A NEW VISUAL SPEECH MODELLING APPROACH FOR VISUAL SPEECH RECOGNITION

Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan

Abstract: In this paper we propose a new learning-based representation that is referred to as Visual Speech Unit (VSU) for visual speech recognition (VSR). The new Visual Speech Unit concept proposes an extension of the standard viseme model that is currently applied for VSR by including in this representation not only the data associated with the visemes, but also the transitory information between consecutive visemes. The developed speech recognition system consists of several computational stages: (a) lips segmentation, (b) construction of the Expectation-Maximization Principal Component Analysis (EM-PCA) manifolds from the input video image, (c) registration between the models of the VSUs and the EM-PCA data constructed from the input image sequence and (d) recognition of the VSUs using a standard Hidden Markov Model (HMM) classification scheme. In this paper we were particularly interested to evaluate the classification accuracy obtained for our new VSU models when compared with that attained for standard (MPEG-4) viseme models. The experimental results indicate that we achieved 90% recognition rate when the system has been applied to the identification of 60 classes of VSUs, while the recognition rate for the standard set of MPEG-4 visemes was only 52%.
Keywords: Visual Speech Recognition, Visual Speech Unit, Viseme, PCA Manifold, HMM, Dynamic Time Warping.

INTRODUCTION

The automated recognition of human speech using only features from the visual domain has become a significant research topic that plays an essential role in the development of many multimedia systems such as audio visual speech recognition (AVSR) [1,2], mobile phone applications, human-computer interaction (HCI) [3] and sign language recognition [4]. The inclusion of the lip visual information is opportune since it can improve the overall accuracy of audio or hand recognition algorithms especially when such systems are operated in environments characterized by a high level of acoustic noise. Visual speech recognition can also be applied in the development of systems for person identification, machine control or game animation.

A review of the literature on VSR indicates that the systems developed can be categorized into two major groups, namely shape-based and appearance-based approaches. The shape-based approaches rely on the extraction of geometrical features from the lips and this information is used to encode a standard set of mouth shapes that are applied to model the lip motions during the speech process. This approach has been applied by Petajan [5, 6] in the development of a lip-reading system where simple shape features such as height, width and mouth area were used to encode the shape of the region described by the lips contour. This approach has been further developed by Luettin et al [7] where they applied a parametric lip template defined by eleven morphometric measurements that were applied to characterize the lips motions. To circumvent the problems related to uneven illumination and noise, other approaches applied Active Shape Models or snakes to extract the lips outlines [8-10], but their application to VSR proved to be problematic since they require a complex initialization procedure.
One limitation associated with the shape-based VSR systems resides in the fact that only geometrical information is used to encode the mouth shapes. In addition these approaches are sensitive to tracking errors and they are not able to efficiently encompass the information contained in consecutive frames. To address these issues, appearance-based approaches have been proposed for VSR and their major advantage is that they use the entire gray-scale (or colour) information available to sample the spectrum of mouth shapes. In this regard, the image area around the lips is extracted for each frame in the video sequence and this information can be compressed to obtain a low-dimensional representation using Principal Component Analysis (PCA) [17], Discrete Cosine Transform (DCT) [16, 17], and Linear Discrete Analysis (LDA) [18]. This representation of the mouth shapes in a low-dimensional feature space proved to be opportune and the performance of these methods is generally better than that attained by the shape-based VSR techniques.

In parallel with feature extraction, significant research efforts were concentrated on the identification of the most discriminative visual speech elements that are able to model the speech process in the continuous visual domain. The early works on visual speech modelling attempted to map the basic linguistic elements such as phonemes in the visual domain. To this end, many authors proposed different modelling strategies where elementary speech units played the central role. The most basic unit that has been employed to describe the visual speech is the viseme. This basic visual speech element can be conceptualised as the interpretation in the visual domain of a phoneme or a group of phonemes and modelling the visual speech with different sets of visemes received a great deal of interest from the research community [19,27,35]. The main motivation behind this interest resides in the fact that only a small number of visemes are required to model more complex speech elements such as words. While the concept behind the application of visemes to model the speech in the visual domain has been embraced by the vast majority of researchers, the selection of the most representative viseme set has been one of the most researched problem in the field of VSR. In this regard, many studies have been conducted using viseme sets with their sizes ranging from 6 [32] to 50 [36]. Although little consensus has been reached in regard to the selection of the most representative visemes, the MPEG-4 viseme set that has been designed to support facial animations has gained the largest acceptance from the research community [35]. While research on the construction of representative viseme-based speech representations is still ongoing, several disadvantages associated with this visual speech representation have recently surfaced. The most important are related to their poor discriminative power since they are defined by a small number of mouth shapes and in particular their vulnerability to the lexical context (distortions that are caused by viseme co-articulation rules that are enforced during the continuous speech process. For instance, the lip shapes associated with the viseme [b] have different visual context during the articulation of the words ‘book’ and ‘but’. In this regard, when the speaker is uttering the word “book” the lip shapes associated with the viseme [b] are described by a tight round shape, whereas the lip shapes associated with the viseme [b] in the word ‘but’ are described by a more elongated shape). To address

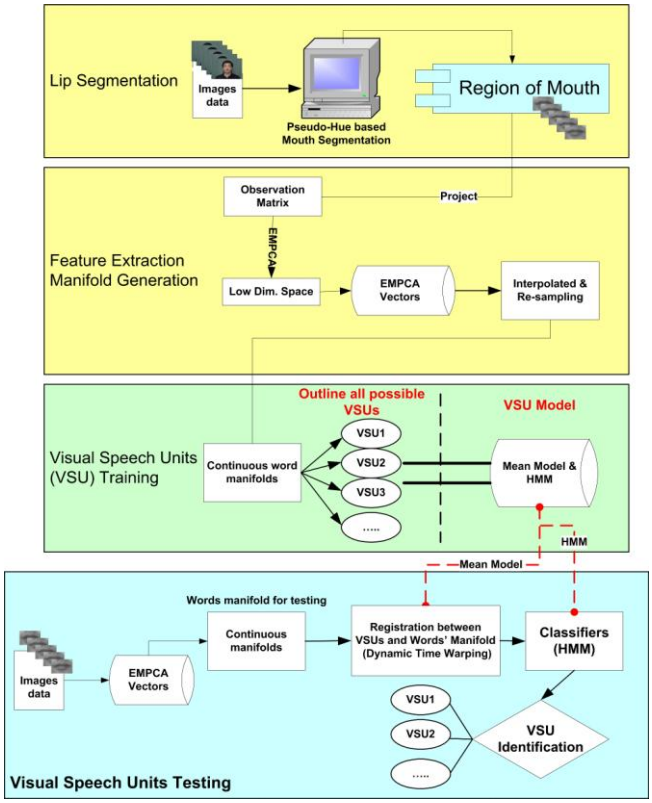
these issues, a number of studies have been recently devoted to evaluate the robustness of new elementary speech units that are modeled based on the representation of the biphones and triphones in the visual domain [37, 40]. Most of the work has been carried out in the context of audio-visual speech recognition [37, 39 & 41], automatic face synthesis [40] and text-to-audiovisual speech synthesis [38]. It is useful to note that the main topic of these papers was focused on the integration of the audio and video information into composite descriptors where the acustic information is used to localize and align the video frames associated with the video speech units. However, this favorable scenario cannot be exploited in the development of lip-reading systems where only the video information is available and to the best of our knowledge no studies that evaluated the stability and performance of the composite-viseme speech models in the context of VSR have been reported so far. The work detailed in the paper by Ezzat and Poggio [38] where the authors describe the development of a text-to-audio-visual synthesis is the most related to the visual speech modelling approach detailed in this paper. In [38] the authors propose a small set of 6 consonant visemes and 7 monophthong viemes that are applied to model the mouth transitions in a smooth and realistic manner. To achieve this goal, the authors employed a viseme morphing strategy to concatenate the viseme prototypes into words using rules that are enforced by an audio-visual synchronization unit. While this approach is opportune when applied to audio-visual speech synthesis, the proposed viseme set lacks the sophistication required in the implementation of video-only speech recognition systems.

From this short literature review we notice that most of the work on VSR has been focused on the robust identification of small independent speech elements (called visemes [19, 28, 31, 32]), while word recognition is viewed as a simple combination between standard visemes. Although words can be theoretically formed using a combination of standard visemes, in practice viseme identification within words is problematic since different visemes may overlap in the feature space, a fact that makes their identification difficult. To address this problem we propose to include additional lexical context by augmenting the standard visemes with the transition information between two consecutive visemes and this new model is referred to as Visual Speech Unit (VSU). Thus, the major aim of this paper is to propose a new visual speech modeling strategy that is able to sample in an elaborate manner the inter-visual context between consecutive visemes. In this paper we will explain in detail the construction of the VSU, and we will demonstrate that the application of this new model to VSR leads to improved performance when compared with the performance offered by the standard set of MPEG-4 visemes [35]. The main contributions associated with this work are located in the areas of feature extraction, visual speech modelling and application of non-rigid data registration to solve the alignment issues (identification of the anchor points) between the video data and the trained VSU models. Another contribution associated with this work resides in the construction and categorization of the proposed VSUs with respect to their lexical properties (or perceptual similarity) in the visual domain.

SYSTEM OVERVIEW

The main computational components of the system described in this paper are shown in Fig. 1. In the first phase, the lips are extracted from input video data. In order to achieve this goal, we calculate the pseudo-hue [12-14] from the RGB colour planes [11, 33] and the lips are segmented by applying a histogram-based thresholding scheme. The image area describing the lips is extracted for each frame from the input video sequence. This region of interest is converted into a matrix form and it is compressed using Expectation Maximization PCA (EM-PCA) [27] into a low dimensional feature space. Then, each image area describing the lips in the input sequence is projected onto the low-dimensional EM-PCA space with a view to obtain a discrete manifold where for each mouth shape a low dimensional vector is assigned. The next step performs a manifold interpolation using a cubic spline function to generate a continuous representation. In the final step, the algorithm attempts to register the VSU models contained in the database with the continuous manifold representation using a Dynamic Time Warping approach. In this manner, the manifold generated from the input image sequence describing visually the spoken word is broken into an ordered sequence of VSUs and the recognition process is carried out using a HMM classification scheme.

Fig.1 An overview of the Visual Speech Recognition system



FEATURE EXTRACTION

Lip Segmentation

In order to extract the lip regions from input video data, we applied a simple procedure based on the calculation of the pseudo-hue component from the RGB colour planes [12-14]. The pseudo-hue component highlights the image areas where strong differences between the red and green colour planes are encountered. This property of the pseudo-hue component is particularly useful in performing a robust

separation of the image regions defined by the lips, where the red component of the RGB data is dominant, from the skin mixture models that are defined by image areas where both the red and green components are dominant. Thus, the pseudo-hue data can be approximated with a bi-modal distribution and the lips segmentation process can be formulated as a two-class clustering problem. Based on this observation the lips segmentation involves a two-step approach. In the first step the image areas characterized by large pseudo-hue values are identified by performing a threshold operation, where the threshold value is automatically detected as the local minima with respect to the second peak in the histogram as illustrated in Fig. 2. The second step of the lip segmentation process involves the identification of the lips in the threshold pseudo-hue data by performing an exhaustive validation of all regions resulting after the application of the threshold operation with respect to anthropometric properties of the human face (for more details about the lips identification procedure the reader can refer to [10]).

The region of interest (ROI) is constructed as the bounding box that encompasses the extreme corners of the upper lips as illustrated in Fig. 3. The grayscale intensity values contained in the ROI are extracted for each frame from the input video data and they are used to generate the EM-PCA manifold. This procedure will be detailed in the next section.

Fig.2 Histogram-based selection of the threshold from the pseudo-hue image

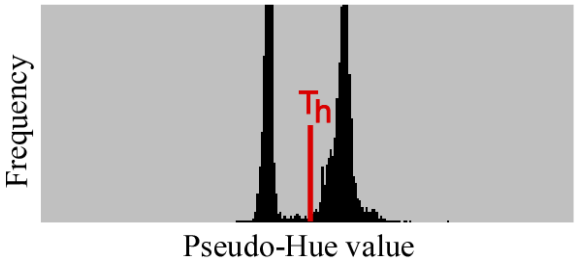


Fig.3 Lip segmentation algorithm

(a) (b) (c) (d) (e)

(a) RGB image. (b) Pseudo-hue image. (c) Image resulting after the application of the histogram-based thresholding. (d) Lips region – pseudo hue. (e) Lips region – gray-scale.

MANIFOLD GENERATION

For visual speech recognition purposes, the information associated with the lips motions in the frames of the input video sequence is of interest. As indicated in the previous section, the lips regions are segmented in each frame and the appearance of the lips is encoded as a point in a feature space that is obtained by projecting the input data onto the low dimensional space generated by the EM-PCA procedure [25]. (A discussion that details the application of the EM-PCA procedure to encode the appearance of the lips into low-dimensional feature vectors is provided in Appendix A.). The feature points obtained after the projection of the lips image data onto the low-dimensional EM-PCA space are joined by a plotline based on the frame order. In this way, we generate a surface in the feature space that is called manifold [8] and this process is illustrated in Fig. 4. (Note that the axes in Figs. 4 to 13 represent the projection of the input feature vector describing the lips data onto the leading three EM-PCA eigenvectors). In the implementation detailed in this paper we used only the first three EM-PCA components since they are able to capture more than 90% of the statistical variation of the 40,000 images that form the training set. (Our results are in line with those reported by Aleksic and Katsagellos [37], where they demonstrated that the first six, two and one leading eigenvectors are able to sample 99.6%, 93% and 81% of the of the total statistical variation of the training data, respectively.) The motivation to use the first three EM-PCA components is also justified by the fact that these components are strongly related to the features that describe the appearance of the mouth shapes. In this regard, the first component captures the texture information around lips, the second component samples more localized information such as the geometry of the mouth shapes, while the third component captures finer details such as the presence of teeth and tongue in the image data.

The manifold determined as illustrated in Fig. 4 is defined by a discrete number of points given by the number of frames in the image data. This discrete manifold representation is inadequate due to factors such as variations in the sampling rate of the video data, inter- and intra- user pronunciation variability and small localization errors that occur during the lips segmentation process. While the problems caused by the variation in the sampling rate of the video data can be controlled during the classification process, the issues caused by the variations in pronunciation and the localization errors in estimating the region of interest around the lips area are more difficult to address, as they have an undesirable effect on the dynamics and the visual context of the visemes. To alleviate these problems, in the proposed implementation the feature points that define the manifold are interpolated using a cubic spline to obtain a continuous representation of the manifold [27]. The process applied to generate the continuous (interpolated) manifolds is illustrated in Fig. 5, where two manifolds constructed from two video sequences representing the same word (‘but’) are plotted.

Fig. 4. EM-PCA manifold representation. Each feature point of the manifold is obtained after the lips image region is projected onto the low-dimensional EM-PCA space.

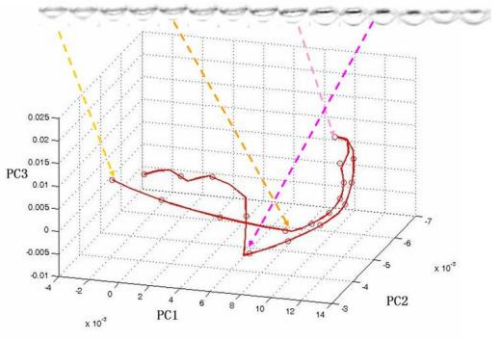
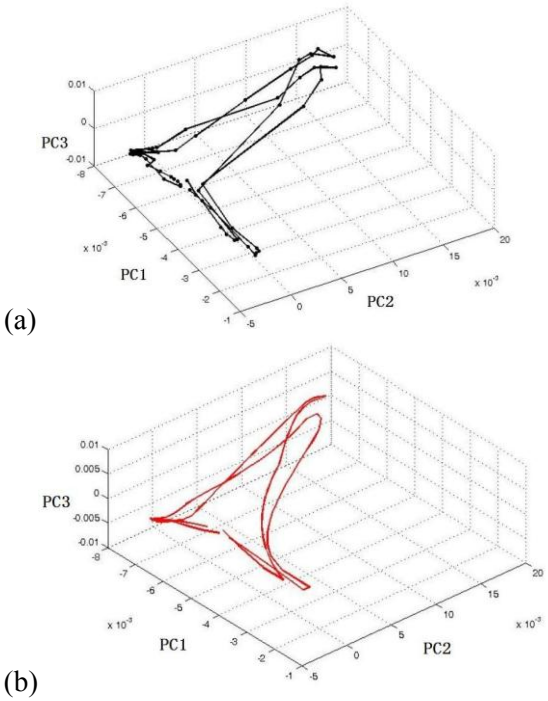


Fig. 5. Manifolds generated from two image sequences representing the word ‘but’.



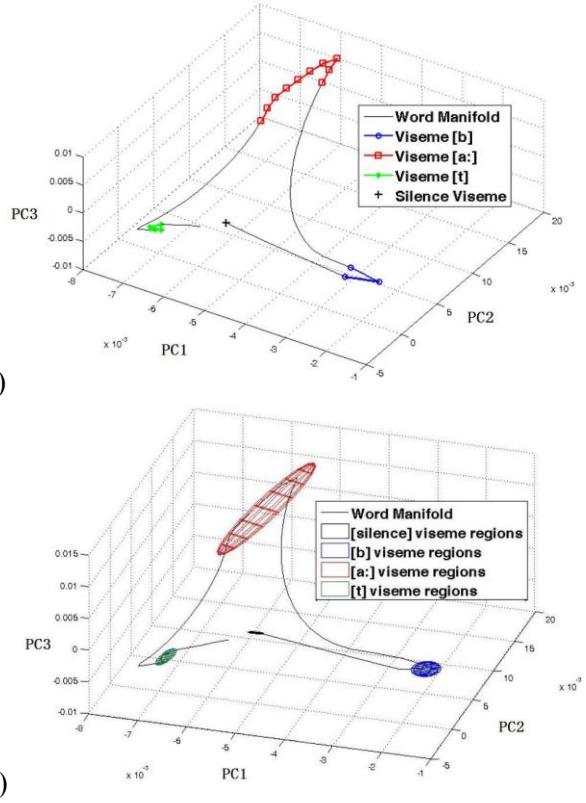
(a) The initial manifolds obtained as illustrated in Fig. 4. (b) The interpolated manifolds

VISEME vs. VISUAL SPEECH UNITS

VISEME REPRESENTATION

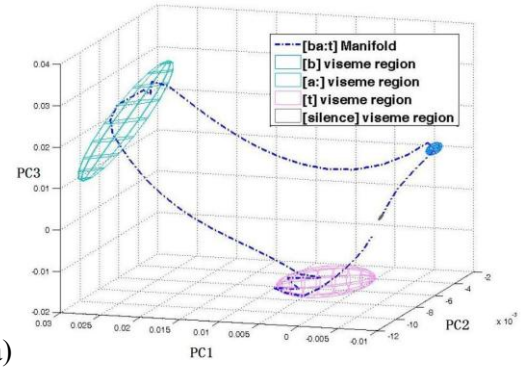
As indicated in the introductory section, a viseme can be regarded as the smallest element that describes a phoneme or a group of phonemes in the visual domain. Visemes play an important role in the development of speech recognition systems and most of the research conducted in the field of VSR has approached the word recognition as a process of sequential viseme recognition. Viseme recognition systems are based on a standard two-step computational scheme [20-22]. Initially, the VSR system is trained either with static visemes generated by the speakers or with visemes that are manually constructed by isolating the frames of interest from the continuous video speech sequence. Visemes are then located and recognized in the visual/feature space domain of the words and this process is usually carried out using HMM classification schemes [15, 19, 20]. In our approach the set of visemes is extracted from input video sequences associated with different words. For instance, the frames describing the viseme [b] are extracted from words such as 'but', 'blue' etc., while the frames describing the viseme [s] are extracted from words such us 'slow', 'snow', etc. The frames describing the standard visemes typically include three independent states, the first state is the initial state of the viseme, the second state describes the articulation process, while the last state models the transition from articulation to the end of the viseme. These frames are projected onto the EM-PCA space and as a result each viseme is defined by a number of feature points as illustrated in Fig. 6 in which the feature points for visemes [b], [a:] and [t] on the EM-PCA manifold are constructed from the video sequence describing the word 'but'.

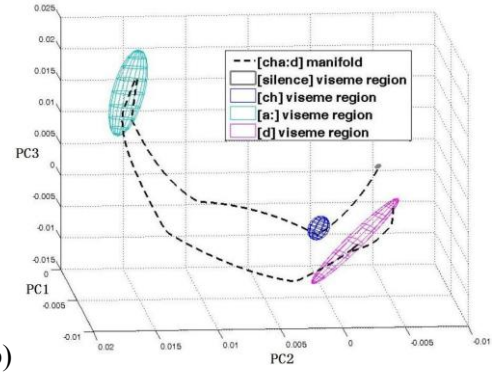
Fig.6 The representation of the visemes [b], [a:] and [t] in the interpolated (continuous) EM-PCA manifold of the word 'but' (represented using a black line).



(a) The feature points are displayed in blue for viseme [b], in red for viseme [a:] and in green for viseme [t]. The initial state of the video sequence (silence state) is shown in the diagram with a black cross. (b) The regions in the EM-PCA feature space for visemes [b], [a:] and [t] constructed from different instances of the word 'but'. The region describing the [silence] state is represented in the diagram with a black ellipsoid.

Fig. 7. The viseme feature space constructed for two different words.





(b)

(a) Word 'but' – visemes [b], [a:] and [t]. (b) Word 'chard' – visemes [ch], [a:] and [d]. Note that in the manifold representation of the word 'chard' the viseme [a:] is distorted and the consonant [r] cannot be distinguished.

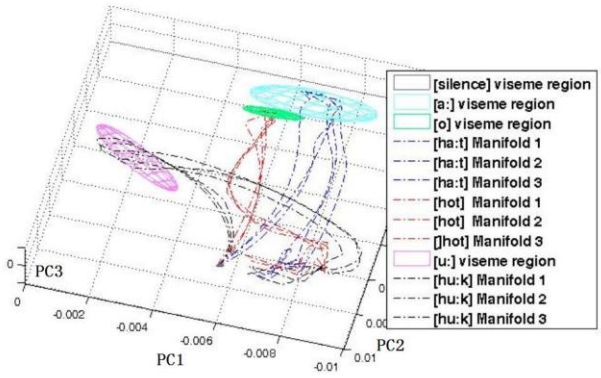
Visual Speech Unit Representation

While the viseme representation detailed in Section 3.1 is intuitive and easy to apply in the development of VSR systems it has several drawbacks. The main shortcoming associated with the viseme representation is given by the fact that a large part of the word manifold (i.e. transitions between visemes) is not used in the recognition process. This approach is inadequate since the inclusion of more instances of the same viseme extracted from different words would necessitate larger regions required to describe the feature space for each viseme (see Fig. 6b) and this will lead to significant overlaps in the feature space describing different visemes. To circumvent this problem most of the developed VSR systems applied the viseme recognition process to a reduced set of visemes and to a relatively small number of words. This problem can be clearly observed in Fig. 7 where the process of constructing the viseme spaces for two different words is illustrated.

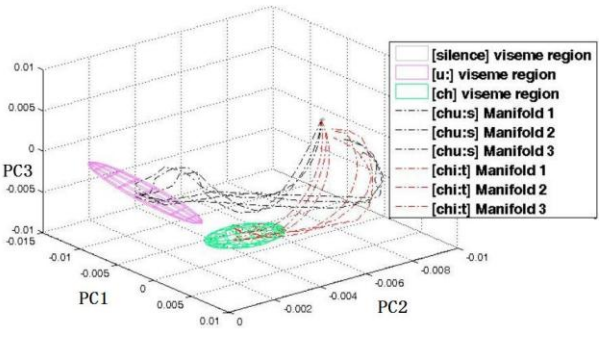
Another limitation of the viseme-based representation resides in the fact that some visemes may be severely distorted and even may disappear in the video sequences that describe visually the spoken words [21, 23, 24]. These problems can be observed in Fig. 8a, where the viseme [h] is silent (cannot be observed) in words 'heart' [ha: t], 'hat' [hæt] and 'hook' [hu: k], while in Fig. 8b we can notice that the viseme [ch] can be clearly located in the manifold of the word 'cheat', but it cannot be located in the manifold of the word 'choose'.

To alleviate the problems associated with the viseme representation, in this paper we propose to extend the viseme model by including the transitions between visemes in a new representation that is called a Visual Speech Unit (VSU). The visual speech unit is also constructed from the word manifolds and it has three distinct states: (a) articulation of the first viseme, (b) transition to the next viseme, (c) articulation of the next viseme. This can be observed in Fig. 9.

Fig.8 Limitations of the viseme-based approach



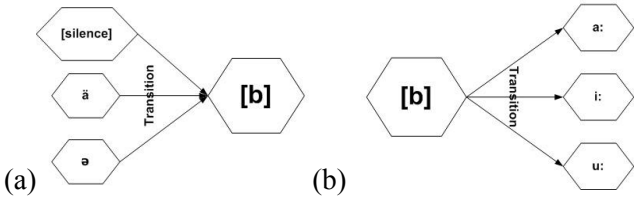
(a)



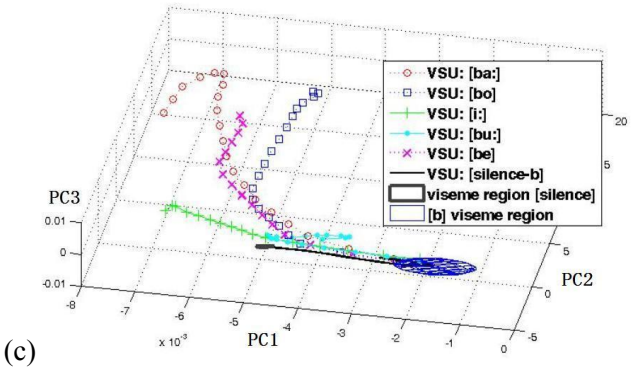
(b)

(a) The EM-PCA manifold for words 'heart' [ha:t] (blue), 'hat' [hæt] (red) and 'hook' [hu:k] (black). The feature space for viseme [a:] is depicted in cyan, for viseme [æ] in green and for viseme [u:] in purple. Viseme [h] cannot be distinguished. (b) The EM-PCA manifolds for words 'cheat' [chi:t] (red) and 'choose' [chu:s] (black). The viseme [ch] displayed in green is visible in the manifold of the word 'cheat', but it cannot be distinguished in the manifold of the word 'choose'.

Fig.9 Examples of Visual Speech Units



(a) (b)



(c)

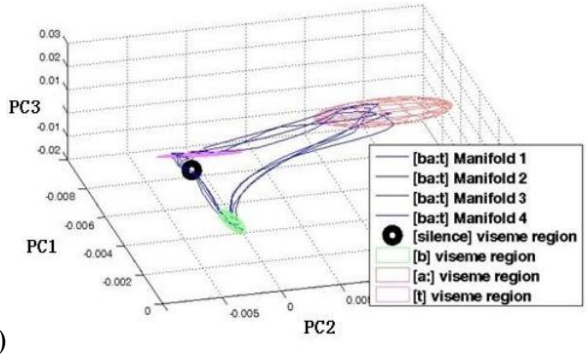
(a) VSUs: [silence –b], [ä-b] and [ə-b] (b) VSUs: [b-a:], [b-i] and [b-u] (c) The EM-PCA manifolds of VSUs: [b-a:], [b-o], [b-i:], [b-u:], [b-e], [silence-b].

VSU TRAINING PROCESS AND VSU REGISTRATION ON THE WORD MANIFOLD

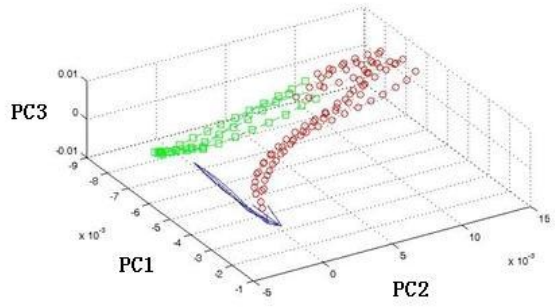
GENERATION OF THE VSU MEAN MODELS

As mentioned in the previous section, the VSUs are manually generated by extracting the frames of interest from the words manifolds and to produce a compact representation we calculate the mean model for each class of VSU. To facilitate this process, the interpolated word manifolds (see Fig. 5) are re-sampled into a fixed number of key-points that are equally spaced and the key-points for the VSUs are manually extracted. This is followed by the calculation of the mean model as illustrated in Fig. 10 (in our implementation the word manifold has been uniformly re-sampled into 50 key-points).

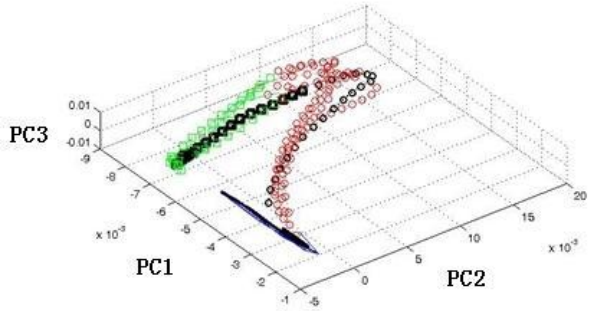
Fig.10 The calculation of VSU Mean Model



(a)



(b)



(c)

(a) Four manifolds of the word [ba:t] displayed in blue, where the four visible visemes are shown as follows: [silence] in black, [b] in green, [a:] in red and [t] in purple. (b) The VSU key-points extracted from the re-sampled manifolds. [silence - b] (blue points), [b-a:] (red points) and [a:-t] (green points). (c) The mean model for all VSUs are marked in black in the diagram ([silence-b] – black line, [b-a:] – black circles and [a:-t] - black squares).

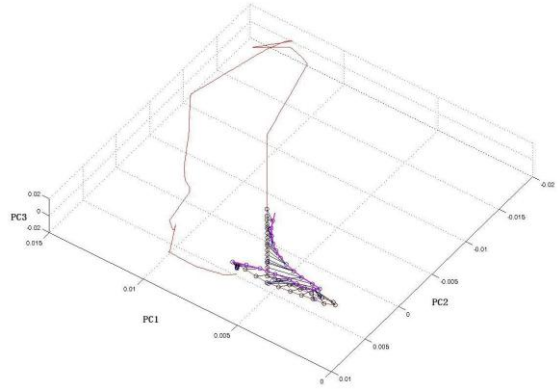
The VSU mean models depicted in Fig. 10 are used to train the HMM classifiers. In our implementation to minimize the class overlap we have trained a classifier for each VSU class. The recognition is viewed as a competitive process where all mean models of the VSUs are fitted to the manifold generated from the input video sequence. In other words we attempt to divide the word manifold into a number of consecutive sections, where each section is registered with the mean models of all VSUs stored in the database. This procedure is detailed in the next section.

REGISTRATION BETWEEN THE VSU MEAN MODELS AND THE WORD'S MANIFOLD

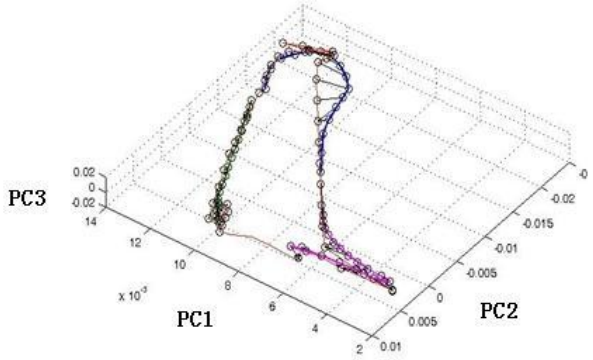
The VSU recognition process is viewed as a two-step approach. In the first step we need to match the VSU's mean models to the word manifold while in the second step we measure the matching cost between the VSU mean models and the registered section of the manifold using HMM classification (in our implementation we have used a three-state HMM classifier. For full details about the HMM classification scheme the reader can refer to [27]). To achieve the registration between the VSU mean models and the manifold constructed from the input video sequence, we applied a Dynamic Time Warping (DTW) procedure [26, 34] to identify the corresponding section of the word manifold and the points of the VSU mean model. The Dynamic Time Warping records the error distances between the

points that define the VSU mean model and the corresponding section of the word manifold and the matching cost is evaluated using HMM classification. This procedure is applied for all VSUs contained in the database and the process is illustrated in Fig. 11.

Fig.11 Registration using Dynamic Time Warping between the VSU mean model and the word



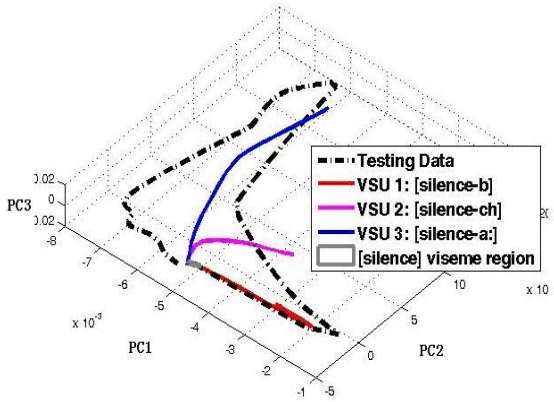
(a)



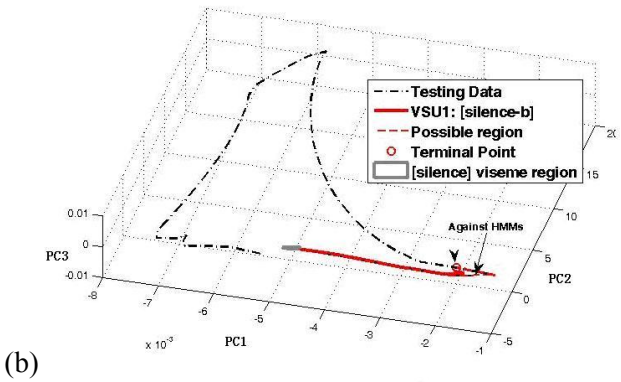
(b)

(a) Registration of the mean model of the [silence-b] VSU (purple line) to the manifold of the word [ba:t] (red line). (b) The complete registration of the word manifold and the mean models of the [silence-b], [b-a:] and [a:-t] VSUs.

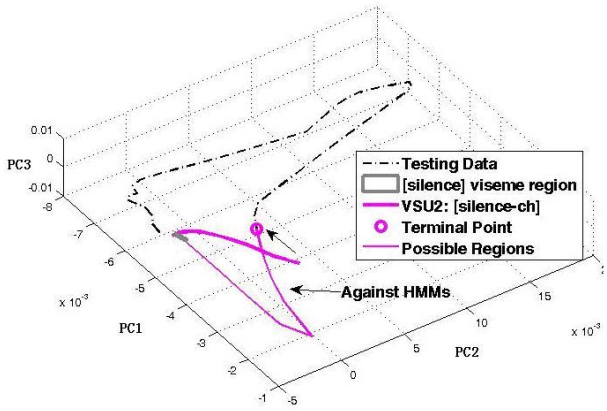
Fig.12 VSU registration and classification



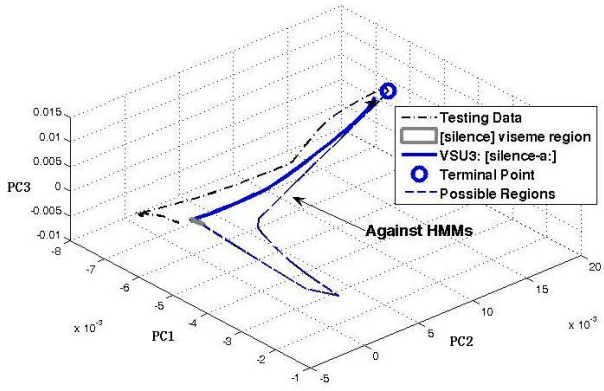
(a)



(b)



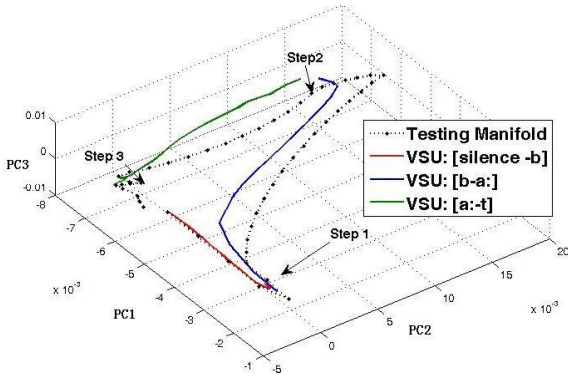
(c)



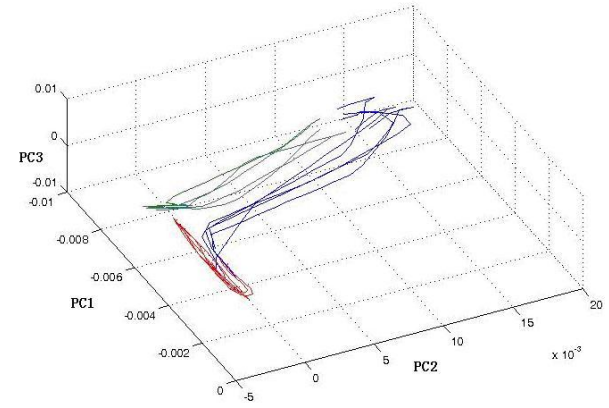
(d)

(a) The registration of three classes of the VSU Class 1: [silence-b] (red line); Class 2: [silence-ch] (purple line); Class 3: [silence-a:] (blue line) to the word manifold (black dotted line). (b) Registration between the [silence-b] VSU mean model and the word manifold. (c) Registration between the [silence-ch] VSU mean model and the word manifold. (d) Registration between the [silence-a:] VSU mean model and the word manifold. The [silence-b] VSU mean model achieved the best matching cost (evaluated using a three-state HMM classification).

Fig.13 The complete registration and matching process between the VSU mean models contained in the database and the word manifold



(a)



(b)

(a) Registration and matching process for a single word. (b) Registration and matching for five instances of the same word.

As illustrated in Fig. 11, the registration between the VSU mean models and the word manifold is iteratively applied until the last section of the manifold ends with the state [silence] that is common for the beginning and the end of the word (mouth closed). This process is illustrated step-by-step in Figs. 12 and 13.

EXPERIMENTAL RESULTS

We have created two databases consisting of 50 words that were generated by two speakers (one male and one female) where each word is repeated 10 times. (The video data has been captured using a SONY DCR-HC19E camera recorder at a sampling rate of 25 frames per second.) These databases include simple words such as 'but', 'heart', 'check', etc. and more complex words such as 'barbie', 'hoover', 'bookman', 'chocolate, etc. In our study we have conducted the experiments to evaluate the recognition rate when 12 classes of MPEG-4 visemes (see Table 1) and 60 classes of VSUs (see Table 2) (speaker 1) and 10 classes of visemes and 30 classes of VSUs (speaker 2) are used as speech elements.

Table 1 The set of MPEG-4 visemes.

Viseme Number	Phonemes	Example Words	Number of samples
1	**[b], [p], [m]**	**but, part, mark**	**300**
2	**[s], [z]**	**zard, fast**	**30**
3	**[ch], [dZ]**	**chard, charge**	**150**
4	**[f], [v]**	**fast, half, hoover**	**80**
5	**[I]**	**beat, heat**	**130**
6	**[A:]**	**but, chard, barbie**	**250**
7	**[e]**	**hat, bet**	**130**
8	**[O]**	**boat, hot**	**100**
9	**[U]**	**hook, choose**	**80**
10	**[t, d]**	**but, bird,**	**190**
11	**[h, k, g]**	**card, hook, bug**	**130**
12	**[n]**	**banana**	**20**
13	[Th]	think, that,	n/a
14	[r]	read	n/a

Note: This table adopts a viseme model established for facial animation applications by MPEG-4, which is an international audiovisual object-based video representation standard [22, 35].

Table 2 60 classes of Visual Speech Units

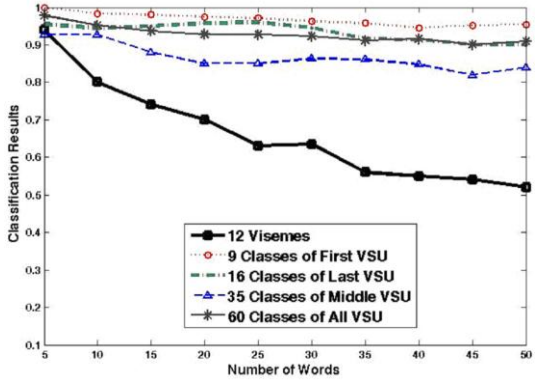
VSU Groups	Number of classes	Example VSUs
Group 1: (Start with [silence])	9	[silence-b], [silence-ch], [silence-z], [silence-f], [silence-a:], [silence-o], [silence-i:], [silence-e], [silence-u:]
Group 2 (End with [silence])	16	[a:-silence], [o-silence], [i:-silence], [u-silence], [k-silence], [i:-silence], [ch-silence], [f-silence], [m-silence], [ing-silence], [ë-silence], [p-silence], [et-silence], [ğ-silence], [s-silence], [ə-silence]
Group 3: (Middle VSU)	35	[b-a:], [b-o:], [b-i:], [b-u:], [b-ə], [b-ë], [a:-t], [a:-b], [a:-f], [a:-ğ], [a:-ch], [o-b], [o-t], [o-k], [i:-f], [i:-p], [i:-t], [u:-t], [u:-k], [u:-f], [ë-t], [f-ə:],[f-o], [k-m], [f-a:], [w-a:], [z-a:], [ə:-t], [ə:-n], [ə:-ch], [n-a:], [a:-n], [ch-a:], [ch-u:], [ch-i:]

The 60 classes of VSUs listed in Table 2 are categorized into three distinct groups. The first group is defined by the VSUs that start from the [silence] state. The second group is formed by the VSUs whose last state is [silence]. The third group consists of "middle" VSUs, which are defined by the articulation of two consecutive visemes and the transitory information between them. We have adopted this segregation of the database in order to speed up the recognition process by using the fact that the VSUs that contain the state [silence] are located either at the beginning or at the end of the word manifold. The database generated by the second speaker consists of a subset of the viseme and VSU classes that are depicted in Table 1 and 2 respectively and it has been employed in our studies to evaluate the robustness of the

proposed VSR system with respect to inter- user pronunciation variability. The first tests were conducted to evaluate the classification accuracy when visemes and VSUs are employed as speech elements and the number of words in the database is incrementally increased.

The classification results are depicted in Figs. 14 and 15 and we note that the correct identification of the visemes in the input video sequence drops significantly with the increase of the number of words in the database. Conversely, the recognition rate for VSUs suffers a minor reduction with the increase in the size of the database. This drop in recognition accuracy when visemes have been used as speech elements was expected due to viseme distortion and the occurrence of silent visemes. For example, in the EM-PCA manifold of the word 'barbie' [ba:bi] we can observe that the second viseme [b] is severely distorted when compared to the first viseme [b]. In the manifold of the word 'beat' [bi:t], the viseme [t] is invisible because the mouth is closing fast and in the manifold of the word 'fast' [fa:st], the transition between visemes [s] and [t] reveals more information than either of the visemes.

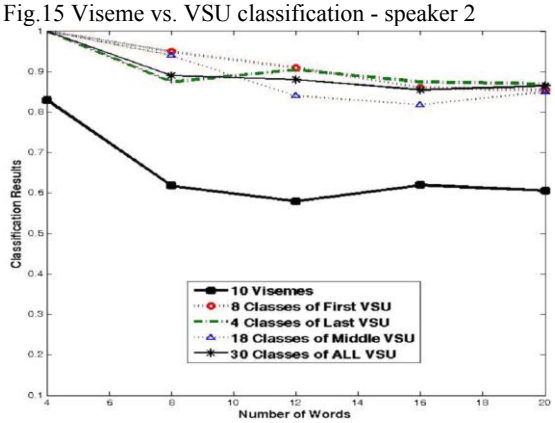
Fig.14 Viseme vs. VSU classification – speaker 1



Viseme	[b,p,m]	[s,z]	[ch]	[f,v]	[I]	[A:]	[e,ə]	[O]
Avg.Rate	69%	40%	80%	72%	57%	68%	30%	42%
	[U]	[t,d]	[k,g]	[n]				
	73%	48%	63%	50%				

The average recognition rate for 12 classes of visemes is 52% while the average recognition rate for 60 classes of VSUs is 90%.

Fig.15 Viseme vs. VSU classification - speaker 2

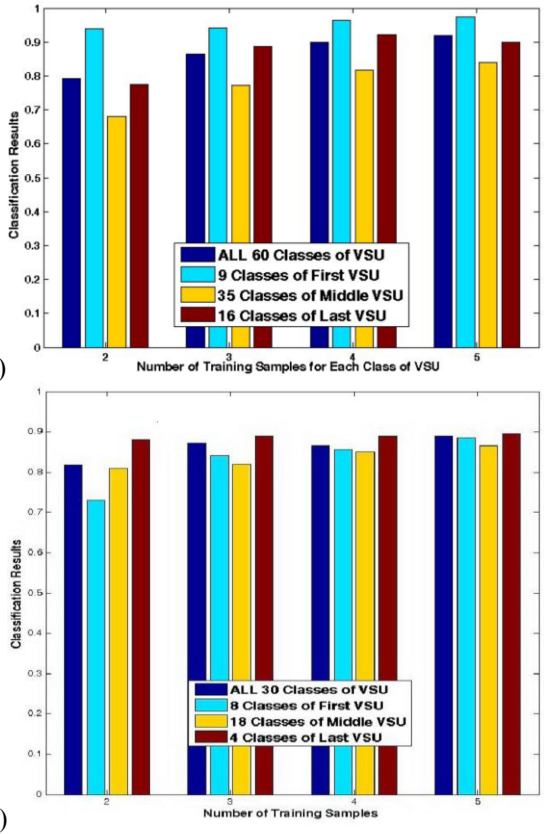


Viseme	[b,p,m]	[ch]	[f,v]	[I]	[A:]	[e,ə]	[O]
Avg.Rate	80%	70%	75%	85%	85%	55%	36%
	[U]	[t,d]	[k,g]				
	90%	43%	90%				

The average recognition rate for 10 classes of visemes is 61% while the average recognition rate for 30 classes of VSUs is 86%.

In the next experiment we evaluate the recognition rate for each class of VSU when the number of samples employed to train the HMM classifiers is varied. In our experiment we have used 2, 3, 4 and 5 samples to train the HMM classifiers for each VSU class and the experimental results are shown in Fig. 16. As expected, the recognition rate is higher when the number of samples used in the training stage is increased. In Fig. 16 we can also observe that the recognition rate for Group 3 (middle VSUs) is lower than the recognition rate for Groups 1 and 2. This is explained by the fact the VSUs contained in Groups 1 and 2 start or end with [silence] and this state can be precisely located in the word's manifold.

Fig.16 Visual Speech Unit classification results with respect to the number of training examples



(a) Speaker 1. (b) Speaker 2. In blue the overall recognition rate for all groups is depicted. In light blue the recognition rate for Group 1, in yellow the recognition rate for Group 3 and in dark red the recognition rate for Group 2 are depicted.

CONCLUSIONS

Visual speech recognition is a difficult task that involves the identification of the visual speech elements based only on the lips movements. The visual speech element is the key component of any VSR systems. In this paper we have described the development of a visual speech recognition system where the main emphasis was placed on the evaluation of the discriminative power offered by a new elementary speech element that is referred to as a Visual Speech Unit. The VSU extends the standard viseme concept by including in this new representation the transition information between consecutive visemes. To fully assess the discriminative power of the new representation we have constructed 60 classes of VSUs and we evaluated their performance when compared with that offered by the standard viseme-based approach. In our experiments we have found that the visemes cannot be robustly recognized in the manifolds of complex words while the recognition rate for VSUs is significantly higher. In our future studies, we will extend the number of VSU classes and test the developed VSR system on larger word databases.

REFERENCES

C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari and J. Zhou, Audio-Visual Speech Recognition, *IBM Research Report*, 2000.

A.V. Nefian, L.H. Liang, X. Liu and X. Pi, Audio-Visual Speech Recognition, *(http://www.intel.com/technology/computing/applications/avcsr.htm)*.

B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu and T. Huang, AVICAR: Audio-visual speech corpus in a car environment, in Proc. of *Interspeech - International Conference on Spoken Language Processing*, Jeju, Korea, 2004, pp. 2489-2492.

A. Shamaie and A. Sutherland, Hand tracking in bimanual movements, *Image and Vision Computing* 23 (2005) 1131-1149.

E.D. Petajan, Automatic lipreading to enhance speech recognition, *Ph.D. dissertation*, Univ. Illinois, Urbana-Champaign, 1984.

E.D. Petajan, B. Bischoff, D. Bodoff and N.M. Brooke, An improved automatic lipreading system to enhance speech recognition, in Proc. *of the SIGCHI Conference of Human Factors in Computing Systems*, 1988, pp. 19-25.

J. Luettin, N.A. Thacker and S.W. Beet, Active shape models for visual speech feature extraction, *University of Sheffield, U.K., Tech. Rep*. 95/44, 1995.

C. Bregler and S.M. Omohundro, Non-linear manifold learning for visual speech recognition, in Proc. of *the International Conference on Computer Vision*, 1995, pp. 494-499.

I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey, Extraction of Visual Features for Lip-reading, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (2002) 198-213.

N. Eveno, A. Caplier and P. Coulon, Accurate and quasi-automatic lip tracking, *IEEE Transactions on Circuits Systems for Video Technology* 14 (2004) 706-715.

A. Hurlbert and T. Poggio, Synthesizing a color algorithm from examples, *Science 239* (1998) 482-485.

N. Eveno, A. Caplier and P. Coulon, A new color transformation for lips segmentation, in *IEEE 4th Workshop on Multimedia Signal Processing*, Cannes, France, 2001, pp. 3-8,

Y.L. Tian, T. Kanade and J. Cohn, Robust lip tracking by combining shape colour and motion, in Proc. *of the Asian Conference on Computer Vision*, 2000, pp. 1040-1045.

C. Bregler and Y. Konig, Eigenlips for robust speech recognition, in Proc. *of the International Conference on Acoustics, Speech and Signal Processing*, Adelaide, Australia, 1994, pp. 669-672.

G. Potamianos, H.P. Graf, and E. Cosatto, An image transform approach for HMM based automatic lipreading, in Proc. of *the International Conference on Image Processing*, vol. 1, Chicago, USA, 1998, pp. 173-177.

R. Harvey, I. Matthews, J.A. Bangham and S. Cox, Lip reading from scale-space measurements, in Proc. of *the IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 582-587.

X.P. Hong, H.X. Yao, Y.Q. Wan and R. Chen, A PCA based visual DCT feature extraction method for lip-reading, in Proc. of *Intelligent Information Hiding and Multimedia Signal Processing*, 2006, pp. 321-326.

Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson and T.S. Huang, Lip-reading by locality discriminate graph, in Proc. of *the IEEE International Conference on Image Processing* (ICIP) 2007, pp. 325-328.

S.W. Foo and L. Dong, Recognition of visual speech elements using Hidden Markov Models, in Proc. of *the IEEE Pacific Rim Conference on Multimedia*, 2002, pp. 607-614.

L. Dong, S.W. Foo and Y. Lian, A two-channel training algorithm for Hidden Markov Model and its application to lip reading, *EURASIP Journal on Applied Signal Processing* 9 (2005) 1382-1399.

W. Yau, D.K. Kumar and A.S. Poosapadi, Visual recognition of speech consonants using facial movement features, *Integrated Computer-Aided Engineering* 14 (2007) 49-61.

W. Yau, D.K. Kumar and H. Weghorn, Visual speech recognition using motion features and HMM, in Proc. *of the 12th International Conference on Computer Analysis of Images and Patterns*, Vienna, Austria, LNCS 4673, 2007, pp. 832-839.

J. Yang, J. Xiao and M. Ritter, Automatic selection of visemes for image-based visual speech synthesis, in Proc. *of the IEEE International Conference on Multimedia and Expo*, 2000, vol. 2, pp. 1081-1084.

B. Rauch, The use of visual information in automatic speech recognition,

Speech Signal Processing Group, *IGK Annual Meeting*, Saarland University, 2005.

S. Roweis, EM Algorithms for PCA and SPCA, *Advances in Neural Information Processing Systems* 10 (1998) 626-632.

C.A. Ratanamahatana and E. Keogh, Everything you know about dynamic time warping is wrong, in Proc. of *the 3rd SIGKDD Workshop on Mining Temporal and Sequential Data*, 2004.

D. Yu, O. Ghita, A. Sutherland and P. F. Whelan, A new manifold representation for visual speech recognition, in Proc. of *the 12th International Conference on Computer Analysis of Images and Patterns*, Vienna, Austria, LNCS 4673, 2007, pp. 374-382.

G. Potamianos, C. Neti, G. Gravier, A. Garg and A.W. Senior, Recent advances in the automatic recognition of audio-visual speech, in Proc. of *the IEEE, 2003, vol. 91, no. 9*, pp. 1306-1326.

K. Yu, X. Jiang and H. Bunke, Sentence lip-reading using Hidden Markov Model with integrated grammar, *World Scientific Series in Machine Perception and Artificial Intelligence Series, Hidden Markov Models: Applications in Computer Vision*, 2001, pp. 161-176.

A. Sagheer, N. Tsuruta, R. I. Taniguchi and S. Maeda, Appearance features extraction versus image transform-based approach for visual speech recognition, *International Journal of Computational Intelligence and Applications* 6 (2006) 101-122.

S. Werda, W. Mahdi and A.B. Hamadou, Lip localization and viseme classification for visual speech recognition, *International Journal of Computing and Information Sciences* 5 (2007) 62-75.

M. Leszczynski and W. Skarbek, Viseme recognition – a comparative study, in Proc. of *the IEEE Conference on Advanced Video and Signal Based Surveillance*, 2005, pp. 287-292.

N. Eveno, A. Caplier and P.Y. Coulon. A new color transformation for lips segmentation, in Proc of *the IEEE 4th Workshop on Multimedia Signal Processing*, 2001, pp. 3-8.

S. Salvador and P. Chan, Fast DTW: Toward accurate dynamic time warping in linear time and space, in Proc. of *the KDD Workshop on Mining Temporal and Sequential Data*, 2004, pp. 70-80.

I.S. Pandzic, R. Forchheimer (Eds.), MPEG-4 Facial Animation – The standard, implementation and applications, John Wiley & Sons Ltd, ISBN 0-470-84465-5, 2002.

K.C. Scott, D.S. Kagels, S.H. Watson, H. Rom, J.R. Wright, M. Lee and K.J. Hussey, Synthesis of speaker facial movement to match selected speech sequences, in Proc. of *the 5th Australian Conference on Speech Science and Technology*, 1994, pp. 620-625.

P.S. Aleksic and A.K. Katsaggelos, Speech-to-video synthesis using MPEG-4 compliant visual features, *IEEE Transactions on Circuits and Systems for Video Technology*, 14 (2004) 682-692.

T. Ezzat and T. Poggio, Visual speech synthesis by morphing visemes, *International Journal of Computer Vision*, 38 (2000) 45-57.

T.J. Hazen, Visual model structures and synchrony constraints for audio-visual speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (2006) 1082-1089.

C. Bregler, M. Covell and M. Slaney, Video rewrite: Driving visual speech with audio, in Proc. of *the 24th Annual Conference on Computer Graphics and Interactive Techniques*, 1997, pp. 353-360.

K. Saenko, T. Darrell and J. Glass, Articulatory features for robust visual speech recognition, in Proc. of *the 6th International Conference on Multimodal Interfaces*, 2004, pp. 152-158.