A Novel Visual Speech Representation and HMM Classification for Visual Speech Recognition

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

Vision Systems Group, School of Electronic Engineering and Computing*
Dublin City University, Dublin, Ireland
dahai.yu2@mail.dcu.ie

Abstract. This paper presents the development of a novel visual speech recognition (VSR) system based on a new representation that extends the standard viseme concept (that is referred in this paper to as Visual Speech Unit (VSU)) and Hidden Markov Models (HMM). The visemes have been regarded as the smallest visual speech elements in the visual domain and they have been widely applied to model the visual speech, but it is worth noting that they are problematic when applied to the continuous visual speech recognition. To circumvent the problems associated with standard visemes, we propose a new visual speech representation that includes not only the data associated with the articulation of the visemes but also the transitory information between consecutive visemes. To fully evaluate the appropriateness of the proposed visual speech representation, in this paper an extensive set of experiments have been conducted to analyse the performance of the visual speech units when compared with that offered by the standard MPEG-4 visemes. The experimental results indicate that the developed VSR application achieved up to 90% correct recognition 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 in the range 62-72%.

Keywords: Visual Speech Recognition, Visual Speech Unit, Viseme, EMPCA, HMM, Dynamic Time Warping.

1 Introduction

Automatic Visual Speech Recognition (VSR) plays an important role in the development of many multimedia systems such as audio-visual speech recognition (AVSR) [1], mobile phone applications, human-computer interaction and sign language recognition [2]. Visual speech recognition involves the process of interpreting the visual information contained in a visual speech sequence in order to extract the information necessary to establish the communication at perceptual level between humans and computers. The availability of a system that is able to interpret the visual speech is opportune since it can improve the overall accuracy of audio or hand recognition systems when they are used in noisy environments.

The task of solving visual speech recognition using computers proved to be more complex than initially envisioned. Since the first automatic visual speech recognition

system was reported by Petajan [7] in 1984, abundant VSR approaches have been reported in the computer vision literature over the last two decades. While the systems reported in the literature have been in general concerned with advancing theoretical solutions to various subtasks associated with the development of VSR systems, this makes their categorization difficult. However, the major trends in the development of VSR can be divided into three distinct categories: feature extraction, visual speech representation and classification. In this regard, the feature extraction techniques that have been applied in the development of VSR systems can be divided into two main categories, shape-based and intensity based. In general, the shape-based feature extraction techniques attempt to identify the lips in the image based either on geometrical templates that encode a standard set of mouth shapes [17] or on the application of active contours [3]. Since these approaches require extensive training to sample the spectrum of mouth shapes, recently the feature extraction has been carried out in the intensity domain. Using this approach, the lips are extracted in each frame based on the colour information and the identified image sub-domain detailing the lips is compressed to obtain a low-dimensional representation.

A detailed review on the research on VSR indicates that numerous methods have been proposed to address the problems of feature extraction and visual speech classification, but very limited research has been devoted to the identification of the most discriminative visual speech elements that are able to model the speech process in the continuous visual domain. Thus, most works on VSR focused on the identification of visemes, but the visemes identification in continuous visual speech proved problematic since visemes have a limited visual support when analysed for continuous lip motions. Consequently, different visemes may overlap in the feature space, a fact that makes their recognition difficult.

To address the problems associated with the standard viseme recognition approach a new set of visual speech elements for VSR, referred to as Visual Speech Units (VSU), is proposed in this paper. This new visual speech representation has been included in the development of a VSR system that consists of four major components:

- Intensity-based lip segmentation.
- Feature extraction using Expectation Maximization PCA (EM-PCA).
- Visual Speech Units speech modelling.
- Visual Speech Units registration and HMM classification.

The main objective of this paper is to demonstrate that the inclusion of this new visual speech representation in the development of VSR leads to improved performance when compared with the performance offered by the standard set of MPEG-4 visemes.

2 Lip segmentation and EM-PCA manifold representation

2.1 Lip Segmentation

To enhance the presence of the skin in the image, the pseudo-hue [5] component is calculated from the RGB representation for each frame in the video sequence. The region around the lips is extracted by applying a histogram-thresholding scheme (the

threshold value is adaptively selected as the local minima between the first and the second peak of the pseudo-hue histogram). The images resulting from the lip segmentation procedure are as shown in Fig. 1. Fig. 1(f) is used as input data to generate the manifold representation. This will be discussed in the next section.

Fig. 1. Lip segmentation process. (a) Original RGB image. (b) Pseudo-Hue component calculated from the RGB image shown in (a). (c) Image resulting after thresholding. (d) Image describing the mouth region. (e) ROI extracted from the original image. (f) Gray-level normalized image shown in (e).

2.2 EM-PCA Manifold Generation

In order to reduce the dimensionality of the data resulting from the lip segmentation process, data compression techniques are applied to extract the lip-features from each frame in the video sequence. To achieve this goal, an Expectation-Maximization Principal Component-Analysis (EM-PCA) scheme is applied to obtain a compact representation for all images resulting from the lip segmentation procedure [6]. The Expectation-Maximization (EM) is a probabilistic framework that is usually applied to learn the principal components of a dataset using a space partitioning approach. Its main advantage resides in the fact that it does not require to compute the sample covariance matrix as the standard PCA technique and it has a complexity limited to $O(knp)$ where k is the number of leading eigenvectors to be learned, n is the dimension of the unprocessed data and p defines the number of vectors required for training.

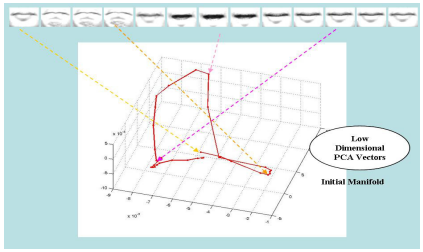
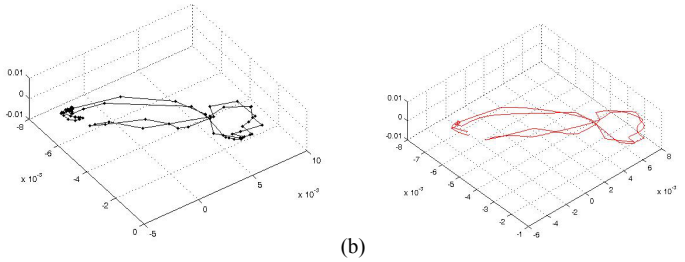


Fig. 2. EM-PCA manifold representation of the word “Bart”. Each feature point of the manifold is obtained by projecting the image data onto the low-dimensional EM-PCA space.

As explained 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. The feature points obtained after data projection on the low-dimensional EM-PCA space are joined by a poly-line by ordering the frames in ascending order with respect to time (Fig. 2) to generate the manifold representation.

2.3 Manifold Interpolation

Since the manifolds encode the appearance of the lips in consecutive frames through image compression, the shape of the manifold will be strongly related to the words spoken by the speaker and recorded in the input video sequence. Fig. 3(a) illustrates the manifolds calculated for two independent image sequences describing the same word. Although the video sequences have been generated by two speakers, it can be observed that the shapes of the manifolds are very similar.



(a) (b)

Fig. 3. Manifold representation and interpolation. (a) Manifold generated from two image sequences of the word "hot". (b) Manifold interpolation results.

While the manifold determined as illustrated in Fig. 3(a) is defined by a discrete number of points that is given by the number of frames in the video data, this manifold representation is not convenient to be used for classification/recognition purposes since the spoken words may be sampled into a different number of frames that may vary when the video data is generated by different speakers. To address this issue, the feature points that define the manifold are interpolated using a cubic-spline to obtain a continuous representation of the manifold [8]. The manifolds resulting from the interpolation procedure are depicted in Fig. 3(b). The main issue related to the identification of the speech elements that define the word manifolds is associated with the generation of a visual representation that performs an appropriate phoneme mapping in the visual domain. This problem will be addressed in detail in the next section of this paper.

3 Viseme Representation

3.1 Viseme Background

The basic unit that describes how speech conveys linguistic information is the phoneme. In visual speech, the smallest distinguishable unit in the image domain is called viseme [4, 14]. A viseme can be regarded as a cluster of phonemes and a model for English phoneme-to-viseme mapping has been proposed by Pandzic and Forchheimer [9].

In 1999, Visser et al [10] developed one of the first viseme-based classification systems where a time-delayed neural network was applied to classify 14 classes of visemes. This work has been further advanced by Foo et al [4, 16], where adaptive boosting and HMM classifiers were applied to recognize visual speech visemes. Yau

et al [11] followed a different approach when they initially examined the recognition of 3 classes of visemes using motion history image (MHI) segmentation and later they increased the number of visemes up to 9 classes. To describe the lip movements in the temporal domain, 2D spatio-temporal templates (STT) were augmented with features calculated using the discrete wavelet transform and Zernike moments. In their approach HMM classifiers were employed to discriminate between different classes of visemes.

Although there is a reasonably strong consensus about the set of English phonemes, there is less unanimity in regard to the selection of the most representative visemes. Since phonemes and visemes cannot be mapped directly, the total number of visemes is much lower than the number of standard phonemes. In practice, various viseme sets have been proposed with their sizes ranging from 6 [12] to 50 visemes [13]. Actually this number is by no means the only parameter in assessing the level of sophistication of different schemes applied for viseme categorisation. For example, some approaches propose small viseme sets based on English consonants, while others propose the use of 6 visemes that are obtained by evaluating the discrimination between various mouth shapes (closed, semi-opened and opened mouth shapes). This paper adopts the viseme model established for facial animation by an international object-based video representation standard known as MPEG-4 [9].

From this short literature review, it can be concluded that a viseme is defined as the smallest unit that can be identified using the visual information from the input video data. Using this concept, the word recognition can be approached as a simple time-ordered combination of standard visemes. Although words can be theoretically formed by a combination of standard visemes, in practice viseme identification within words is problematic since different visemes may overlap in the feature space or they may be distorted by the preceding visemes during the continuous speech process.

3.2 Viseme Representation in the EM-PCA Space

In order to evaluate the feasibility of the viseme representation when applied to continuous VSR, a set of MPEG-4 visemes is extracted from input video sequences associated with different words that are contained in our database. For instance, frames describing the viseme *[b]* are extracted from words such as ‘but’, ’boot’, ‘blue’ etc., while frames describing viseme *[ch]* are extracted from words such as ‘chard’, ‘choose’, ‘chocolate’ etc.

The feature points that define the EM-PCA manifold surface describe particular mouth shapes or lip movements and they are manually selected to represent visemes from spoken words. Fig. 4 shows the correspondence between feature points that form the visemes manifolds and the corresponding images that define visemes in the image domain. From this diagram, it can be observed that frames describing standard visemes include three independent states. The first state is the initial state of the viseme; the second state describes the articulation process and the last state models the mouth actions associated with the relaxed state. These frames are projected onto the EM-PCA space and the resulting manifolds are subjected to spline interpolation, as illustrated in Fig. 5(a). The feature points for visemes *[b], [u:]* and *[t]* are constructed from video sequences describing the word ‘*boot’ [bu:t]*. By analyzing different instances of the same word *[bu:t]*, a group of features points for visemes *[b], [u:]* and *[t]* is constructed to define each viseme in the manifold representation. These

feature points are marked with ellipsoids in the EM-PCA space to indicate the space covered by particular visemes, see Fig. 5(b). Based on these examples, we can observe that visemes are too small entities to fully characterize the entire word information since the transitions between visemes are not used in the standard viseme-based speech representation.

3.3 Viseme Limitations

As indicated in the previous section, the main shortcoming associated with the viseme representation is given by the fact that large parts of the word manifold (i.e. transitions between visemes) are 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 to describe each viseme in the EM-PCA feature space (see Fig. 5b) and this will lead to significant overlaps in the feature space describing different visemes. This problem can be clearly observed in Fig. 6 where the process of constructing the viseme spaces for two different words ('Bart' and 'chard') is illustrated. As illustrated in Fig. 6, a large region is required to describe the viseme *[a:]* in the feature space of the two different words. Viseme *[d]* (green) in word *[cha:d]* and viseme *[t]* (dark green) in word *[ba:t]* are in the same category of visemes and they also require a large region in the feature space.

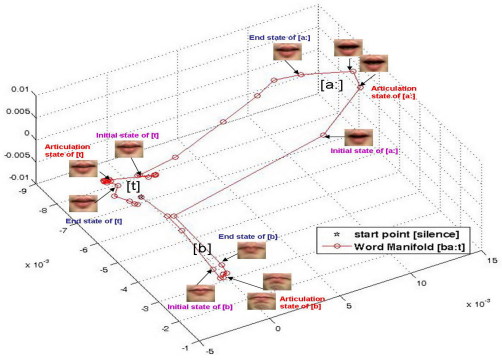
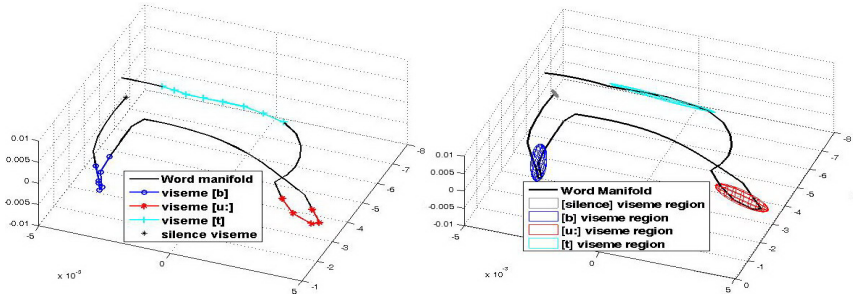


Fig. 4. EM-PCA points generated by the image sequence describing the word [ba:t].



(a) (b)

Fig. 5. Viseme representation. (a) EM-PCA feature points associated with visemes [b] [u:] and [t]. (b) The regions in the feature space for visemes [b], [u:] and [t].

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. For instance, in the manifolds generated for words 'heart', 'hat', and 'hot' the viseme *[h]* cannot be distinguished.

These limitations indicate that visemes do not map accurately the lip motions and they are subjected to a large degree of distortion when evaluated in continuous speech sequences. In conclusion, the viseme model is not optimal when applied to continuous visual speech recognition.

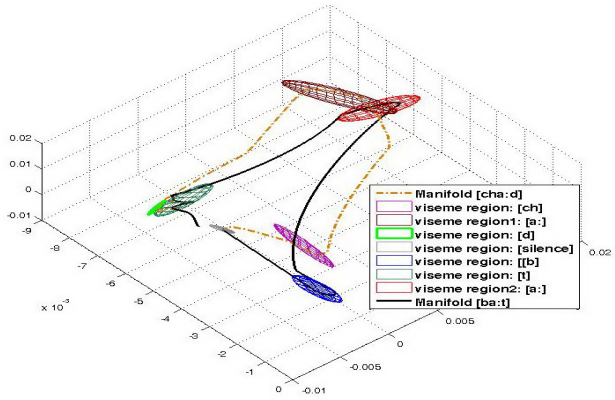
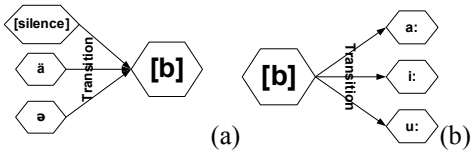


Fig. 6. Viseme feature space constructed for two different words. Word 'Bart'-viseme [b], [a:] and [t]. Word 'chard' – visemes [ch], [a:] and [d].

4 Visual Speech Units

4.1 Visual Speech Units Modelling

The main aim of this paper is to introduce a new representation called Visual Speech Unit (VSU) that includes not only the data associated with the articulation of the visemes but also the transitory information between consecutive visemes. Each VSU is manually 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. The principle behind this new visual speech representation can be observed in Fig. 7 where prototype examples of VSUs are shown.



(a) (b)

Fig. 7. Visual Speech Unit examples. (a) VSU prototypes: [silence-b], [ä-b] and [ə-b]. (b) VSU prototypes: [b-a:], [b-i] and [b-u:].

4.2 Visual Speech Units Training

As mentioned before, the construction of VSUs is based on adjacent "visible" visemes that can be identified in the word manifolds (visible visemes describe the articulation process of lip movements that can be mapped in the visual domain). In the manifold representation, the visible visemes are represented as a unique region in the EM-PCA feature space. Using this approach, the VSUs associated with word 'boot' *[bu:t]* are: *[silence-b], [b-u:]* and *[u:-t]*, they are displayed in Fig. 8(a).

To apply the VSU representation to visual speech recognition it is necessary to construct a mean model for each class of VSU. To facilitate this process, the interpolated word manifolds are re-sampled uniformly into a fixed number of feature-points. In order to generate standard VSU manifolds for training and recognition tasks, the re-sampling procedure will generate a pre-defined number of key-points that are equally distanced on the interpolated manifold surface. This re-sampling procedure ensures the identification of a standard set of feature key-points as illustrated in Fig. 8(b).

Manifolds for each VSU class are extracted from different instances of the same word and they are used to calculate the mean model. This manual procedure is illustrated in Fig. 8(c). The VSU mean models are used to train the HMM classifiers. In the implementation presented in this paper, to minimize the class overlap is has been trained one HMM classifier for each VSU class.

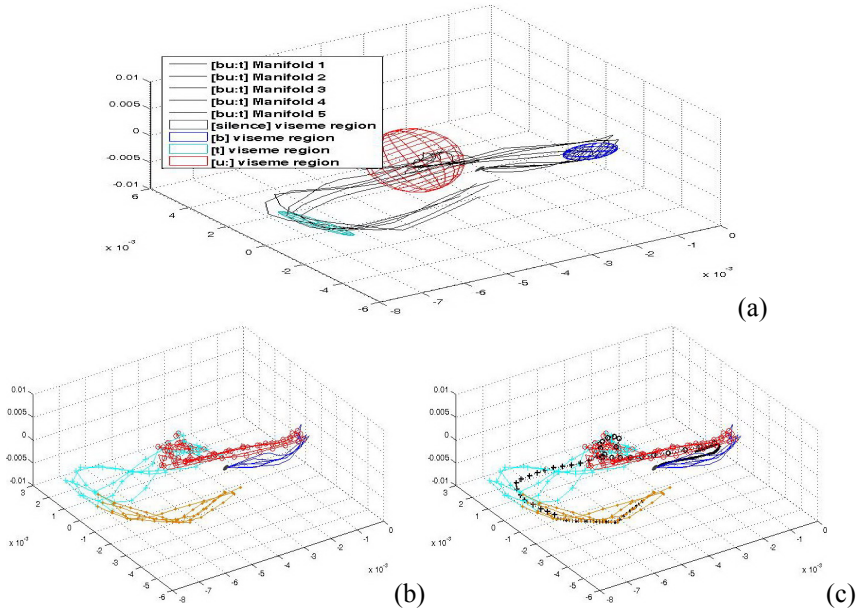
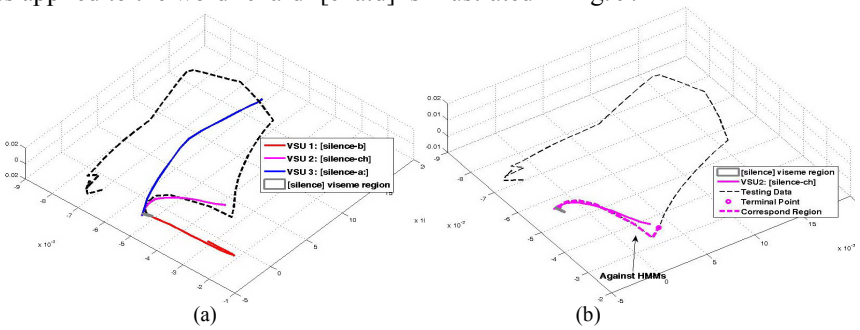


Fig. 8. The VSU training. (a) Five manifolds of the word [bu:t] (black line), four visible visemes: [silence] (gray), [b] (blue), [u:] (red) and [t] (cyan). (b) The VSU manifolds extracted and re-sampled: [silence - b] (blue), [b-u:] (red), [u:-t] (cyan) and [t-silence] (yellow). (c) The mean model for all VSUs are marked in black in the diagram [silence-b] (black line), [b-u:] (black circles), [a:-t] (black cross) and [t-silence] (black dot).

4.3 Registration between VSU and Word Manifolds

The VSU recognition is viewed as a competitive process where all VSU mean models are registered to the interpolated manifold that is calculated from the input video sequence. In this fashion, we attempt to divide the word manifold into a number of consecutive sections, where each section is compared against the mean models of all VSUs stored in the database. To achieve this, we need to register the VSU mean models with the surface of the word manifold. In this work the registration between VSU mean models and the surface of the word manifolds is carried out using the Dynamic Time Warping (DTW) algorithm. DTW is a simple solution that has been commonly used in the development of VSR systems to determine the similarity between time series and to find corresponding regions between two time series of different lengths [15].

The VSU recognition process is implemented as a two-step approach. In the first step we need to register the VSU mean models to the word manifold using DTW, 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. This procedure is applied for all VSUs contained in the database and the registration process applied to the word 'chard' [cha:d] is illustrated in Fig. 9.



(a) (b)

Fig. 9. VSU registration and classification. (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-ch] VSU mean model and the word manifold. The [silence-ch] VSU mean model achieved the best matching cost (evaluated using a three-state HMM classification).

4.4 HMM Classification

The lips motions associated with VSUs can be partitioned into three HMM states using one Gaussian mixture per state and a diagonal covariance matrix. The first state describes the articulation of the first viseme of the VSU. The second state is defined by the transition to the next viseme, while the third state is the articulation of the second viseme.

In the implementation detailed in this paper, we have constructed one HMM classifier for each class of VSU and one HMM classifier for each viseme as well. Each trained HMM estimates the likelihood between the registered section of the word manifold and the VSU mean models stored in the database. The HMM classifier that returns the highest likelihood will map the input visual speech to a particular class in the database.

5 Experimental Results

For evaluation purposes it has been created a database that is generated by two speakers. This database consists of 50 words where each word is spoken 10 times by speaker one and 20 words where each word is spoken 6 times by speaker two. In our database we have included simple words such as 'boat', '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 visemes [9] and 60 classes of VSUs (Table 1) are used as speech elements.

Table 1. The set of MPEG-4 visemes.

Viseme Number	Phonemes	Example Words	No. 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, hoover	80
5	[I]	beat, heat	130
6	[A:]	but, chard,	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,	130
12	[n]	banana	20

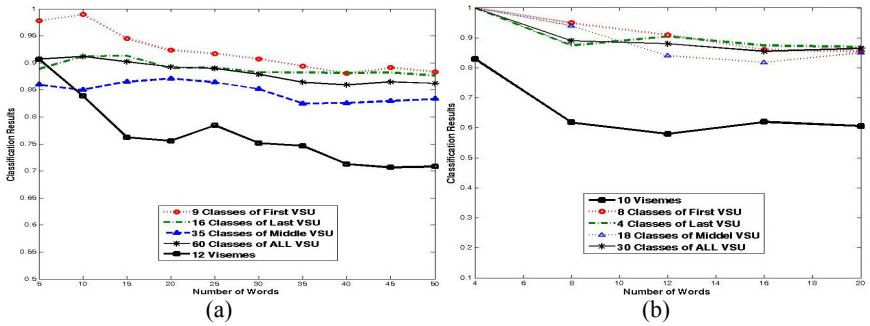
Table 2: 60 classes of Visual Speech Units.

VSU Groups	Numbers	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 experimental tests were divided into two sets. The first tests were conducted to evaluate the classification accuracy when standard MEPG-4 visemes and VSUs are employed as speech elements and the number of words in the database is incrementally increased. The classification results for speaker one is depicted in Fig. 10(a) and for speaker two are depicted in Fig. 10(b). Based on the experimental results, it can be noticed that the correct identification of the visemes in the input video sequence drops significantly with the increase in 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.

The aim of the second set of experiments is to evaluate the performance of the VSU recognition with respect to the number of samples used to train the HMM classifiers. As expected, the recognition rate is higher when the number of samples used in the training stage is increased (see Fig. 11).



(a) (b)

Fig. 10. Viseme vs. VSU classification. (a) Speaker one. (b) Speaker two.

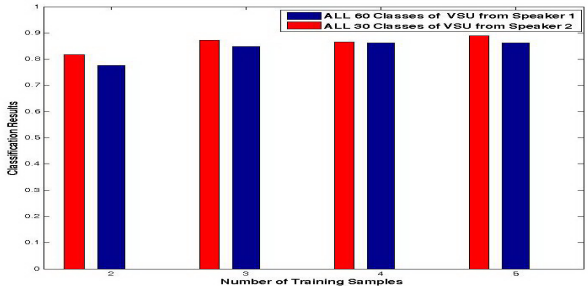


Fig. 11. Visual Speech Unit classification with respect to the number of training examples.

6 Conclusions

In this paper we have described the development of a VSR system where the main emphasis was placed on the evaluation of the discriminative power offered by a new visual speech representation that is referred to as a Visual Speech Unit (VSU). The VSU extends the standard viseme concept by including in this new representation the transition information between consecutive visemes.

To evaluate the classification accuracy obtained for the proposed visual speech representation, we have constructed 60 classes of VSUs that are generated by two speakers and we quantified their performance when compared with that offered by the standard set of MPEG-4 visemes. The experimental results presented in this paper indicated that the recognition rate for VSUs is significantly higher than that obtained for MPEG-4 visemes.

In our future studies, we will extend the number of VSU classes and test the developed VSR system on larger word databases. Future research will be also

concerned with the inclusion of the VSU based visual speech recognition in the implementation of a robust sign language gesture recognition system.

References

1. Potamianos, G., Neti, C., Gravier, G., Garg A., Senior, A.W.: Recent Advances in the Automatic Recognition of Audio-Visual Speech, Proc. of IEEE, 91(9), 1306-1326 (2003).
2. Shamaie, A., Sutherland, A.: Accurate Recognition of Large Number of Hand Gestures, Iranian Conference on Machine Vision and Image Processing, University of Technology, Tehran, ICMVIP Press (2003).
3. Luettin, J., Thacker, N.A., Beet, S.W.: Active Shape Models for Visual Speech Feature Extraction, Speechreading by Humans and Machine: Models, Systems and Applications, NATO ASI Series (1996).
4. Dong, L., Foo, S.W., Lian, Y.: A Two-channel Training Algorithm for Hidden Markov Model and its Application to Lip Reading, EURASIP Journal on Applied Signal Processing, 1382-1399 (2005).
5. Eveno, N., Caplier, A., Coulon, P.: A new color transformation for lips segmentation, 4th Workshop on Multimedia Signal Processing, Cannes, pp. 3-8, IEEE Press (2001).
6. Roweis, S: EM Algorithms for PCA and SPCA, Advances in Neural Information Processing Systems, 10, 626-632 (1998).
7. Petajan, E.D.: Automatic Lip-reading to Enhance Speech Recognition, Ph.D. dissertation, University of Illinois, Urbana-Champaign, USA (1984).
8. Yu, D., Ghita, O., Sutherland, A., Whelan, P.F.: A New Manifold Representation for Visual Speech Recognition, 12th International Conference on Computer Analysis of Images and Patterns, Vienna, Austria, LNCS Press (2007).
9. Pandzic, I.S., Forchheimer, R., (Editors): MPEG-4 Facial Animation – The Standard, Implementation and Applications, John Wiley and Sons Ltd, ISBN 0-470-84465-5 (2002).
10. Visser, M., Poel, M., Nijholt, A.: Classifying Visemes for Automatic Lip-reading, Lecture Notes in Computer Science, vol. 1692, pp. 349-352 (1999).
11. Yau, W., Kumar, D.K., Arjunan, S.P., Kumar, S.: Visual Speech Recognition Using Image Moments and Multi-resolution Wavelet Images, Computer Graphics, Imaging and Visualisation, 194-199 (2006).
12. Leszczynski, M., Skarberk, W.: Viseme Recognition – A Comparative Study, Conference on Advanced Video and Signal Based Surveillance, 287-292 (2005).
13. Scott, K.C., Kagels, D.S., Watson, S.H., Rom, H., Wright, J.R, Lee, M., Hussey, K.J.: Synthesis of Speaker Facial Movement to Match Selected Speech Sequences, 5th Australian Conference on Speech, Science and Technology (1994).
14. Potamianos, G., Neti, C., Huang, J., Connell, J.H., Chu, S., Libal, V., Marcheret, E., Haas, N., Jiang, J.: Towards Practical Deployment of Audio-Visual Speech Recognition, International Conference on Acoustics, Speech and Signal Processing, 3, 777-780 (2004).
15. Ratanamahatana, C.A., Keogh, E.: Everything you know about dynamic time warping is wrong, 3rd SIGKDD Workshop on Mining Temporal and Sequential Data (2004).
16. Foo, S.W., Dong, L.: Recognition of Visual Speech Elements Using Hidden Markov Models, Proc. of the IEEE Pacific Rim Conference on Multimedia, 607-614 (2002).
17. Silveira, L.G., Facon, J., Borges, D.L.: Visual Speech Recognition: A Solution from Feature Extraction to Words Classification, 16th Brazilian Symposium on Computer Graphics and Image Processing, 399–405 (2003).