# A PCA based Manifold Representation for Visual Speech Recognition

**Dahai Yu [1], Ovidiu Ghita [2], Alistair Sutherland [1], Paul F. Whelan [2]**

[1] School of Computing, Dublin City University
dahai.yu2@mail.dcu.ie

[2] Vision Systems Group, School of Electronic Engineering, Dublin City University

**Abstract**

In this paper, we discuss a new Principal Component Analysis (PCA)-based manifold representation for visual speech recognition. In this regard, the real time input video data is compressed using Principal Component Analysis and the low-dimensional points calculated for each frame define the manifold. Since the number of frames that form the video sequence is dependent on the word complexity and speech behavior, in order to use these manifolds for visual speech classification it is required to re-sample them into a fixed pre-defined number of key-points. These key-points are used as input for a Hidden Markov Model (HMM) classification scheme. We have applied the developed visual speech recognition system to a database containing a group of English words and the experimental data indicates that the proposed approach is able to produce accurate classification results.

**Keywords:** Visual speech recognition, PCA manifolds, spline interpolation, Hidden Markov Model.

## 1 Introduction

Visual speech recognition is an active research topic and plays an essential role in the development of many multimedia systems such as audio-visual speech recognition (AVSR) [4], mobile phone applications and sign language recognition [10]. The inclusion of lip visual features to assist the audio or hand recognition is an opportune option while this information is robust to acoustic noise. The aim of this paper is to detail the development of a visual speech recognition system that is able to recognize the visual speech using only the visual features that are extracted from the input video sequence.

For all visual recognition systems the feature extraction is the crucial part and its aim is to encode in a compact model the lip motions revealed in the input head-only video sequence that provides a visual representation of the words spoken by a speaker. The feature extraction techniques applied in the development of visual recognition systems can be broadly categorized into contour based and pixel based. One of the early visual speech recognition systems was developed by Petajan [19] in which the shape of the lips was sampled by simple morphological features including height, width and area of the lips contour. Later in 1995, Luettin et al [6] applied Active Shape Models (ASM) in order to extract the lips outline and this information was used in the recognition of a set standard English phonemes. A different approach was proposed by Harvey et al [17] where they applied a morphological transform called sieve that was applied to calculate simple one-dimensional (1D) and two-dimensional (2D) measurements that are able to sample the lips shapes. The statistics calculated from these 1D and 2D measurements are concatenated into a feature vector that was used to train a standard HMM classifier. Using the same approach, Foo and Lian [9] and Dong et al [16] proposed the use of an adaptive boosted HMM classifier where the input vector was formed by 10 geometric measurements calculated from the lips shapes. It is useful to note that this approach was applied to the identification of standard visual speech elements. More recently, Matthews et al [20] developed a pixel-based method for visual speech recognition that was based on a multi-scale analysis technique and this approach was further advanced by Hong et al [13] where a PCA based technique was employed to reduce the dimensionality of the DCT feature space.

From this short literature review, we can conclude that the pixel based feature extraction techniques [1,3,5,14,17,20] are in general better fitted to encode the lips dynamics in a compact representation than the contour-based feature extraction methods [8,12,15]. Based on this conclusion, we formulated the visual speech recognition as the process of recognizing individual words based on a new manifold representation. This approach to visual speech recognition is appealing since provides a generic framework to recognize words (or standard visual elements) using low-dimensional manifolds that are calculated directly from the input video sequence. Thus, in this paper our main aim is to evaluate the discriminative power offered by the manifold representation when applied to visual speech recognition, while the other important problem consists of designing the classification scheme that returns optimal results. These issues will be addressed in detail in the following sections of this paper.

This paper is organized as follows. Section 2 presents an overview of the proposed system while Section 3 describes the methodology applied to generate the manifolds from input data. This discussion is continued in Section 4 with an introduction on the adopted HMM classification scheme. Section 5 details the experimental performance of the developed system and some concluding remarks are provided in Section 6.

## 2 System Overview

The developed system for visual speech recognition consists of three main steps. In the first step, the lips are extracted from the input video data. In order to achieve this goal we calculate the pseudo-hue component from the RGB data [3] and the lips are segmented by applying a histogram-based thresholding scheme (see Fig. 1). The aim of the second component is to generate the Expectation Maximization PCA (EM-PCA) manifolds and perform manifold interpolation and re-sampling. The re-sampled manifolds are used as inputs to train a HMM classifier (for this implementation each word has a assigned a distinct class). The role of the third component is to classify the manifolds calculated from the input speech sequences into a number of classes that are stored in a database. The block diagram of the proposed visual speech recognition system is depicted in Fig. 2.

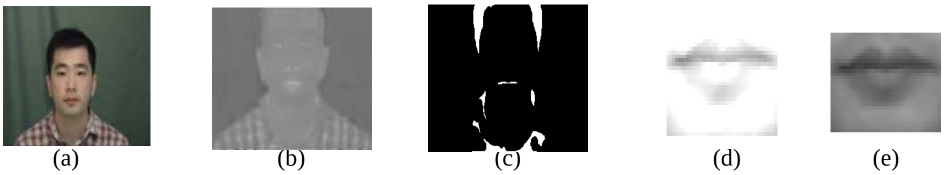

**Fig. 1.** Lip detection algorithm. (a) RGB image. (b) Pseudo-hue image. (c) Image resulting after the application of the histogram-based thresholding. (d) Lips region – pseudo hue. (e) Lips region – grayscale.

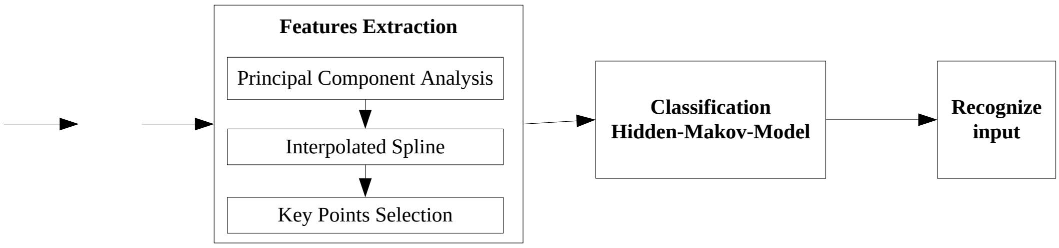

**Fig. 2.** Block diagram of the proposed visual speech recognition system. Step (1) Lip detection from the pseudo-hue component. Step (2) Feature extraction. Step (3) Classification.

# 3 Manifold Representation

## 3.1 EM-PCA Mathematical Background

Expectation-Maximization PCA (EM-PCA) is an extension of the standard PCA technique by incorporating the advantages of the EM algorithm in terms of estimating the maximum likelihood values for missing information. This technique has been originally developed by Roweis [7] and its main advantage over the standard PCA is the fact that it is more appropriate to handle large datasets especially when dealing with sparse training sets. The EM-PCA procedure has two distinct stages, the E-step and M-step:

$$\text{E-step: } W = (V^T V)^{-1}\, V^{-1} A; \quad \text{M-step: } V_{new} = A W^T (W W^T)^{-1} \tag{1}$$

where ***'W'*** is the matrix of unknown states, ***'V'*** is the test data vector, ***'A'*** is the observation data, and $^T$ defines the transpose operator.

## 3.2 Manifold Generation from Input Image Data

As indicated in Section 2, the lips are segmented in each frame by thresholding the pseudo-hue component calculated from RGB data. The grayscale data around the lips region is extracted as illustrated in Fig 1(e) and this information is used to generate the low-dimensional space that is calculated using the EM-PCA procedure. Then, the grayscale data is projected onto the low-dimensional space and for each frame will be calculated a low dimensional point (vector). The feature points obtained after data projection on the low-dimensional EM-PCA space are joined by a polyline by ordering the frames in ascending order with respect to time (see Fig. 3). In this way we generate a surface in the feature space that is called manifold [11,18]. Fig. 4 shows the manifolds calculated from three independent image sequences that describe the same visual speech sequence (word) in the EM-PCA feature space. It can be noted that the shapes of the manifolds are very similar and can be interpreted as a visual "signature" of the word described by the sequence of lips dynamics.

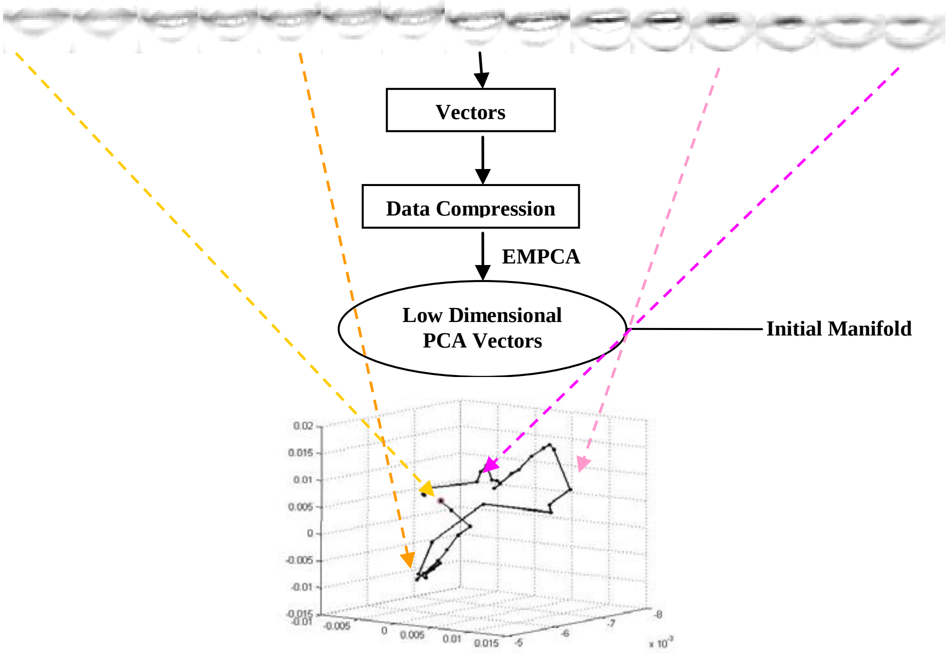

**Fig. 3.** Manifold generation process.

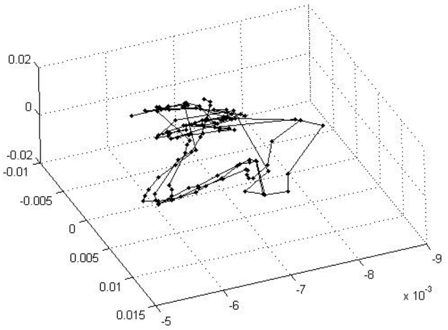

**Fig. 4.** Manifolds calculated from three image sequences that describe visually the same word. The appearances of the manifolds indicate that their shapes are similar and can be regarded as visual signatures of the spoken words.

## 3.3 Manifold Interpolation and Re-sampling

For visual speech recognition purposes, we would like to extract the information associated with the lips motions from all frames that define the input video sequence. In Section 3.2, it has been indicated that the PCA manifolds can be used to generate a compact representation of the lips dynamics that describe the word spoken by a speaker using only the visual information. While the shape of the manifolds can be potentially used to discriminate between different words, they cannot be used directly to train a classifier or to recognize an unknown input image sequence. This is motivated by the fact that the number of frames contained in the input video sequence is variable and depends on the complexity of the word spoken. In this way, short words such as ‘I’, ‘shy’, etc. have a small number of frames. Conversely, the visual articulation for long words such as ‘banana’, ‘another’ and ‘flower’ generates larger image sequences and as a result the number of feature points that describes the manifolds is larger.

This is a real problem when these manifolds are used to train a classifier, as the number of feature points is different. To circumvent this problem, we need to interpolate the manifold to obtain a continuous surface and then re-sampling it uniformly using a pre-defined number of key-points. This procedure is appropriate as it allows us to construct uniform data that can be used to train a classifier or to obtain the classification result for an unknown visual speech sequence.

### 3.3.1 Manifold Interpolation Using a Cubic Spline Function

The application of cubic spline interpolation has two main advantages. Firstly, it allows us to generate a smooth surface for EM-PCA manifolds and secondly it reduces the effect of noise (and the influence of the objects surrounding the lips such as teeth and tongue) associated with the feature points that form the manifolds in the EM-PCA space. This is clearly shown in Fig. 5(a) where are illustrated the manifolds obtained after the application of cubic interpolation. Fig. 5(a) illustrates the interpolated manifolds generated for the words “slow”, “shy”, “art” and “white”.

### 3.3.2 Manifold Re-sampling into a Pre-defined Number of Key-points

As mentioned in the previous section, in order to generate standard data for training/recognition, we need to uniformly re-sample the manifolds into a pre-defined number of key-points. This re-sampling procedure will allow the identification of a standard set of key-points as illustrated in Fig. 5(b). We decided to re-sample the manifold uniformly since this procedure will generate key-points that are equally distanced on the interpolated manifold surface and accurately sample the information associated with the manifold’s shape.

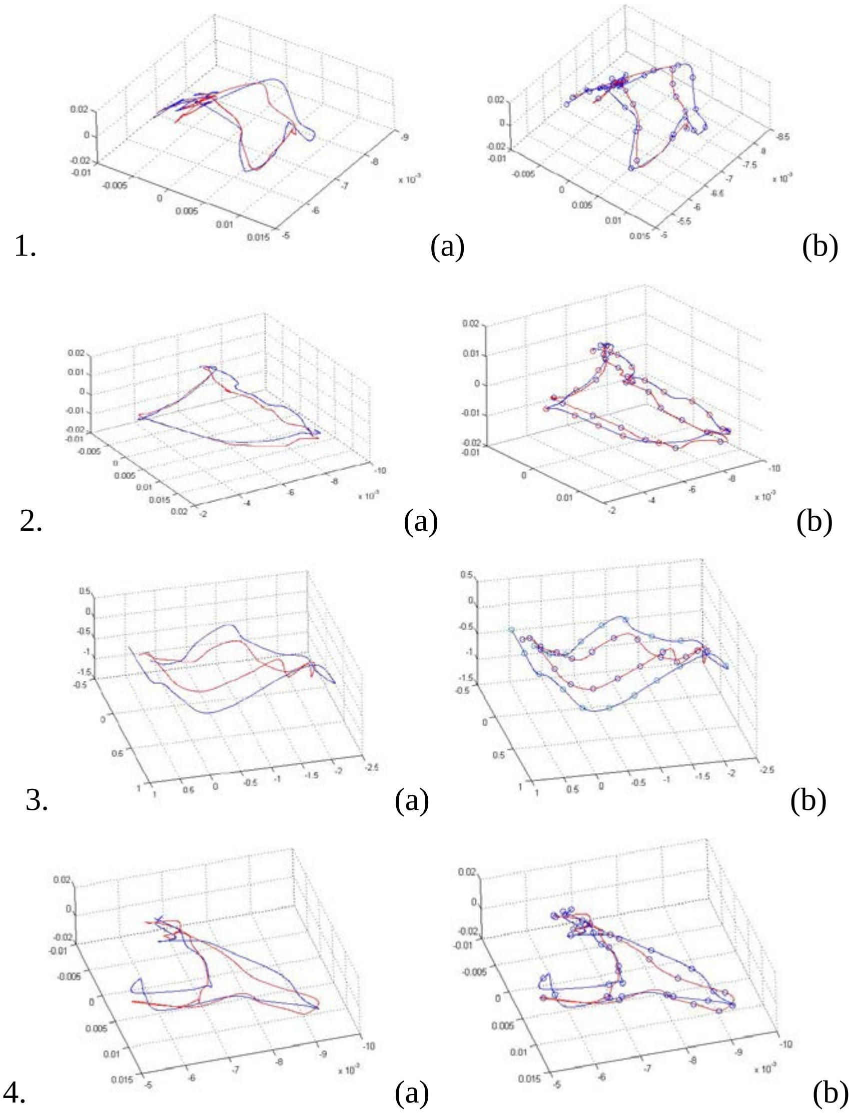

**Fig. 5.** Manifold interpolation and re-sampling procedure. (a) Manifold interpolation results. (b) Uniform re-sampling of the interpolated manifolds (key-points=20). 1. Two image sequences – word "art". 2. Two image sequences – word "shy". 3. Two image sequences – word "slow". 4. Two image sequences – word "white".

# 4 HMM Classification

In this paper the HMM classification scheme is evaluated where the inputs are the key-points obtained after the application of manifold interpolation and re-sampling procedure that is described in Section 3.3.2. Hidden Markov Model classifier performs a division of a process into a number of discrete states [2,9,16]. During the HMM classification process the observation sequence $O_n$ is defined by the key-points of the re-sampled manifold where *n* is the number of key-points. This process is described in Fig. 6 where $O_n$ is assumed to be sampled by a sequence of hidden states $S_t$ ($S_t$ is the state set of the HMM classifier).

Our study on lips dynamics indicates that the lips motions associated with the visual speech can be partitioned into three states. The first state describes the transition from the initial state of the visual speech to articulation state. The articulation state is the part of the visual speech that describes the largest variation in the lips dynamics. The third state is the end part of the speech sequence and defines the transition from articulation to the end cycle of the speech process. Among these states, the articulation state provides the highest level of information in discriminating visual speech words. Fig. 6 illustrates the partition of the visual speech into a sequence of three hidden states.

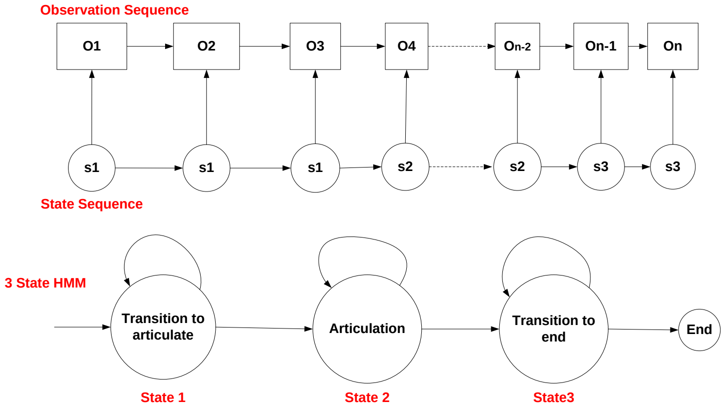

**Fig. 6.** HMM classification.

For this implementation the Baum-Welch algorithm is employed to find the unknown parameters of the HMM classifier and we have constructed a HMM classifier for each class of words (i.e. for *k* words in the database we need to train *k* independent HMMs). Each trained HMM estimates the likelihood between the input visual speech sequence (input manifold) and the words contained in the database. The HMM classifier that returns the best approximation will map the input visual speech to a particular class in the database. During the training process, the number of hidden states is set to three, the length of sequence is set as the number of key-points and the maximum of number of iterations is set to 30.

## 5 Experimental Results

A number of experiments were carried out to assess the performance of the proposed approach. For this study, we have created a database consisting of 10 words (30 examples for each word) generated by one speaker. From these examples of each word 10 examples are used for training and 20 examples are used for testing. The input database is divided into 10 classes and the classification results obtained by the HMM classification scheme discussed in Section 4 are depicted in Fig. 7. Based on the evaluation of the experimental results we can observe that the best results are obtained when the interpolated manifolds are re-sampled to 20-30 key-points. Another important finding resulting from this investigation is the fact that the manifolds offer a good discrimination (average classification success rate is 95%) and we can conclude that the re-sampled manifolds are suitable features to be used for visual speech recognition.

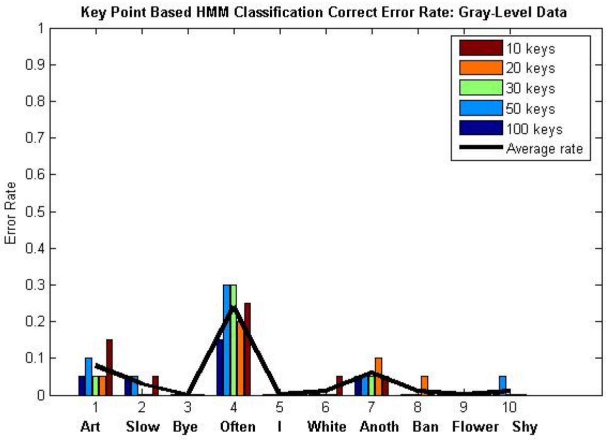

**Fig. 7.** Classification error rate achieved by the three-state HMM classification scheme.

## 6 Conclusions

This paper describes the development of a visual speech recognition system where the main contribution is the evaluation of the discriminative power offered by a new manifold-based feature extraction technique. In this regard, the manifolds are generated from the input image data surrounding the lips and this data is compressed using an EM-PCA procedure. Since the visual speech sequence is defined by a different number of frames, they cannot be used directly as inputs for classification. To address this problem we propose to interpolate the manifolds and then re-sample them uniformly into a predefined number of key-points. This procedure was opportune as it allowed us to compare one manifold with another and to study the variation between manifolds.

It is useful to note that in our experiments we have included image data that was generated by a single speaker and the database was defined only by a small number of words. During experimentation we also observed that the recognition rate for complex words such as "banana" and "another" is generally lower than the recognition rate obtained for simpler words such as "I" and "shy". In the future, we aim to evaluate the proposed approach on a larger number of words that are generated by different persons and to evaluate the temporal information as an additional cue in the recognition process.

## References

[1] N. Eveno, A. Caplier and P. Coulon (2004), "Accurate and Quasi-Automatic Lip Tracking", IEEE Trans. Circuits Syst. Video Techn. 14(5), pp. 706-715.

[2] Z. Ghahramani, Machine Learning Toolbox, Version 1.0 01-04-96, University of Toronto.

[3] Y. L. Tian and T. Kanade (2000), "Robust Lip Tracking by Combining Shape, Colour and Motion", Proc. of the Asian Conference on Computer Vision, pp. 1040 –1045.

[4] A.V. Nefian, L.H. Liang, X. Liu and X. Pi (2002), "Audio-Visual Speech Recognition", Intel Technology & Research.

[5] N. Eveno, A. Caplier, and P.Y. Coulon (2001), "A New Color Transformation for Lips Segmentation", 4th IEEE Workshop on Multimedia Signal Processing, pp. 3-8, Cannes, France.

[6] J. Luettin, N.A. Thacker and S.W. Beet (1995), "Active Shape Models for Visual Speech Feature Extraction", University of Sheffield, U.K., Tech. Rep. 95/44.

[7] S. Roweis (1998), "EM Algorithms for PCA and SPCA", Advances in Neural Information Processing Systems, vol. 10, pp. 626-632.

[8] T. Cootes, G. Edwards and C. Taylor (1988), "A Comparative Evaluation of Active Appearance Model Algorithms" Proc. of the British Machine Vision Conference, pp. 680-689.

[9] S.W. Foo and Y. Lian (2004), "Recognition of visual speech elements using adaptively boosted HMM", IEEE Trans. on Circuits and Systems for Video Technology, 14(5), pp. 693-705.

[10] A. Shamaie and A. Sutherland (2003), "Accurate Recognition of Large Number of Hand Gestures", Proc. of Iranian Conference on Machine Vision and Image Processing, University of Technology, Tehran.

[11] S.R. Das, R.C. Wilson, M.T. Lazarewicz and L.H. Finkel (2006), "Gait Recognition by Two-Stage Principal Component Analysis", Automatic Face and Gesture Recognition, pp. 579-584.

[12] M. Gordan, C. Kotropoulos and I. Pitas (2002), "Application of Support Vector Machines Classifiers to Visual Speech Recognition", Proc. of the International Conference on Image Processing.

[13] X.P. Hong, H.X. Yao, Y.Q. Wan and R. Chen (2006), "A PCA Based Visual DCT Feature Extraction Method for Lip-Reading", Proc. of Intelligent Information Hiding and Multimedia Signal Processing, pp. 321-326.

[14] W.C. Yau, D. K. Kumar, S. P. Arjunan and S. Kumar (2006), "Visual Speech Recognition Using Image Moments and Multiresolution Wavelet Images", Computer Graphics, Imaging and Visualization, pp. 194-199.

[15] H.E. Cetingul, Y. Yemez, E. Erzin and A.M. Tekalp (2006), “Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading“, IEEE Trans. on Image Processing, 15(10), pp. 2879-2891.

[16] L. Dong, S.W. Foo and Y. Lian (2005), “A Two-channel Training Algorithm for Hidden Markov Model and its Application to Lip Reading”, EURASIP Journal on Applied Signal Processing, 2005(9), pp. 1382-1399.

[17] R. Harvey, I. Matthews, J.A. Bangham and S. Cox (1997), “Lip Reading from Scale-Space Measurements”, Proc. of Computer Vision and Pattern Recognition, pp. 582-587.

[18] C. Bregler and S.M. Omohundro (1995), ”Nonlinear Manifold Learning for Visual Speech Recognition”, Proc. of the International Conference on Computer Vision, pp. 494-499.

[19] E.D. Petajan (1984), “Automatic Lip Reading to Enhance Speech Recognition”, in GLOBECOM’84, IEEE Global Telecommunication Conference, 1984.

[20] Matthews, T. Cootes, J.A. Bangham, S. Cox and R. Harvey (2002). “Extraction of Visual Features for Lipreading”, IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(2), pp. 198-213.