Journal on Multimodal User Interfaces manuscript No.
(will be inserted by the editor)

Alin G. Chiţu · Leon J.M. Rothkrantz · Jacek C. Wojdel · Pascal Wiggers

Comparison Between Different Feature Extraction Techniques for Audio-Visual Speech Recognition

Received: date / Accepted: date

Abstract Having a robust speech recognition system that can be relied upon in different environments is a strong requirement for modern systems. However audio-only speech recognition still lacks robustness when the signal to noise ratio decreases. This is especially true when the system is deployed in public spaces or is used for crises situations management where the background noise is expected to be extremely large. The video information is not affected by noise which makes it an ideal candidate for data fusion. The acoustic features have been well defined during the course of the years, the most used features being mel-frequency cepstral coefficiens (MFCCs) or linear predictive coefficients(LPCs). On the visual side, however, there is still much place for improvements. It is still not clear which visual features retain the most speech related information. Until now the visual features used were static features which describe the face of the user at one instance in time only. In the paper [1] the authors have shown that most of the techniques used for extraction of static visual features result in equivalent features or at least the most informative features exhibit this. This means that all techniques describe the same aspect of the visual stream. However the improvement of recognition even though looks promising is still modest. We argue that the main problem of existing methods is that the resulting features contain no information about the motion of the speaker's lips. We present in this paper a new method for extracting useful features from the point of view of speech recognition based on optical flow analysis. The video features extracted using this method are preserving the information about speaker mouth motion. We tested the method on an audio-video database for Dutch language. The Audio-Visual Speech Recognizer(AVSR) used is based on HMMs method and was trained for large vocabulary continuous speech. For completion we also present the method introduced in the paper [2] for extracting static visual features. We will compare these two methods with respect to the induced recognition performance. Another way to recover motion information from static features is to use their first and/or the second derivative as visual features. However this can not always guaranty that the resulted features are physically sound quantities. We will also present for comparison the recognition results based on such features. The evaluation of these methods will be done under different noise conditions. We show that the audio-video recognition based on the true motion features outperforms the other settings in low Signal to Noise Ratio(SNR) conditions.

Keywords audio-visual fusion · speech recognition · automatic lipreading · optical flow · mouth movement

1 Introduction

Over the years much work has been done in the domain of automatic speech recognition. The progress made is significant for small, medium and even large vocabulary systems. However the good results are only valid if the recordings for the database used for testing are performed in similar conditions to the ones present when the training database was created. The level of accuracy of the current Automatic Speech Recognition(ASR) suffers greatly when the background noise increases. Therefore

All authors correspondence:
Man-Machine Interaction Group,
Delft University of Technology
Mekelweg 4, 2628CD Delft,
The Netherlands
Fax: +031-15-2787141

Alin G. Chiţu
Tel.:+031-15-2788543
E-mail: A.G.Chitu@ewi.tudelft.nl

Leon J.M. Rothkrantz
Tel.:+031-15-2787504
E-mail: L.J.M.Rothkrantz@ewi.tudelft.nl

Jacek C. Wojdel
E-mail: J.C.Wojdel@tnw.tudelft.nl

Pascal Wiggers
Tel.:+31-15-2782523
E-mail: P.Wiggers@tudelft.nl

researchers start thinking of using additional reliable information that can account for changes in the quality of acoustic signal, such as visual cues from the face of the speaker.

Stating that combining information from more sources should improve the performance of a system that depends on that kind of knowledge, seems such a common sense statement. However, the case of human speech perception was considered as a purely auditory process. It was the publication in *"Nature"* in 1976 of the paper *"Hearing lips and seeing voices"*, by Harry McGurk and John MacDonald [3] that changed this belief. In this paper the authors present the results of their experiments that proved for the first time that actually human beings use both visual and auditory information processing speech, regardless of the acoustic environment. In the experiments the subjects were presented a film of a person's talking head, in which repeated utterances of the syllable [ba] were dubbed on to lip movements for the syllable [ga]. The results of the experiment showed that in fact the subjects reported hearing [da]. In the reverse experiment a majority reported hearing [bagba] or [gaba]. However, when the subjects listened to the soundtrack from the film without visual input they reported the syllables accurately as repetitions of [ba] or [ga]. This is called McGurk effect. Later studies confirm the reliability of these findings. This discovery proved that the above statement is not at all absurd and encouraged the research in lip reading and audio-visual speech recognition. Hence not only that the visual modality is unaffected by the noise in the acoustic stream, but also if we look at the way humans perceive speech the visual modality is valuable even in noise free situations.

However the research in lip reading domain started only recently, mainly because of reduced computer power. The results of a lip reading system are not influenced by the presence of noise in the acoustic signal. That makes the visual stream extremely useful for speech recognition in noisy conditions.

Since the research in ASR has already a long history, the audio features that best describe the speech gained somewhat maturity. Hence the most popular set of features used to parameterize the speech are the Linear Predictive Coefficients (LPC) and the Mel-Frequency Cepstrum Coefficients (MFCC). This is not the case for visual features, where the search for the best features that capture the most information about speech is still going on.

There were many methods developed to extract visual features. They fit mainly in two broad classes: appearance based methods and geometrical methods; a combination of the two was also used. The methods from the first class consider the raw image or a transformation of it as feature processing [4,5]. The transformation of the image is employed in order to obtain some data reduction. The most popular method for this is Principal Component Analysis (PCA) [6,7]. Other methods which were used as an alternative to PCA are based on discrete cosine transform [8] and discrete wavelet transform. However this approach gives rise to very high dimensionality of the feature vectors. These methods are not trying in anyway to uncover the features that bring the most information about speech, instead they are blindly processing the image. The features are chosen based on the quantity of information they carry in general. Therefore it might happen that the features extracted, even though they carry a large amount of information, are more useful for other types of applications like speaker identification. On the other side the methods from the second class start from the idea that it is better to try to model the speech production apparatus. However not all parts of the speech production system are visible, hence these methods try to model the visual parts of the speech apparatus such as lips, teeth, tongue but also other parts of the face. The algorithms are aiming to detection and tracking of specific points on the face. Usually the detection process is assisted by 2D or 3D geometrical models of the face [9]. One other approach is to use statistical methods based on image filtering algorithms and try to directly extract the shape of the lips. The dimensionality reduction obtained through the latter approaches is very large. Moreover the reduction of dimensionality was done in a direction that is more appropriate for speech recognition. One other classification of the methods used for visual feature extraction can be made based on how much information about the actual motion of the parts of the face is retained by the feature vectors. To understand the implications of the above statement we can think of the following: We see the picture of a person having the mouth half opened. There is no way of telling if the person was closing its mouth or opening it. All the above presented methods consider only one frame as input for the feature extraction module, hence there is no information about the actual motion of the mouth in the resulted feature vector. One way to include motion information starting from static features is to include the first and/or the second derivatives of the features. However there is not always guaranteed that the resulted features have valid physical meaning. Another approach is to perform optical flow analysis on the input video stream. However until now the optical flow was mainly employed as a measure of the overall movement on the face and used for onset/offset detection [10,11,12,13,14,15,16,17,18].

In this paper we will analyze the importance of motion detection for speech recognition. For this we will first present the Lip Geometry Estimation(LGE) method for static feature extraction. This method combines appearance approach with a statistical approach for extracting the shape of the mouth. This method was introduced in [19] explored in detail in [20]. The results, expressed in terms of recognition accuracy, even though they show a promising increase in the case of low signal to noise ratio, are still not spectacular. We will introduce then a new method for visual feature modeling that is based on optical flow analysis. This method is extracting information about the movement on the contour of the mouth. We will show that in this case the improvements are much greater. To analyze the importance of the motion information for speech recognition we will compare the results in the following three settings:

- the visual feature used are obtained using the LGE method, hence no expressed motion information is present.
- the motion information is recovered from the static features by including the first and/or the second order derivatives of the static features.
- the visual features are obtained based on *optical flow analysis*. The motion information is thus present in the features obtained.

2 Related Theory

2.1 Optical flow analysis

The Optical Flow is a concept that is concerned with the notion of motion of objects within a visual representation. A common definition of the Optical Flow is:
"The velocity field which warps one video frame in a subsequent one."

In [21] the optical flow is defined as:
"The distribution of apparent velocities of movement of brightness patterns in an image."

In the latter definition the word "apparent"signals the fact that sometimes the optical flow does not correspond to the true motion field. The most known example is the "rotating barber's pole illusion." The problem of finding the optical flow in an image falls in a broader class of problems called "image registration problem." Data registration in general deals with spatial and temporal alignment of objects within imagery or spatial data sets. Image registration can occur at pixel level (i.e. any pixel in an image can be matched with known accuracy with a pixel or pixels in another image) or at object level (i.e. it relates objects rather than pixels). The domain where the image registration problem is one of the key challenges is medical imaging. In medical imaging the problem of registration arises whenever images acquired from different subjects, at different times, or from different scanners need to be combined for analysis or visualization. In [22] the problem of finding the optical flow seen as an image registration problem is defined as follows:
"We consider that the pixels values in the two images are given by the functions $F(X)$ and $G(X)$ (in 2D $X = (x, y)$). Our goal is to determine the dissimilarity vector h which minimizes some measure of the difference between $F(X + h)$ and $G(X)$, for X in some region of interest $\Re$."

There are quite a few methods for optical flow detection from which we mention Lukas-Kanade method, Horn-Schunck method, phase correlation (i.e. the inverse of normalized cross-power spectrum), gradient constraint-based methods, block correlation methods, etc. However, Lucas-Kanade and Horn-Schunck algorithms are the two most used algorithms for determining the optical flow. The first algorithm was published in [22] by Bruce D. Lucas and Takeo Kanade. This algorithm assumes that the images are roughly aligned and that the optical flow is constant in a small neighborhood. Then it uses a type of Newton-Raphson iteration taking the gradient of the error and assuming that the analyzed function is almost linear and it moves in the direction of this gradient. The second algorithm assumes that the apparent velocity of the brightness pattern varies smoothly almost everywhere in the image. The algorithm minimizes the square of the magnitude of the gradient of the optical flow velocity and the measure of non-smoothness of the optical flow. It was published in [21] by Berthold K.P. Horn and Brian G. Schunck. In [23] the authors explore the possibility of combining the two approaches used in Lucas-Kanade and Horn-Schunck methods, namely local constrains methods and global constraints methods, in order to build a hybrid method that can provide the corroborated strengths of both paradigms.

Other known algorithms are developed by Uras et al. [24], Nagel [25], Anandan [26], Singh[27], Heeger [28], Waxman et al. [29] and Fleet and Jepson [30]. In [31] and [32] a number of nine, respectively eight, different techniques for detection of optical flow were investigated. The performance of these methods was compared on synthetic scenes. The difficulty of comparing different optical flow techniques comes from the fact that is hard to produce ground-truth motion fields. In [32] this problem was overcome with a modified ray tracer that allowed the generation of ground-truth flow maps. The latter study reports that a modified version of Lucas-Kanade algorithm produced the best result, however not very dens flow maps. On the second place was placed Proesmans algorithm which produced very dense flow maps but with less quality.

Consider the function $I(x, y, t)$ which gives the image intensity at location (x, y) at time t. Every optical flow detection method has as goal to compute the motion of every pixel in the image from time t to time $t + \delta t$. If we denote the new position of the pixel (x, y) from time t with $(x + \delta x, y + \delta y)$ at time $t + \delta t$ we get the following constraint equation:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{1}$$

Assuming that the movement inside the image is small enough, the image constraint equation can be rewritten in terms of Taylor series as follows:

$$\begin{aligned} &I(x,y,t) = I(x+\delta x, y+\delta y, t+\delta t) \\ &= I(x,y,t) + \tfrac{\partial I}{\partial x}\delta x + \tfrac{\partial I}{\partial y}\delta y + \tfrac{\partial I}{\partial t}\delta t + \Re \end{aligned} \quad (2)$$

where $\Re$ means higher order terms, which are small enough to be ignored. Using the initial image constraint and ignoring $\Re$ we get the following equation:

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = 0 \quad (3)$$

which can be rewritten as:

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y = -\frac{\partial I}{\partial t} \quad (4)$$

where V_x and V_y are the components of the optical flow. Denoting the partial derivatives of $I(x,y,t)$ with respect to spatial coordinates x and y and time t with I_x, I_y and I_t respectively the new constraint equation reads:

$$I_x V_x + I_y V_y = -I_t \quad (5)$$

Hence the problem of detecting optical flow is equivalent to solving the system (5). However, this system has only one equation but two unknowns making it underdetermined. To be able to solve this system some assumptions need to be taken. Based on these assumptions new equations can be introduced. The resulted flow obtained will carry the marks of these assumptions.

We used for the current research for analyzing the optical flow the algorithms developed by Lucas and Kanade. Lucas and Kanade's algorithm starts with the assumption that optical flow is constant in a small neighborhood of the point (x,y). Assuming that the flow (V_x, V_y) is constant in a small rectangular region of size (n,n) with $n > 1$ (usually $n = 5$ gives sufficient good results), that is centered at point (x,y) and numbering the pixels, we get the following system:

$$\begin{cases} I_{x_1}V_x + I_{y_1}V_y = I_{t_1} \\ \cdots\cdots \\ I_{x_n}V_x + I_{y_n}V_y = I_{t_n} \end{cases} \quad (6)$$

which is an over-determined system. Written in matricial form it reads:

$$\begin{bmatrix} I_{x_1} & I_{y_1} \\ \vdots & \vdots \\ I_{x_n} & I_{y_n} \end{bmatrix} \begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} -I_{t_1} \\ \vdots \\ -I_{t_n} \end{bmatrix} \quad (7)$$

A weighted least squares fit solution of the above system is:

$$V = [A^T W A]^{-1}(-A^T W b) \quad (8)$$

where $A = \begin{bmatrix} I_{x_1} & I_{y_1} \\ \vdots & \vdots \\ I_{x_n} & I_{y_n} \end{bmatrix}$, $V = \begin{bmatrix} V_x \\ V_y \end{bmatrix}$ $b = \begin{bmatrix} I_{t_1} \\ \vdots \\ I_{t_n} \end{bmatrix}$ and W is a weighting function that gives more importance to the center pixel of the window. This means that the optical flow vector can be found only by calculating the derivatives of the image in all dimensions. The Lucas-Kanade optical flow detection algorithm is very fast because it examines only a limited number of possible matches. However its main advantage is the robustness in the presence of noise. One disadvantage of this method is that it does not yield a high density of flow vectors, (i.e. the velocity is only determined close to the boundaries of objects and inside large areas with almost constant brightness the information fades quickly.

2.2 Data Fusion Architecture

After deciding what features should be extracted for each data stream, we came to the question: How should we combine the information from the two modalities, such that to obtain good recognition. In order to tackle this problem researchers look at the methods that were already used for audio-only speech recognition. Over the years several approaches to speech recognition have been explored, including dynamic time warping [33], neural networks [34,35] and support vector machines [36], but by far the most successful and dominant approach is the probabilistic approach based on hidden Markov models. Here the recognition task is phrased as finding the most likely word sequence W given the sequence of audio feature vectors $\mathbf{O}_\mathrm{a}$:

$$\hat{W} = \max_W P(W|\mathbf{O}_\mathrm{a}) \quad (9)$$

Using Bayes' rule and the fact that the observation sequence is constant for a given utterance this can be rewritten as:

$$\hat{W} = \max_W P(\mathbf{O}_\mathrm{a}|W)P(W) \quad (10)$$

the term $P(\mathbf{O}_\mathrm{a}|W)$ is called the acoustic model and is realized by a collection of hidden Markov models, usually one per phoneme when large dictionary speech recognizer is targeted. The a priori probability that the string W occurs in the language $P(W)$ is called the language model and is in general realized using n-gram models.

However we are now having a bi-modal system that takes as input data from two modalities. When combining the two data streams, or for that matter when combining multiple modalities in general, there are a number of issues that a comprehensive model should address:

1. The signals may have *different dynamic ranges*. For example the duration of the sound of a phoneme is usually shorter than the duration of the corresponding lip movement.
2. There may be a *time offset* between the signals. In lip reading the video signal usually starts before the audio signal. This offset is not constant over the utterance and may be as large as 120 ms, almost the duration of a phoneme. The two signals thus evolve asynchronously, within certain limits.
3. There may be a *different number of distinguishable classes* that require different model topologies. For speech phonemes constitute a good set of classes, but for lip reading many of the distinctions between phonemes are not visible. There are for instance 44 different phonemes in English while the number of visemes is only around 12 (e.g for example /t/ and /d/ or /p/, /b/ and /m/ only differ with respect to voicing).
4. The signals may be *sampled at different rates*. In particular, video sample rates are usually slower than audio sample rates (i.d. usually the video frame rate is 25-30fps while the audio signal is sampled at a rate larger than 100fps).
5. The modalities may *not be equally reliable*. Audio contains more information about speech than the visual modality does in clear environment. As a consequence the decisions based upon audio information are generally more reliable than those based upon video information. In some cases the reliability may change dynamically over time. However when the background noise increases (e.g. if people start talking in the background or the system is deployed outside the office) audio recognition becomes less reliable.

One other question that arises when two or more modalities are present is at what processing stage should the modalities be fused. Models on human speech perception [37,38] like the Fuzzy Logical Model of Perception (FLMP) developed by D. Massaro [39] suggest a late integration approach. However, evidence here is not conclusive. Therefore, for automatic audio-visual recognition, models for integration at several levels have been developed.

In the late integration approach the recognition is performed on both modalities separately and in the end the partial results from each subprocess are combined in a final result [40]. As a consequence this approach can easily handle different classes in different channels. Complications introduced by time offsets and limited asynchrony can be dealt with by integrating at the utterance level. On the other hand it does not model any interaction among processes, therefore such information is completely lost in the recognition phase; the two models may base their conclusions about the occurrence of a speech unit at a given time on completely different paths. In fact it is not at all guaranteed that both models will deliver the same hypothesis. As a consequence N-best recognition has to be used for both models, which leads to a considerable increase in processing time. Another major drawback for continuous recognition is that fusion can only take place after a complete utterance has been recognized.

For our current research we used the other end approach, namely early integration. The *feature fusion* approach takes the features extracted for each modality and combine them in a common vector that will be fed to the recognition system. One advantage of this type of fusion is that we can immediately reuse the techniques developed for audio-only speech recognition. For this the different modalities need to be firstly synchronized and if there is a difference in the frame rate then some type of interpolation should be used in order to complete the sparser stream. For audio-visual case the audio features are usually sampled at every 10ms, while a standard video camera will limit the frame rate of video features to 25 or 30Hz. A solution can be to copy a video vector for several frames, arguing that the shape of the mouth will not change much faster than the sampling rate, but this may introduce discontinuities at the boundaries between such vector intervals. The method of choice therefore is linear interpolation between two subsequent video vectors for up-sampling the video signal as depicted in Figure 1. However, the combined vectors become rather large and the interpolation process induces correlations. Therefore, the training of the recognition models can become a burden for the resulted system. To cope with this problem typically a projection of the combined vector to a smaller space is introduced, for example using principal component analysis, possibly followed by an MLLT data rotation to de-correlate the features.

An additional method, feasible for early fusion approach, to cope with the problem of insufficient data for training a multi-modal system is to use a system that was already been well trained for a single modality as an initial step. In our system we used the distributions of our already well-trained speech recognizer to initialize the audio parameters of the bi-modal system thereby reducing the amount of training necessary and ensuring quicker convergence of the video distributions.

The half way solution to the multi-stream fusion problem is to think of a model that can simultaneously cope with both data streams and directly give a unified result. This model should tackle the issues such as the

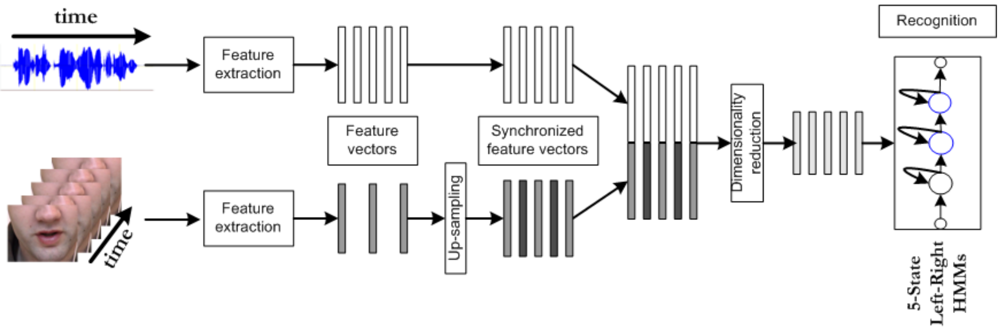

Fig. 1 Feature fusion for audio-visual speech recognition.

different dynamic ranges and different frame rate of the two modalities. In short it should alleviate all problems that appear at the feature level fusion. This is generally called *model fusion*. However, this solution leads most of the time to an explosion in the number of model parameters and hence to slower recognition, overfitting and poor generalization of the models. Based again on the HMM approach there were developed quite a big number of generalizations which can be regarded as model fusion approaches. Probably the first solution that comes to mind to integrate two or more data streams that can be modeled using HMMs is to build one large model having the cartesian product of all the states of the separate models as its states. This is called *Cartesian Product HMM*. Each possible combination of paths through both uni-modal models can be represented in a single path through the product HMM. It hence allows for asynchrony and can deal with different submodel topologies. However, this model is so computationally complex that is only interesting from theoretical point of view. Other models were developed by placing constraints on the states or the transitions in order to make the new models tractable. The *Multi-Stream HMM* [41,42] allows for multiple input feature streams that may have different frame rates and can be asynchronous. It assumes that the model consists of a number of sub-unit models that correspond to the level at which the streams have to synchronize, for example phoneme level or syllable level. Each sub-unit models has parallel HMMs, each of which processes a single stream independently of the other models. The *Factorial HMM* arises by forming a dynamic Bayesian network composed of several independent layers. It was introduced by Ghahramani and Jordan [43] to model time series that can be seen as loosely coupled random processes. Together with the model they also provided several algorithms to efficiently learn the parameters of the model. In case of audio-visual speech recognition the model will have two separate layers, one for each input stream. It thus allows for asynchrony and different numbers of distinguishable classes in both streams. As the streams are combined at the output level in every time step it does however require equal frame rates. Finally the last model created for speech recognition as a generalization of HMM framework is Coupled Hidden Markov Model. Speech recognition and lip reading can be seen as two dependent processes, each having their own dynamics and observations but influencing each other.

3 Features extraction algorithms

3.1 Lip geometry estimation

In this subsection we will describe step by step a feature extraction method called Lip Geometry Estimation (LGE). Using image filtering techniques and based on a statistical interpretation of the results from the filters it directly estimates the geometry of the mouth. However, this technique is unique because it does not rely on any a-priori geometrical lip model. The overview of the signal processing described in this section is depicted in Figure 2.

As the first step of the processing pipeline we have to locate the face and then the mouth of the speaker. The detection of the Region of Interest(ROI) removes unnecessary areas from the image which is very important from at least two reasons: firstly the processing time is greatly reduced and secondly many possible unwanted artifacts can be avoided. For this we use the Viola-Jones algorithm for object detection [44]. This classifier uses a new method for detecting the most representative Haar like features using a learning algorithm based on AdaBoost. It combines the weak classifiers using a "cascade"approach which corroborated with a fast method for computing the Haar-like features allows for high speed and very low false-negative rates. In order to increase the reliability of the ROI extraction process in the following frames we use a combination detection/tracking processing model. Hence in a first step the ROI is detected using a mouth detector and then in the next frames we use a tracking algorithm which is trained using the last extracted ROI. The object tracking algorithm uses a Gaussian Mixture Model to model the color distribution of the object and of the background. Then it uses a deformable template to optimally fit the tracked object.

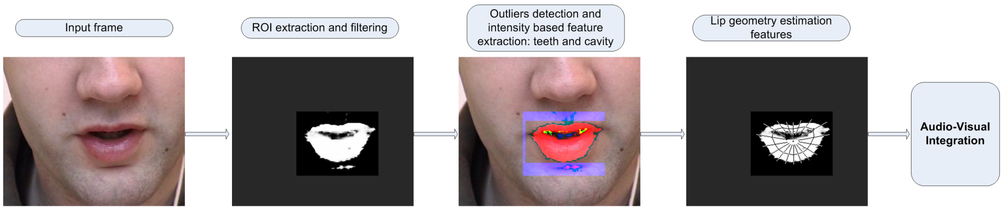
Input frame
ROI extraction and filtering
Outliers detection and intensity based feature extraction: teeth and cavity
Lip geometry estimation features
Audio-Visual Integration

Fig. 2 Signal processing pipeline for the audio-visual speech recognition system based on LGE feature extraction technique.

The next step in the process is to somehow detect which pixels belong to the lips. Fortunately, now, because the input image contains only the mouth area and since the lips have a distinct coloring we can extract the lip's pixels without the need for complicated object recognition techniques. In order to utilize this fact, we need to apply some sort of *lip-selective* filter to the image. In our current research we use several different filters depending on the illumination conditions and the quality of the recorded video sequences.

The simplest way for doing this segmentation is to use a thresholding technique based on the appropriate color channel. For color input images the Hue channel is most used. In the case of grayscale images, when the Hue value is not defined the thresholding can be performed directly on the gray channel. As we present further this method requires that the result for each pixel be in terms of degrees of "belongness", so binary segmentation is not enough. Therefore we use for our current research parabolic thresholding. This method for image segmentation was first proposed in [45]. The results of the filter are given according with the following parabolic shaped function:

$$F(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2}, & |h-h_0| \leq w \\ 0, & |h-h_0| > w \end{cases} \tag{11}$$

Attention should be paid to the dynamic range of the variable h, for instance in the case of Hue-based filtering since Hue is a circular variable. The filter is defined by the center of the interval h_0 and by its half width w. Both values should be calibrated in advance in order to obtain sufficient accuracy. We may also combine a series of such parabolic shaped filters for more robust lip detection. Using the product of the Hue-based filter and a Value-based filter can for example remove some of the noise in the dark or bright areas of the image where the hue values behave rather randomly.

A parabolic shaped filter is very simple and computationally very effective. Unfortunately, during our experiments we found that in many cases, if the illumination of the face is not perfect, the hue component itself is not sufficient for proper lip selection. Instead of putting additional constraints on the filtered color (such as a threshold value of saturation or value component), we decided to use a black-box approach. We trained a simple feed-forward neural network and used it as a filter. The network that was used had only 5 neurons in a hidden layer and one output neuron. It was fed with the RGB values of the processed pixel of the image. This filter achieved extremely accurate results. The second image in Figure 2 shows an example. Some artifacts are still visible in the result.

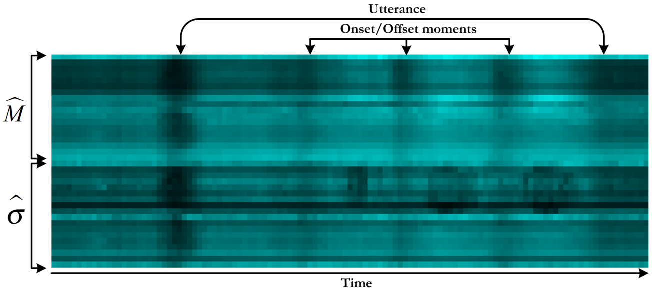

Fig. 3 Pairs of $\widehat{M(\alpha)}$ and $\widehat{\sigma(\alpha)}^2$ vectors extracted from a video sequence. The periods of silence around two spoken sequences can be seen clearly.

The filtered image is then treated as a bivariate distribution $I(X,Y)$ (after normalization). The mean of this distribution: $[EX, EY]$ accurately approximates the center of the mouth. Using this value, we transform the image into polar coordinates:

$J(a,r){=}I(EX{+}r\cos(a), EY{+}r\sin(a))$

We then compute the means and variances for the conditional distributions, conditioned on the direction. We therefore define the following functions of mean and variance values for any angle:

$$M(\alpha) = \frac{\int_r J(a,r)rdr}{\int_r J(a,r)dr} \tag{12}$$

$$\sigma^2(\alpha) = \frac{\int_r J(a,r)(r-M(\alpha))^2)dr}{\int_r J(a,r)dr} \tag{13}$$

As the image is discrete rather that continuous, all of the values are obtained from summation rather than integration, so we only operate on estimations of those values, namely $\widehat{M(\alpha)}$ and $\widehat{\sigma(\alpha)}^2$. The vectors resulting from sampling of those functions for one of the video sequences can be seen in Figure 3.

As can be seen in Figure 2 the value of $\widehat{M(\alpha)}$ for a specific angle relates directly to the distance from the center of the mouth to the center of the lip in that given

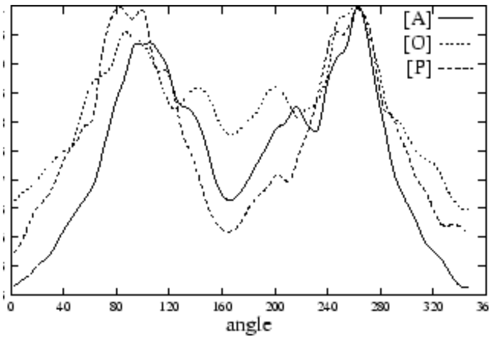

Fig. 4 Average $\widehat{M(\alpha)}$ values for specific visemes in a video sequence.

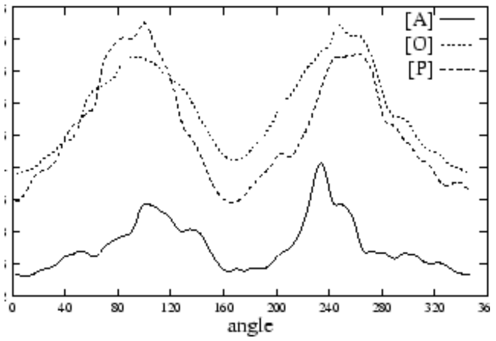

Fig. 5 Average $\widehat{\sigma(\alpha)}^2$ values for specific visemes in a video sequence.

direction. Therefore a vector constructed from those values for a number of angles describes the shape of the mouth on a given image. Such shape profiles were gathered from a video sequence of spoken Dutch and grouped based on the visemes. Figure 4 shows three profiles for visemes [A], [o u O y Y 2: 9:] and [p b m]. In order to obtain scale independence, the vectors were scaled so that all their values fit in the interval. It can be seen that the [ou] viseme can be distinguished from the others by its relatively high value in the middle. There is however no feasible way of distinguishing between [A] and [p b m] visemes. The lip shapes that correspond to those visemes are completely different (mouth closed versus mouth opened), but scaling of feature vectors removes all of those differences.

Therefore the values of $\widehat{M(\alpha)}$ alone are not sufficient for extracting useful data from the video sequence. However using the additional information about the variance of conditional distributions, as can be seen in Figure 5 we can clearly discriminate the viseme [A] from [P]. The values of $\widehat{\sigma(\alpha)}^2$ are also scaled in order to obtain size independence, but the scaling factor is directly determined by the mean values rather than the variances. So obviously using the two sets of values we get an accurate shape of the mouth. This can also be seen from Figure 2, the 95% confidence interval clearly describes how thick is the lip in that specific direction. The lips of a wide-stretched mouth, appear thinner than those of a closed mouth when related to the overall size of the mouth.

As can be seen in the second image in Figure 2 even after reducing the area of interest and even with optimal filtering of the mouth in some cases the filtered image still contains unwanted artifacts. In order to reduce the impact of such occurrences a process of outliers deletion can be used before the actual feature extraction. The blue area showed in Figure 2 superimposed on the input image was cleaned by the outlier deletion process.

The last stage of feature vector extraction is choosing which direction to be used. Obviously, the chosen vector dimension is a compromise between accuracy and processing efficiency. The longer the vectors, the more information on the original distribution they contain but the longer it takes to extract and process them. Also higher dimensionality generally makes it more difficult to train the recognition modules. After some experiments with the data we chose to use the 18-dimensional vectors for both features(see [2]).

3.2 Mouth motion estimation based on optical flow analysis

Until now in the domain of lipreading and audio-visual speech recognition the optical flow analysis was used as raw data[9,13,16], or as a method to measure the global movement[10,11,12,14,15,17,18] of the speaker face. The variances of the horizontal and vertical components of flow vectors were used as visual features for silence detection, in the cases when the noise in the audio modality was not allowing for an accurate decision. Even though the results were very promising we argue that using the optical flow only as a global measure much of the information about speech is discarded. We propose here a method that based on the optical flow better describes the actual speech. Our method measures the lip movement on the contour of the mouth. The first step of the method tries to accurately detect the center of the speaker mouth. Since the LGE method provides a good way for detecting the center of the mouth we used that approach again for the present method. Also the detection of the appropriate region of interest is highly appreciated. Since the optical flow vectors can be computed in every region of the image we need to restrict the searching space in order to exclude unnecessary regions. The same approach in two steps detection/tracking was employed here. The Viola-Jones classifier and the Gaussian Mixture based tracking algorithm were used.

The optical flow is computed inside the area of interest. We use the algorithm developed by Lucas and Kanade, since this algorithm has the best accuracy. We are only interested in the mouth area, therefore we do not need a very dense vector flow. The processing time is another aspect that was considered when taking the decision to use Lucas-Kanade. We will be interested to

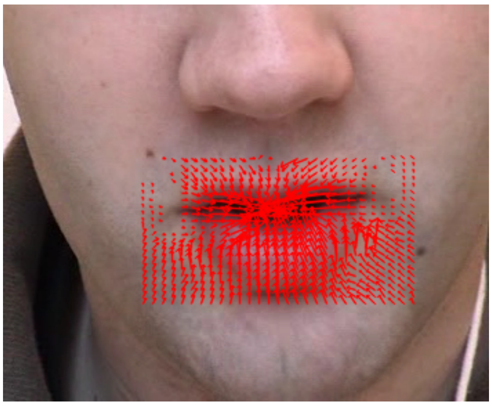

Fig. 6 Optical flow sample result for lipreading.

have a real time recognizer. A sample of the computed optical flow is shown in Figure 6.

Hence considering the filtered image as the mass function of a bi-variate distribution we can compute its mean vector $[EX, EY]$. The point represented by the mean vector can be safely considered as the center of the mouth.

Having identified the center of the mouth and computed the optical flow in the area of interest, we can then start to extract the visual features. The selected features should describe the movement of the mouth along its contour. Hence we divide the $2D$ space originating in the center of the mouth into 18 equally wide directions. The number of directions was taken such that to be equivalent with the number of features extracted by the LGE method. The features extracted should capture as much information as possible, while keeping the dimensionality in manageable limits.

The visual features are obtained by computing statistics about the optical flow, hence the movement of the mouth, in all chosen directions respectively. Even though the global variance of the optical flow can be a valuable feature for onset/offset discrimination, the information about movement in certain parts of the mouth is canceled out by averaging, hence is not suitable for our settings. For a small enough angle the contour of the lip should deform in the same way on the entire distance considered. Thus, the variance of the flow computed only in a certain direction should always be close to zero.

We computed therefore as visual feature only the mean displacement on the horizontal and vertical respectively. Figure 7 shows the features extracted based on the optical flow seen in Figure 6.

Figure 8 a) shows the horizontal mean of the optical flow in all 18 directions for a certain utterance. The utterance spreads over 119 video frames. While it is clear from this picture that these visual features carry significant amount of information we can not tell anything about what triggers these values. This can be somehow clarified by looking to the cumulative sum of the features in time, which is shown in Figure 8 b).

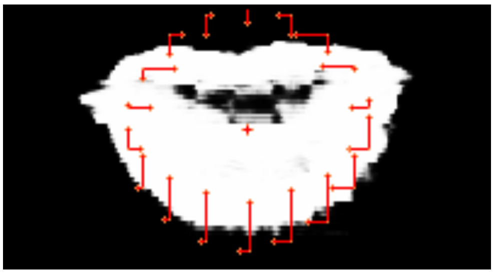

Fig. 7 Optical flow based extracted features.

The same analysis for the same utterance but this time for the vertical mean of the optical flow is shown in Figure 9. Above all the remarks made for the horizontal case, which also stands here, we have to notice that the amount of movement is much grater in the vertical direction. Hence we can conclude that the vertical movement should carry more information about what is being said than the horizontal movement. The cumulative sum in this case makes a lot more sense for a human observer than before. We can track most of the onset/offset moments by just looking at this image (for instance around frames 75, 90 and 105).

3.3 Intensity based features

The shape of the lips is not the only determinant of a spoken utterance. There are some other important factors such as the position of the tongue, teeth etc. Some of them can be observed in the video sequence, the others not. It is essential in the case of lipreading to extract from the visual channel as much information as possible about the utterance being spoken. We propose therefore to augment the visual features extracted until now with a few simple intensity related features. It would probably be possible to track the relative positions of the teeth and tongue with respect to the lips. The tracking accuracy would be limited by the fact that the visibility of lips and tongue is normally very poor. Such a task would also be too complex and therefore infeasible for a lipreading application. There are however some easily traceable features that can be measured in the image which relate to the positions and movements of the crucial parts of the mouth. The teeth for example are much brighter than the rest of the face and can therefore be located using a simple filtering of the image intensity. The visibility and the position of the tongue cannot be determined as easily as in the case of the teeth, because the color of the tongue is almost indistinguishable from the color of the lips. We can however easily determine the amount of mouth cavity that is not obscured by the tongue. While teeth are distinctly bright, the whole area of the mouth behind the tongue is usually darker than the rest of the face. So we can apply an intensity based thresholding filter for both cases. The teeth and cavity areas are both highlighted in Figure 2. In order to use

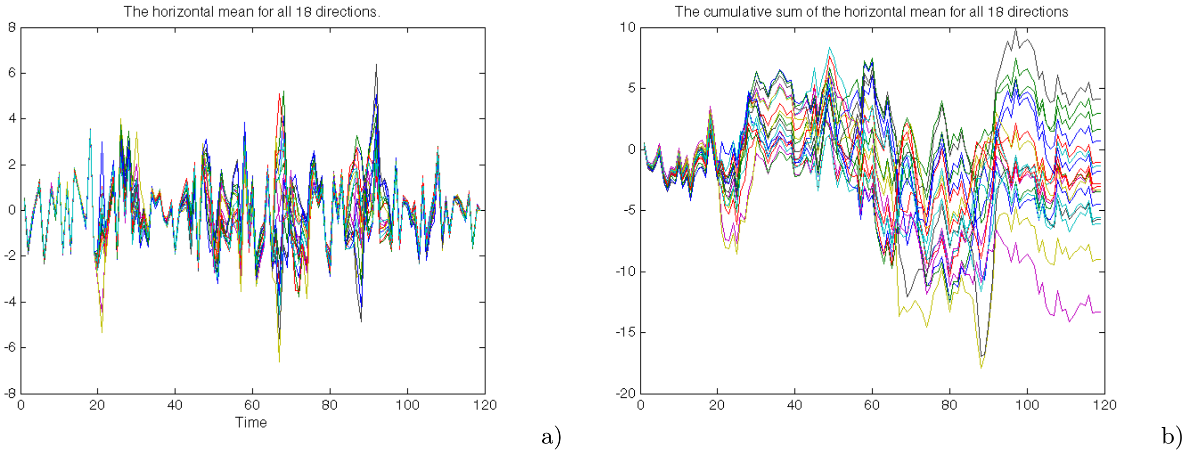

Fig. 8 The distribution over time of the horizontal mean of the optical flow for a fan with 18 distinct directions: a) shows the actual values, while b) shows the cumulative sum.

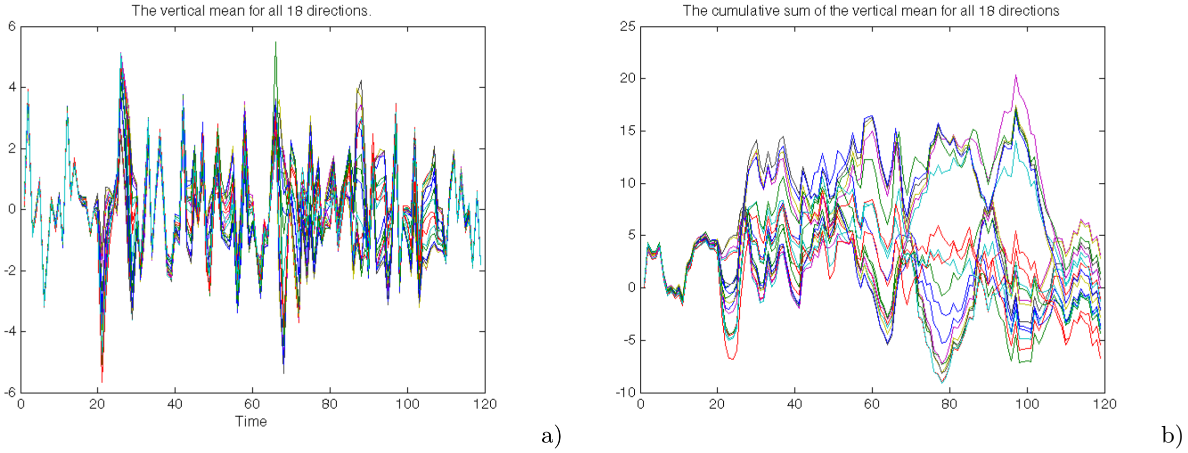

Fig. 9 The distribution over time of the vertical mean of the optical flow for a fan with 18 distinct directions: a) shows the actual values, while b) shows the cumulative sum.

the information presented in the filtered images, we need to extract from them some quantitative values. We chose to use the total area of the highlighted region and the position of its center of gravity relative to the center of the mouth.

4 Data Corpus

Data corpora are an important part of any audio-visual speech recognition research. However, partly because the field is still young, or partly because the time and resources it takes to record a multi-modal data corpus can be overwhelming, the number of existing multi-modal data corpus is small compared to the number of uni-modal datasets. Having a good data corpus, (i.e. well designed, capturing both general and also particular aspects of a certain process) might be of great help for the researchers in this field. Data corpora are also developed to be shared between different researchers in order to have the means for comparison of their results, so a greater level of reusability is required. There are a number of limitations that an audio-visual dataset has, such as:

- The recordings contain only a small number of respondents. This greatly reduces the generality of the results, since it generally generates highly undertrained systems.
- The pool of utterances is usually very small.
- They usually contain only isolated words or digits or even only the letters of the alphabet rather than continuous speech.
- They have a poor coverage of the set of phonemes and visemes in the language. This problem is related with the previous ones since the phonemes/visemes coverage is correlated with the set of words/sentences used. So using utterances that are rich in phonemes and visemes should be a strong requirement especially for the cases when the dataset is intended for speech recognition.

- The last and definitely not the least important issue is that usually the datasets are collected with some specific applications in mind, which lessens their reusability [46].

One of the first datasets that used for lipreading was TULIPS1.0 [47]. This database was assembled in 1995 and consists of very short recordings of 12 subjects uttering the first 4 digits in English. Later other datasets were compiled, which are larger and have a greater degree of usability. These datasets were built having as target mainly audio-visual speech recognition, speaker detection and person identification. The most well know are ValidDB [48], XM2VTSDB [49], BANCA [50], AVOZES, CUAVE [51] and VidTIMIT.

For this research we used an audio-visual data corpus for Dutch. The Delft University of Technology Audio-Visual Speech Corpus (DUTAVSC) was intended only for lip reading and audio-visual speech recognition hence the video only shows the lower part of the face. The recordings were made during 5 sessions with 8 respondents and contain continuous speech. The utterances contain phonetically rich sentences. The sentences were gathered from Dutch newspapers and grouped in sets of five sentences such that each set contained all phonemes found in Dutch. To these sets were added separate words as spelling samples. Special attention was also paid to the phoneme distribution in the chosen sentences, such that to result in a natural distribution over them. All respondents were native Dutch speakers. The audio recordings were sampled at 44kHz with 16-bit resolution and were stored in uncompressed form so that no signal degradation resulted during storage. The video recordings have 384x288 pixel resolution and were saved at a frame rate of 25fps. Since the video shows only the lower part of the face, then using this resolution, similar with the resolution used in some well known databases such as XM2VTS [49] or BANCA [50], a very fine detail of the mouth region was achieved. Such a restricted view can be achieved in real application by using a smart camera which can detect the face of the speaker and zoom in on it. The video clips were stored using MPEG1 compression with a high bit rate in order to make an optimal trade between image quality and file size. The database contains approximately 7 hours of recordings. Some sample video frames from this database are shown in Figure 10.

5 Test results

5.1 Setup of the experiments

The task of our system was continuous speech recognition under different levels of background noise for Dutch (see section 4). From the audio stream we extracted the MFCCs at every 10ms using a Hamming window of 25ms. Each feature vector contained 12 MFCCs plus log energy, plus the corresponding deltas and acceleration values, hence a total of 39 features in each vector. All features were scaled around zero by subtracting the cepstral mean on a per utterance basis. The video data was recorded at 25fps. Each of the two algorithms is extracting a number of 36 features for the contour of the mouth plus 6 intensity based features. In order to reduce the dimensionality of the resulted vectors we applied PCA on the visual features (excluding the intensity features) and saved the first 5 features corresponding to the most informative directions. In the end a total of 50 features were fed to the recognizer. For the case where the motion information has to be recovered from static features additional 10 features represented by deltas and accelerations values in time were taken.

The Cambridge Hidden Markov Model Toolkit [52] was used for actual training and recognition. The recognition units were established at phonemes level. Each phoneme was modeled by Gaussian mixtures continuous density left-right HMM with 5-states, with only three emitting states. The model is shown in Figure 1. The models were trained iteratively, using embedded Baum-Welch re-estimation and Viterbi alignment. We used 44 models for phonemes and 16 models for visemes recognition. However since there is no direct mapping from phonemes set to visemes set we chose to use the phonemes as basic recognition units and to define the visemes by clustering the corresponding phonemes together. This was obtained by sharing the distributions in the visual stream among phonemes models from the same set (i.d. by tying the states of the models). Since our audio-visual database is relatively small for training a robust continuous speech recognizer, let alone a bi-modal recognizer, we had to think of a way to better guess the initial values for the parameters in order to improve the convergence during training. For the parameters related with the audio stream we used as initial guess the values of the parameters from an already trained speech recognizer. Therefore the training on our database will only induce some adaptation on the audio side. For the additional visual parameters we used as initial values the global means and variances computed over the entire video training set. As all models initially had the same parameters for their visual features the distribution of the feature vectors during Baum-Welch re-estimation was guided by the speech features. In this way a continuous multi-modal recognizer can be obtained in a few training cycles with a limited amount of training data. The combined models were re-estimated three times, using the bi-modal training data. We also included one short pause model and a silence model. The short pause model had only one emitting state which was tied to the center-state of the silence model.

To check the performance of the recognizer we used two widely used performance measures. The word error rate (WER) is defined as follows:

$WER = 100 - \frac{N-D-S}{N} x100\%$, where N is the total number of words, D is the number of deletion error

Fig. 10 DUTAVSC Database.

and S is the number of substitutions. However in the next figures and tables we will use the Word Recognition Rate(WRR) which is defined as 100 - WER. To take into account also the errors by insertion, the percentage accuracy(ACC) is also frequently used. ACC is computed as follows:

$ACC = \frac{N-D-S-I}{N} x100\%$, where I represents the number of insertion errors.

We also investigated the performance of the system when the two data streams are weighted. If the weight for the audio stream is w_a then the weight of the video side is computed as $1 - w_a$.

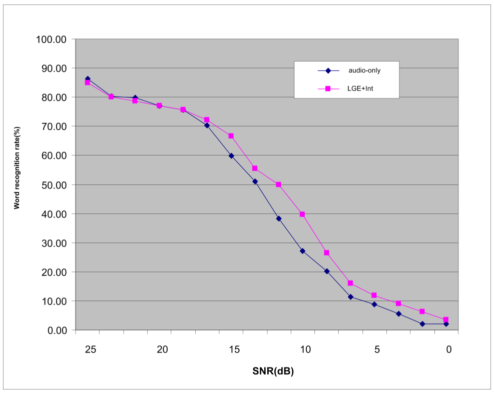

Fig. 11 Recognition results at several noise levels in terms of word recognition rate. The visual features used were extracted using the LGE method.

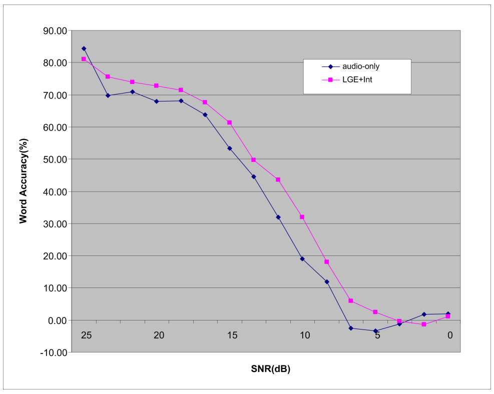

Fig. 12 Recognition results at several noise levels in terms of word recognition accuracy. The visual features used were extracted using the LGE method.

5.2 Static visual features

Table 1 shows the results of the recognition for static visual features. The results for several weighting settings are also shown. The conclusion that emerges from this table is that as soon as the noise level increases the weight of the audio channel should be reduced, which is exactly what one expects. We can see that when the SNR is around 20dB the audio modality is still reliable as the maximum of performance was attained for relatively high audio weight.

Since in noisy free environment the visual modality does not bring much improvement to the overall recognition accuracy it is more interesting to look at the accuracy of the system under different levels of noise. Figure 11 shows the performance of the bi-modal recognizer in terms of WRR as a function of SNR. The audio signal was disturbed by adding white noise of different intensity such that the SNR lies in the interval 25 to 0dB. Only the samples used for evaluation were modified by adding noise. For comparison the results of the audio-only speech recognizer are presented along with the results from the bi-modal recognizer. The training of the audio recognizer was done using the same settings. In Figure 11 is clearly visible that when the level of noise increases, in this case when the SNR goes below 15dB, using the combined information gives better results. For instance around level 10dB the increase in performance is more than 10%. The same trend can be seen in the word accuracy levels as it is shown in Figure 12. However since the accuracy takes into account also the errors made by insertion of words the overall levels are lower and we can have even negative values. Because the two channels have the same weight in recognition, the accuracy of the bi-modal recognizer approaches also 0% in extreme noise conditions.

5.3 Deltas and accelerations

The question we want to answer is what happens if we recover some of the motion information starting from static visual features. The motion can be recovered by using the first and/or second derivatives of the visual features. We will consider the derivatives only of the geometry features and not for the intensity based ones. The analysis

Table 1 Word recognition percentage for static visual features.

Recognition system	WRR(%)	WAcc(%)
clear audio; equal weights	84.89	81.11
20dB SNR; audio weight: 0.50	71.17	68.59
20dB SNR; audio weight: 0.72	76.74	70.78
20dB SNR; audio weight: 0.92	78.33	73.96
20dB SNR; audio weight: 1.20	76.54	71.77
10dB SNR; audio weight: 0.50	38.80	31.61
10dB SNR; audio weight: 0.72	41.75	38.97
10dB SNR; audio weight: 0.92	38.97	33.40
10dB SNR; audio weight: 1.20	33.60	22.07
0dB SNR; audio weight: 0.50	4.57	3.38
0dB SNR; audio weight: 1.10	5.57	-1.79

proceeds into four distinct settings. Firstly we consider the del features as additionally to the static features, hence the total number of features will be 55. A second round takes into account also the acceleration features, making the feature vector dimension to increase to 60. Adding more features provides more information about what is being said. However, it also increases the dimensionality of the problem. This, on the basis of a constant training dataset, makes the resulted models to be poorer trained. Therefore a comparison between them, or with the initial bi-modal system is not entirely fair. For this reason we also investigated the situations when the delta features or when delta and acceleration features are used instead of unmodified geometrical features. Figure 13 shows pair comparisons with respect to the resulted systems' performances in some of the four settings.

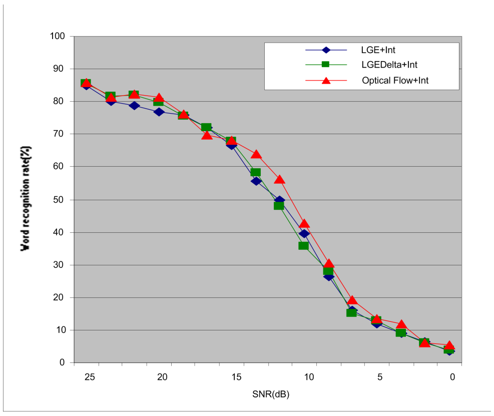

Fig. 14 Comparison among the performances of the AVSR when three different approaches are employed for computing the visual features.

5.4 Optical flow based features

This section presents the results obtained when the features extracted contain real information about the motion on the speaker face. In order to keep the dimensionality of the problem in manageable limits and also to make the comparison of the results fair we applied PCA to the visual data and saved only the features corresponding to the first 5 directions. The first 5 directions accounted for 95.82% of the total variance.

Figure 14 shows for comparison the results of the system in the three cases: the static visual features obtained by using the LGE method, the visual features obtained by computing the delta values of the geometrical features and finally the visual features obtained by performing optical flow analysis. In all cases the intensity features are added. The results show that the optical flow features perform slightly better than the first two methods. Again when the SNR has high values the audio modality is driving the recognition process.

6 Conclusions

The research in the domain of audio-visual speech recognition is still going on. In contrast with audio feature extraction technology, in the case of the visual modality it is still not very clear what features are more appropriate for robust lipreading and audio-visual speech recognition. It has been shown [1] that the visual features used until now describe more or less the same aspects of the mouth during speech, or at least this is the conclusion with respect to direction that accounts for the largest part of the variance in the dataset. However, since the

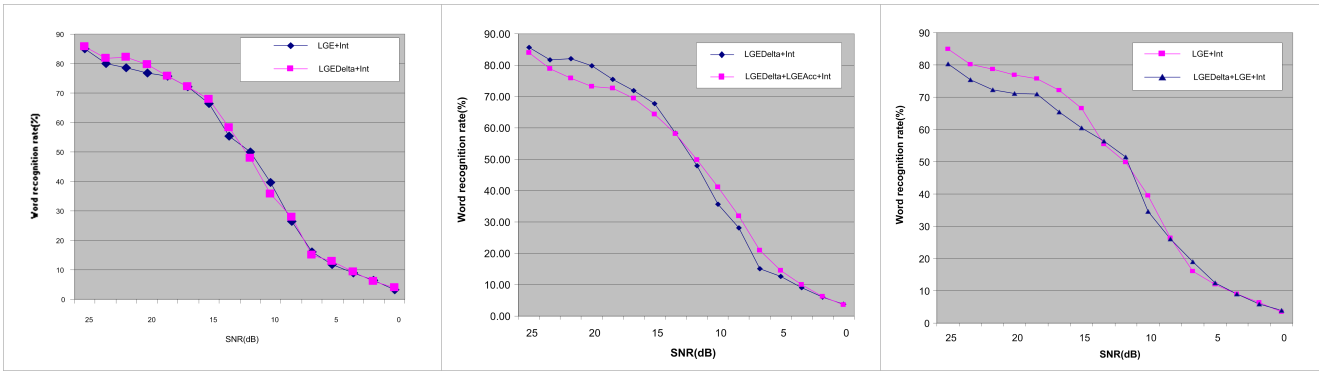

Fig. 13 Recognition results at several noise levels in terms of word recognition rate. The visual features used were extracted using the LGE method. The motion information is recovered by computing the delta and acceleration values of the geometrical features.

methods developed until now only use one video frame from the input visual data for computing the visual features, the resulted features described only one instance in time. Therefore these methods might fail to account for the actual movement in the mouth area. We introduced in this paper a new method for extracting speech related visual features based on the optical flow. The visual features computed by the presented method capture real information about the motion of the speaker's mouth. We showed that the performance of the recognition system build based on the new features is much greater than the performance of a system trained on static features. We also compared the results based on the new method with the results based on a middle way of recovering motion information, namely through delta and acceleration computation.

We also showed that it is very important to set the appropriate weights to the two data channels. The best results are visible when the noise level increases which was an expected behavior.

The results shown here are very promising, but we could argue that the early fusion approach was not enough to accurately capture the dependencies between the two data streams and to perfectly cope with problems as asynchronism and asymmetrical sample rate. Hence we plan to investigate the use of Dynamic Bayesian Networks(DBNs) for audio-visual fusion. Many inference routines have been developed for DBNs [53, 54] including methods for calculating the probability of an observation sequence given the model and for finding the most likely path through the network. These new developments make the DBNs models more tractable, hence their use becomes more attractive [55].

References

1. L. J. M. Rothkrantz, J. C. Wojdel, and P. Wiggers, "Comparison between different feature extraction techniques in lipreading applications", in *Specom'2006*, SpIIRAS Petersburg, 2006. 1, 13
2. J. C. Wojdel and L. J. M. Rothkrantz, "Visually based speech onset/offset detection", in *Proceedings of 5th Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Application (Euromedia 2000)*, (Antwerp, Belgium), pp. 156–160, 2000. 1, 8
3. H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices", *Nature*, vol. 264, pp. 746 – 748, December 1976. 2
4. N. Li, S. Dettmer, and M. Shah, "Lipreading using eigen sequences", in *Proc. International Workshop on Automatic Face- and Gesture-Recognition*, (Zurich, Switzerland), pp. 30–34, 1995. 2
5. N. Li, S. Dettmer, and M. Shah, "Visually recognizing speech using eigensequences", *Motion-based recognition*, 1997. 2
6. X. Hong, H. Yao, Y. Wan, and R. Chen, "A pca based visual dct feature extraction method for lip-reading", *iihmsp*, vol. 0, pp. 321–326, 2006. 2
7. C. Bregler and Y. Konig, ""eigenlips" for robust speech recognition", in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94 IEEE International Conference on*, 1994. 2
8. P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel, "Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition", in *International Conference on Acoustics, Speech, and Signal Processing, 1995 (ICASSP-95)*, vol. 1, pp. 109–112, 1995. 2
9. I. A. Essa and A. Pentland, "A Vision System for Observing and Extracting Facial Action Parameters", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 76–83, IEEE, June 1994. 2, 8
10. S. Tamura, K. Iwano, and S. Furui, "A Robust Multi-Modal Speech Recognition Method Using Optical-Flow Analysis", in *Extended summary of IDS02*, (Kloster Irsee, Germany), pp. 2–4, June 2002. 2, 8
11. T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, "Audio-Visual Speech Recognition Using Lip Movement Extracted from Side-Face Images", in *AVSP2003*, pp. 117–120, September 2003. 2, 8
12. T. Yoshinaga, S. Tamura, K. iwano, and S. Furui, "Audio-Visual Speech Recognition Using New Lip Features Extracted from Side-Face Images", in *Robust 2004*, August 2004. 2, 8
13. K. Mase and A. Pentland., "Automatic Lipreading by Optical-Flow Analysis", in *Systems and Computers in Japan*, vol. 22, pp. 67–76, 1991. 2, 8
14. K. Iwano, S. Tamura, and S. Furui, "Bimodal Speech Recognition Using Lip Movement Measured By Optical-Flow analysis", in *HSC2001*, 2001. 2, 8

15. D. J. Fleet, M. J. Black, Y. Yacoob, and A. D. Jepson, “Design and Use of Linear Models for Image Motion Analysis”, *International Journal of Computer Vision*, vol. 36, no. 3, pp. 171–193, 2000. 2, 8
16. A. Martin, “Lipreading by Optical Flow Correlation”, tech. rep., Compute Science Department University of Central Florida, 1995. 2, 8
17. S. Tamura, K. Iwano, and S. Furui, “Multi-modal speech recognition using optical-flow analysis for lip images”, *J. VLSI Signal Process. Syst.*, vol. 36, no. 2-3, pp. 117–124, 2004. 2, 8
18. S. Furui, “Robust Methods in Automatic Speech Recognition and Understanding”, in *EUROSPEECH 2003 - GENEVA*, 2003. 2, 8
19. J. C. Wojdel and L. J. M. Rothkrantz, “Using Aerial and Geometric Features in Automatic Lipreading”, in *Proceedings Eurospeech 2001*, (Scandinavia), September 2001. 2
20. L. J. M. Rothkrantz, J. C. Wojdel, and P. Wiggers, “Fusing Data Streams in Continuous Audio-Visual Speech Recognition”, in *Text, Speech and Dialogue: 8th International Conference, TSD 2005*, vol. 3658, (Karlovy Vary, Czech Republic), pp. 33–44, Springer Berlin / Heidelberg, September 2005. 2
21. B. K. Horn and B. G. Schunck, “Determining optical flow.”, *Artificial Intelligence*, vol. 17, pp. 185–203, 1981. 3
22. B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision.”, in *Proc. Seventh International Joint Conference on Artificial Intelligence*, p. 674679, 1981. 3
23. A. Bruhn and J. Weickert, “Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods”, *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005. 3
24. S. Uras, F. Girosi, A. Verri, and V. Torre, “A computational approach to motion perception”, in *Biological Cybernetics*, vol. 60, pp. 79–87, December 1988. 3
25. H.-H. Nagel, “On the estimation of optical flow: relations between different approaches and some new results”, *Artificial Intelligence*, vol. 33, no. 3, pp. 298–324, 1987. 3
26. P. Anandan, “A Computational Framework and an Algorithm for the Measurement of Visual Motion”, *International Journal of Computer Vision*, vol. 2, pp. 283–310, 1989. 3
27. A. Singh, “Optic flow computation. a unified perspective”. IEEE Computer Society Press,, 1991. 3
28. D. J. Heeger, “Model for the extraction of image flow”, *Journal Opt. Soc. Amer.*, vol. 4, pp. 1455–1471, August 1987. 3
29. A. Waxman, J. Wu, and F. Bergholm, “Convected activation profiles and receptive fields for real time measurement of short range visual motion”, in *Proceedings of Conference Computational Visual Pattern Recognition*, pp. 771–723, 1988. 3
30. D. J. Fleet and A. D. Jepson, “Computation of Component Image Velocity from Local Phase Information”, *International Journal of Computer Vision*, vol. 5, pp. 77–104, August 1990. 3
31. J. L. Barron, D. J. Fleet, and S. S. Beauchemin, “Performance of optical flow techniques”, *International Journal of Computer Vision*, vol. 12, pp. 43–77, February 1994. 3
32. B. Galvin, B. McCane, K. Novins, D. Mason, and S. Mills, “Recovering Motion Fields: An Evaluation of Eight Optical Flow Algorithms”, in *Proceedings of the British Machine Vision Converence (BMVC) ’98*, September 1998. 3
33. L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. N.J, USA: Prentice Hall, 1993. 4
34. R. P. Lippmann, “Review of neural networks for speech recognition”, *Neural Comput.*, vol. 1, no. 1, pp. 1–38, 1990. 4
35. L. J. Rothkrantz and D. Nollen, “Automatic speech recognition using recurrent neural networks”, in *Neural Network World*, vol. 10, pp. 445–453, July 2000. 4
36. A. Ganapathiraju, *Support vector machines for speech recognition*. PhD thesis, 2002. Major Professor-Joseph Picone. 4
37. T. S. Andersen, K. Tiippana, and M. Lampien, “Modeling of audio-visual speech perception in noise”, in *In Proceedings of AVSP 2001*, (Aalborg, Denmark), September 2001. 5
38. P. Smeele, *Perceiving speech: Integrating auditory and visual speech*. PhD thesis, Delft University of Technology, 1995. 5
39. D. Massaro, “A fuzzy logical model of speech perception”, in *Proceedings of XXIV International Congress of Psychology. Human Information Processing: Measures, Mechanisms and Models* (D. Vickers and P. Smith, eds.), (Amsterdam, North Holland), pp. 367–379, 1989. 5
40. G. Meyer, J. Mulligan, and S. Wuerger, “Continuous audio-visual digit recognition using n-best decision fusion”, *Information Fusion*, vol. 5, pp. 91–101, 2004. 5
41. S. Dupont, H. Bourlard, and C. Ris, “Robust speech recognition based on multi-stream features”, in *Proc. of ESCA/NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, (Pont- Mousson, France), pp. 95–98, 1997. 6
42. J. Luettin and S. Dupont, “Continuous audio-visual speech recognition”, in *IDIAP, Dalle Molle Institute for Perceptual Artificial Intelligence*. 6
43. Z. Ghahramani and M. I. Jordan, “Factorial hidden markov models”, in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 1997. 6
44. P. Viola and M. Jones, “Robust Real-time Object Detection”, in *Second International Workshop On Statistical And Computational Theories Of Vision Modeling, Learning, Computing, And Sampling*, (Vancouver, Canada), July 2001. 6
45. T. Coianiz, L. Torresani, and B. Caprile, “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.), vol. 150 of *NATO ASI Series F: Computer and Systems Sciences*, Berlin and New York: Springer, 1996. 7
46. J. Millar, M. Wagner, and R. Goecke, “Aspects of Speaking-Face Data Corpus Design Methodology”, in *Proceedings of the 8th International Conference on Spoken Language Processing ICSLP2004*, vol. II, (Jeju, Korea), pp. 1157–1160, oct 2004. 11
47. J. R. Movellan, “Visual Speech Recognition with Stochastic Networks”, in *Advances in Neural Information Processing Systems*, vol. 7, (Cambridge), MIT Pess, 1995. 11
48. N. A. Fox, *Audio and Video Based Person Identification*. PhD thesis, Department of Electronic and Electrical Engineering Faculty of Engineering and Architecture University College Dublin, 2005. 11
49. K. Messer, J. Matas, and J. Kittler, “Acquisition of a large database for biometric identity verification”, 1998. 11
50. E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Pore, B. Ruiz, and J.-P. Thiran, “The BANCA Database and Evaluation Protocol”, in *Audio and Video Based Biometric Person Authentication*, vol. 2688, pp. 625–638, Springer Berlin / Heidelberg, 2003. 11
51. E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, “CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research”, in *Proceedings of*

the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. 11

52. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. 2005. 11
53. K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002. 14
54. J. Pearl, *Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988. 14
55. A. V. Nefian, L. Liang, X. Pi, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition", *EURASIP Journal on Applied Signal Processing*, vol. 11, p. 12741288, 2002. 14