C. Riss. Multi-stream speech recognition. Technical Report IDIAP-RR 96-07, IDIAP, 1996.

[4] H. Bourlard, and S. Dupont. Sub-band-based Speech Recognition. In *IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 1251–1254, 1997.

[5] L. Braida. Crossmodal integration in the identification of consonants. *Quarterly Journal of Experimental Psychology*, 43A(3):647–677, 1991.

[6] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *IEEE International Conference on Computer Vision*, pages 494–499. IEEE, Piscataway, NJ, USA, 1995.

[7] G. Chollet, J. L. Cochard, A. Constantinescu, and P. Langlais. Swiss French Polyphone and Polyvar : Telephone speech databases to study intra and inter speaker variability. Technical report, IDIAP, Martigny, 1995.

[8] T. Coianiz, L. Torresani, and B. Capril. 2D deformable models for visual speech analysis. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 391–398. Springer Verlag, Berlin, 1996.

[9] R. Cole, L. Hirschmann, L Atlas, and et al. The challenge of spoken language processing: research directions for the nineties. *IEEE Trans. on Speech and Audio Processing*, 3(1):1–20, 1995.

[10] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. Use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12:355–365, Jul-Aug 1994.

[11] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, Jan 1995.

[12] N. P. Erber and C. L. De Filippo. Voice-mouth synthesis of tactual/visual perception of /pa, ba, ma/. *Journal of the Acoustical Society of America*, 64:1015–1019, 1978.

[13] I. A. Essa and A. P. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proc. 5th Int. Conf. on Computer Vision*, pages 360–367. IEEE Computer Society Press, July 1995.

[14] H. Fletcher. *Speech and Hearing in Communication.* Krieger, New York, 1953.

[15] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16:261–291, 1995.

[16] M. S. Gray, J. R. Movellan, and T. J. Sejnowski. Dynamic features for visual speechreading: A systematic comparison. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, Cambridge, MA, 1997.

[17] K. P. Green and J. L. Miller. On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, 38(3):269–276, 1985.

[18] W. J. Hardcastle. *Physiology of Speech Production.* Academic Press, New York, NY, 1976.

[19] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[20] B. K. P. Horn. *Robot Vision.* McGraw-Hill, New York, 1986.

[21] M. I. Jordan and Z. Ghahramani and L. K. Saul". Hidden Markov Decision Trees. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, Cambridge, MA, 1997.

[22] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[23] J. Luettin and N. A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, February 1997.

[24] J. Luettin. *Visual Speech and Speaker Recognition.* PhD thesis, University of Sheffield, 1997.

[25] K. Mase and A. Pentland. Automatic lipreading by optical flow analysis. *Systems and Computers in Japan*, 22(6), 1991.

[26] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *IEEE International Conference on Computer Vision*, pages 786–793. IEEE, Piscataway, NJ, USA, 1995.

[27] E. D. Petajan. Automatic lipreading to enhance speech recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 40–47, 1985.

[28] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database. In *Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication*, Lecture Notes in Computer Science. Springer Verlag, 1997.

[29] M. U. Ramos Sanchez, J. Matas, and J. Kittler. Statistical chromaticity models for lip tracking with B-splines. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 69–76. Springer Verlag, 1997.

[30] P. L. Silsbee and A. C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 1996.

[31] A. Q. Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London, Series B*, 335:71–78, 1992.

[32] M. J. Tomlinson, M. J. Russell, and N. M. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, volume 2, pages 821–824, 1996.

[33] A. Varga and R. Moore. Hidden markov model decomposition of speech and noise. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 845–848, 1990.

[34] 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*, 78(10):1658–1668, October 1990.